University of Verona

Department of Computer Science

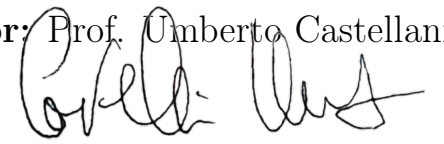
Doctoral Program in Computer Science

With the financial contribution of
Humatics, a SYS-DAT Group company

S.S.D. INF/01, Cycle XXXVIII°, 2022

Introducing and Refining Language in Vision Models

Coordinator: Prof. Umberto Castellani

Signature: 

Supervisor: Prof. Marco Cristani

Signature: 

Ph.D. Candidate: Federico Girella

Signature: 

Introducing and Refining Language in Vision Models
Federico Girella

PhD thesis

Verona, 26 03 2026

La borsa di dottorato è stata cofinanziata con le risorse del PNRR:

• per il DM 351 nell'ambito della Missione 4 ("Istruzione e ricerca") – Componente 1 ("Potenziamento dell'offerta dei servizi di istruzione: dagli asili nido all'Università"), Investimento 3.4. ("Didattica e competenze universitarie avanzate") e Investimento 4.1 ("Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale") - progetto M4C1 –Inv. 3.4 e progetto M4C1 – Inv. 4.1

• per il DM 352, nell'ambito della Missione 4 ("Istruzione e Ricerca") – Componente 2 ("Dalla Ricerca all'Impresa"), Investimento 3.3 ("Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori da parte delle imprese") – progetto M4C2 Investimento 3.3

THESIS COMMITTEE

Thesis Supervisor

Marco Cristani
Professor at University of Verona
Department of Engineering for Innovation Medicine

Thesis Reviewers

Angelo Porrello
Researcher and Professor at University of Modena and Reggio Emilia
Department of Engineering "Enzo Ferrari"

Pietro Moreiro
Researcher at Italian Institute of Technology

by

Federico Girella

Submitted to the Department of Computer Science
on March 10, 2026 in partial fulfillment of the requirements for the

PH.D. IN COMPUTER SCIENCE

ABSTRACT

This thesis investigates how to extend the role of language within multimodal artificial intelligence, moving from its initial grounding in visual systems toward improved usability in representation, evaluation, and generative control. While vision-and-language models (VLMs) have achieved impressive capabilities, they still struggle with two key challenges: integrating language into domains that traditionally rely on purely visual or sensory information, and effectively leveraging the full expressive power of language, particularly when dealing with abstract concepts, compositional semantics, and user-driven creative tasks.

The first part of this thesis focuses on *introducing language* into vision-centric systems. We enhance spatial perception through a Language-enhanced Renderable Neural Radiance Map (Le-RNR-Map), enabling natural language queries and affordance-based navigation within learned visual environments. In industrial inspection, we demonstrate how linguistic knowledge can guide anomaly detection and data generation via diffusion-based methods. Through text-guided defect synthesis and human-in-the-loop feedback, these contributions show that language can improve interpretability, robustness, and human-AI collaboration in practical, domain-specific systems.

The second part addresses the challenge of *improving the usability of language* in multimodal models. We propose a training-free latent adaptation method that strengthens the representation of abstract language in VLMs, enabling better retrieval and semantic alignment. We then introduce a localized evaluation metric for text-to-image models, designed to assess fine-grained compositional correctness between entities and attributes. Finally, we develop a multimodal generation framework that combines textual and sketch-based conditioning for controllable, stepwise diffusion-based fashion design.

Together, these contributions outline a trajectory from grounding linguistic meaning in visual tasks to enhancing its expressivity and operational value. The thesis demonstrates that language is not merely an annotation layer for visual data, but a powerful interface that enables richer reasoning, interaction, and creativity in multimodal AI systems.

Thesis supervisor: Marco Cristani
Title: Professor

Acknowledgments

Completing a PhD is a journey that extends far beyond the confines of research and academia, and I am deeply grateful to all those who have walked alongside me during these years.

First and foremost, I thank my family and my brother for their unwavering support throughout this journey. Your belief in me, even during the most challenging moments, has been my anchor. The sacrifices you made and the encouragement you provided have made this achievement possible, and for that, I am forever grateful.

To my dear friends, thank you for reminding me that life exists beyond work. Your presence, laughter, and the moments we shared have been essential in maintaining my balance and perspective. You reminded me to celebrate the small victories, to find joy in the everyday, and that work, no matter how important, is not the only thing that matters.

To my colleagues and friends who accompanied me throughout this PhD journey, I am grateful for your camaraderie, the stimulating discussions, the shared frustrations, and the mutual encouragement. Whether it was discussing ideas over coffee, anxiously waiting for results on a wednesday night over pizza, or simply sharing the ups and downs of doctoral life, your companionship made this experience richer and more meaningful. Thank you for creating an environment where collaboration and friendship flourished.

Finally, I extend my deepest gratitude to my supervisors for their invaluable guidance and support. Your mentorship, patience, and expertise have shaped not only this research but also my growth as a researcher. Thank you for challenging me to think critically, for providing direction when I felt lost, and for supporting me through the inevitable challenges that accompany a PhD. Your dedication to my development has been instrumental in bringing this work to completion.

To all of you, and to everyone who has contributed to this journey in ways both big and small: thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Vision-and-Language Models (VLMs) have emerged as a central paradigm in modern Artificial Intelligence, enabling systems to jointly reason over visual and textual information. The first popular example of VLMs can be attributed to the contrastive family of models introduced by CLIP [12], published in 2021. This work introduced a contrastive learning framework, training two text and image encoders to project the two modalities into a shared representation space. While the initial publication primarily focused on the generalization of these models as zero-shot classifiers, the shockwave in the research community sparked the beginning of numerous other research works that jointly unify vision and language representations. An example of this can be seen in text-to-image generative models. Earlier works in the image generation field focused on learning the distribution of images either without any conditioning [13] or with fixed class conditioning [14], limiting the expressivity of such models to a pre-defined set of concepts, *i.e.*, the classes contained in the training dataset. With the introduction of CLIP, text-conditioning became the de facto standard for image generation, popularized by the Stable Diffusion [15] family of text-to-image generative models, first published in 2022, which have now become a staple in the community.

Despite this remarkable progress, two major challenges remain largely open. The first concerns the **introduction of language into traditionally non-linguistic AI systems**. In many applied domains, such as industrial manufacturing, perception, and decision-making, sensory and visual data dominate [2], while language is rarely integrated as an input modality. This prevents users from interacting with these systems through natural communication, limiting accessibility and adaptability [8].

The second challenge concerns the **usability of language within multimodal models**. Even when language is incorporated, current systems often struggle to handle the richness of human linguistic expression [11, 10]. Abstract or affective terms are underrepresented in training data [11]; evaluation metrics fail to capture fine-grained compositional relationships between entities and their attributes [10, 16]; and generative models frequently lack controllability when conditioned on complex textual or multimodal prompts [17, 18, 9].

This thesis addresses both challenges by following a progressive research trajectory: from *introducing language* into domain-specific visual systems to *improving the usability of language*

in advanced multimodal architectures. The overarching goal is to advance multimodal AI towards richer, more interpretable, and more human-centered forms of understanding and generation.

1.2 Contributions and Structure of the Thesis

This thesis advances the understanding and design of Vision-and-Language Models through two complementary parts.

Part I: Introducing Language into Vision tasks

In the first part, we investigate whether it is possible to introduce natural language inside traditional visual tasks in order to improve human-machine interaction and overall performance. In Chapter 3 we will discuss *Le-RNR-Map* [5], a Language-enhanced Renderable Neural Radiance Map with which we introduce natural language querying into spatial-visual maps, enabling affordance-based search and navigation. Afterwards, we turn to a second family of problems in which vision dominates the pipeline but where practitioners naturally reason in language: industrial inspection. Here, the central question is how to inject linguistic knowledge (*e.g.*, expert-written defect descriptions) into diffusion models so that generation and detection can be steered by semantics rather than only by pixel-space cues.

Chapter 4 introduces *In&Out* [6], a diffusion-based augmentation strategy designed to make anomaly detection systems more robust when only scarce defect data are available. Beyond improving performance through synthetic data, this chapter serves a narrative role in the thesis: it is the first step in making language actionable in a non-linguistic setting, by making use of industrial semantic knowledge to compensate for the lack of training data.

Chapter 5 then builds on this direction with *DIAG* [8] (Diffusion-based In-distribution Anomaly Generation), which systematically generates in-distribution anomalies guided by expert knowledge. In the broader research trajectory, DIAG acts as a bridge between Part I and Part II: it highlights that introducing language is not only a matter of adding text as an input, but of designing models and objectives that can *use* linguistic structure to produce targeted, interpretable, and practically useful outputs.

Part II: Enhancing Language in Vision-and-Language Models

In the second part of this thesis, we shift our focus towards improving the existing interaction and representation of language within Vision-and-Language Models, making them more expressive and human-aligned. Starting with the *Abstract-to-Concrete Translator (ACT)* in Chapter 6, we improve the treatment of abstract language in pre-trained VLMs, showing that abstract linguistic cues can be effectively aligned with visual representations without retraining. The *Localized VQA-Score (L-VQAScore)* in Chapter 7 introduces a novel metric for evaluating compositional text-to-image generation by localizing and probing entity-attribute alignments through Visual Question Answering. Finally, the *Localized Text and Sketch (LOTS)* adapter in Chapter 8 enables fine-grained control in multimodal diffusion models

by conditioning generation on paired text and sketch information, bridging human creative workflows and generative AI.

1.3 Thesis Outline

The remainder of this thesis is organized as follows:

Chapter 2: Background of the technologies used throughout this thesis.

Chapter 3-5: Part I: Introducing Language into Vision tasks. These chapters present methods that bring natural language into vision-dominated tasks, including spatial navigation and industrial anomaly detection.

Chapter 6-8: Enhancing Language in Vision-and-Language Models. These chapters explore the representation of abstract language, the evaluation of compositional multimodal understanding, and controllable text-and-sketch-based generation.

Chapter 9: Conclusion and Future Directions. This chapter summarizes the findings, highlights the thematic connections between works, and outlines limitations and future research directions in the language-image multimodal field.

Chapter 2

Background

This chapter provides a technical overview of the principal models and learning paradigms that underpin the methods developed throughout this thesis. Rather than offering a broad survey of multimodal learning, we focus on the specific components that are repeatedly used across the proposed approaches: contrastive vision-language representation learning, diffusion-based generative modeling, and parameter-efficient adaptation strategies. These elements form the methodological backbone for grounding, retrieval, evaluation, and controllable generation.

2.1 Vision–Language Representation Learning

Vision-Language Models (VLMs) aim to learn a joint semantic space in which visual and textual inputs can be compared, combined, and used for conditioning [12, 19]. Given paired image-text data, the model learns correspondences between visual patterns (objects, attributes, spatial configurations) and linguistic tokens. This shared representation enables a wide range of cross-modal tasks, including zero-shot classification, image-text retrieval [11], visual question answering [10], and text-to-image generation [15, 16].

A key property of these models is *semantic transfer*: knowledge acquired in one modality can be leveraged in another without task-specific supervision. For instance, a model trained on image-caption pairs can recognize novel categories at inference time by matching an image embedding to textual prompts describing candidate classes. This capability is particularly relevant in data-scarce domains, where annotated training data may be limited.

Despite these advantages, the learned alignment is often biased toward frequently occurring, concrete visual concepts. Abstract descriptors, compositional relationships, and fine-grained attribute binding remain challenging [11, 10]. These limitations motivate several of the contributions in this thesis, which aim to improve the expressivity and controllability of language within multimodal systems.

2.2 Contrastive Vision–Language Models: CLIP

Contrastive Language–Image Pretraining (CLIP) [12] is a dual-encoder architecture composed of an image encoder f_v and a text encoder f_t. Each encoder maps its input to a normalized

embedding vector in a shared latent space. Training is performed on large-scale image–caption datasets using a symmetric contrastive objective that maximizes similarity for matching pairs and minimizes it for mismatched pairs.

Let $\mathbf{v}_i = f_v(I_i)$ and $\mathbf{t}_i = f_t(T_i)$ denote the normalized embeddings of image I_i and text T_i. The similarity between image i and text j is computed as

$$s_{ij} = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\tau}, \tag{2.1}$$

where τ is a learnable temperature parameter. The loss is defined as the average of two cross-entropy terms, one treating images as queries over texts and the other treating texts as queries over images.

This formulation produces a globally aligned embedding space that supports:

- **Zero-shot classification**, by comparing image embeddings with textual class descriptions;
- **Image–text retrieval**, by ranking candidates according to cosine similarity;
- **Semantic grounding**, by mapping natural-language queries to visual regions or objects.

CLIP plays a central role in this thesis as:

- a grounding mechanism for language in vision-centric tasks,
- a feature extractor for evaluation metrics,
- a text encoder for diffusion-based generation pipelines.

However, the dual-encoder design imposes limitations. Since image and text representations are computed independently at inference time, the model lacks explicit cross-modal reasoning. Moreover, the global contrastive objective encourages coarse semantic alignment and can struggle with compositional structure, localized attributes, and abstract language [11, 10]. These shortcomings directly motivate the development of methods for abstract-language handling and localized evaluation in Part II.

2.3 Diffusion Models for Conditional Image Generation

Diffusion models have emerged as a dominant paradigm for high-quality image synthesis. They learn a generative process by reversing a predefined noising trajectory. In the forward process, Gaussian noise is gradually added to a data sample x_0 over T steps:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}\big), \tag{2.2}$$

resulting in a noisy sample x_T that approximates a standard normal distribution. A neural network ϵ_θ is trained to predict the noise component at each timestep, enabling the reverse denoising process:

$$p_\theta(x_{t-1} \mid x_t, c), \tag{2.3}$$

where c denotes an optional conditioning signal.

Latent Diffusion Models (LDMs) [15] perform diffusion in the latent space of a pre-trained variational autoencoder (VAE), reducing computational cost while maintaining image fidelity. The denoising network is typically a U-Net equipped with cross-attention layers that inject conditioning information, most commonly text embeddings produced by a frozen encoder such as CLIP.

2.3.1 Text Conditioning and Guidance

Text-to-image generation is achieved by conditioning the denoising process on textual embeddings. Classifier-free guidance interpolates between conditional and unconditional predictions to control the strength of the conditioning signal, enabling a trade-off between image diversity and prompt fidelity.

In addition to global text prompts, modern diffusion models support structured conditioning signals, including:

- spatial masks for inpainting,
- edge maps or sketches for structural guidance,
- region-specific prompts for localized control.

These mechanisms enable fine-grained manipulation of generated content and form the basis for the controllable generation strategies explored in Chapters 4–8. In particular, adapter-based conditioning allows multiple modalities (*e.g.*, sketches and text) to be integrated without retraining the full model.

2.3.2 Diffusion for Data Augmentation

Beyond image synthesis, diffusion models can be used for data augmentation by generating realistic variations of existing samples. This is particularly valuable in data-scarce domains such as industrial inspection, where collecting labeled anomalies is costly. Conditioning on textual descriptions of defects enables semantic control over the generated samples, introducing language as a mechanism for guiding visual data generation.

2.4 Parameter-Efficient Adaptation of Large Models

State-of-the-art VLMs and diffusion models are typically trained on internet-scale datasets and contain hundreds of millions to billions of parameters. Full fine-tuning is often infeasible due to memory and computational constraints. Parameter-efficient techniques address this challenge by adapting only a small subset of parameters while keeping the backbone frozen.

2.4.1 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [20] introduces trainable low-rank matrices that approximate weight updates:

$$W' = W + BA, \tag{2.4}$$

where A and B are low-rank matrices. This approach drastically reduces the number of trainable parameters and memory requirements while preserving most of the performance gains of full fine-tuning. LoRA is widely used for adapting diffusion models to new domains or conditioning signals.

2.4.2 Adapter-Based Conditioning

Adapter modules are lightweight neural components inserted into intermediate layers of a pre-trained network. In diffusion models, adapters can be used to inject additional conditioning signals, such as sketches or region-specific embeddings, without modifying the original weights. This strategy enables modular control and facilitates multi-modal conditioning, as explored in the localized generation framework proposed in Chapter 8.

2.4.3 Training-Free Approaches

An alternative to parameter-efficient training is *training-free* adaptation, where the model parameters remain frozen and only the input embeddings or conditioning signals are modified. This paradigm is particularly useful when computational resources are limited or when generalization across multiple models is desired. The Abstract-to-Concrete Translator (Chapter 6) follows this approach by transforming textual embeddings to better align with the visual representation space.

2.5 Summary

The methods presented in this thesis are built upon three core components:

1. contrastive vision–language models for shared multimodal representations and grounding,
2. diffusion-based generative models for controllable image synthesis and data augmentation,
3. parameter-efficient and training-free adaptation strategies for operating under computational constraints.

Together, these foundations enable the integration of language into vision-centric tasks (Part I) and the enhancement of language usability within multimodal models (Part II), supporting grounding, retrieval, evaluation, and controllable generation within a unified multimodal framework.

Part I

Introducing Language into Vision tasks

Introduction. The first half of this thesis investigates how linguistic information can be embedded into visual pipelines to enhance perception, understanding, and decision–making. Instead of treating language merely as an output or annotation, we examine how it can serve as an active conditioning signal for visual tasks in both interactive and industrial settings. Beginning with human–robot navigation in complex environments (Le-RNR-Map, Chapter 3) and extending to surface–defect detection in industrial quality control (In&Out and DIAG, Chapters 4–5), these works demonstrate how even simple forms of language grounding can introduce semantic structure into high–dimensional vision problems. Natural language becomes not just a description of what is seen, but a tool to steer perception, express intent, and improve task performance.

Across these contributions, we find that language can serve as an additional modality for reasoning about visual scenes when used deliberately as a control interface or generative conditioning input. Prompting and instruction–based interaction make visual systems more interpretable and flexible; text–guided editing and augmentation strategies produce data distributions that better reflect real scenarios, improving generalization and training efficiency.

These findings will then motivate a shift in perspective that guides the second half of this thesis: if language can meaningfully enhance visual perception when used as guidance, then expanding the expressivity and interpretability of language within multimodal models may unlock even greater potential.

Chapter 3

Language-enhanced RNR-Map: Querying Renderable Neural Radiance Field maps with natural language

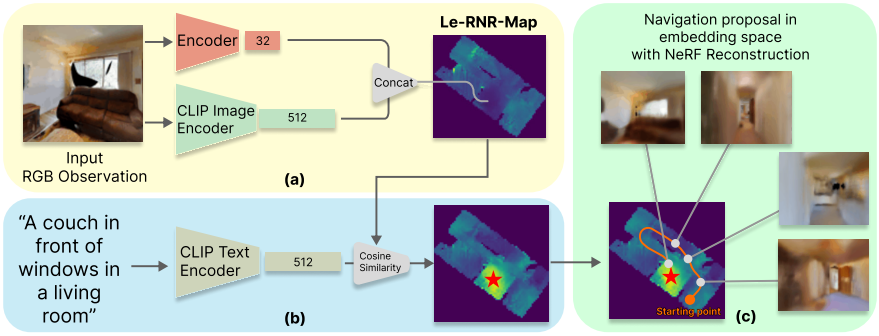


Figure 3.1: Overview of Le-RNR-Map. (a) shows Le-RNR-Map map construction in which we embed visual and visual-language-aligned features, (b) shows the query search, at inference time, with natural language, and (c) shows the reconstruction of the images on the path, from the starting position (orange circle) to the goal (red star) using NeRF.

3.1 Abstract

In this first work, we tackle the challenge of injecting language information in vision-based neural features to enable multimodal interaction between humans and navigation-based agents. Concretely, we present Le-RNR-Map, a Language-enhanced Renderable Neural Radiance map for Visual Navigation with natural language query prompts. In a previous work, RNR-Map presented a Neural Map containing Neural Radiance Fields of the environment of a building. They employed a grid structure comprising latent codes, positioned at each pixel, derived from image observation, enabling: *i)* image rendering given a camera pose, thanks to the Neural Radiance Field; *ii)* image navigation and localization with astonishing accuracy. On top of this, we enhance RNR-Map with CLIP-based embedding latent codes, allowing natural language search without additional label data. We evaluate the effectiveness of this map in single and multi-object searches. We also investigate its compatibility with a Large Language

Model as an "affordance query resolver".

3.2 Introduction

Embodied AI is receiving a lot of attention in recent years, with interesting yet very challenging tasks such as Embodied Question Answering [21], Image-Goal Navigation [22], Visual-Language Navigation [23], and zero-shot Object-Goal navigation [24, 25]. At the time, some areas of research focused on creating explicit map representations that could improve the performance on those tasks, such as semantic segmentation maps [26, 27], occupancy maps, and top-down semantic map prediction [28]. More recently, some works attempted to embed latent vectors in the explicit map to perform navigation [29]. However, very few tried to combine Visual-Language-aligned and NeRF [30] latent codes into the map itself to solve different tasks simultaneously.

In this preliminary work, we expand RNR-Map [31] by enhancing it with visual-language-aligned features. The result is a *Language enhanced, Renderable Neural Radiance Map* (Le-RNR-Map) for visual navigation that is *visually descriptive*, thanks to the Neural Radiance Field embeddings, *generalizable*, since it uses off-the-shelf models requiring no training, and queryable with *both text and images* thanks to the addition of visual-language-aligned features. To the best of our knowledge, this was the first work in the literature to create a map representation that allows to solve three distinct tasks *at the same time*: *i)* Localization of objects given a query image, by using the RNR-Map [31] embeddings; *ii)* Open Vocabulary localization of objects through natural language, using Visual-Language-aligned embeddings; *iii)* rendering of the path to the target, highlighting the searched object with Visual-Language-aligned features, without the need to physically move the agent.

3.3 Related Work

3.3.1 Maps For Navigation

There is an extensive research area that focuses on how to build maps to aid navigation in indoor environments [27, 29, 31–34]. An occupancy grid map is a $m \in \mathbb{R}^{M \times M \times C}$ matrix, where $M \times M$ is the spatial size and C is the number of channels storing information about a corresponding region. Originally, the authors of RNR-Map [31] introduced a novel type of map, in which the latent codes are embedded from visual observations and can be converted to a neural radiance field, which enables image rendering given a camera pose, thus being visually descriptive. Moreover, the author showed that this novel type of map can be useful for visual localization and navigation.

In AutoNeRF [27], the authors introduce a method to collect data required to train NeRFs using autonomous embodied agents, and use the experience to build an implicit map representation of the environment. Moreover, they augment the NeRF rendering procedure with a segmentation head over S predefined classes. In VLMaps [29], the authors ground language information to visual observation fusing pre-trained visual language features [35] into a 2D spatial representation. A similar approach is presented in [36]. In contrast to [29],

Le-RNR-Map also allows us to perform Image-Goal Navigation, and NeRF rendering given a camera pose. Additionally, Le-RNR-Map can be built faster, in around 60 seconds, for $\sim$1000 RGB-D images (128×128).

3.3.2 Nerf

Neural Radiance Fields, introduced in [30], address the problem of view synthesis, that is generating scene views from unseen novel viewpoints. NeRF scenes are modelled by a multilayer perceptron network which outputs the radiance emitted by a 3D point, given a spatial location (x, y, z) and a viewing direction (θ, ϕ). The original NeRF formulation [30] can only represent small scenes, and does not generalize to new scenes/objects. To solve these limitations, other works take the challenge of learning a distribution over complex scenes. Generative Scene Networks (GSNs) [37] can be used to learn a rich scene prior in order to generate new scenes or fill the given one, decomposing the scene in local radiance fields that can be rendered from a moving camera. Moreover, they can be used to render images from latent codes, which in our case are stored in Le-RNR-Map, like the original RNR-Map formulation [31]. Some recent works have extended the principle goal of NeRF with some other tasks. LERF [38] proposed a method for grounding CLIP representations in a dense, multi-scale 3D field, which can render dense relevancy maps given textual queries. However, LERF is still limited to small scenes and requires 45 minutes for a capture.

3.3.3 Language and Vision

Mapping text and images is the problem of estimating a function that maps images to the desired text (*e.g.*, captioning [39]) and vice versa (*e.g.*, image generation with text-conditioning [15, 40]). Introduced in [12], CLIP is a model composed of an image-encoder and text-encoder trained to map embeddings from images and their description close to each other in the feature space, with a contrastive loss that enforces non-related pairs to be mapped further from each other. The authors showed with extensive experiments that CLIP achieves competitive zero-shot performance and thus can be used as a foundational model in a variety of task, such as scene segmentation [35, 41], Open-Vocabulary object detection [42] and Image-Generation [15]. Moreover, [43] introduces MaskCLIP, a framework for obtaining scene segmentation with indirect supervision from language.

3.4 Method

3.4.1 Map creation

We create a Le-RNR-Map by first extracting visual and Visual-Language-aligned features from RGB frames, while a subsequent feature registration process projects them in the map (Fig. 3.1a).

Visual embeddings. Inspired by [31] we consider a robot agent exploring the scene with a random walk, using RGB-D data and its on-board sensors, *i.e.* odometry information, to build Le-RNR-Map. An encoder-decoder architecture performs the creation of the RNR-Map.

To allow an effective encoding of the 3D environment, the authors of [31] perform training of the encoder-decoder as follows: *i)* the encoder takes in input the RGB-D image and extracts the pixel features; *ii)* the decoder uses pixel features, along with the current pose (*i.e.* position of the agent), to sample latent information along each camera ray corresponding to each pixel, and tries to render the corresponding images. The latent codes extracted from the encoder, then, represent the pixel-level visual information from the current view. In our experiments, the encoding features are $F_{rnr} \in \mathbb{R}^{32}$. These features allow image rendering using Neural Radiance Field, and image localization in the map. For a more in-depth description, we refer the reader to [31].

Natural language. To include language features, we use a pre-trained CLIP [12] image-encoder to get $F_{clip} \in \mathbb{R}^{512}$ from the current RGB-D frame.

Le-RNR-Map. The final embedding space for the current observation RGB-D frame is then composed as $F_{le-rnr} = F_{clip} \oplus F_{rnr}$. $F_{le-rnr} \in \mathbb{R}^{544}$ with the first 32 channels generated from RNR-Map and the 512 remaining from CLIP. Both the features from RNR and CLIP are then projected to the 2D map using the depth information, as in [31]. This allows us to keep the exact performances of RNR-Map for the Image-Goal navigation task, with the addition of being able to query the navigation through natural language thanks to the CLIP features.

3.4.2 Language-vision object search

The vision-language object search is performed by providing a natural language query (Fig. 3.1b). This query should indicate the objects required to find (either big furniture or small objects) that the navigation module has to handle as goal objects. Once the query is provided, the text embeddings are extracted with a pre-trained CLIP [12] text-encoder, and the cosine similarity with each cell of the Le-RNR-Map is computed. We select the location of maximal similarity as the goal location. The 3D end-goal predicted location, given the (x, y) indices expressed in Le-RNR-Map coordinate system, is obtained through inverse projection as in [31]. To ensure the correct maximal similarity is found, prompt engineering is performed with negative prompting following [12, 38]. Together with the query prompt, this empirically shows a more fine-grained similarity on the maps and gives better localization as shown in Fig. 3.3. An extension of GSN [37] is then used to synthesize novel views and render a possible pathway that leads from any point starting point of the map to the required goal. Once the navigation reaches the proximity of the goal (*i.e.* the location of maximal similarity in the Le-RNR-Map), we estimate the camera orientation towards the goal by rotating the camera by 360°, computing the CLIP features for each degree and choosing the one with maximal cosine similarity with the target query. Finally, the visual saliency of the goal (Fig. 3.2c) is extracted from the RGB reconstructed by GSN [37] using pixel-wise CLIP features [43]. When multiple target objects are requested, the navigation is performed sequentially for each object following the same procedure, using as starting location the previous target location.

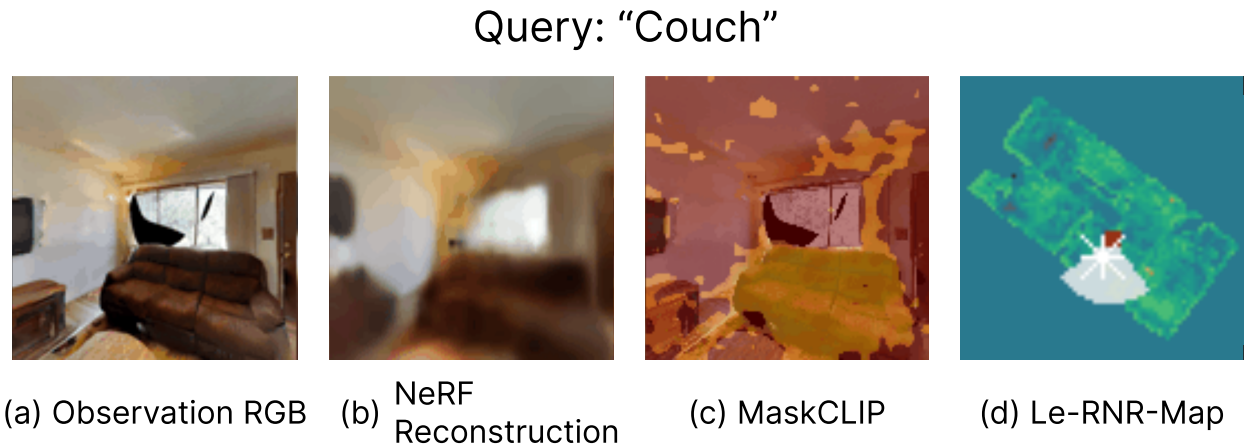


Figure 3.2: (a) Observation from Habitat-sim [44]. (b) Reconstruction using latent code from Le-RNR-Map using Neural Radiance Field (c) Feature visualization of the text query using MaskCLIP [43]. (d) Top-down view of Le-RNR-Map.

3.5 Experiments

To highlight the benefits that Le-RNR-Map brings to the Object-Goal Navigation task, we designed a set of targeted experiments. In Sec. 3.5.1 we evaluate the ability of Le-RNR-Map to correctly locate items using only Natural Language prompts. In Sec. 3.5.2 we show that the Renderable Neural Radiance map allows the user to properly select the correct item of interest in case of ambiguity. Finally, in Sec. 3.5.3, we explore the possible collaboration between a Large Language Model and Le-RNR-Map to find items and locations based on contextual prompts, called *Affordance queries* (*e.g.* the query *“Find me a drink to wake me up”* results in the agent looking for a cup of coffee). All of our experiments were conducted inside the Habitat-Sim [44] using the Gibson [45] dataset.

3.5.1 Searching items by Language prompts

The goal of Le-RNR-Map is to provide the user with a Natural Language interface with the agent. Such an interface would allow the user to prompt the agent with low effort, even in a constrained environment where traditional interactions may be unavailable (*e.g.* the user is holding something and can’t physically interact with the agent) or improbable (*e.g.* asking an agent to look for a particular object by showing it a picture of the object itself). With this goal in mind, we test the ability of Le-RNR-Map to locate different common items and/or locations in the environment, using only the CLIP features embedded in the navigation map.

In Tab. 3.1 we report the Success Rate and Distance To Success (DTS), as defined in [26], on some scenes of the validation split of the Gibson tiny dataset [45]. The dataset provides a textual label for the target object and its location in the scene as ground truth. We use the label as a text prompt and compute the metrics by comparing our predicted location with the ground truth. Additionally, as explained in Sec. 3.4.2, we define a series of unwanted objects or general/background elements (*e.g.* `stuff, wall, floor`) as negative prompts. These

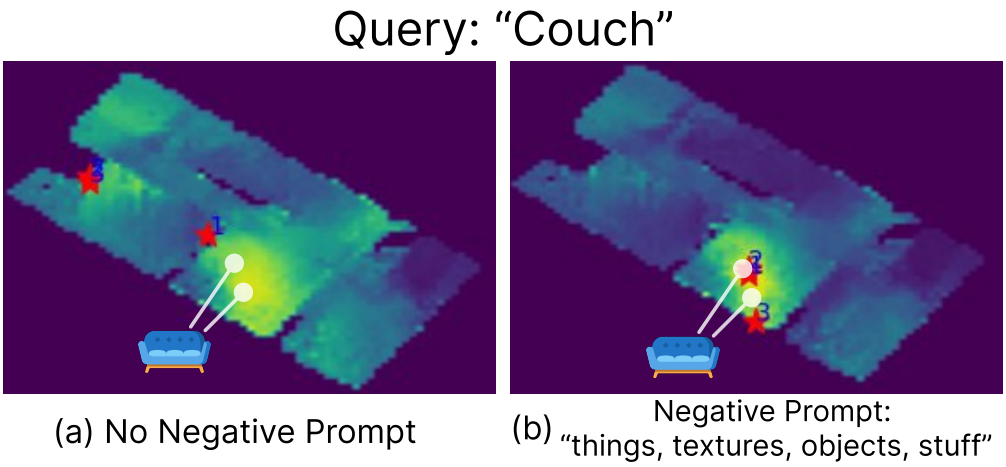


Figure 3.3: Similarity heatmap between the prompt `Couch` and Le-RNR-Map. (a) is without negative prompts. (b) shows that negative prompts result in cleaner maps with more concentrated similarity zones. The stars indicate the maximum similarity locations.

prompts may be specific to each scene. Together with the target prompt, we compare their similarity results with the map embeddings (as seen in [12]) resulting in similarity maps with more consistent areas of interest as shown in Fig. 3.3. In general, Le-RNR-Map enables us to obtain a decent success rate and DTS without additional training required. Moreover, we investigate the low Success Rate for the *Darden* scene. We found that the observations, given to the agent during the map creation, contain several artifacts, such as mirror reflections, missing walls, and mesh holes, leading to incorrect embeddings into the map. Further study will analyze this problem.

Table 3.1: Results on the `val` split of Gibson tiny dataset. Note that our setup is *known* since we generate the map beforehand. Each scene has a different negative prompt.

Scene name	**Negative Prompts**	**Success ↑**	**DTS (m) ↓**
Corozal	✗	0.69	1.19
	✓	0.69	**0.65**
Darden	✗	0.37	4.12
	✓	**0.59**	**2.59**
Markleeville	✗	0.81	**1.38**
	✓	**0.83**	1.50
Wiconisco	✗	0.76	0.50
	✓	0.76	0.50
Average	✗	0.66	1.79
	✓	**0.72**	**1.31**

3.5.2 Solving prompt ambiguities

One of the advantages of having a Renderable Neural Map is the possibility for the agent *to explore the scene without actually moving in the real world.* This is particularly useful in scenarios where the user prompt may be ambiguous and refer to multiple objects or locations in the scene (*e.g.* `window` may refer to different windows). When this happens, the RNR-Map can be used to provide the user with visual previews of the paths it would take to get to all the possible solutions. In this section, we explore this case and provide a solution using Le-RNR-Map.

First, as in Sec. 3.5.1, we compute similarities between the CLIP features extracted from the prompt and the ones embedded in Le-RNR-Map by also using negative prompts. After finding the maximum similarity location, we suppress all the similarities in the adjacent cells of the map. We then look for the new maximum similarity above a threshold $th = 0.6$. For each target found, we render the path that the agent would follow to reach the target using a shortest path algorithm. The process ends when there is no longer a similarity score greater than th. We consider it a success when this process finds the target item in at least one of the N predicted paths, simulating the user "selecting" the desired item. While this evaluation is heavily reliant on hyper-parameters, such as the number N of predicted paths allowed and the value T of the threshold, we still believe it to be an interesting use case, and propose our work as a first informal approach to this problem.

3.5.3 Affordance search

We argue that Natural Language alone is not enough to achieve natural interaction with the user and the agent, especially if we restrict the user to a limited set of words (classes) or a rigid sentence structure. The idea comes from the following observation: what if we want to search for some specific location of an indoor environment, but we are unable to express the query in a direct way? As an example, consider the scenario where we want to search for a location "*that can be relaxing after a long day at work*". We define this use case as "affordance search", and propose to take advantage of Large language model to translate the query and output a set of possible target descriptions, using the GPT-3.5 chat-completions API available at the time. After retrieving the descriptions, we sequentially search for each target with the procedure presented in Sec. 3.4.2.

3.6 Conclusions & Future works

In this work, we introduced natural language as a functional layer on top of the RNR-Map framework [31], enabling semantic search and navigation within a visual map that was originally purely perception-driven. Using an off-the-shelf language model, we demonstrated that linguistic queries, ranging from single-object prompts to multi-object specifications, can be grounded into the map's latent representation, allowing the system to identify targets and generate visualizations of the shortest paths through the underlying Neural Radiance Field. We further showed that large language models can act as an effective "affordance query resolver," translating abstract user intentions into spatially meaningful navigation commands.

Although preliminary, these results illustrate a key theme of this thesis: language can enrich visual systems with semantic structure, interpretability, and more human-aligned interaction mechanisms. Several promising directions follow from this work. Future efforts could enhance rendering fidelity through a language-driven grounding head for NeRFs, explore end-to-end agents capable of zero-shot object-goal navigation, and investigate how online map generation can support improved generalization. Additionally, we believe updating this methodology with recent advancements in the Neural Radiance Field, as well as mechanisms for updating both the radiance field and linguistic embeddings in dynamic environments, is a promising and relevant direction.

By introducing language into the navigation pipeline, this chapter lays the foundation for the next one, where we explore how linguistic knowledge can similarly enhance visual intelligence in industrial anomaly detection.

Chapter 4

Diffusion-based Image Generation for In-distribution Data Augmentation in Surface Defect Detection

4.1 Abstract

Following our exploration of language-guided navigation in Le-RNR-Map (Chap. 3), this work shifts the focus to industrial visual inspection and examines how modern generative models can strengthen surface-defect detection pipelines. A persistent challenge in this domain is the severe imbalance between normal samples and defective ones: defect classifiers are typically trained with abundant negative data but only a small number of positive examples. Existing augmentation strategies attempt to compensate by superimposing artificial defects onto normal samples, but these overlays are often unrealistic and out of distribution, encouraging models to recognize artifacts rather than the true appearance of defects.

In this study, we demonstrate that diffusion models provide a powerful alternative for generating in-distribution synthetic defects, producing realistic variations that help classifiers learn the genuine visual structure of anomalies. We introduce a hybrid augmentation strategy, *In&Out*, which combines traditional out-of-distribution artifacts with diffusion-generated in-distribution defects. The method operates effectively across multiple regimes: zero-shot, where no real defect samples are available; few-shot, where only a small number are provided; and full-shot, where defects are abundant.

We evaluate our approach on the most challenging benchmark to date, the Kolektor Surface-Defect Dataset 2, and establish a new state-of-the-art weakly supervised AP score of 0.782.

4.2 Introduction

Building on the previous chapter, which explored how modern generative and multimodal models can enhance robot perception and navigation, we now turn to a different but equally demanding application domain: industrial visual inspection. Surface defect detection is a long-standing and challenging problem in manufacturing, defined as the task of identifying

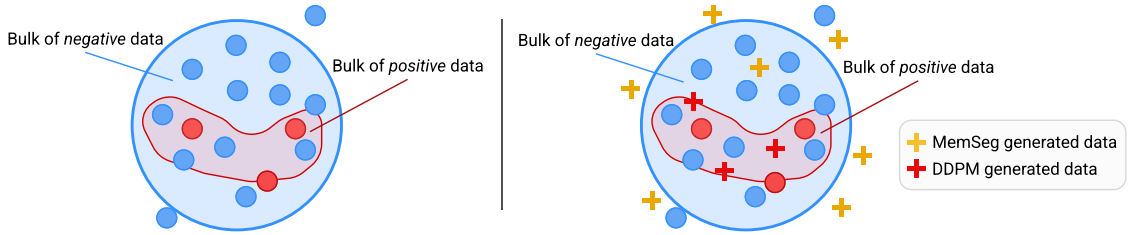


Figure 4.1: Idea underlying our *In&Out* data augmentation approach. (*Left*, blue dots) The blue dots outside the bulk of negative data could be wrongly classified as anomalies (false positives), being slightly different from most of the negative data. (*Right*, yellow crosses) State-of-the-art per-region data augmentation methods (for example, MemSeg [46]) add positive synthetic samples in that zone, which helps in deciding what is certainly not anomalous data. (*Left*, red dots) On the other hand, the red dot partially outside the bulk of positive data could be, in principle, understood as a negative sample, leading to a false negative. (*Right*, red crosses) Diffusion-based generated data is capable of producing defects very similar to the ones in the bulk of positive data, helping the classifier not produce false negative classifications.

samples that contain a defect [47]. The most direct solution relies on human experts visually inspecting each product and removing defective items. However, this approach is costly, inconsistent, and highly sensitive to human fatigue and bias.

Automated defect detection systems [48, 49] address these limitations by learning classifiers from collections of normal (negative) and defective (positive) samples. Yet acquiring such datasets is difficult in practice: labeling requires significant human effort, and defects naturally occur far less frequently than normal samples. As a result, the positive class is severely underrepresented, and the resulting imbalanced training distribution strongly limits downstream performance.

To compensate for this imbalance, various data augmentation techniques have been proposed [46, 50, 51]. The prevailing strategy consists of per-region augmentation [46], in which artificial texture artifacts are overlaid onto normal images to produce synthetic defect samples. While this out-of-distribution augmentation helps reduce false positives by teaching the model what clearly "does not look normal," it fails to capture the subtle, fine-grained deviations that characterize real defects. Consequently, it provides limited improvements in recall, which remains a primary challenge in surface anomaly detection.

Recent progress in generative modeling, particularly diffusion models [15, 52], presents a promising alternative. Their ability to sample rich latent spaces enables the synthesis of realistic, high-fidelity imagery, making them well-suited for producing in-distribution anomalous samples that more accurately reflect the appearance of genuine defects. In this work, we explore the use of Denoising Diffusion Probabilistic Models (DDPMs) to generate such fine-grained synthetic positives.

We consider two practical augmentation scenarios: *i)* zero-shot data augmentation, where no defect samples are available; *ii)* few-shot or full-shot augmentation, where a limited or large number of defects exist.

In the zero-shot case, language becomes the only information available for defect generation:

starting from only negative (healthy) samples, we adopt a human-in-the-loop paradigm in which operators guide the generation process of positive (defective) samples using short textual descriptions that reflect their domain knowledge (*e.g.*, "scratches," "holes"). On the other hand, when defect samples are available (*i.e.*, few-shot or full-shot scenario), fine-tuning can be performed directly on them, eliminating the need for human input and allowing the diffusion model to learn the underlying distribution of defects.

A central contribution of this chapter is the observation that diffusion-based in-distribution defects and traditional out-of-distribution overlays are highly complementary. The former enriches the true structure of the positive class, improving recall, while the latter reinforces boundary cases that help avoid false positives. We therefore propose a combined augmentation strategy, *In&Out*, which integrates both forms of synthetic data to maximize their strengths. As illustrated in Figure 4.1, this fusion leads to more robust classifiers by balancing realism and diversity in the augmented dataset.

We evaluate our approach on the challenging Kolektor Surface-Defect Dataset 2 (KSDD2) [53], achieving a new state-of-the-art weakly supervised AP score of .782 using only 120 augmented samples.

This chapter represents an initial step toward integrating language into industrial visual pipelines. Here, language is introduced in a targeted setting (zero-shot data augmentation) where textual descriptions provide expert semantic knowledge about expected defects and compensate for the scarcity of annotated visual data. This exploratory study establishes the conceptual and methodological foundations for the subsequent DIAG chapter (Chapter 5), in which language becomes the primary driver of the augmentation process.

More broadly, the chapter demonstrates how generative models, when guided by structured textual input, can extend beyond natural-image domains to support perception in data-constrained manufacturing environments. By embedding domain knowledge through language, we show that multimodal conditioning can improve both the diversity and the relevance of synthetic training data. These results foreshadow the central theme of the thesis: language is not only a descriptive modality but also a mechanism for control, generalization, and knowledge injection. In this sense, the present work lays the groundwork for the deeper exploration of multimodal guidance, controllability, and semantic alignment developed in the chapters that follow.

4.3 Related Work

At the time of this work, one of the most adopted frameworks for automated quality control was defect detection, where the goal is to find images that contain defects. Specifically, we focused on weakly supervised approaches [53, 54], in which positive and negative training images were labeled at the image level, that is, without per-pixel masks. This represented the cheapest and most widely used annotation approach in industrial contexts. Despite its importance and wide usage, the practice of data augmentation for defect detection had received little attention in the literature at that time, and our work was one of the first to focus on it entirely. The most commonly adopted pipeline for generating anomalous synthetic samples consisted of a series of random standard augmentations on the input image, such as mirror symmetry, rotation, brightness, saturation, and hue changes, followed by the

super-imposition of noisy patches on the image [46, 55]. Notably, in [50], an ablation study focused on the generation of synthetic anomalies led to the following findings: *i)* adding synthetic noise images was never counterproductive, though it could diminish effectiveness in percentage terms; *ii)* few generated anomaly images (in the order of tens) were sufficient to increase performance substantially; *iii)* textural injection in the anomalies was important, or equivalently, adding uniformly colored patches was not effective. In all of these contemporary works, it was evident that the synthetically generated images were simply out-of-distribution patterns, which did not have to represent the target-domain anomalies faithfully. We improved upon this setup by being the first to focus on genuine in-distribution defect data. A modest improvement had been made in [51], where the authors introduced the concept of "extended anomalies," in which the specific anomalous regions of the seen anomalies were placed at any possible position within the normal sample after applying random spatial transformations. Unfortunately, this approach required segmenting the training data, which we sought to avoid.

4.4 Background

We organize this section into four different parts, each one providing an overview of a topic related to our work: *i)* DDPMs; *ii)* Dreambooth fine-tuning; *iii)* Low-Rank Adaptation (LoRA), and *iv)* per-region data augmentation.

Denoising Diffusion Probabilistic Models. DDPMs are probabilistic models inspired by the non-equilibrium statistical physics phenomenon of diffusion [56, 57]. In recent years, diffusion models have gradually become state-of-the-art in image synthesis, surpassing GANs in performance [52]. One of the main advantages of such models is the ability to guide the sampling steps with additional input data with a technique called conditioning. The most common form of conditioning is a text that describes what the expected image should look like [15]. However, recent developments have explored other forms of conditioning, such as images, segmentation maps, or logic formulas [2].

Dreambooth fine-tuning. Dreambooth [58] is a procedure for DDPMs that allows fine-tuning the model with a small number N of images. During the fine-tuning steps, each of the N images is associated with a prompt defining the identification token and the subject class. At the same time, regularization images (images of the same class but without the subject identification token) are used to prevent the fine-tuning model from forgetting the subject class learned during the original (non-fine-tuning) training, thanks to a prior preservation loss. This allows the DDPM to learn a new specialized concept, represented by the identification token, with fewer iterations and without overwriting its prior knowledge.

Low-Rank Adaptation (LoRA). In recent years, fine-tuning Large Language Models (LLMs) has become prohibitively expensive due to the huge number of parameters. In [20], the authors introduced Low-Rank Adaptation (LoRA), a model-agnostic method of fine-tuning models in an efficient way. LoRA has the following advantages: *i)* many small LoRA modules

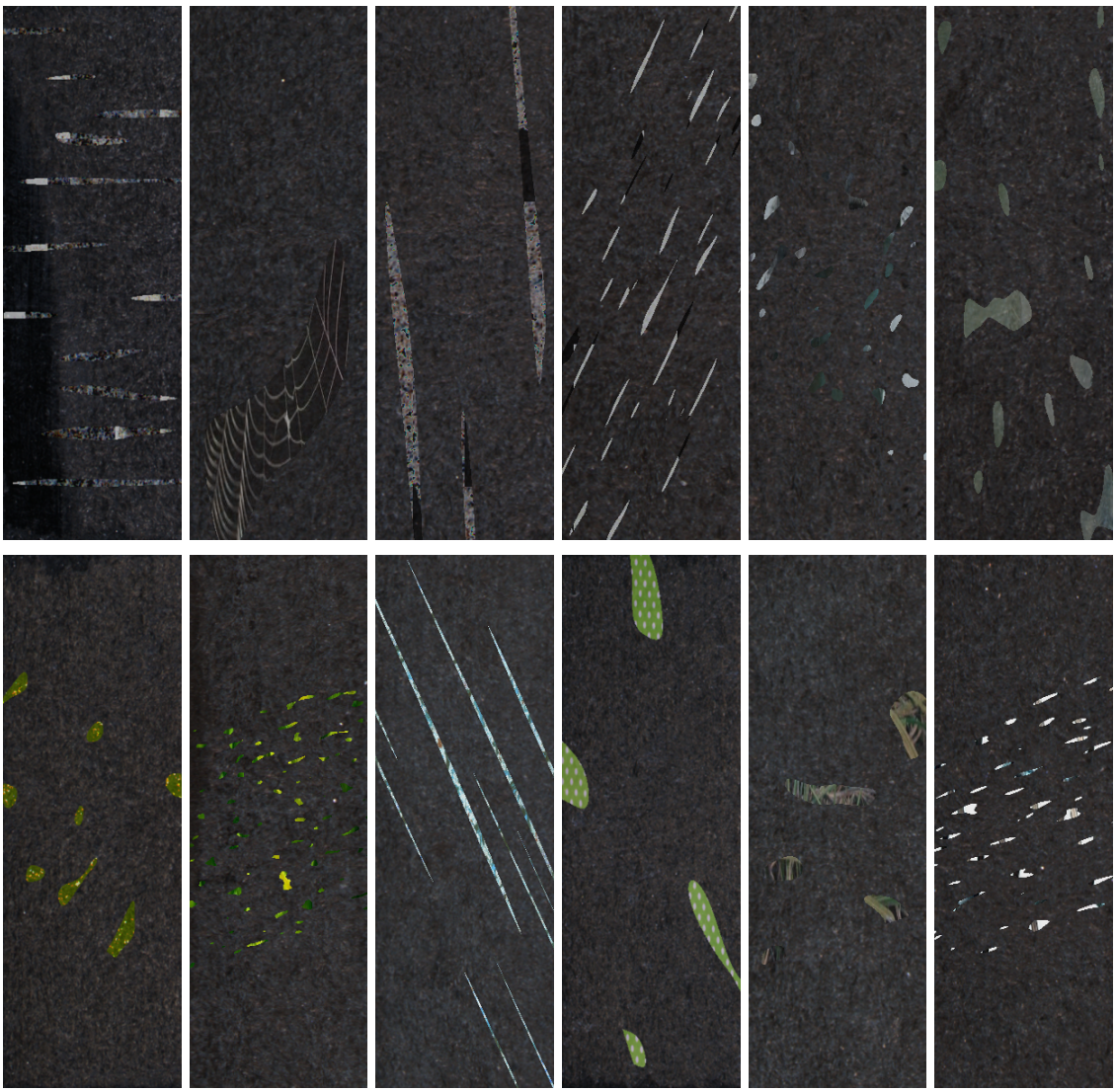


Figure 4.2: Augmented images generated by the MemSeg [46] pipeline. It is evident how it provides out-of-distribution positive samples.

for different tasks can be built by a single pre-trained model; *ii)* optimizes only the injected, much smaller low-rank matrices, lowering the hardware requirements barrier; *iii)* the final model, obtained by merging the original pre-trained model and the low-rank matrices, has no additional inference latency.

Per-region data augmentation. With *per-region* data augmentation, we refer to out-of-distribution data augmentation procedures that superimpose noise regions on the original image. In our study, we will use MemSeg [46] as our out-of-distribution data augmentation. Some examples of images generated by the MemSeg pipeline are reported in Figure 4.2.

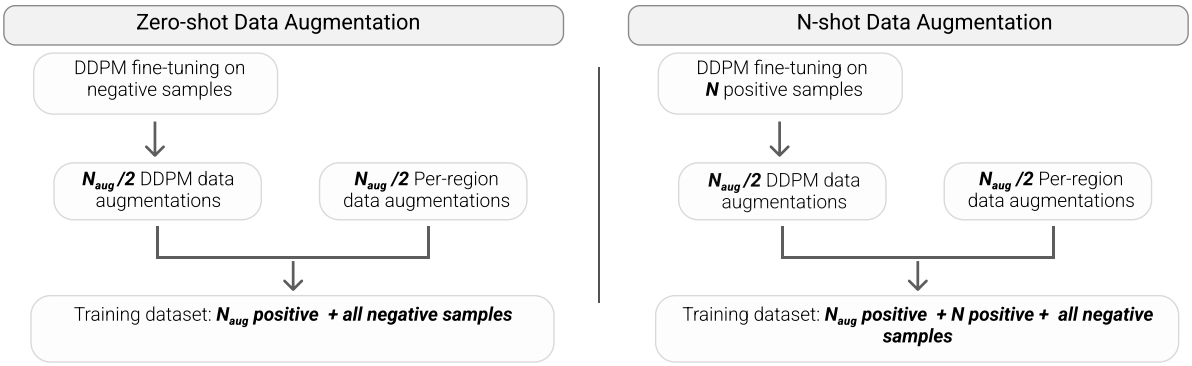


Figure 4.3: General schema of our *In&Out* method.

4.5 Method

The *In&Out* data augmentation aims at producing N_{aug} additional positive images. The approach can be applied, with slightly different pipelines, on two scenarios: *i)* when no positive samples are available (zero-shot data augmentation) and *ii)* when positive samples are available (N-shot data augmentation, where N can be small or large). In the following, the two pipelines are detailed; a graphical sketch is presented in Figure 4.3.

4.5.1 Zero-shot data augmentation

In this scenario, we simulate that no positive samples are available in the training set. Thus, our aim is a zero-shot data augmentation procedure in which two steps are performed: fine-tuning and data augmentation.

Fine-tuning step. Dreambooth is adopted to perform fine-tuning on a DDPM. To reduce training time and lower computation requirements, we only train low-rank update matrices by employing LoRA. These update matrices are then summed to the original weights, completing the fine-tuning procedure. Specifically, we control the weight of the LoRA update matrices during the merge with a parameter α: a value close to 0 results in no fine-tuning, while a value close to 1 results in the strongest fine-tuning.

In the zero-shot data augmentation, we perform fine-tuning with a portion of randomly chosen negative samples from the training set. The number of samples depends on the complexity of the data we want to manipulate: the larger the intra-class variance, the larger the number of elements to sample. In this preliminary study, we select the number of samples heuristically (see Section 4.6 for details).

Data augmentation. In this step, we create the N_{aug} augmented images generating $N_{aug}/2$ in-distribution images and $N_{aug}/2$ out-of-distribution images. The $N_{aug}/2$ in-distribution images are obtained by exploiting the fine-tuned DDPM through natural language prompts, describing the desired anomalies. To define the types of defects in natural language and verify how well text expressions are suited to generate a genuine defect for the data at hand, it is reasonable to perform some human-in-the-loop cycles, exploiting the expert's domain

knowledge to evaluate the augmentation quality. Specifically, the operator prompts textual expressions and evaluates the generated data (total of $N_{aug}/2$), certifying reasonable defects or revising expressions for improved generations. The $N_{aug}/2$ out-of-distribution images are obtained by the per-region data augmentation, detailed in Section 4.4.

This ensures that half of the augmented data will be in-distribution, describing the visual appearance of the defects (the diffusion-based one), while the other half of the data will focus on specifying what is certainly not a perfect sample (the per-patch images). After the augmentation, the final training dataset will be formed by N_{aug} augmented positive images plus all the original negative samples.

4.5.2 N-shot data augmentation

In this scenario, we assume to have N images from the positive pool of dataset images on which we perform Dreambooth fine-tuning with LoRA. We refer to the cases where $N \sim 5$ as few-shot data augmentation. After the fine-tuning, $N_{aug}/2$ in-distribution positive samples are generated. As for the zero-shot data augmentation scenario, the additional $N_{aug}/2$ out-of-distribution images are obtained by the per-region data augmentation, detailed in Section 4.4.

After the augmentation, the final training dataset will be formed by N_{aug} augmented positive images + N original positive images plus the negative samples.

4.6 Experiments

In this study, we explore the efficacy of our *In&Out* data augmentation approach for defect detection on the KSDD2 dataset.

Dataset. The KSDD2 contains RGB images of defective production items, provided and annotated by Kolektor Group d.o.o. The defects vary in shape, size, and color, ranging from small scratches and minor spots to large surface imperfections.

Since the images are of different sizes, we standardize the dataset resolution by center-cropping and resizing all the images to 200×600 pixels. The dataset is split into train and test subsets, with 2085 negative and 246 positive samples in the training set, and 894 negative and 110 positive samples in the test set. At the moment of writing, the state-of-the-art AP on this dataset stands at .733 [53]. We show several normal and anomalous samples in Figure 4.4.

4.6.1 Implementation details

In this section, we specify all the implementation details for the sake of reproducibility. All training and inferences have been carried out on an NVIDIA RTX 4090 GPU.

DDPM fine-tuning. In our experiments, we use Stable Diffusion [15] as DDPM. The fine-tuning process follows the Dreambooth procedure (see Section 4.4 for details). We used the prompt "`skt background`", where "`skt`" is the identification token. As written in

Figure 4.4: Normal (top row) and anomalous (bottom row) samples from the KSDD2 dataset. Note that some defects are very difficult to find.

Section 4.4, the string "`skt`" has no semantic meaning, and was selected to define an ID code for a new visual class. On the other hand, "`background`" is the subject class, identified as the most suited to obtain images with a homogeneous background. The regularization images have been generated using the prompt "`background`". The weight of the prior preservation loss is set to 1.0 as in the original paper. For faster training time and lower computation requirements, we also employ the LoRA-c3Lier low-rank adaptation, a modified version of LoRA that also applies low-rank approximations to 3×3 convolutional kernels and linear layers.

The code is implemented in PyTorch. We used AdamW8bit [59] as an optimizer, with a learning rate of $1e-5$. We kindly direct the reader's attention to our configuration file for a more comprehensive exploration of the various hyperparameters involved.

DDPM data augmentation. After training Stable Diffusion, we use it to generate $N_{aug}/2$ augmented images. In the zero-shot scenario, we use the prompts "`skt background cracked`" and "`skt background scratched`" to induce the generation of anomalous samples. These prompts have been chosen after a series of tests and result in images containing plausible anomalies like the ones shown in Figure 4.5. These generated images are then added to the training set, which will be used to train the anomaly detection model. We train and evaluate this model with four different seeds for each of our experiments, generating $N_{aug}/2$ new images each time to provide the most statistically relevant results.

ResNet-50 training and testing. We use the PyTorch implementation of the ResNet-50 [60] as our anomaly detection model, in which we substitute the fully connected layers after the backbone to make it a binary classifier. The network is trained for 50 epochs with

Figure 4.5: Anomalous samples generated by DDPM. It is evident how it provides in-distribution positive samples.

an SGD optimizer, a learning rate of 0.01, and a batch size of 5.

To keep consistency with the training and evaluation procedures of the KSDD2, we modify their official implementation to accommodate our ResNet-50 model. In particular, our setup is similar to the weakly supervised one presented in [53], where only the images and ground truth labels are used to train the model. For each scenario, i.e., zero-shot data augmentation and N-shot data augmentation, we will train three versions of our ResNet-50 model: *i)* using only MemSeg to generate N_{aug} images; *ii)* using only our DDPM to generate N_{aug} images; and *iii)* using *In&Out* as data augmentation, resulting in $N_{aug}/2$ images generated by MemSeg and $N_{aug}/2$ generated by our DDPM.

4.6.2 Zero-shot data augmentation

Table 4.1: Results between MemSeg and DDPM when ***no*** anomalous samples are available.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑	$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
MemSeg 80	.514 (.026)	**.733** (.113)	.436 (.033)	DDPM 80	**.547** (.086)	.427 (.301)	.695 (.194)
MemSeg 100	.388 (.066)	.633 (.129)	.432 (.054)	DDPM 100	.532 (.028)	.387 (.277)	**.714** (.286)
MemSeg 120	.511 (.050)	.683 (.054)	.470 (.091)	DDPM 120	.445 (.186)	.465 (.329)	.591 (.274)
Average	.471 (.047)	**.683** (.099)	.446 (.059)	Average	**.508** (.100)	.426 (.302)	**.667** (.251)

In these experiments, we emulate a situation where ***no*** positive samples are available in the training set. With this premise, we train our diffusion model with only 50 randomly chosen negative samples from the training set. We chose this number empirically and deemed

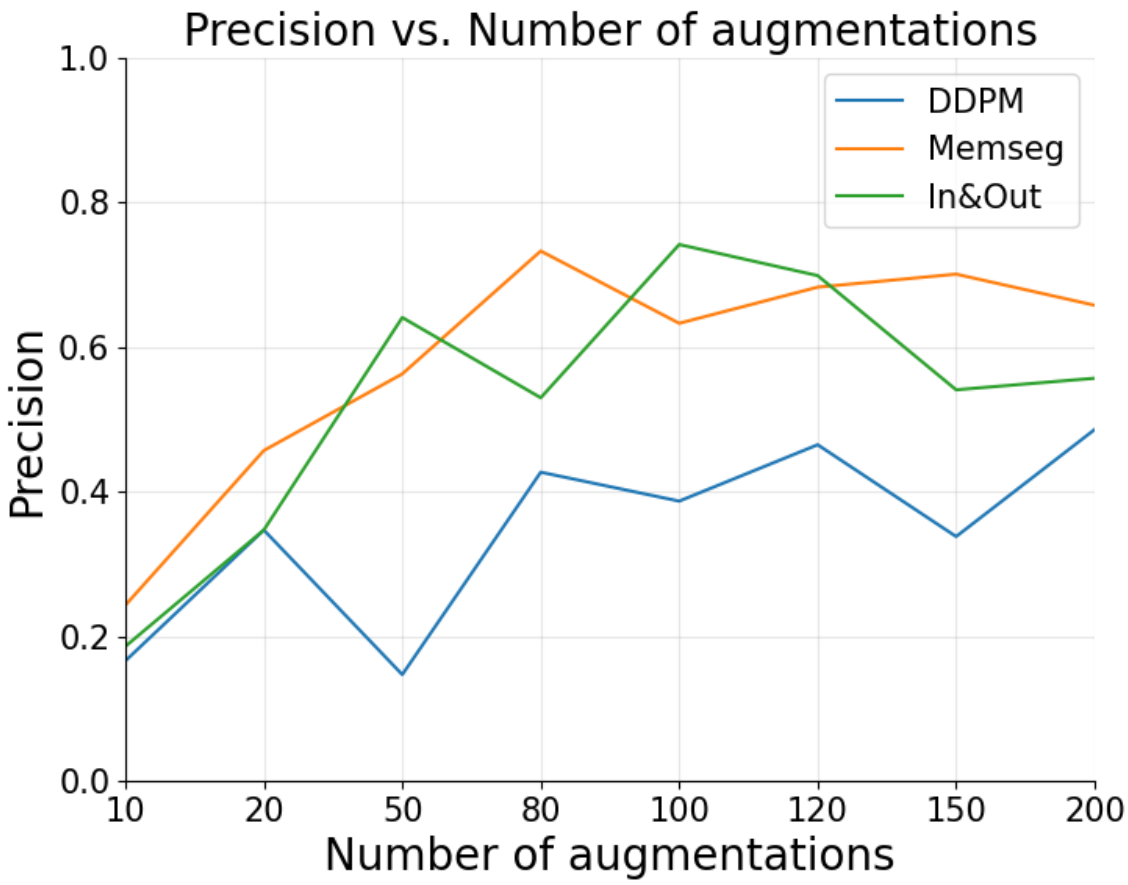


Figure 4.6: Precision of the methods as a function of the number of augmentations. Note that MemSeg has higher overall precision. *In&Out* balances this metric.

it sufficient to represent the intra-class variance of the negative samples. We train the DDPM for 5 epochs, using as guiding prompt "`skt background`" and $\alpha = 0.60$.

Once the diffusion model is trained, we generate $N_{aug}/2$ augmented positive samples using prompts specific to the dataset. In our case, we used prompts such as "`skt background cracked`" and "`skt background scratched`", resulting in images like the ones shown in Figure 4.5. Therefore, we produce $N_{aug}/2$ out-of-distribution images by MemSeg, obtaining the N_{aug} of our *In&Out* approach. We also experiment with fully-MemSeg and fully-DDPM augmentation pipelines for comparison.

We train the ResNet-50 model on different values of N_{aug} and evaluate it on the original test set. For each number of data augmentation, four different seeds have been used to report the most statistically relevant results. We report the comparison between MemSeg and DDPM in Table 4.1, where the numbers outside the parenthesis indicate the average results over the four seeds, while the numbers between parenthesis indicate the standard deviation. As we can see, DDPM achieves the highest AP (.547), recorded at 80 augmented images, while also resulting in an overall higher mean AP when compared to the MemSeg pipeline (.508 vs. .471).

We want to highlight the difference between the precision and recall scores of MemSeg

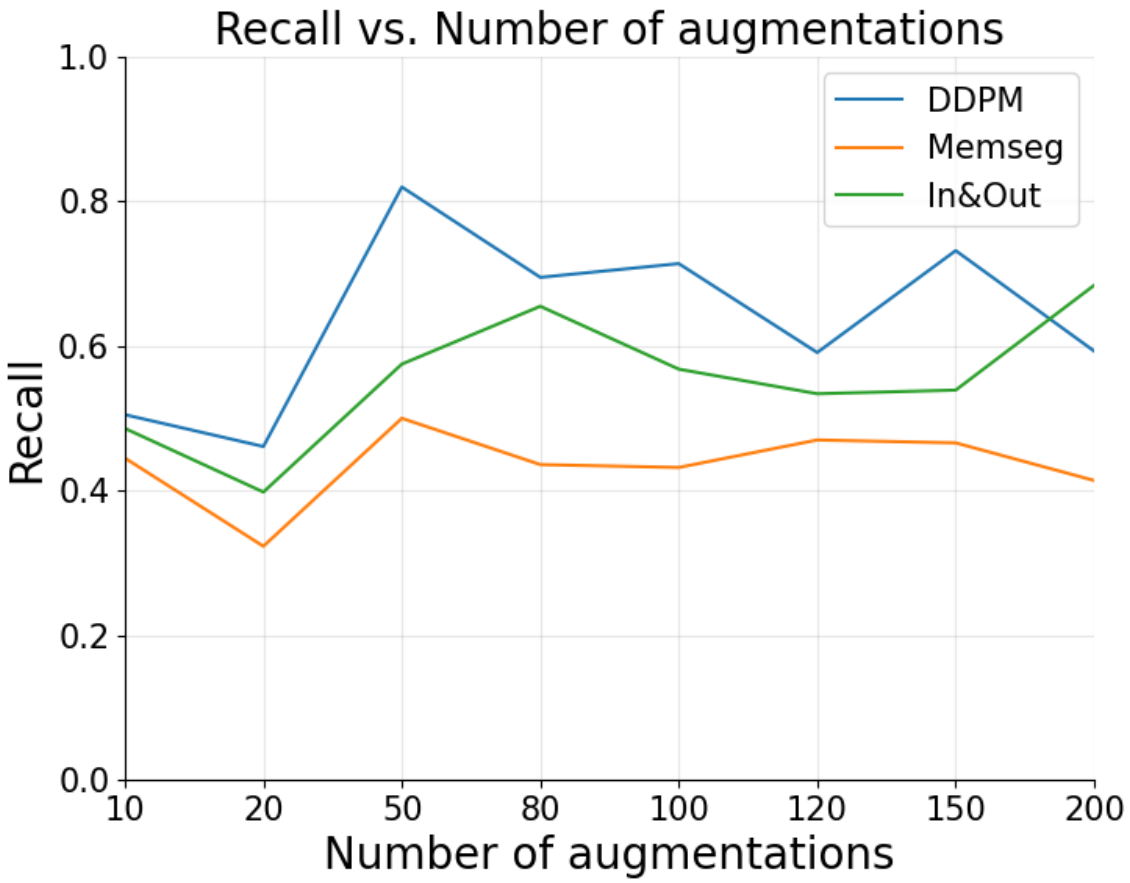


Figure 4.7: Recall of the methods as a function of the number of augmentations. Note that DDPM has a higher overall recall. *In&Out* balances this metric.

and DDPM. While DDPM achieves a higher recall (.714), the MemSeg pipeline results in a higher precision (.733). This behavior is clearly shown in Figure 4.6 and 4.7, where we plot the values of precision and recall of the two methods for different N_{aug}.

Table 4.2: Results when ***no*** anomalous samples are available using *In&Out*. Thus, $N_{aug}/2$ samples generated with DDPM and $N_{aug}/2$ with MemSeg.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
In&Out 80	.556 (.085)	.530 (.219)	**.655** (.065)
In&Out 100	**.626** (.059)	**.742** (.109)	.568 (.029)
In&Out 120	.536 (.023)	.699 (.085)	.534 (.086)
Average	**.573** (.056)	.657 (.138)	.586 (.060)

When combined in the *In&Out* pipeline, where half of the augmented positive samples are provided by DDPM and the other half is provided by MemSeg, we obtain a huge performance boost in maximum (.626) and average (.573) AP, with balanced precision and recall metrics. These results, reported in Table 4.2, suggest how combining in-distribution (DDPM) and out-of-distribution (MemSeg) data ameliorates precision and recall scores, helping the model

better understand what an anomalous sample is.

4.6.3 N-shot data augmentation, N small

Within manufacturing environments, organizations strive to minimize the occurrence of defects, resulting in a generally restricted number of anomalous samples. In this sub-section, we put ourselves in this situation, i.e., only a minimal amount of ground truth positive samples are available in the dataset.

Table 4.3: Results between MemSeg and DDPM when ***few*** anomalous images are available. Each training set contains $N = 5$ anomalous samples, plus N_{aug} augmented images.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑	$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
MemSeg 80	.582 (.018)	**.836** (.101)	.466 (.049)	DDPM 80	.580 (.045)	.542 (.270)	**.634** (.212)
MemSeg 100	.511 (.086)	.686 (.082)	.527 (.069)	DDPM 100	.526 (.075)	.610 (.063)	.477 (.081)
MemSeg 120	**.593** (.044)	.801 (.065)	.507 (.053)	DDPM 120	.535 (.063)	.659 (.127)	.491 (.046)
Average	**.562** (.049)	**.774** (.083)	.500 (.057)	Average	.547 (.061)	.604 (.153)	**.534** (.113)

To simulate this challenging setup, we randomly select only $N = 5$ anomalous samples from the KSDD2 training dataset and use them to fine-tune the DDPM for 49 epochs with $\alpha = 0.95$. Following the procedure introduced in Section 4.5.2, we generate several training sets induced by the different N_{aug} of new samples, plus the N images on which we trained the DDPM. For the classifier, we use the same ResNet-50 architecture. The findings of this experiment are documented in Table 4.3. As we can see, the MemSeg method slightly outperforms DDPM, resulting in an average AP of .562 and .547, respectively. Moreover, MemSeg produces a maximum AP of .593 at $N_{aug} = 120$, while DDPM records a maximum AP of .580 at $N_{aug} = 80$. The precision and recall have similar behavior as seen in 4.6.2, with DDPM having a higher recall (.634 vs. .527) and lower precision (.659 vs .836) w.r.t. MemSeg.

Interestingly enough, in Table 4.4, we can see that the *In&Out* pipeline does not seem to increase the performance, achieving an average AP on par with MemSeg (.561) while recording a slightly lower maximum AP (.578 vs. .593). We hypothesize that, in this setup, DDPM overfits the minimal number of anomalous images and cannot generalize the anomalous samples properly. This is a problem if the samples on which we fine-tune the model are a subset of all the anomalies and, thus, are not representative enough of the entire anomalous distribution.

4.6.4 N-shot data augmentation, N large

Finally, to showcase *In&Out* as a general data augmentation technique, we explore the scenario with more positive samples in the training set. To this aim, we make all 246 positive samples available to the anomaly detection model during training, in addition to the usual N_{aug} augmented anomalous images. Following the procedure in Section 4.5.2, we use all the $N = 246$ positive samples from the training set to fine-tune our diffusion model for 25 epochs with $\alpha = 0.80$. Finally, we define a baseline by training the ResNet-50 with $N_{aug} = 0$ (*In&Out* 0), achieving an average AP of .747. The results are reported in Table 4.5.

Table 4.4: Results when ***few*** anomalous images are available using *In&Out*. Each training set contains $N_{pos} = 5$ anomalous samples, plus N_{aug} augmented images, where half samples are generated by DDPM and half by MemSeg.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
In&Out 80	.531 (.041)	.507 (.220)	.655 (.126)
In&Out 100	**.578** (.041)	.450 (.343)	**.761** (.245)
In&Out 120	.575 (.025)	**.635** (.316)	.636 (.189)
Average	.561 (.036)	.531 (.293)	**.684** (.187)

Table 4.5: Results when ***all*** the anomalous samples are available using *In&Out*. Each training set contains all the anomalous KSDD2 samples, plus N_{aug} augmented images, where half of the samples are generated by DDPM and half by MemSeg. Additionally, *In&Out* 0 indicates the performance achieved without data augmentation. Note that MixedSegdec [53] indicates the results reported under the weakly supervised setting.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
MixSegdec	.733 (-)	- (-)	- (-)
In&Out 0	.747 (.055)	.826 (.081)	.723 (.058)
In&Out 80	.747 (.022)	.764 (.046)	**.734** (.032)
In&Out 100	.775 (.013)	.868 (.050)	.720 (.026)
In&Out 120	**.782** (.030)	**.906** (.064)	.689 (.030)
Average	**.768** (.022)	**.846** (.053)	.714 (.029)

Table 4.6: Results between MemSeg and DDPM when ***all*** the anomalous samples are available.

$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑	$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
MemSeg 80	.744 (.007)	.851 (.055)	.691 (.058)	DDPM 80	.758 (.007)	.808 (.056)	**.768** (.043)
MemSeg 100	**.774** (.016)	.814 (.038)	.752 (.028)	DDPM 100	.763 (.008)	.829 (.059)	.725 (.034)
MemSeg 120	.734 (.032)	.772 (.107)	.707 (.031)	DDPM 120	.772 (.034)	**.858** (.084)	.725 (.061)
Average	.751 (.018)	.812 (.067)	.717 (.039)	Average	**.764** (.016)	**.832** (.066)	**.739** (.046)

The results of the two separate data augmentation procedures are reported in Table 4.6. In this scenario, the anomaly detection model trained with DDPM augmented images achieves a maximum AP of .772, outperforming both the baseline (.747) and resulting in a higher average AP than MemSeg (.764 vs. .751). As we can see in Table 4.5, *In&Out* achieves the highest average AP yet (.768) while balancing the precision and recall metrics, confirming our intuition. Notably, with 120 augmented images, the maximum AP classification score is .782, beating the previous .733 [53] and setting the new state-of-the-art.

4.7 Conclusion

In this work, we introduced *In&Out*, a hybrid data augmentation strategy that combines diffusion-generated in-distribution defects (DDPMs) with traditional per-region out-of-distribution overlays. Our experiments on the Kolektor Surface-Defect Dataset 2 (KSDD2) demonstrate that this approach achieves a new state-of-the-art weakly supervised classification AP score of 0.782, highlighting the effectiveness of generating realistic positive samples to complement existing augmentation methods.

These results highlight two key insights that support the broader thesis narrative. First, modern generative models can meaningfully enrich visual datasets, enabling classifiers to better capture fine-grained anomalies even in the face of severe data scarcity. Second, integrating human guidance in the augmentation process, through textual prompts or domain-specific cues, can further improve the relevance and diversity of synthetic defects, particularly in zero- or few-shot settings.

Building on these findings, the next chapter extends this idea by fully exploring human-in-the-loop, text-conditioned anomaly generation. In DIAG, we investigate how multimodal guidance can be systematically leveraged to generate interpretable and controllable defect examples, enabling both higher performance and more intuitive interaction with industrial AI systems.

Chapter 5

Leveraging Latent Diffusion Models for Training-Free In-Distribution Data Augmentation for Surface Defect Detection

5.1 Abstract

Defect detection in industrial production is inherently challenging due to the scarcity of positive samples relative to abundant normal data. Traditional data augmentation techniques attempt to mitigate this imbalance by superimposing synthetic artifacts onto normal samples, but these out-of-distribution approaches often fail to capture the true appearance of real-world defects, limiting model performance. Building on the hybrid augmentation strategy introduced in Chapter 4, we present DIAG, a training-free Diffusion-based In-distribution Anomaly Generation pipeline. Unlike conventional augmentation methods, DIAG incorporates a human-in-the-loop paradigm in which domain experts provide multimodal guidance through textual descriptions and localized annotations of potential anomalies. This approach enhances the realism and relevance of generated defect samples while fostering interpretability and iterative refinement via human feedback. Remarkably, DIAG operates in a zero-shot manner, eliminating the need for time-consuming fine-tuning while delivering strong performance. Experiments on the challenging KSDD2 dataset demonstrate that DIAG achieves substantial improvements over state-of-the-art augmentation strategies, increasing average precision by approximately 18% when positive samples are available and 28% in the absence of real defect data. This work advances the language-based interaction paradigm from the previous chapter, demonstrating that human-in-the-loop frameworks can effectively replace expensive fine-tuning procedures. By leveraging the language capabilities of modern generative models, we significantly enhance the quality, interpretability, and practical usability of industrial AI systems.

5.2 Introduction

Surface Defect Detection (SDD) remains a challenging problem in industrial scenarios, defined as the task of identifying samples containing defects [47], i.e., products whose texture or structure deviates from a prototypical pattern. In traditional setups, human experts inspect every item and remove defective pieces. While effective, this approach is expensive, slow, and prone to errors related to stress and fatigue, which limits its scalability.

Automated defect detection systems [48, 49] alleviate many of these limitations by learning classifiers on defective and nominal samples. However, collecting sufficient data for training remains a major bottleneck: defective items (positive samples) are rare relative to nominal items (negative samples), and the rise of Industry 4.0 with flexible manufacturing [61] has increased demand for systems that can quickly adapt to new product lines or small-batch production. Traditional data collection and labeling approaches struggle to meet these requirements.

Recent work has addressed this challenge by framing SDD as an unsupervised learning problem [62, 63], training models exclusively on nominal samples [64], or applying few-shot learning strategies [65]. In these cases, anomalies are identified as deviations from a learned nominal distribution. While effective in some scenarios, these approaches often generate false positives, particularly on datasets with complex textures or structures [66].

Importantly, in real industrial settings, anomalies are typically not random but arise from predictable production issues. Expert operators can anticipate the types, locations, and frequencies of defects. Consequently, generative AI offers a promising avenue for SDD: by producing realistic defect images conditioned on domain knowledge, models can be trained on richer, more representative positive samples. In recent years, Denoising Diffusion Probabilistic Models (DDPMs) [57] have emerged as a powerful tool for high-fidelity image generation, making them particularly suitable for this task.

Building on the previous chapter's exploration of diffusion-based augmentation (In–Out), we propose DIAG, a training-free Diffusion-based In-distribution Anomaly Generation pipeline for SDD. Our approach leverages pre-trained DDPMs with multimodal conditioning, allowing domain experts to guide the generation of realistic defect images through textual prompts. This enables the creation of plausible anomalies without requiring real positive samples, and the generated images can then be used to train anomaly detection models. Empirical results demonstrate that classifiers trained with DIAG achieve notable improvements over prior state-of-the-art augmentation methods. Figure 5.1 illustrates the overall workflow, detailed in Section 5.4.

The main contributions of this work are:

- We introduce DIAG, a complete pipeline for generating in-distribution defect samples using textual prompts and nominal images, demonstrating superior anomaly detection performance compared to existing augmentation strategies.

- We present methods for spatially controlled defect generation, enabling human-in-the-loop guidance to produce region-specific anomalies and improve the plausibility and relevance of synthetic defects.

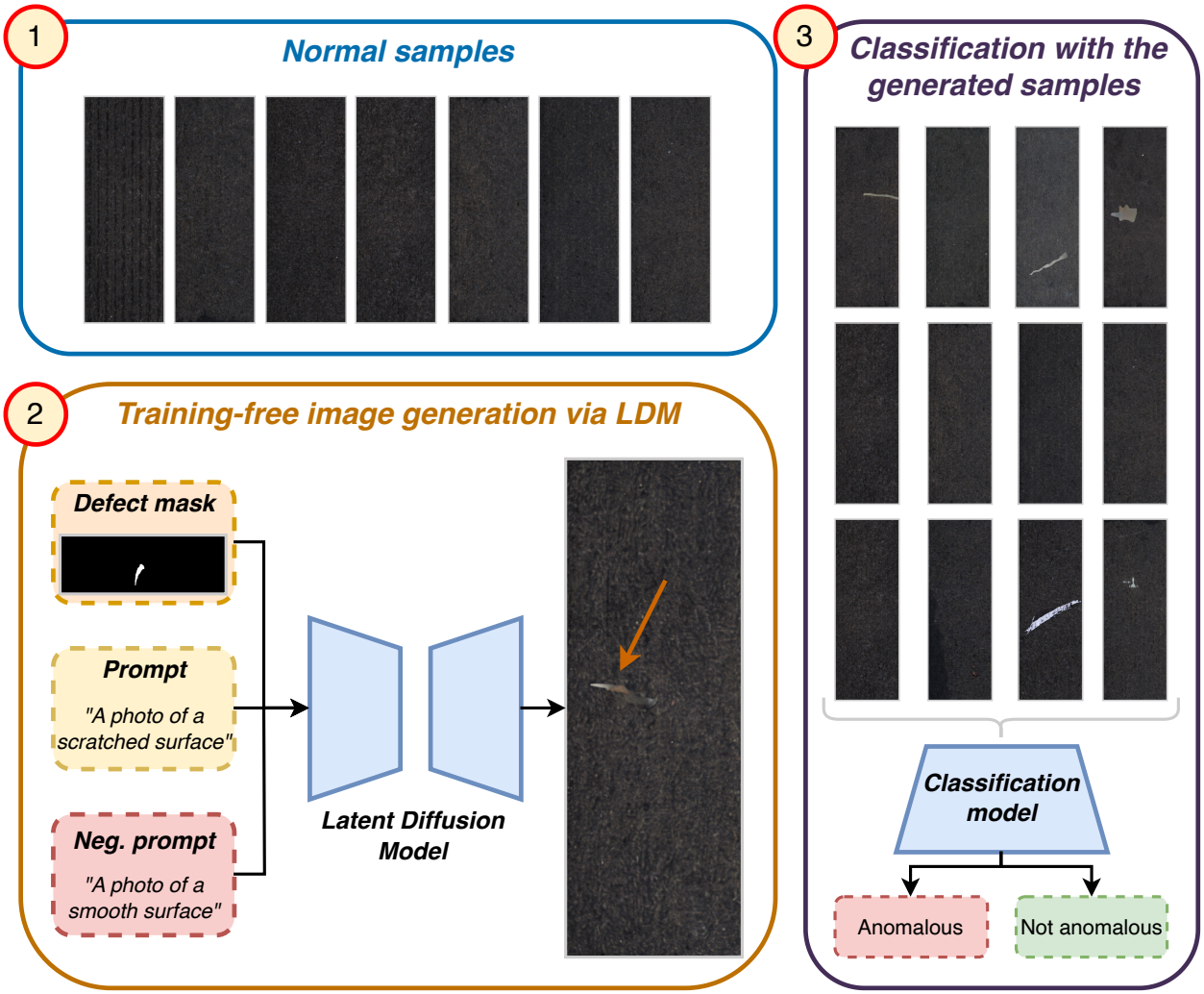


Figure 5.1: The DIAG pipeline. Starting from positive samples, we leverage a Latent Diffusion Model (LDM) to synthesize novel in-distribution high-quality images of defective surfaces based on defect localization and textual prompts. These synthetic images are then used as anomaly samples to train a binary classifier for anomaly detection.

5.3 Related Work

Surface defect detection refers to identifying and categorizing irregularities, flaws, or imperfections on the surface of materials or objects. These defects include scratches, cracks, discolorations, and any other anomaly that deviates from the expected surface quality. At the time of this work, research had been conducted according to different setups: unsupervised approaches [67] used a mixture of unlabelled positive and negative sample images for training; supervised approaches required labeled samples in the form of binary masks representing the defects (full supervision)[68] or simply as a tag for the whole image (weak supervision)[53]. Supervised methods have demonstrated superior accuracy in the identification of anomalies. Nevertheless, the effort required to provide good annotations was not always justified. Collecting positive samples could be time and resource-consuming due to the low rate of defective products generated by industrial lines. Thus, many contemporary approaches adopted a "*clean*" setup, where the training set consisted of only nominal samples. Two strategies could be adopted in clean setups: model fitting and image generation. Model fitting approaches aim to generate an accurate model of the nominal distribution, considering every sample with a likelihood lower than, or a distance from the nominal prototype higher than, a predefined threshold [64, 69]. On the contrary, data augmentation approaches leveraged generative methods to synthesize images of defects and used these images as positive samples for training a supervised model. Specifically, this work focused on generation-based data augmentation under clean setups. The most popular data augmentation pipeline for SDD at that time consisted of a series of random standard transformations of the input image, such as mirroring, rotations, and color changes, followed by the superimposition of noisy patches [46]. In [50], an ablation study focused on the generation of synthetic anomalies led to the following findings: *i)* adding synthetic noise images was never counterproductive, it just diminished the effectiveness in percentage; *ii)* few generated anomaly images (in the order of tens) were enough to increase the performance substantially; *iii)* textural injection in the anomalies was essential, or, equivalently, adding uniformly colored patches was ineffective. In MemSeg [46], the pipeline for the generation of the abnormal synthetic examples was divided into three steps: *i)* a Region of Interest (ROI) indicating where the defect would be located was generated using Perlin noise and the target foreground; *ii)* the ROI was applied to a noise image to generate a noise foreground ROI; *iii)* the noise foreground ROI was super-imposed on the original image to obtain the simulated anomalous image. However, all these approaches were based on generating out-of-distribution patterns that did not faithfully represent the target-domain anomalies. In [51], the authors introduced the concept of "extended anomalies", where the specific abnormal regions of the seen anomalies were placed at any possible position within the normal sample after applying random spatial transformations. Unfortunately, this required an accurate segmentation of the training images, an operation we sought to avoid. The first work that drew attention to in-distribution defect data was In&Out [6], discussed in Chapter 4, where we showed that diffusion models provided more realistic in-distribution defects. In this paper, we significantly improved the generation of in-distribution anomalous samples by incorporating domain knowledge provided by an expert user through textual prompts and localization of salient regions. We used state-of-the-art MemSeg [46] and In&Out [6] methods as competitors for our augmentation pipeline. With

DIAG, we produced a distribution of defective images closer to the real one, resulting in a more precise identification of the decision boundaries in the classification step.

5.4 Method

In this section, we provide detailed explanations of DIAG. In particular, Section 5.4.1 covers techniques for diffusion-based image generation, Section 5.4.2 showcases the anomalous image generation pipeline, and Section 5.4.3 outlines the anomaly detection model training procedure.

5.4.1 Multimodal diffusion-based image generation

DDPMs [56, 57] are a class of deep latent variable models that work by modeling the joint distribution of the data over a Markovian inference process. This process consists of small perturbations of the data with a variance-preserving property [70], such that the limit distribution after the diffusion process is approximately identical to a known prior distribution. Starting with samples from the prior, a reverse diffusion process is learned by gradually denoising the sample to resemble the initial data by the end of the procedure.

Formally, the data distribution $q(x_0)$ is modelled through a latent variable model $p_\theta(x_0)$:

$$p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T} \ , \tag{5.1}$$

$$p_\theta(x_{0:T}) := p_\theta(x_T) \prod_{t=1}^{T} p_\theta^{(t)}(x_{t-1}|x_t) \ , \tag{5.2}$$

where $x_1, \ldots, x_T$ are latent variables of the same dimensionality as x_0.

The parameters θ are learned by maximizing an ELBO of the log evidence, *i.e.*:

$$\max_\theta \mathbb{E}_{q(x_0)}[\log p_\theta(x_0)] \leq$$
$$\max_\theta \mathbb{E}_{q(x_0, x_1, \ldots, x_T)} \left[\log p_\theta(x_{0:T}) - \log q(x_{1:T}|x_0)\right] \ , \tag{5.3}$$

where $q(x_{1:T}|x_0)$ represents a fixed inference process defined as the following as a Markov chain:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \ , \tag{5.4}$$

$$q(x_t|x_{t-1}) := \mathcal{N}\left(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}} x_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right) I\right) \ , \tag{5.5}$$

where $\alpha_{1:T} \in (0, 1]^T$ is a predefined variance schedule, and the covariance matrix is ensured to have positive terms on its diagonal. Specifically, this parametrization has the property:

$$q(x_t|x_0) = \int q(x_{1:t}|x_0) dx_{1:(t-1)} =$$
$$\mathcal{N}(x_t; \sqrt{\alpha_t} x_0, (1 - \alpha_t) I) \ , \tag{5.6}$$

therefore we can write x_t as a linear combination of x_0 and a noise variable ϵ.

When we set α_T sufficiently close to 0, $q(x_T|x_0)$ converges to a standard Gaussian for all x_0, so it is natural to set $p_\theta(x_T) := \mathcal{N}(0, \mathbf{I})$. Given that all the conditionals are modeled as Gaussians with fixed variance, the objective in Equation equation 5.3 can be greatly simplified. In particular, [57] shows that the following (further simplified) lower bound provides optimal generative performance:

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \mathbb{E}_{x_0,\epsilon_t} \left[\|\epsilon_\theta^{(t)}(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t) - \epsilon_t\|_2^2 \right] , \tag{5.7}$$

where $x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0, I)$, $\epsilon_\theta = \{\epsilon_\theta^{(t)}\}_{t=1}^T$ is a set of T functions, with each $\epsilon_\theta^{(t)} : X \to X$ having trainable parameters $\theta^{(t)}$.

In practice, these functions are approximated by a neural network conditioned on the diffusion time t. After the model is trained, we can generate new samples by first sampling x_T from the known prior $p_\theta(x_T)$, and then iteratively reversing the diffusion process, thereby sampling $\{x_{T-1} \dots x_0\}$.

In addition, we leveraged the natural ability of DDPMs to incorporate multimodal conditioning in the generation process, taking inspiration from [15, 2, 17, 71]. Specifically, we will use prompts, *i.e.*, textual descriptions of the anomaly, and negative prompts, i.e., prompts that guide the image generation "away" from its concepts. This results in high-quality images that comply with the given descriptions [40, 72, 73].

In particular, we opt to utilize an inpainting model, as demonstrated in [15, 56]. Given an image with a masked region, inpainting seamlessly fills it with content that harmonizes with the surrounding image. Although typically employed to eliminate undesired artifacts, the inpainting process ensures that the masked area incorporates the provided prompt, effectively merging textual and visual content.

5.4.2 The DIAG pipeline

To generate an anomalous image i_a, the process starts by sampling a random negative image, an anomaly description, and a mask, forming the triplet (i_n, d_a, m_a). Instead of directly operating on the image pixels using DDPM, we use a Latent Diffusion Model (LDM) to work in a lower-dimensional latent space [15]. Thus, the above information will be fed to a text-conditioned LDM to perform inpainting on image i_n using the mask m_a.

The anomaly description d_a guides the generation, filling the masked region of i_n with an anomaly that complies with the prompt. To generate images resembling real anomalous samples, domain knowledge from industrial experts is exploited, providing textual descriptions of the potential anomalies' type, shape, and spatial information.

The LDM is then conditioned on this information to inpaint plausible anomalies on defect-free samples. Formally, given pictures of defect-free (negative) samples I_n, domain experts will provide textual descriptions D_a of what different anomalies may look like. At the same time, regions where these anomalies may appear on the defect-free samples will be designated. We define this set of regions as a set of binary masks M_a of possible anomalies, shapes, and locations. The result of this operation is i_a, an anomalous version of i_n, where an anomaly has been inpainted in the masked region m_a. Due to the stochastic nature of LDMs,

this process can be repeated multiple times to generate an augmented set of anomalous sample images I_a. Finally, the set I_a can be used as data augmentation for training anomaly detection models, as presented in the following section.

5.4.3 Anomaly detection task

We approach the anomaly detection problem as a binary classification problem, where the objective is to predict whether a sample belongs to one of two classes. Specifically, we utilized a ResNet-50 [60] backbone trained with a binary cross-entropy loss function denoted as $\mathcal{L}_{\text{BCE}}$. The binary cross-entropy loss measures the dissimilarity between the predicted probability distribution and the actual distribution of the labels. Mathematically, it is defined as:

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right] \ , \tag{5.8}$$

where, y represents the ground truth labels, $\hat{y}$ represents the predicted probabilities, and N is the number of samples. In detail, y_i denotes the true label for sample i, which can be either 0 or 1, while $\hat{y}_i$ signifies the predicted probability that sample i belongs to class 1.

5.5 Experiments

In this section, we show the efficacy of our data augmentation approach for defect detection from a quantitative and qualitative point of view.

5.5.1 Experiment setup

Datasets. We use the Kolektor Surface-Defect Dataset 2 (KSDD2) [53], one of the most recent, complex, and real-world surface defect detection datasets. This dataset comprises 246 positive and 2085 negative images in the training set and 110 positive and 894 negative images in the testing set. Positive images are images with visible defects, such as scratches, spots, and surface imperfections.

Since the images have different dimensions, we standardize the dataset resolution, resizing all the images to 224×632 pixels while keeping the number of normal and anomalous samples unchanged.

Evaluation metrics. The anomaly detection performance was evaluated based on Average Precision (AP), Precision, and Recall, following the evaluation protocol defined in [6].

Additionally, to evaluate the visual similarity between generated images and the original dataset images, we employ the Fréchet Inception Distance (FID) [74], a popular metric in the image generation field which computes the distance between the distribution of two sets of images. More specifically, the Fréchet distance calculates distance $d(.,.)$ between a Gaussian with mean (m, C) obtained from $p(.)$ and a Gaussian with mean (m_w, C_w) obtained by $p_w(.)$, where $p_w(.)$ represents real world data and $p(.)$ represents generated data. In practice, these distributions are two sets of data: the "world" data (*i.e.*, the images in a dataset) and the

"generated" data (*i.e.*, the generated images). These sets are then fed to an Inception model pre-trained on ImageNet to extract deep features from each sample of the distributions. The resulting two sets of features represent the Gaussians with mean (m_w, C_w) and (m, C) for the "world" and "generated" data, respectively. Specifically, [74] shows that a lower FID score matches a human's higher perceived visual similarity (a lower perceptual distance), meaning that similar sets of images will have a lower FID than dissimilar sets. Formally, the FID score is:

$$d((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + \mathrm{T}(C + C_w - 2(CC_w)^{\frac{1}{2}}) \tag{5.9}$$

where T refers to the trace linear algebra operation.

5.5.2 Implementation details

In this section, we specify all the implementation details for reproducibility. All training and inferences were conducted on an NVIDIA RTX 3090 GPU.

Inpainting via DIAG. We use the pre-trained implementation of SDXL [73] from Diffusers [75] as our text-conditioned LDM. In particular, SDXL shows drastically improved performance compared to the previous versions of Stable Diffusion [15] and achieves results comparable to commercial state-of-the-art image generators.

Following the procedure outlined in Section 5.4.2, we use the negative images of KSDD2 as the set I_n. As the set of anomaly descriptions D_a, we used the prompts "`white marks on the wall`" and "`copper metal scratches`". Instead, "`smooth, plain, black, dark, shadow`" were used as a negative prompt to improve the performance further. These prompts were selected through a human-in-the-loop iterative pipeline, until the resulting images resembled plausible anomalies. We used the segmentation masks of positive samples in the KSDD2 dataset to simulate the domain experts' definition of plausible anomalous regions.

Then, these data are fed to the pre-trained SDXL model to perform inpainting on the negative images in a training-free process, generating the set of augmented anomalous images I_a as described in Section 5.4.2.

Finally, the generated images I_a are added to the training set, which will be used to train the anomaly detection model.

ResNet-50 training and testing. For a fair comparison with [6], we use the same PyTorch implementation of the ResNet-50 [60] as our anomaly detection model, in which we substitute the fully connected layers after the backbone to make it a binary classifier. The network is trained for 50 epochs with Adam [76] as an optimizer, a learning rate of 0.0001, and a batch size of 32.

To maintain consistency with the training and evaluation procedures of KSDD2, our setup is the same as presented in [6, 53], where only the images and ground truth labels are used to train the model. For our comparison, we use the official code of In&Out [6] to fine-tune a DDPM in the *full-shot* scenario (on all the positive images of the KSDD2 dataset) and generate their augmented images. Likewise, we follow the procedure of MemSeg [46] to generate the "per-region" augmented images. Finally, we generate DIAG augmented images, following the inpainting methodology outlined in Section 5.4. The set of images used for

Table 5.1: Results between MemSeg, In&Out and DIAG when ***no*** anomalous samples are available. In **bold**, the best results. Underlined, the second best.

Model	N_{aug}	AP ↑	Precision ↑	Recall ↑
MemSeg [46]	80	.514	.733	.436
MemSeg [46]	100	.388	.633	.432
MemSeg [46]	120	.511	.683	.470
In&Out [6]	80	.556	.530	.655
In&Out [6]	100	.626	.742	.568
In&Out [6]	120	.536	.699	.534
DIAG (ours)	80	.769	.851	**.673**
DIAG (ours)	100	**.801**	.924	.664
DIAG (ours)	120	.739	**.944**	.609

training changes depending on the experiment and the pipelines being tested, but in general, it can be seen as a combination of the original negative images I_n, an optional set I_p of original positive images, and the set of generated positive images I_a.

5.5.3 Quantitative results

Zero-shot data augmentation. Here, we emulate the situation where ***no*** original positive samples are available in the training set. This scenario makes generating augmented positive samples necessary and restricts the users to augmentation procedures that do not rely on positive images. To do this, we build the set of augmented anomalous images I_a by generating N_{aug} augmented positive samples with different pipelines, *i.e.*, MemSeg [46], In&Out [6] and DIAG. Then, we train a ResNet-50 on a dataset that includes the original negative samples I_n and the augmented positive samples I_a. Finally, we evaluate the model on the original test set.

Table 5.1 reports the comparison between the models trained with MemSeg, In&Out, and DIAG augmented data at different values of N_{aug}. As we can see, our proposed method achieves the highest AP (.801), recorded at 100 augmented images, while also resulting in a consistently higher AP when compared to the MemSeg and In&Out pipelines. These impressive results highlight how, through domain expertise in the form of anomaly descriptions and segmentation masks, it is possible to generate in-distribution images able to meaningfully guide an anomaly detection network, even in a complicated scenario where no real anomalous data is available.

Surprisingly, the DIAG performance with $N_{aug} = 120$ augmented images is lower than using a smaller number of augmented images. We hypothesize this is due to the stochastic nature of the LDMs image generation. While it allows the generation of various images given the same guidance, it can also lower, in some cases, the predictability of the quality of the generated samples, which sometimes may not faithfully comply with the prompt. Future works will focus on studying quality consistency in the image generation pipeline.

Table 5.2: Results between MemSeg, In&Out and DIAG when ***all*** the anomalous samples are available. In **bold**, the best results. Underlined, the second best.

Model	$\mathbf{N_{aug}}$	**AP** ↑	**Precision** ↑	**Recall** ↑
MemSeg [46]	80	.744	.851	.691
MemSeg [46]	100	.774	.814	.752
MemSeg [46]	120	.734	.772	.707
In&Out [6]	80	.747	.764	.734
In&Out [6]	100	.775	.868	.720
In&Out [6]	120	.782	.906	.689
DIAG (ours)	80	.869	<u>.912</u>	.755
DIAG (ours)	100	<u>.911</u>	**.978**	<u>.800</u>
DIAG (ours)	120	**.924**	.896	**.864**

Full-shot data augmentation. To showcase DIAG as a general data augmentation technique, we also explore the scenario where real positive samples are available in the training set. To this aim, we include all the 246 real positive samples I_p in the training set, together with the real negative images I_n and the N_{aug} augmented positive images I_a.

As we can see from Table 5.2, DIAG achieves the highest average AP yet (.924), surpassing the .782 set by the previous state-of-the-art data augmentation pipeline [6]. When comparing these results to the ones obtained in the "zero-shot data augmentation" scenario, it is clear how more in-distribution images improve model performance during training. This is highlighted by the improvement in performance of all the models when adding the real positive images I_p to the training set. At the same time, the inclusion of DIAG augmented images allows the model to explore the anomaly distribution further, resulting in the difference in performance between the different data augmentation pipelines.

5.5.4 Qualitative results

The main goal of our data augmentation pipeline is to generate in-distribution synthetic positive images, meaning images that closely resemble the real ones. Figure 5.2 shows qualitative results. It's evident that the images produced by DIAG are markedly more realistic compared to those generated by MemSeg [46] and In&Out [6].

In addition, we provide a numeric evaluation of the similarity between the generated images and the real ones by employing FID [74]. It is worth noting that due to the limited number of anomalous images in the original dataset, we are forced to calculate FID on a different network layer, precisely the second max pooling layer. This is a common procedure in cases where the number of images is low, as the calculation requires the number of samples (images) to be higher than the number of features. Note that this only changes the magnitude of the values obtained, not the metric's general behavior. In the specific case of KSDD2, we choose the first and second max-pooling layers with 64 and 192 features, respectively. Specifically, we compare the images generated with MemSeg, In&Out, and DIAG with the ones available in the KSDD2 dataset and compute the FID scores between the positive images

Figure 5.2: First row displays some negative samples from the KSDD2 dataset. Instead, the second row shows some images of positive samples from the same dataset. In the third row, we show the MemSeg-generated defect samples. The fourth row shows In&Out generated defect samples. Lastly, the final row showcases images generated with DIAG. Notably, the defect images that DIAG generated are more realistic and in-distribution.

Table 5.3: FID scores between the real positive images of KSDD2 and the images generated by MemSeg, In&Out and DIAG. The scores are calculated using the first and second max pooling layers of the Inception network, having 64 and 192 features, respectively. In **bold**, the best results.

Augmentation procedure	**FID 64** ↓	**FID 192** ↓
MemSeg [46]	0.834	4.376
DDPM [6]	0.334	1.520
DIAG (ours)	**0.096**	**0.411**

of the KSDD2 and the previously mentioned sets of augmented images.

The results, reported in Table 5.3, highlight how DIAG can generate images that are very similar to the ones originally present in the dataset, resulting in the lowest FID out of all the other methodologies.

Another interesting observation is how both the generative-model-based procedures (DDPM and DIAG) result in images that are more in-distribution (lower FID) than the "per-region" augmentation techniques such as MemSeg, which records the highest FID out of all the tested methodologies.

5.6 Conclusions

In this work, we presented DIAG, a training-free, language-conditioned data augmentation pipeline for surface defect detection. By incorporating domain experts into the generation process, textual prompts describe both the type and location of defects, guiding a pre-trained Latent Diffusion Model (LDM) to produce realistic in-distribution positive samples. These synthetic images are then used to train a binary classifier for anomaly detection.

Our experiments on the KSDD2 dataset demonstrate that DIAG establishes a new state-of-the-art for data augmentation in this context, achieving AP scores of 0.801 and 0.924 in zero-shot and full-shot scenarios, respectively. These results confirm that realistic, in-distribution defect generation significantly improves downstream classification performance, particularly in situations where real positive samples are scarce or entirely unavailable.

Beyond performance metrics, this work highlights the broader thesis theme: integrating language and human guidance into visual perception pipelines enhances both the usability and effectiveness of AI systems. By enabling domain experts to directly steer the generative process, DIAG not only produces higher-quality data but also fosters interpretability and control, bridging the gap between human expertise and automated anomaly detection.

These promising results open several directions for future research, including testing robustness to noisy or imprecise textual prompts and exploring the application of human-guided generative augmentation across different industrial datasets. This chapter sets the stage for the second half of this thesis, which focuses on enhancing the expressivity and controllability of vision-and-language models in multimodal tasks, continuing the overarching narrative of human-centered, language-enhanced AI in industrial and creative domains.

Part II

Enhancing Language in Vision-and-Language Models

Introduction. The first half of this thesis has explored how language and generative models can enhance visual perception in practical applications, ranging from robot navigation in complex environments (Le-RNR-Map Chap. 3) to industrial surface defect detection (In–Out and DIAG, in Chapters 4-5). Across these studies, we observed a common theme: introducing language, whether through textual prompts, human-in-the-loop guidance, or natural-language queries, can provide semantic structure, improve interpretability, and enrich data representations for downstream tasks.

While these results demonstrate the value of language as a guiding signal, they also reveal a limitation in current multimodal systems. In many existing Vision-and-Language Models, the language input is often treated as a static, secondary modality. Models are trained primarily to align with concrete visual content, such as objects, colors, and spatial arrangements, while more abstract, descriptive, or creative aspects of language remain underutilized. This gap limits the ability of VLMs to fully leverage textual information for reasoning, retrieval, or generative tasks, especially when the language contains abstract properties, stylistic cues, or nuanced descriptions that go beyond literal visual features.

The second half of this thesis addresses this challenge by shifting focus from language as a guidance signal for visual perception to enhancing the expressive and functional capacity of language within multimodal models. Our goal is to develop methods that allow VLMs to: (i) Better capture abstract or underrepresented language concepts; (ii) Improve cross-modal retrieval and reasoning using richer textual embeddings; (iii) Support more controllable and interpretable generative or analytic outcomes.

The first contribution in this direction is the Abstract-to-Concrete Translator (ACT) in Chapter 6. ACT is a training-free, model-agnostic method designed to bridge the gap between abstract textual expressions and the concrete representations that VLMs are typically optimized to understand. By transforming abstract language into semantically richer, visually grounded embeddings, ACT enables VLMs to perform more accurate retrieval, exhibit stronger generalization across datasets, and leverage textual information that would otherwise be ignored.

This chapter marks the transition in the thesis narrative: from exploring how language enhances visual perception in applied scenarios to investigating how we can enhance the utility, expressivity, and interpretability of language itself within multimodal models, laying the foundation for the subsequent ACT study and further works on evaluation and controllable generation.

Chapter 6

Seeing the Abstract: Translating the Abstract Language for Vision Language Models

6.1 Abstract

The first half of this thesis explored how language can guide visual perception in applied settings, from language-augmented navigation to human-in-the-loop industrial anomaly detection. Building on these insights, we now focus on the complementary problem of enhancing the *expressivity and utility of language itself* within Vision-and-Language Models (VLMs).

Natural language extends beyond describing concrete visual content: it conveys abstract concepts, stylistic properties, and subtle attributes that are not directly perceivable in images. Despite their prevalence, current VLMs largely overlook these abstract-oriented expressions, limiting their effectiveness in retrieval, reasoning, and generative tasks. We investigate this issue in the fashion domain, a field rich in abstract descriptions, and find that abstract terms are widespread, provide novel information, and play a critical role in cross-modal retrieval.

To address this limitation, we introduce a *training-free* and *model-agnostic* method, Abstract-to-Concrete Translator (ACT), which maps abstract textual representations to well-represented concrete concepts in the VLM latent space using existing pre-trained models and multimodal datasets. On the text-to-image retrieval task, ACT consistently outperforms fine-tuned VLMs in both same- and cross-dataset evaluations, demonstrating strong generalization. Furthermore, its improvements are consistent across multiple VLM architectures, providing a plug-and-play solution for enhancing abstract-oriented language understanding.

This work illustrates that systematically improving the representation of abstract language can substantially enhance multimodal model capabilities, completing the thesis narrative transition from *language-guided perception* to *language-enhanced representation and controllability*.

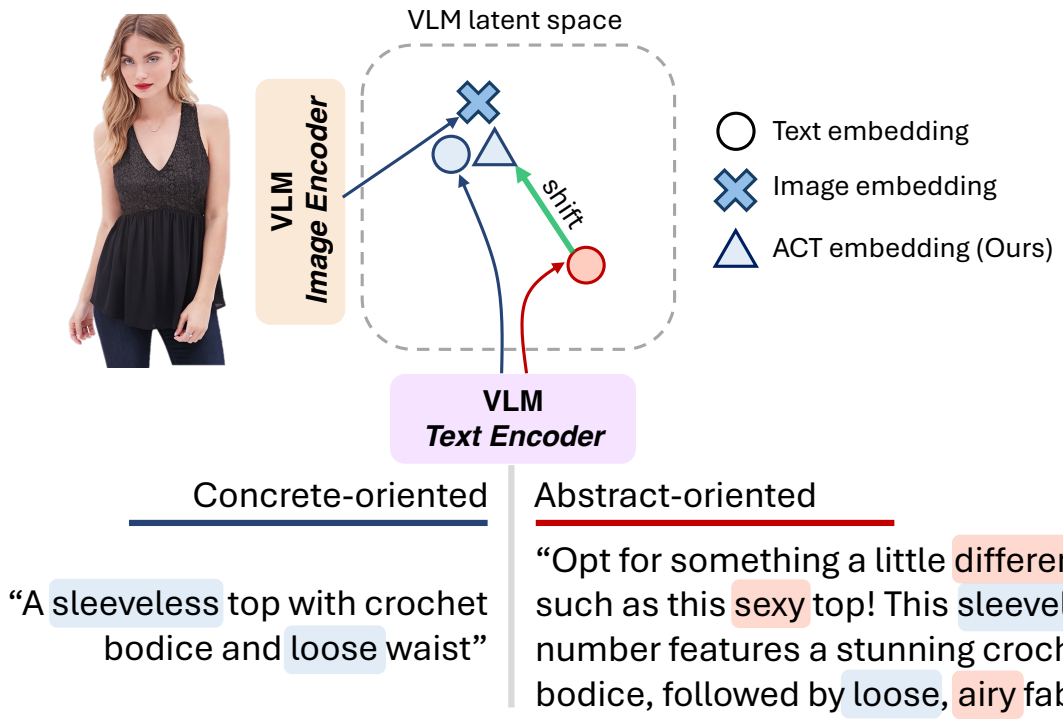


Figure 6.1: Human language can exhibit both *abstract* and *concrete* words to express feelings, desires, and properties together with perceivable elements, *e.g.* when describing a fashion item. However, Vision Language Models (VLMs) are mostly pre-trained with *concrete* -oriented web-image texts, thus under-representing the abstract-oriented ones. When encoding the abstract-oriented description with pre-trained VLMs, there exists a noticeable representation shift from the concrete-oriented description, hindering the performance in downstream tasks, *e.g.* text-to-image retrieval in fashion. Our proposal Abstract-to-Concrete Translator (ACT) can effectively compensate for such representation shift in a training-free manner, bringing the representation of the abstract-oriented language towards the concrete-oriented one in the latent space of existing VLMs, thereby improving the downstream task performance.

6.2 Introduction

Human language extends far beyond the straightforward depiction of visual elements and physical attributes. Over millennia, humans have developed the ability to express complex feelings, stylistic qualities, and creative concepts through a nuanced interplay of abstract and concrete language [77, 78]. In many domains where aesthetics and creativity are central, abstract language provides richer, more expressive descriptions than concrete terms alone. One of the most illustrative domains is fashion, where experts, critics, and consumers employ terms such as "sexy," "airy," or "different" to convey impressions or moods, complementing concrete descriptors like "strapless," "loose," or "black" (Fig. 6.1). Abstract-oriented language allows users to evoke emotions or stylistic qualities without relying on strictly perceivable attributes, enhancing expressivity and flexibility in human-computer interaction.

This work focuses on abstract-oriented language in the context of fashion, a domain well-represented in large-scale multimodal datasets such as FACAD [79] and DeepFashion [80]. Through statistical analysis of fashion descriptions, we find that abstract adjectives are as prevalent as, or even more frequent than, concrete ones. Our experiments show that abstract attributes convey novel and complementary information, improving the effectiveness of text-to-image retrieval. Practically, leveraging abstract language enables users to query visual content in natural, human-friendly terms, such as "an oversized summer-mood t-shirt" or "a sporty top with a fresh collar."

Despite its importance, abstract language is under-represented in existing Vision-and-Language Models (VLMs) [12, 81, 82]. Pre-trained general-purpose and fashion-specific VLMs struggle to capture abstract expressions, leading to poor retrieval performance for abstract-rich queries (Fig. 6.3-right). This limitation primarily arises from the pre-training data, as web-scale corpora such as LAION-400M [83] are heavily biased toward concrete descriptors, with abstract adjectives appearing infrequently. While fine-tuning on fashion datasets could potentially address this issue, dataset sizes are orders of magnitude smaller than web-scale corpora, and proprietary constraints limit the availability of abstract-rich examples. Consequently, naïve fine-tuning is insufficient to close the representation gap.

To overcome this challenge, we introduce a training-free, model-agnostic method, Abstract-to-Concrete Translator (ACT), which effectively shifts under-represented abstract language toward well-represented concrete concepts in the latent space of existing VLMs (Fig. 6.1). ACT operates in two phases:

In the *Preparation Phase*, we construct a paired Abstract–Concrete multimodal database by identifying abstract-oriented descriptions in a dataset and augmenting them with concrete captions generated via image captioning models. Dimensionality reduction is then applied to extract the primary representation differences, defining the abstract-to-concrete shift.

In the *Inference Phase*, Abstract-oriented queries are first rephrased using a Large Language Model (LLM) to approach a concrete formulation. The textual representation is then further adjusted according to the shift learned in the preparation phase, yielding embeddings that better align with the VLM latent space.

Evaluations on text-to-image retrieval tasks show that ACT outperforms both zero-shot and fine-tuned VLMs, achieving up to +12.6% improvement in H@1 over zero-shot models and +2.0% over fine-tuned models. Furthermore, the method is model-agnostic, providing consistent gains across multiple VLM architectures without requiring retraining.

6.2.1 Our contributions

The main contributions of this chapter are:

- We demonstrate the prevalence of abstract language in multimodal fashion datasets and its role in expressing complementary, discriminative information for retrieval tasks.
- We empirically show that existing VLMs under-represent abstract-oriented language, leading to degraded performance in abstract-rich queries.
- We propose ACT, a training-free and model-agnostic method to bridge the abstract-to-concrete representation gap in VLM latent spaces.

- We validate ACT through extensive text-to-image retrieval experiments, showing substantial improvements over zero-shot and fine-tuned VLMs across same- and cross-dataset settings.

6.3 Related Work

6.3.1 Fashion datasets and text descriptions

Fashion image datasets saw a shift in textual descriptions starting in the mid-2010s. Early datasets like FashionStyle [84] focused on labeling entire outfits with broad styles ("goth", "retro"). Others like [85, 86] used noisy tags from a limited vocabulary. Subsequently, datasets like iMaterialist [87], FashionMNIST [88], Fashion200K [89] adopted more fine-grained text descriptions, including attributes, class labels, and multiple weighted tags. Over the past few years, the trend has been to employ natural language descriptions, with human-annotated captions being considered in datasets such as Fashion-Gen [90], Fashion IQ [91], and KAGL [92]. However, these descriptions are aimed towards image generation [90] or pair-wise comparison tasks [91], and mainly capture fine-grained concrete properties, with retrieval not being their primary goal. Consequently, textual representations focus on visually grounded properties, namely, colors, fabrics used, and the presence of specific details or accessories. Coarser information about the garments, usually associated with abstract properties such as emotive qualities (*e.g.* "aggressive", "minimal"), is typically overlooked. More recently, fashion applications have taken advantage of aligned text-image representations and have started considering more complex descriptions [79, 80]. In DeepFashion [80], data samples are crawled from fashion e-commerce platforms. It contains descriptions involving a range of fashion trends, moods, and abstract attributes of various garment items. Similar approach characterize FACAD [79], while it features a larger scale and generally more compact descriptions.

6.3.2 Retrieval in the Fashion domain

Early fashion retrieval systems were initially used for matching real-world garments to shop advertisements [80, 93–97]. In more recent literature, some systems rely on feedback-based techniques, where users provide an image and text describing desired changes, and the system retrieves similar images reflecting those modifications [98–102]. Currently, Vision-Language Models (VLMs) like CLIP [12] have revolutionized retrieval tasks with their strong generalization power. These models leverage web-scale text-image data through contrastive learning. Recent research builds on CLIP-like models for text-only retrieval, adapting general-purpose models to the fashion domain [81, 82]. Starting from a pre-trained model on large-scale data [83], F-CLIP [81] adapts to the specialized domain by employing proprietary data. Nonetheless, from the qualitative inspection available in their publication, the textual descriptions in the training data mainly contain concrete attributes. Similarly, OF-CLIP [82] focuses on open-source data for fine-tuning to the fashion domain. However, these data only involve concrete-oriented descriptions, such as FASHION200K [89], iMATERIALIST [87], Fashion Gen [90] and Fashion IQ [91].

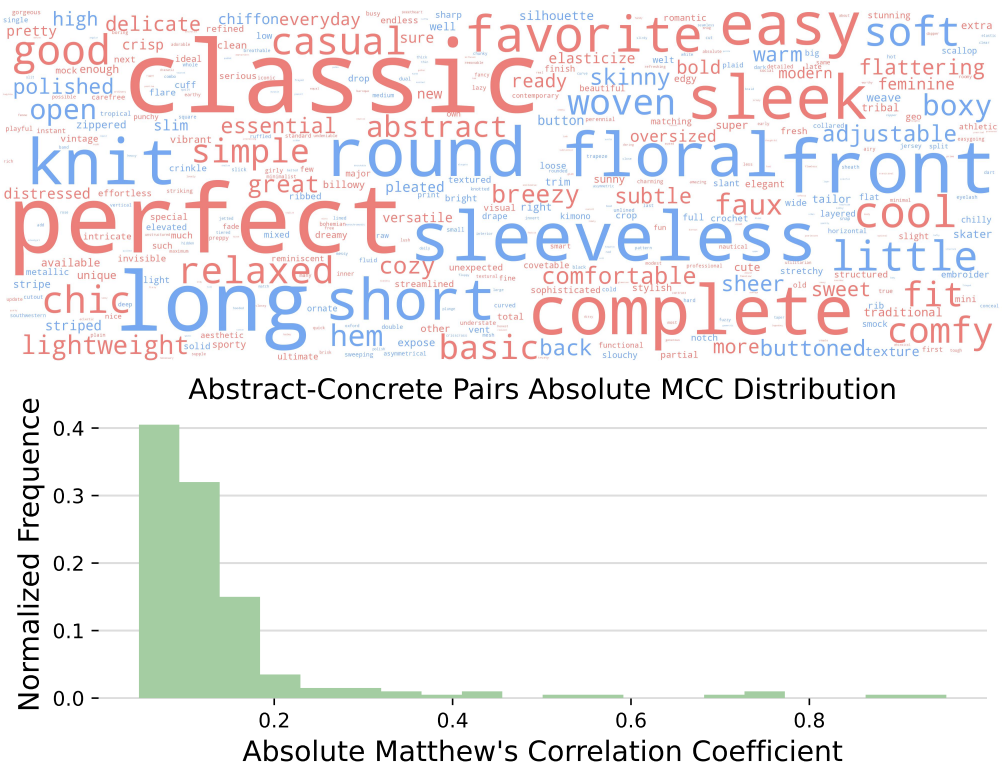


Figure 6.2: **Top:** Wordcloud of the abstract and concrete adjectives in the DeepFashion dataset. The larger font represents a higher frequency. **Bottom:** Distribution of maximum absolute Matthews's Correlation Coefficient (MCC) between each abstract attribute and concrete ones in DeepFashion. The peak near 0 reveals that the majority of abstract attributes have a low correlation with the concrete ones.

6.4 Analyzing the Language of Fashion

In this section, we analyze recent large-scale multimodal datasets in the fashion domain, and highlight four main findings: i) abstract attributes are abundantly used in the textual descriptions of fashion images; ii) abstract attributes convey novel information; iii) abstract attributes could boost retrieval performance; iv) current VLMs under-represent abstract-oriented language.

6.4.1 Datasets

We study DeepFashion [80] and FACAD [79] as popular multimodal fashion datasets, and LAION 400M [83] as one of the main multimodal datasets used to pre-train large VLMs [103, 104]. Specifically, both *DeepFashion* [80] and *FACAD* [79] contain data collected from real fashion e-stores. The items cover different garment classes, from pants, shirts, and skirts to shoes and jewelry. Each item is paired with a multi-sentence description (median 37 words) originally used by the sellers on the platform, reflecting the exact vocabulary used in the fashion domain. On the contrary, *LAION 400M* [83] is composed of web-crawled images,

Dataset	Concrete occurrences			Abstract occurrences		
	Unique	Total	Per desc.	Unique	Total	Per desc.
DeepFashion	625	10,905	3	716	14,045	3
FACAD	2795	202,883	2	2019	194,528	2
LAION 400M	14,217	26M	0	12,844	19M	0

Table 6.1: Occurrences for each adjective in different datasets. In the "Per description" column (Per desc.), the median is reported. It is clear to see how, in abstract-oriented datasets, abstract adjectives are as common (if not more) as concrete ones. In concrete-oriented datasets (100M subset of LAION 400M), abstract adjectives are the minority.

along with short descriptions, usually coming from the image metadata.

6.4.2 Attribute extraction and categorization.

We focus on *attributes* that describe certain property of an item, composed of an adjective with an associated noun (if present) (*e.g.* "relaxed neckline"). We develop an attribute extraction and categorization pipeline that leverages spaCy [105], a popular NLP tool, to localize the attributes present in the item descriptions. Finally, we categorize them into abstract and concrete based on the adjectives using a well-established *concreteness lexicon* [78]. We classify attributes to be abstract if the ratings of their adjectives are below 3.0, following the threshold strategy of [78].

i) Fashion descriptions are abstract-oriented. Tab. 6.1 reports the number of occurrences of adjectives in different datasets. The ratio between total abstract and concrete adjectives differ greatly among DeepFashion, FACAD, and LAION 400M. In FACAD and DeepFashion, abstract adjectives are as frequent (or more frequent) than concrete adjectives, while descriptions in LAION 400M are heavily skewed towards concrete adjectives.

Fig. 6.2-top presents the cloud of words of the extracted concrete and abstract adjectives of DeepFashion. The most frequent concrete adjectives are "long", "sleeveless", "front", and "round", and all of them are clearly effective in describing the fashion items. Nonetheless, fashion language isn't about objectively describing a product, it also expresses feelings, reflects preferences, and unveils desires, *i.e.* stylistic nuances that all have a very specific translation in terms of appearance. The cloud of words exhibits exactly this function of the fashion language. Abstract adjectives like "classic", "sleek", "casual" paint a picture in the reader's mind of the feeling the item conveys, while others like "relaxed" or "comfy" can be used to focus on the details of the silhouette, *e.g.* "a comfy neckline".

ii) Abstract attributes convey novel information. We aim to demonstrate that the abstract and concrete attributes carry different information when describing fashion items. Intuitively, if an abstract attribute always co-occurs with a concrete one, then they are highly correlated and convey similar information. Instead, if the abstract attribute is rarely bound

to other concrete ones, then it conveys diverse information. We thus conduct a correlation study, showing a low correlation between abstract and concrete attributes.

We limit our analysis to the set of 200 most frequent concrete $\mathcal{A}^c$ and abstract $\mathcal{A}^a$ attributes, resulting in $K = 400$ attributes considered in the correlation study. Leveraging the Matthew Correlation Coefficient (MCC) [106], we define, for each abstract attribute $\mathcal{A}^a_i$, its correlation Φ_i against concrete attributes as the maximum absolute correlation ϕ between the abstract attribute i and all the concrete ones.

$$\Phi_i = \max_{\mathcal{A}^c_j \in \mathcal{A}^c} \left| \phi(\mathcal{A}^a_i, \mathcal{A}^c_j) \right|, \tag{6.1}$$

Fig. 6.2-bottom shows the distribution of the maximum absolute MCC Φ of $\mathcal{A}^a$ on DeepFashion. The distribution is highly skewed towards left, *i.e.* most abstract attributes have a low correlation to the concrete ones. A small set of abstract-concrete pairs exists with a high correlation. By manual inspection, we notice that those pairs often refer to specific attributes jointly used. For instance, ("faux belt", "leather belt") is common because of "faux leather belt", and similarly for "invisible side zipper".

iii) Abstract attributes are useful for fashion retrieval. We investigate whether, given the same number of attributes in the text query, the description containing abstract attributes is more discriminative than that of those containing concrete ones. To limit ambiguity induced by text/image representation in the retrieval performance, we perform the retrieval task with an oracle retrieval system that is purely based on attribute matching.

Given the set D of descriptions that contain at least one of the K investigated attributes, we build the matrix P, where the rows are the descriptions $d \in D$, and the K columns report the binary presence of each of the K attributes in that description. An ideal retrieval system, once queried with a set of attributes q, will retrieve a set of items R_q where each item contains all (but not only) the attributes present in q. Each query will be a set of attributes present in the matrix P, meaning that we use each item attribute set to query the matrix itself. Let $N_q = |R_q|$ be the number of retrieved items when querying with attributes indexed by q. We can quantify how discriminative the query is by computing the retrieval precision $p_q = 1/N_q$, with a $p_q = 1$ indicating the query contains a *unique* set of attributes.

Fig. 6.3-left reports the average precision of sentences when grouping them by the number of attributes in the sentence and categorizing them based on the most present attribute categories (where mixed indicates the ratio between the number of abstract and concrete attributes is between 0.33 and 0.66). We consider sentences containing at most 5 attributes, consisting of $\sim$ 90% of the data. The figure shows that abstract attributes consistently provide more or equal information to concrete ones, as highlighted by the higher precision across a varying number of attributes in both mixed and abstract descriptions. In particular, with smaller numbers of attributes, *i.e.* one-three, we observe that abstract attributes can achieve much higher precision than concrete attributes. This suggests that abstract attributes encapsulate more specialized information than concrete ones, allowing for more faithful descriptions with fewer words.

iv) Current VLMs under-represent abstract language. We probe the capability of different state-of-the-art VLMs to represent and handle abstract-oriented descriptions. In the

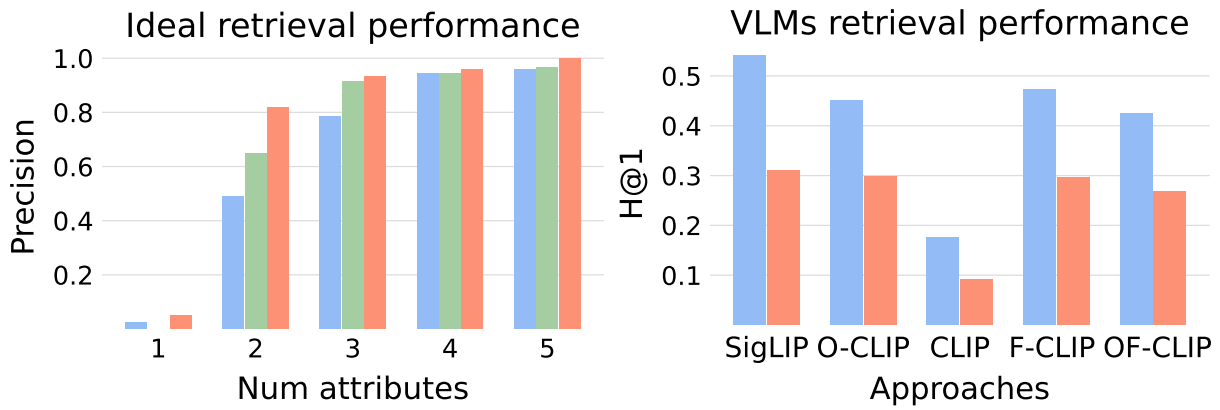


Figure 6.3: **Left:** Retrieval performance of an ideal system on Deepfashion original descriptions when the majority of present attributes are concrete, abstract or mixed. Abstract attributes allow for better retrieval performance. **Right:** performance of current VLMs on DeepFashion when using original abstract descriptions or concrete VLM generated ones. Current VLMs achieve better performance with concrete-oriented descriptions.

first step, we generate concrete descriptions from visual data using state-of-the-art captioning models. As highlighted by the large presence of concrete adjectives in Tab. 6.2, generated descriptions mainly rely on concrete attributes to convey visual information. In the second step, we evaluate the retrieval performance of a model when queried with original abstract descriptions *vs* concrete-oriented ones. We align with the retrieval literature [81, 82] and use Hit-Rate@1, which considers the retrieval successful if the image most similar to the description is correct. From Fig. 6.3-right, it is clear how both general-purpose and fashion-specialized VLMs (F-CLIP, OF-CLIP) strongly favor concrete descriptions over abstract ones, suggesting an inability to properly encode abstract attributes.

6.5 Abstract-to-Concrete Translator

The proposed Abstract-to-Concrete Translator (ACT) augments the latent representation of abstract-oriented descriptions to process fashion items textual data more effectively. Our approach is guided by the intuition that aligning these representations closer to the well-modeled concrete representations within the VLM latent space can enhance the model's performance on downstream tasks. As illustrated in Fig. 6.4, the method comprises two phases. During the *first preparation phase*, the approach characterizes the shift from abstract to concrete representations. To this end, the *database construction step* constructs an abstract-concrete (A-C) database from existing abstract-oriented multimodal data: we leverage an image captioning model to produce captions describing the images, providing paired descriptions of the same visual data. As shown in Tab. 6.2, such generated captions are more visually concrete-oriented. Later on, the *representation shift analysis step* leverages a dimensionality reduction strategy to identify the main representation shift among the paired A-C descriptions. These shift directions capture the discrepancies in the latent space between concrete-oriented and abstract-oriented descriptions. Once the shift has been characterized,

Model	Concrete occurrences			Abstract occurrences		
	Unique	Total	Per desc.	Unique	Total	Per desc.
Captioning Models						
Qwen2-VL	245	12,819	3	98	2,623	0
CogVLM2	351	19,042	5	265	10,774	2
Large Language Models						
Phi3	425	7,737	2	281	3,742	1
LLaMa 3.1 8B	405	7,778	2	199	2,675	0

Table 6.2: The number of occurrences for each adjective category in the captioned train split of DeepFashion. As we can see, both captioning models and LLMs are biased towards concrete properties.

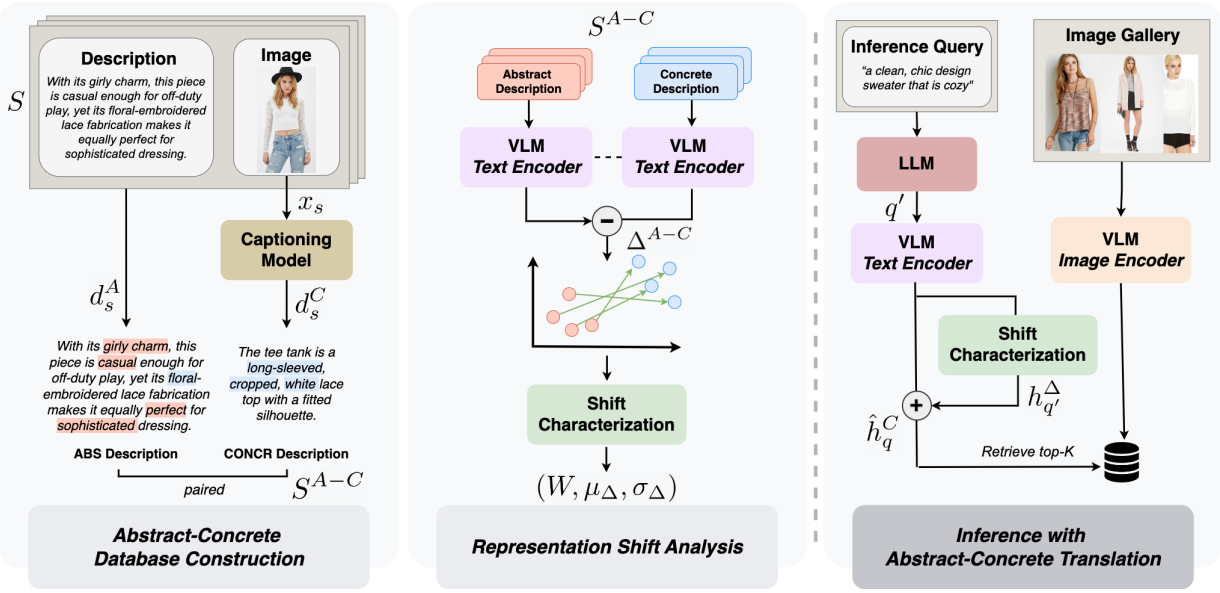


Figure 6.4: Overview of our two-phase Abstract-to-Concrete Translator (ACT). During the *preparation phase*, ACT conducts a first *database construction step*, processing an Abstract-Concrete (A-C) database by using an image captioning model to produce concrete-oriented captions describing the images. Then, in the *representation shift analysis step*, ACT analyzes the main representation shifts among the paired A-C descriptions with a dimensionality reduction strategy. During the *inference phase*, ACT first prompts a frozen LLM to rephrase the abstract-oriented description to convert the abstract-oriented language into a more concrete-oriented expression. Then, ACT enhances the VLM textual representation by compensating with the main shifts extracted from the A-C multimodal database. This allows ACT to perform better on downstream multimodal tasks with abstract-oriented language, *e.g.* text-to-image retrieval, without any training.

the later *second inference phase* can leverage it to augment all provided queries. Here, we start by utilizing a frozen LLM to rephrase the abstract-oriented description, an important step to convert the abstract-oriented language into a more concrete-oriented expression. Then, we further enhance the textual representation by shifting it towards concrete ones along the dimensions extracted from the A-C multimodal database.

6.5.1 A-C Database Construction

Starting from a multimodal database in the fashion domain containing abstract-oriented textual descriptions, we leverage a pre-trained image captioning model to construct the Abstract-Concrete (A-C) database, aiming to pair abstract and concrete descriptions associated by their images (the left block of Fig. 6.4).

Consider a multimodal database $S = \{(d_s^A, \mathcal{I}_s)\}_s$ where each clothing item s has an abstract-oriented description d_s^A and a set of corresponding images $\mathcal{I}_s$. For each item, we consider the visual sample $x_s \in \mathcal{I}_s$, and use the frozen captioning model, denoted as ψ, to output a concrete-oriented description d_s^C:

$$d_s^C = \psi(x_s, p_v), \tag{6.2}$$

where p_v is the text prompt given to the captioning model for describing the clothing item s.

Formally, we construct our A-C database $S^{\text{A-C}}$ as:

$$S^{\text{A-C}} = \{(d_s^A, d_s^C) \mid s \in S\}. \tag{6.3}$$

6.5.2 A-C Representation Shift Analysis

The paired abstract-concrete descriptions allow us to analyze the main differences between the textual representations of abstract-oriented and concrete-oriented descriptions anchored by the same visual content (the middle block of Fig. 6.4).

For each item s, we use the text encoder of a VLM f_T to encode its abstract-oriented and concrete-oriented descriptions (d_s^A and d_s^C), obtaining:

$$\begin{aligned} h_s^A &= f_T(d_s^A), \quad s \in S, \\ h_s^C &= f_T(d_s^C), \quad s \in S. \end{aligned} \tag{6.4}$$

Let H^A and H^C denote the vectorized abstract-oriented and concrete-oriented representations of all items in the database, respectively, *i.e.* the s-th row of H^A is h_s^A, and the s-th row of H^C is h_s^C. We aim to identify the dominant differences among H^A and H^C. To this end, we perform subtraction between H^C and H^A as:

$$\Delta^{\text{A-C}} = H^C - H^A. \tag{6.5}$$

We then standardize $\Delta^{\text{A-C}}$ to have zero-mean and unit variance by rescaling using statistics $\mu_\Delta, \sigma_\Delta$ denoting the shift mean and standard deviation, respectively. To extract the main

directions that best capture the differences between H^C and H^A, we apply a Principal Component Analysis (PCA) [107]:

$$W = \text{PCA}(\Delta^{\text{A-C}}, k), \tag{6.6}$$

obtaining the shift projector $W \in \mathbb{R}^{l \times k}$, where l is the dimension of the textual embeddings and k is the number of considered components. W models the main directions along which lies the information that H^A struggles to represent with respect to the concrete counterpart H^C.

6.5.3 ACT at inference

ACT converts an abstract-oriented textual representation to a more concrete-oriented one, in order to enhance the downstream tasks that require multimodal alignment (the right block in Fig. 6.4). In the context of text-to-image retrieval task, given an abstract-oriented text query q^A, we first leverage a LLM, denoted by g, to rephrase it to a more concrete version q':

$$q' = g(q^A, p_r), \tag{6.7}$$

where p_r is the rephrasing prompt.

We then embed q' with the VLM text encoder f_T, obtaining its representation $h_{q'}$. As demonstrated in Tab. 6.2, while the LLM can improve language concreteness, the rephrased text q' may still contain abstract words. Thus, we further leverage the characterization of the representation shift $(W, \mu_\Delta, \sigma_\Delta)$ obtained from the previous step to shift the representation of $h_{q'}$ towards a more concrete-oriented representation $\hat{h}_q^C$. Specifically, we compute $h_{q'}^\Delta$ from $h_{q'}$ through the representation shift characterization, as expressed below:

$$\begin{aligned} h_{q'}^\Delta &= N(h_{q'})\, WW^T * \sigma_\Delta + \mu_\Delta, \\ \hat{h}_q^C &= h_{q'} + h_{q'}^\Delta, \end{aligned} \tag{6.8}$$

where $N(\cdot)$ indicates the standardization operation using the mean and standard deviation among query representations. After the projection, the representation is then re-scaled with the Δ^{A-C} statistics (μ_Δ, σ_Δ). Intuitively, the projection into the main components along W extracts the meaningful differences between the abstract-oriented and concrete-oriented textual representations. Hence, re-scaling using Δ^{A-C} statistics projects the representation back into the VLM manifold, allowing the downstream cross-modal task. The final $\hat{h}_q^C$ combines the representation of the rephrased text $h_{q'}$ and the representation shift. Notice that all queries are augmented during the inference phase using the $(W, \mu_\Delta, \sigma_\Delta)$ that characterize the major variations from abstract to concrete representations. Due to $(W, \mu_\Delta, \sigma_\Delta)$ being pre-processed during the preparation phase, the A-C database is not required in the inference phase.

6.6 Experiments

The experiments focus on the text-to-image retrieval task, serving as a practical scenario to demonstrate how VLMs can benefit from ACT. Here, we first introduce the evaluation

protocol and implementation details. Then, we discuss the main comparison conducted on DeepFashion [80], together with ablation studies to prove the benefit of the Language Rewriting and Representation Shift modules.

Retrieval task. We follow common practice [12, 81, 82] and simulate the user querying the q-th item by using the textual description d_q, and expect the system to retrieve the images within the set $\mathcal{I}_q$. Specifically, in the first step, the visual encoder extracts embeddings for all the images in the queried collection, creating a gallery of image embeddings H_I. Then, given a query d_q, the text encoder extracts embeddings h_q, which are used to retrieve the most similar images $\hat{\mathcal{I}}_q$ by selecting the corresponding top-K most similar image embeddings from H_I under cosine similarity. Similarly, given a query d_q, ACT shifts abstract-oriented queries to obtain the final embedding $\hat{h}_q^C$, which is used to retrieve the top-K most similar images with the same procedure.

Metrics. We quantitatively evaluate the retrieval performance using commonly-used metrics [81, 82], *i.e.* Recall@K (R@K) and Hit-Rate@K (H@K). Recall@K measures the percentage of all relevant images corresponding to item q presented in the K retrieved ones $\hat{\mathcal{I}}_q$. Hit-Rate@K evaluates the number of successful queries out of all those performed. Formally, given a description d_q, we consider the query a success if at least one of the images in $\mathcal{I}_q$ is present in $\hat{\mathcal{I}}_q$. Following the protocol presented in [82], we report the results at different values of K, with $K \in \{1, 5, 10\}$, and report the average metric over all the queries.

Baselines. To assess the benefit of ACT we consider various families of VLMs as retrieval models, including SigLIP [19], CLIP [12], O-CLIP [103] and EVA-CLIP [104]. In particular, for each family of retrieval models, we consider a zero-shot evaluation of the general purpose model to the retrieval task. We further compare with their fine-tuned version to available abstract-oriented data, denoted as `<model>-ft-<data>`. As representative of state-of-the-art approaches, we compare with fashion-specialized retrieval approaches, namely F-CLIP [81] and OF-CLIP [82]. Finally, we also evaluate the change in performance when scaling to a larger number of parameters.

Implementation Details. As ACT is architecture-agnostic, we integrate the proposed strategy across several CLIP-like pre-trained models. For primary evaluations, we use `SigLIP (ViT-B-16)` [19] as the retrieval model unless stated otherwise. The A-C database builds upon an Int4-quantized `Qwen2-VL-7B` [108] as the captioning model, prompted to generate captions that focus on the fashion item of interest, specifically by querying its item class. The representation shift analysis step retains the top $k = 600$ principal components, which is an empirical choice. During inference, ACT utilizes `Llama-3.1-8B` to generate an initial concrete version of the original query.

6.6.1 Results

Zero-shot + same-dataset retrieval. In the first two blocks of Tab. 6.3, we report the results of models evaluated on the DeepFashion evaluation set. In the same-dataset setting,

Model	R@1	R@5	R@10	H@1	H@5	H@10
Zero-Shot						
SigLIP	.062	.228	.322	.311	.536	.639
CLIP	.019	.061	.091	.092	.201	.278
O-CLIP	.061	.204	.291	.298	.527	.637
EVA-CLIP	.045	.151	.218	.225	.412	.517
F-CLIP	.060	.205	.292	.297	.512	.618
OF-CLIP	.054	.181	.260	.268	.491	.606
Same-Dataset (DeepFashion → DeepFashion)						
SigLIP-ft-df	.083	**.307**	**.421**	.417	.659	.753
CLIP-ft-df	.029	.095	.142	.140	.299	.393
O-CLIP-ft-df	.080	.272	.378	.398	.640	.740
EVA-CLIP-ft-df	.063	.219	.316	.317	.539	.654
F-CLIP-ft-df	.074	.254	.350	.363	.603	.705
OF-CLIP-ft-df	.072	.245	.344	.358	.607	.712
ACT-df (Ours)	**.089**	.303	.409	**.437**	**.665**	**.756**
Cross-Dataset (FACAD → DeepFashion)						
SigLIP-ft-facad	.070	.259	.364	.352	.568	.681
CLIP-ft-facad	.023	.079	.122	.114	.265	.358
O-CLIP-ft-facad	.065	.230	.324	.324	.562	.667
EVA-CLIP-ft-facad	.049	.169	.249	.246	.457	.561
F-CLIP-ft-facad	.060	.206	.295	.295	.526	.631
OF-CLIP-ft-facad	059	.209	.298	.298	.542	.649
ACT-facad (Ours)	**.087**	**.302**	**.407**	**.428**	**.661**	**.754**

Table 6.3: Results on Deepfashion. In **bold** the best results, while underlined are the second best. ACT proves to be the best (or second best) method *w.r.t.* zero-shot and fine-tuned models in both the same-dataset and cross-dataset settings.

the fine-tuned models are trained on the DeepFashion training set, while ACT leverages the DeepFashion training set to construct the A-C database (`ACT-df`). Firstly, we note that shifting the representations from abstract to concrete allows for a large improvement with respect to the zero-shot approaches (the top block): the result holds for both general purpose models (up to +12.9% H@5 *w.r.t.* the second-best SigLIP) and fashion-specialized ones (+2.9% R@1 *w.r.t.* F-CLIP). Interestingly, despite requiring no training, ACT proves to be the best or the second best model even when compared to fine-tuned models (second block). While the higher precision and recall at $k = 1$ shows the ability of the model to select the closest sample, the benefit is especially evident in the H@K metric, where ACT achieves the

Model	Backbone	R@1	R@5	R@10	H@1	H@5	H@10
SigLIP	ViT-B-16	.062	.228	.322	.310	.536	.639
SigLIP-ACT	ViT-B-16	**.089**	**.303**	**.409**	**.437**	**.665**	**.756**
SigLIP	ViT-L-16-384	.094	.348	.473	.458	.679	.775
SigLIP-ACT	ViT-L-16-384	**.108**	**.385**	**.513**	**.527**	**.736**	**.823**
O-CLIP	ViT-B-32	.061	.204	.290	.298	.527	.637
O-CLIP-ACT	ViT-B-32	**.062**	**.210**	**.298**	**.312**	**.537**	**.655**
O-CLIP	ViT-L-14	.085	.301	.411	.412	.660	.755
O-CLIP-ACT	ViT-L-14	**.091**	**.313**	**.422**	**.447**	**.676**	**.771**
O-CLIP	ViT-H-14	.091	.319	.432	.447	.673	.772
O-CLIP-ACT	ViT-H-14	**.098**	**.336**	**.449**	**.483**	**.696**	**.789**
EVA-CLIP	EVA02-B-16	.045	.151	.218	.225	.412	.516
EVA-CLIP-ACT	EVA02-B-16	**.052**	**.175**	**.249**	**.258**	**.463**	**.564**
EVA-CLIP	EVA02-L-14	.063	.217	.307	.308	.521	.618
EVA-CLIP-ACT	EVA02-L-14	**.072**	**.252**	**.345**	**.352**	**.574**	**.676**
EVA-CLIP	EVA02-E-14	.090	.312	.423	.440	.667	.762
EVA-CLIP-ACT	EVA02-E-14	**.096**	**.334**	**.451**	**.471**	**.689**	**.787**

Table 6.4: Retrieval performance of ACT on DeepFashion when integrated on different zero-shot models. Shifting towards concrete representations consistently provides a performance boost.

largest gap (+2% H@1) *w.r.t.* the strongest fine-tuned model.

Cross-dataset retrieval. As real-life applications of fashion-domain text-to-image retrieval models should be deployed in the wild, we consider a cross-dataset evaluation where the model is required to generalize beyond its training distribution. In the bottom block of Tab. 6.3, we report the cross-dataset performance of the different models fine-tuned on FACAD [79] (`<model>-ft-facad`), and compare them to ACT with the A-C database built on FACAD. ACT's gain is consistent over previous state-of-the-art approaches on all retrieval metrics with an average improvement of 8.6% with respect to the second-best zero-shot approach (SigLIP). Notably, our cross-dataset version achieves almost identical results to the same-dataset scenario (`ACT-df`). On the other hand, we see a consistent performance drop in the fine-tuned backbones when compared to their same-dataset versions. This confirms our hypothesis, highlighting the generalization capability of our representation shift and simultaneously showcasing, once again, the difficulties of popular VLMs when faced with concepts not seen during training.

Model scaling. With the rapid growth of the VLM field, increasingly powerful models are released daily. We explore how ACT can be adapted for integration across different model families and scales. Tab. 6.4 reports the retrieval metrics on a same-dataset setting compared to zero-shot models on DeepFashion. Our results show that ACT consistently provides a

Representation Shift	Language Rewriting	R@5	H@1
✗	✗	.228	.311
✗	Llama 3.1-8B	.294	.411
Qwen2-VL	✗	.246	.347
CogVLM2	Llama 3.1-8B	.302	.433
Img-embeddings	Llama 3.1-8B	.266	.406
Qwen2-VL	Llama 3.2-3B	.259	.365
Qwen2-VL	Phi3-mini-4K	.295	.424
Qwen2-VL	Llama 3.1-8B	**.303**	**.437**

Table 6.5: Ablation analysis on the components of ACT, with performance on the retrieval task on DeepFashion. In green, the parameters used for our ACT. Both the Language Rewriting and Representation Shift steps are needed to get the best performance. Furthermore, ACT is robust to different component choices.

performance gain across all approaches (with an average of +4.9% H@1 over all backbones) and model sizes (*e.g.* on `EVA-CLIP`, ACT provides an average of +3.6% H@1 across all model dimensions). We hypothesize that, due to the concreteness of data used during model pre-training highlighted in Sec. 6.4, larger pre-trained models still under-represent abstract concepts, justifying ACT as a necessary approach to deal with the representation shift.

Ablation analysis. We conduct ablation experiments to evaluate the contribution of ACT pipeline components. Tab. 6.5 reports the retrieval performance on DeepFashion. As can be noted from the first section, both the Representation Shift and the Language Rewriting provide a benefit to the original model retrieval performance with a +3.7% and 10% H@1 gain, respectively. While properly prompted LLMs can provide an effective first estimate of the concrete counterpart of an abstract-oriented description, ACT further boosts the performance by explicitly accounting for the representation shift. In the second section of Tab. 6.5, we experiment with alternative sources of concrete information, such as captions from `CogVLM2` [109] and the direct use of visual embeddings (`Img-embeddings`) coming from the VLM image encoder. Results show that on one side ACT is robust to captions coming from another state-of-the-art VLM, while on the other, moving from textual to visual concrete information leads to performance degradation. We hypothesize the modality gap [110] between visual and text-embedding may dominate the shift. Finally, we ablate the language-rewriter choice with different open-source instructed LLMs, including `Llama-3.1-8B`, `Llama-3.2-4B` and `Phi3-mini-4k`. ACT proves robust to the choice of the language rewriter, achieving consistently higher performance over the pre-trained backbones.

6.7 Conclusions

In this work, we explored the role of abstract-oriented language in Vision-and-Language Models (VLMs), focusing on the fashion domain and leveraging multimodal datasets. Our analysis demonstrated that abstract language is both prevalent and informative, yet underrepresented in standard pre-training corpora, which limits VLM performance on abstract-rich retrieval tasks.

To address this limitation, we proposed Abstract-to-Concrete Translator (ACT), a training-free and model-agnostic approach that shifts abstract textual representations toward well-represented concrete concepts in the VLM latent space. Empirical evaluations on text-to-image retrieval show that ACT consistently outperforms zero-shot and even fine-tuned VLMs in same- and cross-dataset settings, highlighting the potential of language-aware representation enhancement in multimodal models.

These results motivate the next stage of this thesis: investigating how to better evaluate and control VLM outputs, particularly in complex multimodal generation tasks. In the following chapter, we introduce L-VQAScore, a fine-grained evaluation framework that combines localization and VQA-based probing to measure entity-attribute alignment in generated images. This work builds directly on ACT's findings by providing tools to quantify the effective utilization of abstract and concrete language in VLM outputs, enabling more interpretable, reliable, and human-aligned evaluation metrics.

Limitations & Future work. Our analysis primarily focuses on adjectives as indicators of abstract expressions, providing a first step toward systematic exploration. Extending this work to other domains such as art, movies, or interior design, where abstract language is also prevalent, would further validate and generalize our approach.

Broader Societal Impacts. Working with fashion datasets requires careful attention to ethical considerations, including anonymization, bias mitigation, and fair representation. These practices are crucial for ensuring that language-guided multimodal AI tools remain socially responsible and inclusive.

Chapter 7

Evaluating Attribute Confusion in Fashion Text-to-Image Generation

7.1 Abstract

Building on the findings of the previous chapter, which highlighted the importance of abstract-oriented language in Vision-and-Language Models (VLMs), we now turn our attention to *evaluating how effectively multimodal models utilize textual information* in complex generative tasks. Despite the rapid advances in Text-to-Image (T2I) generation, existing evaluation methods remain limited, particularly in domains like fashion, which require precise compositional alignment between entities and their attributes. Standard automated metrics often fail to capture *attribute confusion*, where visual attributes are correctly depicted but associated with the wrong entities, and rely on coarse cross-modal alignment rather than fine-grained semantic understanding.

To address this, we introduce Localized VQAScore (L-VQAScore), a novel evaluation framework that combines *visual localization with VQA probing*, measuring both correct attribute generation (reflection) and mislocalized attributes (leakage) on a per-entity basis. We complement this with a *localized human evaluation protocol* to validate fine-grained attribute alignment in generated images. Experiments on a newly curated fashion dataset with challenging compositional scenarios show that L-VQAScore *outperforms state-of-the-art T2I evaluation methods* in terms of correlation with human judgments. Our results demonstrate that L-VQAScore provides a *reliable, scalable, and fine-grained measure of entity-attribute alignment*, enabling systematic evaluation of how effectively multimodal models leverage both concrete and abstract language in generation tasks.

This work extends the thesis narrative from enhancing language representation (ACT in Chapter 6) to *measuring and quantifying its effective use in multimodal outputs*, laying the foundation for subsequent studies on controllable generation and interactive language-guided synthesis.

7.2 Introduction

Building on the previous chapter on ACT, which highlighted the importance of abstract-oriented language in Vision-and-Language Models (VLMs), we now turn to the challenge of evaluating how effectively multimodal models leverage textual inputs in complex generative tasks. In particular, we focus on Text-to-Image (T2I) generation, where models must produce images that not only align with the conditioning text but also correctly associate attributes with multiple entities.

Recent T2I models [15, 73] can generate highly detailed and semantically rich images from natural-language prompts. Concurrently, evaluation methods have evolved to measure *alignment between text and generated image* [16, 111–117]. Most automated metrics leverage pre-trained VLMs, either through cross-modal embedding similarity (*e.g.*, CLIPScore [111]) or via Visual Question Answering (VQA) [16, 115–117], as illustrated in Fig. 7.1. However, prior work has revealed that VLMs often behave like *bag-of-words models* in cross-modal understanding [118, 119], limiting their ability to evaluate compositional semantics with complex entity-attribute bindings. This limitation is particularly relevant in domains like fashion, where multiple garments (entities) and attributes (colors, patterns, styles) co-occur.

VQA-based metrics partially address this problem by probing whether attributes are correctly reflected on their corresponding entities [115–117]. Yet, as our preliminary study shows, these methods struggle with attribute confusion, where attributes are generated correctly but applied to the wrong entity.

Human evaluation remains the gold standard [112–114], but traditional global Likert-scale ratings are inconsistent: our study found that, on average, $\sim 40\%$ of the time, evaluators disagree about image-text alignment. Interestingly, localized assessments—asking about specific entity-attribute pairs—substantially increase agreement, highlighting the importance of localized evaluation for capturing attribute confusion.

Motivated by these insights, we propose a human evaluation protocol and an automatic, localization-aware T2I metric, Localized VQAScore (L-VQAScore), specifically designed to measure attribute confusion. For human evaluation, we frame each entity-attribute binding as a binary *reflection question*, *e.g.*, `Is the blazer floral?`", and additionally pose *leakage questions* to detect misassigned attributes, *e.g.*, `Is the blazer gold?`". For automated evaluation, L-VQAScore applies the same principles using segmentation masks to localize visual regions and VQA models to answer both reflection and leakage questions. This approach provides a fine-grained, reliable measure of attribute alignment in T2I outputs.

7.2.1 Contributions

The contributions of this work are four-fold:

- We identify and validate the overlooked attribute confusion problem in T2I evaluation, providing a dataset and evaluation protocol covering both automated metrics and human assessment.
- We show that visual localization and attribute-centric VQA are effective strategies for addressing attribute confusion.

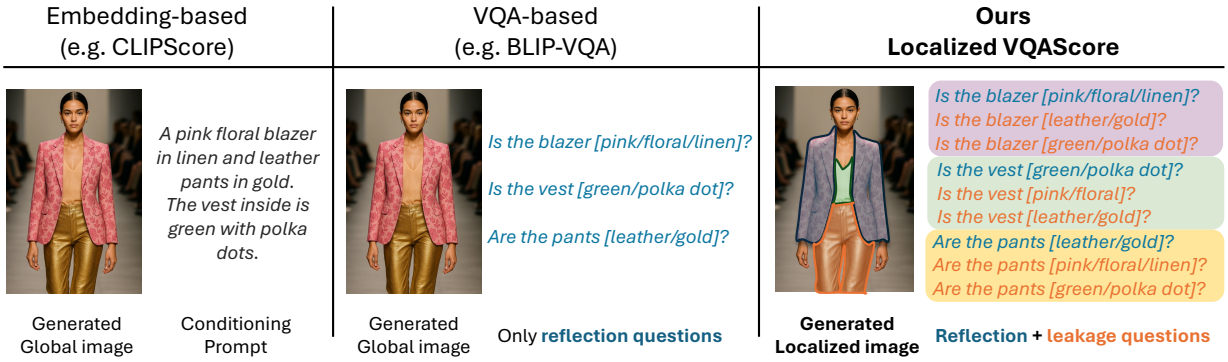


Figure 7.1: Text-to-Image evaluation in compositional prompts, particularly in fashion. Existing embedding-based metrics (*e.g.* CLIPScore [111]) struggle with entity-attribute bindings. Recent VQA-based methods (*e.g.* BLIP-VQA [115], VQAScore [16]) improve compositional understanding by probing attribute reflection globally, but fail to capture *attribute confusion*, where attributes are misassigned. Localized VQAScore addresses this via localized reflection and *leakage questions*, explicitly checking attribute assignment at the entity level.

- We introduce a new human evaluation protocol and an automated T2I evaluation method, L-VQAScore, leveraging reflection and leakage questions on localized visual regions.
- L-VQAScore mitigates attribute confusion, achieving higher correlation with human judgments than state-of-the-art metrics, providing a scalable and interpretable evaluation framework for compositional T2I tasks.

7.3 Related Work

7.3.1 Text-to-image Generation

Recent advances in T2I generation have been largely powered by diffusion models [71], which synthesize high-quality visuals by simulating a noise-injection (forward process) followed by iterative denoising (backward process) [15, 72, 120]. Various conditioning strategies have been introduced to improve control over generation, with the most popular being text-guidance. In particular, GLIDE [120] introduces classifier-free guidance; Stable Diffusion (SD) [15] further introduces cross-attention-based conditioning and performing diffusion in latent space. More recently, Diffusion Transformers (DiT) [121], a novel architecture based on Vision Transformers [122], improves generation quality through effective parameter scaling. [121] popularizes the modeling of a Rectified Flow training objective, substantially increasing generation quality, speed, stability, and prompt adherence. In this work, we will use recent pre-trained state-of-the-art text-to-image models [73, 121, 123–125] to investigate the evaluation of attribute confusion.

7.3.2 Text-to-image Evaluation Metrics

The evaluation of T2I models has been a longstanding challenge due to the complex interplay between semantic fidelity, visual quality, and compositionality.

Visual quality. Classical perceptual metrics such as FID [74] has been extensively used for assessing image realism, but fail to capture semantic alignment with the input text.

Global text-image alignment. More recent approaches leverage contrastive Vision-Language Models (VLMs), notably CLIPScore [111], which computes cosine similarity between text and image embeddings. However, such metrics are sensitive to lexical variations and often act as "bag-of-words" encoders, thus fail in capturing nuanced relations and attribute bindings [118, 119].

Compositional text-image alignment. To address these limitations, proposed metrics investigate human-feedback strategies [112–114], or leveraging Visual Question Answering (VQA) models for automatized global [16, 126] or localized [115–117] scoring. However, they fail to assess information leakage, and remain prone to attribute confusion, as the visual backbone still processes the full image.

Human evaluation. Human evaluation remains a crucial yet inconsistently implemented component in assessing text-to-image generation models [127, 128]. Key details such as inter-rater agreement are frequently under-reported, and evaluation criteria are often inconsistent and subjective, raising concerns about the reliability and reproducibility of evaluations. Most implemented user studies focus on overall visual quality [15, 120, 129] and global text relevance [120, 129], while rarely any assess localization or compositional correctness, overlooking finer-grained issues such as attribute confusion [127]. These emphasize the need for a more standardized and consistent protocol to capture both global quality and fine-grained localization in human evaluation practices.

7.4 On T2I Evaluation of Attribute Confusion

This section first formally defines the attribute confusion problem and introduces the data used in our T2I evaluation analysis. Then, we present preliminary studies showing that: (i) localization enables less subjective evaluation when complex attributes are present in the T2I evaluation; and (ii) recent T2I evaluation methods focusing on semantic alignment fail to catch attribute confusion.

7.4.1 Attribute confusion

We formally define attribute confusion, a critical problem limiting accurate semantic visual-text alignment [72, 130, 131]. Let P be the textual prompt expressed in natural language and let S denote its structured version, $S = \{(e_i, A_i) : i = 1, .., N\}$ a set of $N \geq 0$ entities with each entity e_i associated to a set A_i of K_i attributes. We refer to attribute confusion

when an attribute $a \in A_i$ is associated to a different entity $e_j, j \neq i$ in the generated image. Attribute confusion occurs when a visuo-textual model misassigns attributes to irrelevant regions within an image, resulting in semantically inaccurate results. For instance, when P is expressed as "`a pink blazer and gold pants`", attribute confusion occurs when the T2I model generates the image with pink pants instead of gold pants. Note that while the attribute confusion problem impacts T2I generative models, its automated evaluation requires metrics effectively recognizing correct entity-attribute associations.

7.4.2 Evaluation data

To conduct human studies and investigate the effectiveness of T2I metrics on attribute confusion, we construct our evaluation data based on Fashionpedia [132], a multimodal dataset with fashion images paired with structured item-attribute annotations, suitable for our analysis on attribute confusion. We select images containing at least two large garments to facilitate easy visual inspection. Recognizing specific fashion attributes, *e.g.*, "`notched lapel`", "`set-in`", might not be straightforward for non-expert users. Thus, we ensure that for each large garment, at least one pattern attribute that is easily recognizable, *e.g.* "`striped`" or "`dotted`", is included in the attribute list. Moreover, we ensure each pattern attribute appears only on a single clothing item to enable easy identification of attribute confusion. For each image, we then concatenate the garments' attributes with their class labels (*e.g.*, "`a striped, notched lapel, long-sleeve shirt`") and build a conditioning prompt by appending all garment descriptions (*e.g.*, "`a striped [...] shirt. a pair of dotted [...] pants`"). With the textual description as conditioning prompt, we then generate images using 5 state-of-the-art T2I models, namely `FLUX.1-dev` [123], `SD-3-medium` [121], `SD-3.5-large` [124], `SDXL` [73], and 4-bit quantized `HiDream-I1-Full` [125]. Overall, our evaluation data contains 50 outfit descriptions, each featuring at least two entities with unique patterns, and 250 generated images (5 per description).

7.4.3 Localized Assessment Improves Agreement in Human Evaluation

As our human study serves as the reference in T2I evaluation, we first investigate how an existing human evaluation protocol handles attribute confusion. We then explore effective strategies exploiting visual localization to improve human evaluation, serving as important inspiration for devising an automated T2I evaluation method.

Baseline Likert human study. We first conduct a baseline user study following Likert 1-5 evaluation, a representative protocol widely adopted in T2I evaluation [127]. Specifically, a user evaluates how closely the presented global image aligns with the complete conditioning prompt, on a scale from 1 (weak alignment) to 5 (strong alignment).

Localized human study. Following the intuition that attending to localized regions could encourage accurate attribute-level assessment, we devise a new human study protocol requiring users to focus on specific entities in the generated image. For an entity, we probe

User Study	**Agreement** (↑)
Likert [16, 115, 127]	63.5
Localized (Ours)	**93.2**
Method	**Failure** (↓)
CLIPScore [111]	46.1
PickScore [112]	29.2
HPSv2Score [113]	29.2
ImageReward [114]	6.15
BLIP-VQA [115]	4.62
VQAScore[16]	4.62
L-VQAScore (Ours)	**0.00**

Table 7.1: Pilot study on current evaluation. **Top:** Agreement rates for user human evaluation studies. **Bottom:** Failure rate of current T2I evaluation metrics, measured as the percentage of test cases where attribute-swapped pairs receive higher scores.

the users with both *reflection questions* and *leakage questions*, where *reflection questions* verify if an attribute is correctly depicted on the entity, while *leakage questions* check if attributes belonging to other entities are miss-localized on the investigated entity, explicitly indicating occurrences of attribute confusion (as demonstrated in Fig. 7.1 (rightmost)).

Result discussion. On our evaluation data, we collected 1042 and 833 Question-Answer pairs for the baseline Likert and the proposed localized human study, respectively. Then, we measure the agreement among users. Specifically, for each image-question pair, we identify the majority choice among annotators and define user agreement as the average ratio of annotators selecting the majority option.

As shown in Tab. 7.1 (Top), the classic Likert protocol yields lower agreement, indicating that assessing the alignment between global description and generated image can be challenging and subjective for human evaluators. Complex compositional prompts with multiple attributes often introduce confusion. Moreover, each evaluator has individual preferences and interpretations of the rating criteria, resulting in high variations. Differently, the new localized user study achieves a higher agreement, reaching approximately 93%, a 30% improvement with respect to Likert. *Localized assessment helps in reducing the complexity and subjectivity in user answering.* Attending to one specific entity at a time allows users to provide more accurate and consistent evaluation on whether an attribute is reflected or leaked.

7.4.4 Existing T2I Metrics Fail on Attribute Confusion

We further examine how state-of-the-art T2I evaluation metrics handle attribute confusion. We consider both embedding-based metrics (CLIPScore [111], PickScore [112], HPSv2Score [113] and ImageReward [114]) and VQA–based metrics (VQAScore [16] and BLIP-VQA [115]), as detailed in Sec. 7.6.1. We conduct a controlled attribute-swapping test. Specifically, for

each image-description pair, we swap the attributes belonging to different entities in the description, forming a negative description with swapped attributes, *e.g.*, "`a dotted dress and a striped shirt`" becomes "`a striped dress and a dotted shirt`". The negative description maintains the same entities and attributes, but with incorrect entity-attribute associations.

We evaluate the metrics of all compared methods using both the correct description and the negative description. We define a metric failure to address attribute confusion when it yields higher alignment scores to images paired with negative descriptions than the correct descriptions. We report the failure rate in Tab. 7.1 (Bottom). As expected, embedding-based metrics such as CLIPScore [111], PickScore [112] and HPSv2Score [113] inherit the bag-of-words problem from the underlying CLIP [12] backbone, yielding a significantly high failure rate of 46.1%, 29.2%, and 29.2%, respectively. VQA-based metrics, while mitigating attribute confusion, still present a noticeable ratio of attribute confusion failures. By adopting a localized VQA strategy, inspired by our localized human evaluation, the proposed L-VQAScore can avoid failures due to attribute confusion completely on our evaluation data. We describe the metric in detail in the following section.

7.5 Localized VQAScore

We present Localized VQAScore, a novel VQA-based T2I evaluation metric, evaluating attribute reflection and confusion on localized visual content. L-VQAScore scoring can be summarized in three key steps: (i) *localizing the queries* via automatic segmentation, blurring, and cropping of the area of interest, effectively enhancing localization focus (Sec. 7.5.1); (ii) *scoring the presence of attributes* by querying a state-of-the-art VQA model with both *reflection* and *leakage* questions (Sec. 7.5.2); and (iii) *metric computation* accounting for both attribute reflection and confusion (Sec. 7.5.3). We refer to Fig. 7.2 for an illustration.

7.5.1 Localizing the Queries

To enable more fine-grained evaluation of visual elements and mitigate the problem of attribute confusion evaluation, we introduce a *query localization* approach that explicitly links each attribute to its expected visual region. Unlike prior methods that operate over the entire image, our method enforces spatial localization by leveraging entity-level masks obtained from pre-trained segmentation models.

Consider a conditioning textual prompt P, describing the desired output image, and its structured version S. For each entity $i, i = 1, .., N$, a segmentation mask $M_i \in \{0, 1\}^{H \times W}$ is generated by a pre-trained segmentation model prompted with the entity class e_i. Formally, a segmentation model ϕ segments the generated image $x \in \mathbb{R}^{3 \times H \times W}$:

$$M_i = \phi(x, e_i) \tag{7.1}$$

where M_i is the output segmentation mask localizing the class of interest e_i in the visual space. Then, to minimize the influence of irrelevant visual context on the prediction, we apply a *spatially localized* blurring operation outside mask M_i:

$$\hat{x_i} = M_i \odot x + (1 - M_i) \odot \tilde{x} \tag{7.2}$$

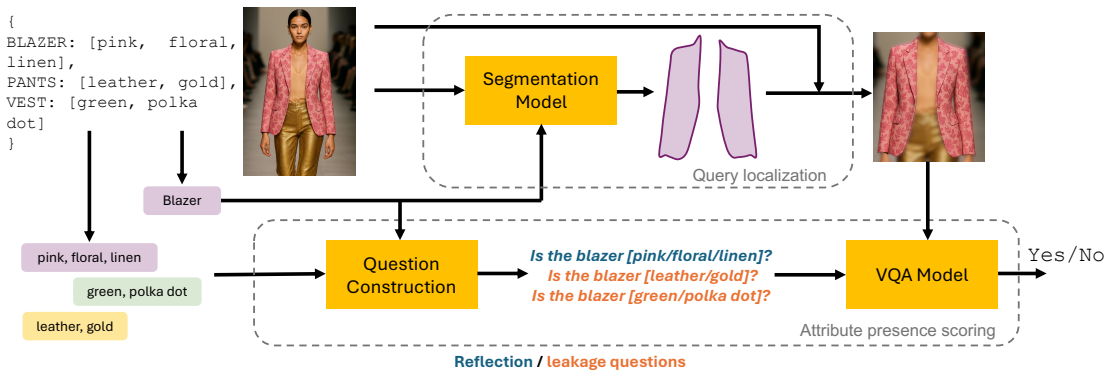


Figure 7.2: The pipeline of the proposed L-VQAScore in measuring the alignment between the conditioning prompt and the generated image. We represent the conditioning text into structured entity-attribute pairs. L-VQAScore localizes regions of interest leveraging entity categories via a semantic segmentation module. Then reflection and leakage questions are composed to evaluate the presence of desired and leaked attributes in the localized regions, accounting for both attribute depiction and localization.

where $\odot$ denotes the Hadamard product and $\tilde{x}$ is the processed version of the original image, *i.e.*, its blurred version. Finally, we calculate the bounding box b_i containing mask M_i and crop the image around it with some extra margin to accommodate for small segmentation errors. We maintain its original resolution by re-shaping the crop:

$$x_i = \text{Resize}(\text{Crop}(\hat{x}_i, b_i), H, W) \tag{7.3}$$

where x_i is the final image, $\text{Crop}(\cdot, \cdot)$ is a function that returns image $\hat{x}_i$ cropped around the bounding box b_i, and $\text{Resize}(\cdot, \cdot, \cdot)$ is a function that resizes an image to have the longest side matching the given dimension (H, W) and applies white padding to achieve the desired ratio.

7.5.2 Scoring the Presence of Attributes

Building on the localization strategy, we then evaluate whether each attribute is both correctly *reflected* and *localized* in the generated image, without attribute confusion. To this end, we employ a Visual Question Answering (VQA) model, and ask visual questions on localized regions.

The set of reflection questions Q_r is constructed from the structured prompt S to verify that attributes explicitly mentioned in the conditioning prompt are correctly reflected within their expected spatial regions:

$$Q_r = \{t(e_i, a, x_i) : \forall a \in A_i, i = 1, .., N\}, \tag{7.4}$$

where $t(e_i, a, x_i)$ denotes the templated question construction asking about attribute a in entity e_i using the spatially localized visual input x_i.

Leakage questions are designed to assess the presence of undesired attributes leaked from other entities, revealing attribute confusion. Formally, the set of leakage questions is constructed as:

$$Q'_l = \{t(e_i, a, x_i) : \forall a \in A_j, i, j = 1, .., N, j \neq i\}. \tag{7.5}$$

We then remove conflicting questions when the same attribute appears for different entities, resulting in $Q_l = Q'_l \setminus Q'_l \cap Q_r$. Finally, building on prior work [16], for all questions $q \in Q = Q_l \cup Q_r$, we evaluate the VQA model *predicted label* in terms of the probability of positive answer to each existence question, given its associated image region as $Prob(\text{“Yes”} \mid q)$.

7.5.3 Metric Computation

To quantitatively assess T2I alignment, we assign a *target label* to each question based on its semantics, *i.e.*, positive for reflection and negative for leakage questions. We then compare the predicted labels to the target ones: a positive reply to a reflection question is counted as a True Positive, whereas a negative answer is counted as a False Negative, *i.e.*, the attribute should have been generated, but it was not. Similarly, for leakage questions, a negative reply from the VQA model is counted as a True Negative, while a positive one is counted as a False Positive, *i.e.*, the attribute was not supposed to be generated in this location. The Positive and Negative answers are the presence/absence of the attributes inside the generated image, thus, only dependent on the generative model used to create x. The final metric evaluates classic *Precision, Recall and F1 Score*, with precision reflecting the model's ability to avoid hallucinated or misattributed features, and recall capturing the model's ability to faithfully realize and localize intended attributes. The F1 Score offers a balanced assessment that jointly penalizes omissions and incorrect insertions. With little abuse of notation, in the following, we generally refer to L-VQAScore as the collection of both Precision, Recall, and F1 Score metrics, with the specific metric clarified from the context.

7.6 Experiments

We first compare L-VQAScore against existing T2I evaluation methods by measuring their correlation to subjective evaluation, following the localized human study protocol and the evaluation data introduced in Sec. 7.4. Next, we present ablation studies to analyze the major designs of L-VQAScore.

7.6.1 Comparative Evaluation

Baselines. We consider embedding-based methods including CLIPScore [111] and its human preference–aligned variants, PickScore [112] and HPSv2Score [113], as well as ImageReward [114]. For VQA–based methods, we consider VQAScore [16], evaluating the global alignment through a single visual question, and BLIP-VQA [115], which instead assesses alignment by focusing on individual attributes.

Performance measures. We employ rank correlation measures to quantify the agreement between automatic metric rankings and the one from the proposed localized human study. The human study images are randomly divided into 25 groups and ranked based on their group F1 Score(Precision/Recall). Similarly, automated metrics rank groups according to their average image-level scores. Results are averaged over 5 different random seeds. We

Metric	Spearman's Rho (↑)	Kendall's Tau (↑)
Localized Study F1 Score		
CLIPScore [111]	.460	.326
PickScore [112]	.433	.293
HPSv2Score [113]	.215	.141
ImageReward [114]	.494	.349
VQAScore [16]	.704	.536
BLIP-VQA [115]	.636	.492
L-VQAScore (Ours)	**.818**	**.650**
Localized Study Precision		
VQAScore [16]	.658	.504
L-VQAScore Precision (Ours)	.722	.567
Localized Study Recall		
VQAScore [16]	.547	.413
L-VQAScore Recall (Ours)	.768	.670

Table 7.2: Performance in T2I alignment regarding the localized study F1 Score, Precision and Recall. L-VQAScore consistently surpasses existing state-of-the-art methods.

report both Spearman's Rho (↑) to capture global ranking patterns and Kendall's Tau (↑) to capture the pairwise ranking consistency.

Results. Tab. 7.2 presents the performance of L-VQAScore and state-of-the-art approaches. The comparison specifically evaluates the extent to which each method aligns with human study results based on model rankings regarding F1 Score, Precision and Recall. Accounting for F1 Score, we observe that, overall, global metrics such as CLIPScore and HPSv2Score exhibit lower performance. Among these embedding-based methods, ImageReward, which incorporates fine-tuning of a regressor on human preference data, demonstrates relatively improved performance. Notably, L-VQAScore consistently outperforms the strongest VQA-based metrics, VQAScore and BLIP-VQA, by around 16% and 26% respectively. The results demonstrate the effectiveness of L-VQAScore building on localized assessment to probe attribute confusion. Similarly, the improved correlation is observed with Localized Study Precision/Recall metrics, where VQAScore fails to catch the nuances of reflection and leakage questions.

7.6.2 Ablation Study

Localization strategy. We investigate how different localization strategies affect L-VQAScore correlation with the localized human study: i) without localization, ii) segmentation-based *masking* by blacking out the non-target regions, iii) *blurring* of surrounding context, iv) *cropping* according to the mask bounding box, as well as v) sequentially *Masking and Cropping*, and vi) the proposed *Blurring and Cropping*. We further experiment with *Blurring and Cropping w/ OV-SEG*, relying on OV-SEG [135] for segmentation, as opposed to our adopted Grounded-SAM-2 [136, 137]. We report results in Tab. 7.3 (Top). Segmentation-based *Blurring* consistently outperforms hard exclusion techniques such as *Masking* and *Cropping*, while our *Blurring and Cropping* yields further improvements. We hypothesize that this combination effectively balances the trade-off between context preservation and spatial focus. Crucially, the choice of the segmentation model affects the metric alignment to

Localization Strategy	Spearman's Rho (↑)	Kendall's Tau (↑)
✗	.549	.416
Masking	.697	.527
Blurring	.742	.579
Cropping	.682	.519
Masking and Cropping	.723	.553
Blurring, Cropping w/ OV-SEG	.682	.546
Blurring, Cropping **(Ours)**	**.818**	**.650**

VQA Model	Spearman's Rho (↑)	Kendall's Tau (↑)
LLaVA-v1.5-Vicuna-7b [133]	.660	.500
InstructBLIP-Flan-T5-xl [134]	.715	.570
CLIP-Flan-T5-xxl [16]	**.818**	**.650**

Table 7.3: Ablation analysis on L-VQAScore. **Top:** the effect of localization strategy. **Bottom:** the choice of VQA model.

user evaluation. Grounded-SAM-2 [136, 137] is superior to OV-SEG in localization accuracy. Qualitative investigation suggests that OV-SEG struggles to segment fashion-related entities.

VQA model. As shown in Tab. 7.3 (Bottom), we examined the impact of the VQA models. Consistent with other VQA-based methods [16, 115], the underlying VLM model has a notable impact on results. Our approach is orthogonal to the choice of VQA model, allowing flexible integration with any backbone and implying that future stronger VQA models can further enhance L-VQAScore.

7.7 Conclusions

In this work, we identified and addressed a critical gap in Text-to-Image (T2I) evaluation: the attribute confusion problem, where attributes are correctly generated but misassigned to entities. By introducing a localized evaluation framework based on spatially-aware Visual Question Answering (VQA), our proposed L-VQAScore metric provides fine-grained, entity-level assessment of generated images. Our experiments show that both existing automated metrics and traditional human evaluation protocols fail to capture these subtle semantic mismatches, especially in multi-entity compositional scenarios. To overcome this, we designed a novel human evaluation protocol specifically targeted at reflection and leakage of attributes. L-VQAScore achieves stronger correlation with human judgments than prior metrics, providing a reliable, automated solution for detecting mislocalized attributes.

These findings parallel the development in the next chapter: controllable image generation conditioned on structured and compositional inputs. While L-VQAScore allows us to measure whether a model correctly associates attributes and entities, the following chapter focuses on guiding T2I generation through both global textual descriptions and localized, structured conditioning, such as sketch-text pairs, to produce fully controllable and semantically aligned outputs. This progression of evaluation and control enables not only the measurement of compositional fidelity but also its direct enforcement in generative models.

Chapter 8

LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing

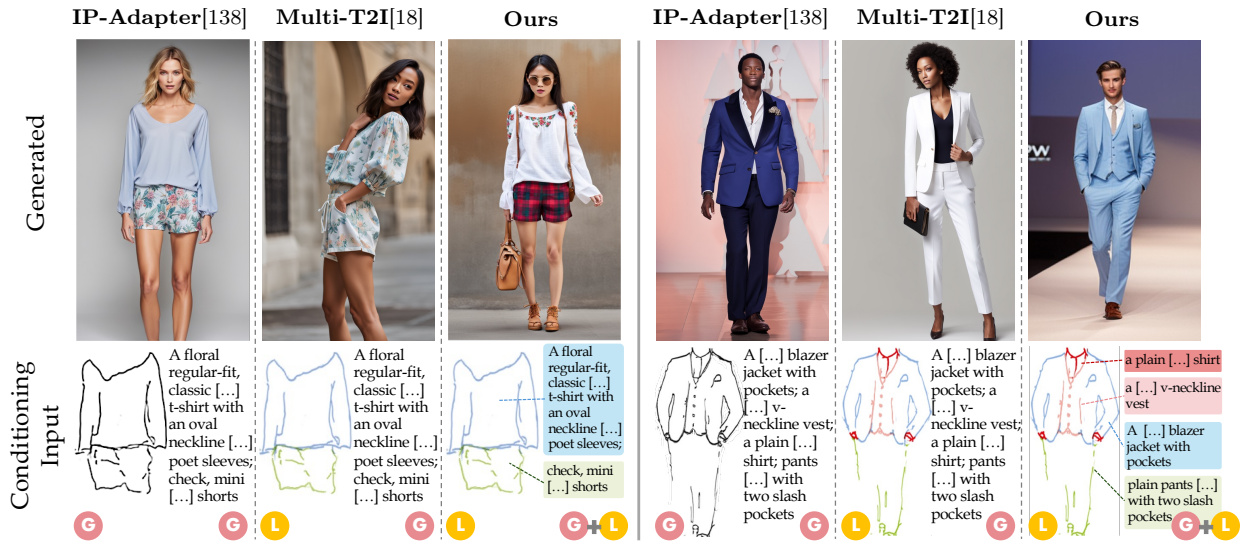


Figure 8.1: We present LOTS, enabling fashion image generation with an unprecedented level of control. LOTS represents the natural evolution of fashion design methodologies, progressing from global text and sketches (IP-Adapter [138]) to localized sketches with global text (Multi-T2I [18]). Our approach leverages a global description (omitted here for brevity) alongside a set of localized sketch-text pairs (the coloured boxes), effectively defining both the layout and appearance of individual garment items.

8.1 Abstract

Fashion design is a highly creative process that integrates both visual and textual expressions. Designers convey ideas through sketches, which define spatial structure and design elements, and through textual descriptions, which capture material, texture, and stylistic nuances. Building on the emphasis on abstract-oriented language (ACT) and compositional evaluation (L-VQAScore), we introduce LOcalized Text and Sketch for fashion image generation (LOTS),

a framework for compositional sketch-text based generation of complete fashion outfits. LOTS combines a global textual description with paired localized sketch + text information for conditioning, employing a step-based merging strategy to adapt diffusion models. First, a Modularized Pair-Centric representation encodes sketches and text into a shared latent space while preserving independent localized features. Then, a Diffusion Pair Guidance phase integrates local and global conditioning via attention-based guidance during the multi-step denoising process. To support evaluation, we extend Fashionpedia with Sketchy, the first fashion dataset providing multiple text-sketch pairs per image. Quantitative results demonstrate that LOTS achieves state-of-the-art performance on both global and localized metrics, while qualitative examples and a human evaluation study reveal unprecedented control and customization in generated fashion designs.

This work, together with L-VQAScore in Chapter 7, highlights parallel directions in the second half of the thesis: evaluating compositional fidelity and enabling controllable, semantically aligned generation in multimodal fashion modeling.

8.2 Introduction

In fashion design, designers often need to translate abstract inspirations into forms that are naturally understandable, such as sketches or natural language descriptions. For instance, a designer might sketch a t-shirt, complemented with a description like *a floral, regular-fit, classic t-shirt with an oval neckline and wrist-length poet sleeves*", or shorts as *check, mini, symmetrical, gathering shorts with a regular fit*" (Fig. 8.1). Sketches and textual descriptions convey complementary information: sketches define spatial layout and silhouette, while textual details provide rich style and accessory information. Complete outfits consist of multiple garments, each with localized descriptions, enabling fine-grained control over design attributes.

We address the task of converting localized sketch-text conditions into coherent fashion images, framing it as a conditional image generation problem. Existing methods for multi-localized conditioning [17, 18, 138, 139] often rely on a single global textual description, which can lead to attribute confusion, where properties of one garment are incorrectly assigned to another (e.g., a floral t-shirt pattern appearing on the shorts in Fig. 8.1).

To overcome this, we introduce LOcalized Text and Sketch for fashion image generation (LOTS), a framework that leverages multiple localized sketch-text pairs. First, the Modularized Pair-Centric Representation encodes sketches and text independently into a shared latent space, preserving localization and limiting inter-pair information leakage. Then, the Pair-former merges text and sketch information within each pair, ensuring that single-item attributes are spatially grounded. During the Diffusion Pair Guidance phase, these localized embeddings condition a pre-trained diffusion model, while a global textual representation encodes overall style and background. Unlike prior methods, our approach defers the merging of conditions to the diffusion process, breaking down the generation across multiple denoising steps via cross-attention.

For evaluation, we introduce Sketchy, a dataset built on Fashionpedia [132], providing multiple sketch-text pairs per outfit to support localized generation. Experiments demonstrate that LOTS achieves state-of-the-art performance in image quality and attribute localization,

as measured by quantitative metrics (+3.4% GlobalCLIP vs. fine-tuned ControlNet), human evaluation (+3.1% F1 Score over SDXL), and qualitative analysis.

8.2.1 Contributions

- We advance localized sketch-text generation, improving conditioning for multi-garment fashion images.
- We propose LOTS, which mitigates attribute confusion through modularized attention per sketch-text pair and multi-step diffusion merging.
- We introduce Sketchy, a new fashion dataset supporting multi-localized sketch-text generation and evaluation.
- We demonstrate state-of-the-art performance in sketch-text conditioning and attribute localization through metrics, human evaluation, and qualitative results.

8.3 Related Work

In this section, we focus on text-to-image, sketch-to-image, and controllable diffusion-based generation.

8.3.1 Text-to-Image Generation

Recent advances in Text-to-Image (T2I) generation have been driven by diffusion models [57, 71, 140], which generate high-quality images from textual prompts [15, 40, 72, 120] through a noise-based forward and denoising reverse process. Conditioning techniques enhance control: GLIDE [120] employs classifier-free guidance, DALLE-2 [72] uses a two-stage CLIP-based approach, and Imagen [40] integrates large-scale language models for improved realism and semantic alignment. Stable Diffusion (SD) [15] refines conditioning via cross-attention while optimizing efficiency through latent-space diffusion. Building on SD-like models, we extend control beyond textual descriptions, focusing on multiple sketch-text pairs for localized and fine-grained conditioning.

8.3.2 Sketch-to-Image Generation

Sketch-to-Image generation has evolved from GAN-based methods [141–145] to pre-trained diffusion models [146–148]. On this line, PITI [146] maps sketches to semantic latents of a large-scale diffusion model, while SDEdit [148] employs sketch perturbation and denoising to guide generation. Finally, LGP [147] enforces consistency between noisy features and spatial sketch guidance. Recent approaches for sketch-to-image generation have implemented different techniques to control the downstream diffusion model that address the problem of spatial conditioning [17, 18, 138], sketch abstraction [149, 150], or professional sketches [151]. Unlike prior work [17, 18, 138, 146–151], which conditions on global sketches, we introduce localized sketch control.

8.3.3 Controllable diffusion-based generation

While textual prompts enable high-quality image generation in T2I models, they often lack fine-grained control. To enhance controllability, various methods integrate additional conditioning elements [17, 115, 138, 139, 152, 153], including bounding boxes [152], blobs [154], and segmentation masks [155, 156]. GLIGEN [152] conditions the diffusion model with bounding box coordinates to localize textual information, but does not allow for paired sketch-text localization. ControlNet [17] introduces zero-convolution modulation in a frozen diffusion model, while subsequent works propose multi-modal control [157] and all-in-one control adapters [153], though both rely on fixed-length input channels. AnyControl [139] enables multi-control conditioning but requires a trainable copy of the diffusion model. Alternative methods, such as T2I [18] and IP [138] adapters, employ residual feature maps and cross-attention, respectively, to aggregate multiple control signals before conditioning. However, these approaches depend on global textual prompts and are limited by the 77-token constraint of text encoders. In contrast, we couple localized textual descriptions with their corresponding sketches to improve fine-grained generation. Our adapter enables a pre-trained T2I diffusion model to condition on a variable number of sketch-text pairs while remaining lightweight to train. Furthermore, recent approaches [158, 159] employ image editing in the Fashion domain. Multimodal Garment Designer[158] requires a starting image as input for the edit, while LOTS performs image generation from scratch. HieraFashDiff [159] is a concurrent work presenting a two-stage pipeline performing generation and iterative editing. Differently, LOTS performs one-shot generation and allows for additional sketch conditioning. Since direct comparison would require non-trivial modifications, risking unfair or misleading results, we will not compare to editing-based approaches.

8.4 Method

In this section we present LOTS: *LOcalized Text and Sketch for fashion image generation*. We start with a formalization of the task, defining the input-output of the problem, and later introduce the key modules of our adapter strategy.

8.4.1 Localized sketch-to-image generation

The input and desired output of the problem are as follows. **Input:** Let $\mathcal{C}$ be a set of sketch-text pairs $\mathcal{C} = \{C_1, .., C_{N_i}\}$ where $C_i = (S_i, T_i)$ denotes the i-th sketch-text pair with sketch conditioning $S_i \in \{0,1\}^{H \times W}$ and textual description T_i, with N_i the number of pairs associated with the i-th sample and H, W denoting the width and height of the desired generated image, respectively. We assume that the provided sketches have a globally coherent spatial layout. In other words, local sketches associated with an item should satisfy the constraint $\sum_i^N S_i = S$, where S is the global sketch of the desired image. To enable global information conditioning in natural language, we further allow for the global textual representation T_g to prompt the model with general appearance information that the desired image should have, *e.g.*, the fashion style or the background. **Output:** Localized sketch-to-image generation aims to synthesize with the generative model ϕ an image $X \in \mathbb{R}^{3 \times H \times W}$

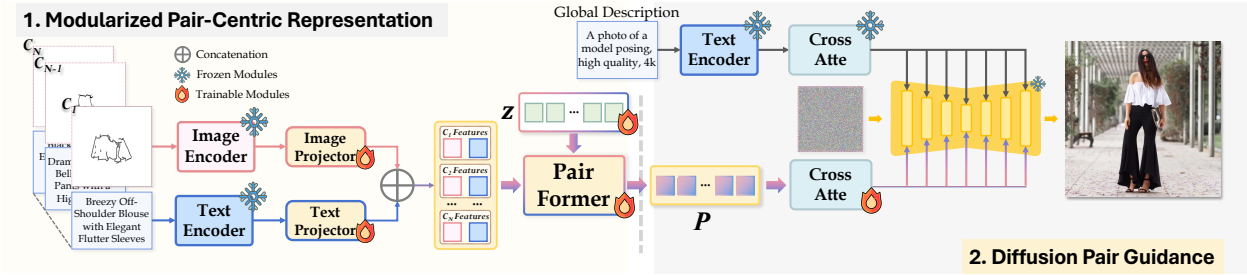


Figure 8.2: LOTS mitigates attributes confusion building on paired sketch-text conditioning for image generation. **1.** In an initial phase, the modularized Pair-Centric Representation (Sec. 8.4.3) independently processes available pairs by first embedding the different modalities with pre-trained modality-specific encoders, and later localizing the semantic textual information according to the associated sketch structure in the Pair-Former. **2.** In the second Diffusion Pair Guidance phase (Sec. 8.4.4), pair representations are directly injected into the downstream diffusion model. By breaking down the merge task within the denoising diffusion steps, LOTS avoids explicit merge of pair representations that lead to attribute confusion.

based on the conditioning input $\mathcal{C}$ as $X = \phi(\mathcal{C}, T_g)$. The resulting generation should faithfully preserve the spatial constraints of sketches and the semantic details of text descriptions while ensuring global coherence between all localized pairs and global description. In particular, the specified properties of the textual description T_i associated with the i-th item should be correctly reflected in the location described by its associated spatial information S_i, while not leaking to other local parts of the image S_j, with $i, j = 1..N, j \neq i$.

8.4.2 Method overview

LOTS enables sketch-text conditioned image generation through a two-phase process. An illustration of the proposed approach is presented in Fig. 8.2. In the *Modularized Pair-Centric Representation* phase (Section 8.4.3), sketch and text inputs are independently encoded using pre-trained modality-specific encoders and then projected into a multi-modal shared latent space via the *Pair-Former module*, ensuring localized feature extraction without interference between pairs. To address *attribute confusion* when merging multiple conditioning inputs, the *Diffusion Pair Guidance* phase (Section 8.4.4) employs attention-based conditioning within the multi-step denoising process, integrating both local and global conditioning effectively.

8.4.3 Modularized Pair-Centric Representation

Localized sketch-to-image generation requires generating an image starting from the set of local conditionings $\mathcal{C} = \{C_1, .., C_2\}$ while ensuring no semantic information from C_i is leaking to unrelated parts of the image. To address this, we propose to independently process the input information in a pair-centric fashion: each pair is considered independent, or in other words, pairs do not influence or see each other. Consider a single localized representation in the form of a sketch-text pair $C_i = (S_i, D_i), i = 1..N$. We embed conditioning information

via modality-specific encoders:

$$h_i^T = f^T(T_i) \tag{8.1}$$
$$h_i^S = f^S(S_i) \tag{8.2}$$

where f^T and f^S denote respectively the text and sketch encoders, and h_i^T, h_i^S their associated output latent representations, $i = 1..N$.

Pair-Former

Starting from the modality-specific representations, for each pair, we integrate the sketch spatial guidance and the text semantic information into a shared feature space, where textual information is localized according to the sketch structure. Inspired by recent vision-language advancements [160], we devise a new self-attention strategy aiming to (i) compress the sparse sketch representation in a limited number of tokens and (ii) merge multi-modal information. Let $z \in \mathbb{R}^{k \times d}$ be a set of k d-dimensional learnable tokens prepended to the concatenation of visual and textual embedding representations. We obtain the pair tokens by first computing the self-attention with sketch and text representations:

$$h_i = \texttt{Self-attn}([z; h_i^S; h_i^T]), \tag{8.3}$$

and restricting p_i to be the first k tokens of h_i, *i.e.*, the output tokens associated to z, where $[\cdot\ ;\ \cdot]$ denotes the concatenation operation along the token dimension. Hence, self-attention allows the prepended token z to pool the informative content from both the sketch and textual modalities in a coordinated manner. We remark that while the z learnable tokens are shared between the different pairs, the Pair-former is independently processing the N input pairs.

8.4.4 Diffusion Pair Guidance

Merging information from multiple conditioning pairs is challenging, as it requires effective coordination while avoiding information leakage. Existing approaches encode multiple guidance signals in a single pooling step, which may cause interference among pairs. In contrast, we defer the merging process to the pre-trained diffusion model, leveraging its iterative denoising steps to progressively integrate conditioning pairs, rather than combining them in a single step. Given $P \in \mathbb{R}^{N \times k \times d}$ as the concatenated sequence of the N pooling tokens $p_i, i = 1..N$ extracted by our Pair-Former, we avoid explicitly merging the pairs. Instead, we inject the entire sequence into the diffusion model, allowing the merge operation to occur throughout the entire diffusion process. This allows for more dynamic interaction between the multiple given conditionings during the reverse diffusion process. Inspired by previous work [138], we rely on cross-attention layers to guide diffusion generation and introduce an additional learnable cross-attention layer $\hat{w}$ after each pre-existing cross-attention layer w in the frozen denoising model. These new layers inject the conditioning sequence P into the model features at each diffusion timestep, allowing for the iterative merging of information. Let x denote the input features of a global text-conditioning cross-attention

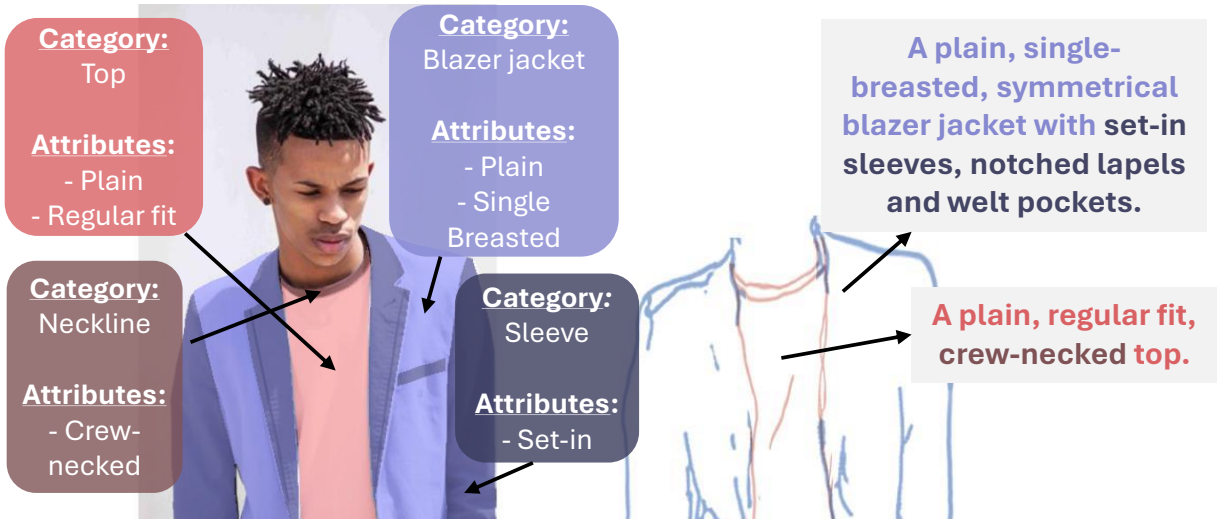


Figure 8.3: Example of the hierarchical structure of Sketchy. Starting from whole-body item (light colors) and garment parts (dark shades) annotations, we build a hierarchical structure by pairing the garment part annotations to their related whole-body garment. Then, we use this structure to generate garment-level sketches and natural language descriptions by relying on off-the-shelf models.

layer in the denoising network, the conditioned output x' of the paired cross-attention layers is computed as:

$$x' = w(x, h^{T_g}) + \alpha \hat{w}(x, P), \tag{8.4}$$

where $w(\cdot, \cdot)$ represents the cross-attention between the two input token sequences and h^{T_g} denotes the embedding representation of the global text prompt guiding the model with semantic information that should be globally represented, *e.g.*, the style or the background. Here, α is a scaling hyper-parameter, constrained to the range $[0, 1]$, that regulates the strength of the guidance from P. During training, we set α to 1 to fully enable the cross-attention layers $\hat{w}$ to learn the appropriate merging behavior. Notably, the attention-based nature of these adapter blocks allows an arbitrary number of conditioning tokens, enabling LOTS to work with a variable number of conditioning pairs.

8.5 Experiments

We assess LOTS by comparing its performance with state-of-the-art sketch-to-image adapters in the fashion domain. First, in Sec. 8.5.2, we present our newly proposed dataset, detailing its construction and structure, along with the evaluation metrics and implementation details. In Sec. 8.5.3, we outline the experimental setup and baseline models, and analyze the quantitative and qualitative results. Finally, in Sec. 8.5.4, we conduct ablation studies to highlight key design choices of LOTS.

8.5.1 The Sketchy dataset

We construct a novel dataset, named Sketchy, to support training and evaluation on the task of localized sketch-to-image generation, with localized text-sketch pairs associated with high-quality ground truth images. In the following, we explain in detail how we organize the dataset based on garments and the localized text-sketch creation, as well as the dataset statistics.

Local garments organization. We build Sketchy on Fashionpedia [132], a dataset composed of 46k images for training and 1.2k for testing, where fashion experts annotate garments with fine-grained attributes and segmentation masks. While these masks include detailed part annotations (*e.g.*, pockets, zippers, sleeves), they lack a hierarchical structure linking garment components. To improve compositionality, we introduce a two-level hierarchical organization based on segmentation mask overlaps. Specifically, following Fashionpedia taxonomy, the 330k item annotations are first categorized into 14 "whole-body items" (*e.g.*, tops, shirts, and skirts) and 32 "garment parts" (*e.g.*, sleeves, pockets, and necklines). To ensure high-quality compositional annotations, we retain all whole-body categories, along with 21 sub-item categories from Fashionpedia, while removing 11 categories (31k annotations) that are rare, seldom overlap, or lack consistent overlap with any whole-body items, such as umbrellas, bags, and glasses. Then, for each image, we determine the overlap between each garment part's mask and every whole-body item mask. Whole-body annotations are considered as top-level annotations, *i.e.*, a garment in the image, while part annotations are assigned to the whole-body item with which they have the greatest overlap, *i.e.*, they are considered sub-garment annotations referring to a property, such as sleeves, necklines, and pockets.

Localized text-sketch creation. While the Fashionpedia annotations are rich in attributes, they lack a coherent natural language description, which is essential for our text conditioning. Thus, we generate the textual description for each garment in the image by prompting a pre-trained Large Language Model [161] with the hierarchical annotation structure of each garment, along with some in-context learning examples of the desired reply format. We further augment the dataset by including localized garment-level sketches, generated from the ground-truth images, using a pre-trained Image-to-Sketch model [162]. We remove background information via masking to ensure each sketch contains only information about the associated item. We provide a global composition of all the garment sketches, which depicts the sketch of the entire outfit in the original image. Finally, we pre-process the images by resizing them to 512 pixels. We preserve their aspect ratio with white padding into a square format to maintain consistency across samples.

Dataset statistics. Our Sketchy extends Fashionpedia, providing a total of 47k images and 79k garment-level annotations, resulting in an average of 1.7 annotations per image (min 1, max 6). As shown in Fig. 8.3, each annotation contains the associated sketch, hierarchical attributes, and natural language description of the item. The average word length of the descriptions is 16 words.

8.5.2 Experimental protocol

Quantitative evaluation and metrics. Following prior works [139, 156, 163–165], we adopt Fréchet Inception Distance (FID) [74] to assess the global fidelity of the generated images distribution, relative to ground truth images. Lower values of FID indicate better perceptual quality and stronger correspondence. To measure the semantic alignment, we rely on GlobalCLIP [12] score, calculated as the cosine similarity of image embeddings, in line with [165]. To further capture fine-grained and localized alignment, we adopt LocalCLIP score, building on [155, 163]. Specifically, we first use the masking annotations to obtain a crop for each garment in the ground truth and generated images. Then, we apply the CLIP [12] visual encoder to compute the cosine similarity between each pair of ground truth and generated image crops, averaging their score across all garments. Higher scores correspond to better semantic alignment. In addition, we utilize the state-of-the-art compositional semantic alignment metric, VQAScore [16] that captures how well-generated images reflect complex text descriptions by relying on Visual Question Answering approaches to query the presence of desired properties. A larger VQAScore suggests improved compositional alignment to the provided prompt. Finally, we follow previous work [151, 156] and evaluate the proposed LOTS based on the Structural Similarity Index Measure (SSIM) assessing the sketch structural alignment and edge fidelity (the higher, the better).

Human evaluation. In line with prior works [17, 153], we conduct a user study involving 14 participants with an average of 57 answers each, targeting the evaluation of localized control, rather than visual perception. We generate images ensuring that attributes appear only once in the entire outfit. We leverage a questionnaire to assess whether a specific attribute associated with the i-th garment is correctly localized in the desired garment of the image and if it leaks to other ones. We quantitatively evaluate considered models in terms of Precision (↑), Recall (↑), and F1 Score (↑) metrics with respect to localized conditioning. A high F1 Score indicates the model's overall performance in correctly reflecting and localizing the attribute, showing its capacity to balance accurate placement with less attribute confusion.

8.5.3 Main Comparisons

Baselines. We compare LOTS with baselines and state-of-the-art sketch-to-image approaches. Specifically, for text-only approaches, we evaluate Stable Diffusion 1.5 (SD) [15] and Stable Diffusion XL (SDXL) [73], which generate images solely from a global text prompt, and GLIGEN [152], which enables localized textual conditioning. We then compare to sketch-to-image adapter methods, including SD-based ControlNet [17], SDXL-based T2I-Adapter [18] and SDXL-based IP-Adapter [138], which all incorporate both a global text prompt and a global sketch as inputs, offering enhanced spatial conditioning. Additionally, we examine compositional modifications of ControlNet (Multi-ControlNet) and T2I-Adapter (Multi-T2I-Adapter) that were modified to allow multiple local sketches conditioning alongside a single global text prompt. Furthermore, we evaluate the most recent localized control method, AnyControl [139], that allows for local sketches and a single global text description as inputs. Finally, we also assess the performance of the adapter-based approaches with fine-tuning on our Sketchy dataset.

Experimental setup. We adapt conditioning to different generative models based on their input requirements. For global descriptions, we concatenate all garment descriptions, while for models requiring a single image guide, we create a composite global sketch by merging individual garment sketches. When localized control is supported, we use garment-specific sketches and/or descriptions as input. For consistency, all images are generated at `(512x512)` resolution using each model's default inference setup. Additionally, to ensure fairness, our global description T_g remains fixed across samples as "A picture of a model posing, high-quality, 4k".

Implementation Details. We adopt `DINOv2 vits14` [166] as sketch encoder. As text encoder, we follow the findings in [73] and use a combination of `OpenCLIP ViT-bigG` [167] and `CLIP ViT-L` [12] by concatenating the penultimate text encoder outputs along the channel-axis.

Quantitative results. Tab. 8.1 reports the quantitative evaluation across global quality, semantic alignment, and structural similarity metrics on the test split of Sketchy. Our method demonstrates state-of-the-art performance in GlobalCLIP, LocalCLIP, and VQAScore, while ranking third in SSIM, indicating overall superiority and strong alignment both semantically and structurally. Specifically, LOTS attains the second lowest FID, demonstrating a high perceptual fidelity. A similar pattern is observed in GlobalCLIP and LocalCLIP scores, with our method surpassing all baselines and state-of-the-art models (+3.4% and +1.2%), indicating strong semantic alignment and feature similarity with the ground truth. For compositional semantic alignment, T2I-Adapter and our method achieve the highest VQAScore, outperforming all alternatives in textual prompt following. While some models outperform LOTS in some metrics, their improvements come with trade-offs: the fine-tuned IP-Adapter achieves a lower FID but sacrifices both text and sketch guidance, as evidenced by its significantly lower Compositional Alignment metrics. Similarly, for structural similarity, Multi-T2I-Adapter achieves the highest SSIM, followed by IP-Adapter, while our method ranks third. As opposed to LOTS, however, both IP-Adapter and Multi-T2I-Adapter emphasize sketch-guidance over prompt adherence and image coherence, as evidenced by their subpar LocalCLIP, VQAScore, and FID scores, which figure among the lowest. In conclusion, these results highlight how, thanks to its novel pairing strategy, LOTS strikes an optimal balance between image quality and prompt adherence, surpassing prior approaches and setting the new state-of-the-art performance in the fashion localized sketch-to-image task.

Results with human evaluation. In this evaluation, we aim to measure whether an attribute is correctly localized in a garment of the generated image. In Tab. 8.2, we present the results of our human study for attribute localization and confusion across different models. Our LOTS achieves the highest F1 score among all models, indicating superior performance in accurately localizing attributes on the desired items while minimizing unintended leakage. Furthermore, it performs the best in Precision, demonstrating its effectiveness in preventing attribute confusion. In contrast, models such as SDXL [73] and Multi-T2I-Adapter [18], while strong in Recall, exhibit lower Precision scores, suggesting that attributes may inadvertently leak to unintended items, as shown in Fig. 8.4 (first row, the striped trousers). On the

Model	Conditioning Visual/Textual	Global Quality		Compositional Alignment		
		FID (↓)	GlobalCLIP (↑)	LocalCLIP (↑)	VQAScore (↑)	SSIM (↑)
SD [15]	-/G	1.11	.603	.745	.719	.663
SDXL [73]	-/G	1.77	.529	.701	.660	.544
GLIGEN [152]	-/L	0.93	.568	.704	.395	.614
ControlNet [17]	G/G	1.08	.622	.789	.733	.674
Multi-ControlNet [17]	L/G	1.10	.615	.780	.730	.672
IP-Adapter [138]	G/G	2.80	.537	.682	.611	<u>.715</u>
T2I-Adapter [18]	G/G	2.16	.534	.705	.635	.482
Multi-T2I-Adapter [18]	L/G	1.14	.583	.766	.697	**.723**
AnyControl [139]	L/G	0.99	.602	.777	.712	.544
GLIGEN [152]	-/L	1.70	.564	.713	.419	.514
ControlNet [17]	G/G	0.80	<u>.645</u>	<u>.801</u>	.717	.574
Multi-ControlNet [17]	L/G	0.84	.638	.799	.720	.572
IP-Adapter [138]	G/G	**0.69**	.621	.787	.714	.631
T2I-Adapter [18]	G/G	1.03	.570	.753	**.749**	.612
Multi-T2I-Adapter [18]	L/G	1.11	.559	.744	<u>.734</u>	.605
LOTS **(Ours)**	L/L	<u>0.79</u>	**.679**	**.813**	**.749**	.678

Table 8.1: Comparisons between LOTS and state-of-the-art sketch-to-image approaches. In the Conditioning column, L and G indicate whether the model accepts Local or Global inputs as Visual or Textual conditioning. We divide the table into three sections: zero-shot approaches, fine-tuned approaches on Sketchy, and our approach LOTS. We highlight the best performance in bold and underline the second best

Model	Attribute Localization		
	Precision (↑)	Recall (↑)	F1 (↑)
SDXL [73]	.636	**.754**	<u>.690</u>
ControlNet [17]	.596	.449	.512
Multi-ControlNet [17]	.487	.365	.418
IP-Adapter [138]	.625	.139	.227
T2I-Adapter [18]	.409	.170	.240
Multi-T2I-Adapter [18]	.370	.270	.312
AnyControl [139]	.281	.134	.182
ControlNet [17]	<u>.667</u>	.516	.582
Multi-ControlNet [17]	.541	.417	.471
IP-Adapter [138]	.559	.384	.455
T2I-Adapter [18]	.463	.397	.427
Multi-T2I-Adapter [18]	.551	.692	.614
LOTS **(Ours)**	**.813**	<u>.650</u>	**.722**

Table 8.2: Results of qualitative user study of attribute localization and confusion conducted between LOTS and other models. We highlight the best results for each metric in bold and underline the second best.

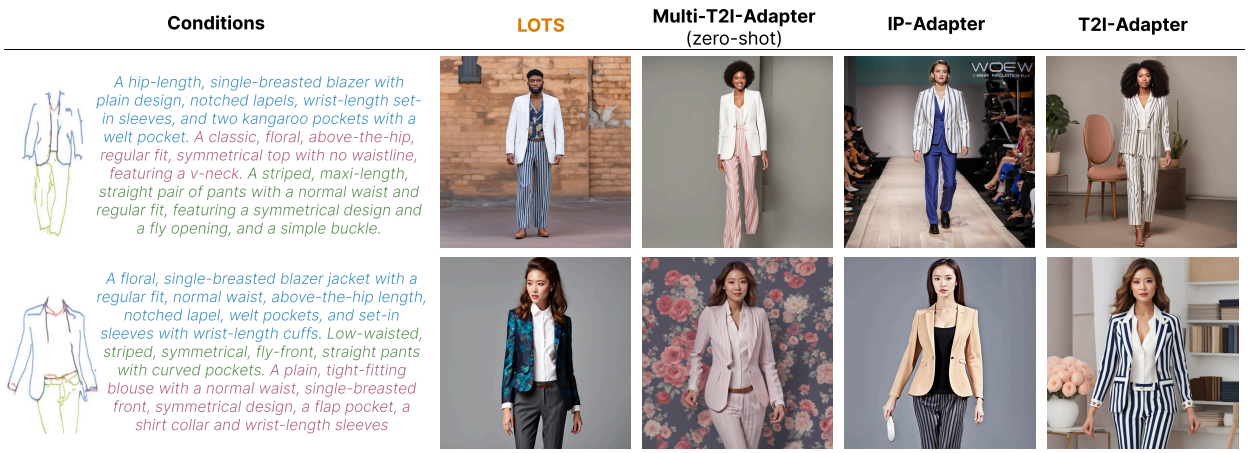


Figure 8.4: Qualitative results of LOTS in comparison with Multi-T2I-Adapter [18], IP-Adapter [138], and T2I-adapter [18]. Given paired localized text-sketch pairs as conditioning inputs, LOTS can better reflect fine-detailed attributes in the intended local region of the generated images, effectively mitigating attribute confusion.

other hand, T2I-Adapter [18], ControlNet [17], and AnyControl [139] show relatively lower performance in both Precision and Recall. Models that underwent fine-tuning also improved performance, with Multi-T2I-Adapter [18] attaining the second-highest F1 score following LOTS. Finally, we measure the statistical relevance of these results and record a high Krippendorff's $\alpha = 0.81$, testifying to a high inter-annotator agreement.

Qualitative results. We qualitatively analyze the attribute localization ability of different models, evaluating whether models accurately associate attributes with the intended garment, ensuring that details such as patterns, silhouettes, and structural elements are correctly rendered in their designated locations. As shown in Fig. 8.4 (top), the input description specifies three main items associated with their attributes and garments. In particular, a plain single-breasted blazer jacket, a floral top, and striped pants. In this case, a proper generation should be able to faithfully follow the description, placing the floral pattern on the top while keeping the jacket's plain pattern. As shown, LOTS is the only approach that effectively generates a plain jacket with a floral top and striped pants. In contrast, multiple baselines and state-of-the-art models exhibit errors. For instance, Multi-T2I-Adapter successfully follows the sketch, but fails to capture the floral pattern of the top. On the other hand, both IP-Adapter and T2I-Adapter wrongly localize the "striped" attribute to the jacket, failing to maintain semantic alignment with the prompt. These trends are not limited to isolated cases but are consistently observed across multiple examples, as denoted by the results in Tab. 8.1 and Tab. 8.2.

8.5.4 Ablation Analysis

In this section, we present an ablation analysis of the key components used in LOTS, and evaluate the impact of different design choices, such as the image encoder, diffusion guidance strategy, and the number of pooling tokens used in the Pair-Former module.

Sketch Encoder. In the first section of Tab. 8.3, we analyze different encoders for our sketch guidance. Training a dedicated sketch encoder end-to-end simultaneously to LOTS (Trained) is ineffective, resulting in the lowest SSIM recorded (.615). We hypothesize this is due to the pre-trained text features dominating the initial training phase, causing back-propagation to neglect the sketch encoder optimization. Thus, to enhance sketch guidance, we test different frozen pre-trained image encoders, namely CLIP [12], BLIP2 [160], and DINOv2 [166]. DINOv2 (LOTS) results in the highest SSIM (.678), with both CLIP and BLIP2 achieving subpar SSIM performance (.623 and .630 respectively).

Diffusion Guidance. The second section of Tab. 8.3 showcases different diffusion guidance strategies. In the first experiment (No Pooling), we compose our conditioning sequence P by concatenation of all the N paired features coming from the projectors without any pooling operation, *i.e.*, $P = [h_1^S; h_1^T; \dots; h_N^S; h_N^T]$. While this approach keeps all the encoder information available for guidance, the model ignores the sketch information, focusing only on textual conditioning, as testified by the high VQAScore (.724) and low SSIM (.623). We hypothesize this is due to the large number of uninformative tokens coming from the image encoder, *i.e.*, the tokens where no sketch is present, pushing the model to ignore their guidance and focus only on the feature-rich textual features. Averaging all the pairs into a single unified representation $P = \frac{1}{N}\sum_{i=1}^{N}[h_i^S; h_i^T]$, improves SSIM, but is unsatisfactory in textual adherence with a low .676 VQAScore. By compressing the pairs' features and deferring the pair-merging operation to the diffusion process as presented in Sec. 8.4.3, LOTS achieves the best performance in both compositional semantic alignment (.749 VQAScore) and sketch-guidance (.678 SSIM).

Pair-Former. We first evaluate the impact of our Pair-Former by replacing it with standard cross-attention layers (Cross-attn), resulting in lower performance across all metrics. We believe this is due to cross-attention layers limiting interactions to cross-modality only, while Pair-Former's self-attention enables both intra-/cross- modality interactions between text and sketch. Additionally, to validate the feature compression performed by our Pair-Former, we augment the number of learning pooling tokens and find that, interestingly enough, a higher number of pooling tokens (64 vs. 32) lowers the performance of LOTS by .08 in VQAScore. We hypothesize this is due to both the increased number of trainable parameters and the data redundancy introduced by the larger number of tokens, causing issues similar to the No Pooling ablation experiment.

8.6 Conclusions

In conclusion, we tackled the challenging task of image generation with localized sketch-text pairs, reflecting a realistic scenario in the fashion design process. This work continues the thesis narrative on enhancing multimodal understanding through richer language modeling, following our investigation of abstract-oriented language in ACT and precise entity-attribute evaluation in L-VQAScore. We proposed LOTS, a novel framework that mitigates attribute confusion between input sketch-text pairs through modular pair-wise processing and defers the fusion of multiple conditions to the downstream denoising stage. By integrating LOTS

Component	Choice	Compositional Alignment		
		LocalCLIP ↑	VQAScore ↑	SSIM ↑
Sketch Encoder	Trained	.762	.689	.615
	CLIP	.802	.685	.623
	BLIP2	.797	.678	.630
Diffusion Guidance	No Pooling	.813	.724	.623
	Mean Pooling	**.819**	.676	.659
Pair-Former	Cross-attn	.804	.681	.626
	64 Tokens	.804	.669	.628
	LOTS	.813	**.749**	**.678**

Table 8.3: Ablation over different components of LOTS

with Stable Diffusion via an adapter-based approach, we effectively aggregate localized visual and textual information, echoing the thesis's overarching theme of leveraging language and vision jointly to improve downstream tasks.

To support model training and evaluation, we introduced Sketchy, a dataset of high-quality fashion images paired with localized sketch-text conditions, enabling systematic assessment of both visual grounding and text-image alignment. Experiments demonstrate that LOTS achieves state-of-the-art performance, confirming that careful treatment of localized language in multimodal settings improves both compositional control and generative fidelity.

Future directions include refining evaluation protocols for localized text-image alignment with fine details, and extending LOTS to better handle non-visually grounded or abstract attributes (*e.g.*, "summer-vibe"), building on insights from ACT regarding the under-representation of abstract language in current models. Incorporating more powerful textual embeddings could further strengthen the interaction between textual semantics and visual outputs, reinforcing the thesis's overarching goal of bridging language richness and multimodal understanding.

This work also highlights natural connections with previous chapters: while ACT addressed abstract-to-concrete language transformations for retrieval, and L-VQAScore focused on fine-grained evaluation of entity-attribute alignment, LOTS demonstrates how localized, compositional textual conditioning can enhance generative models, effectively completing the second half of the thesis exploration.

Chapter 9

Conclusions and Future Directions

9.1 Summary of Contributions

This thesis explored how to make artificial intelligence systems not only see and perceive, but also understand and use language effectively. Through a progression of studies, we examined two interconnected challenges: how to introduce language into visual systems that traditionally operate without it, and how to improve the usability of that language once integrated.

In the first part of this thesis (Chap. [3-5]), we focused on **introducing language into vision-centric systems**. We proposed methods that connect linguistic concepts to spatial, perceptual, and symbolic representations. In the context of visual navigation, the introduction of a Language-enhanced Renderable Neural Radiance Map (Le-RNR-Map) [5] (Chap. 3) demonstrated that language can serve as an intuitive querying and interaction tool, bridging natural communication with precise spatial reasoning. Through the In&Out [6] and DIAG [8] approaches (Chap. [4-5]), language became a medium for guiding visual synthesis, integrating expert descriptions and multimodal feedback into the learning process of the Industrial Anomaly Detection field. These works collectively showed that language can be successfully grounded in visual and sensory domains, expanding the interpretability, controllability, and human alignment of AI systems in mission-critical environments such as production lines.

The second part of this thesis (Chap. [6-8]) addressed **improving the usability of language within multimodal models**. Here, we examined the limitations of current Vision-and-Language Models (VLMs) in representing abstract concepts, evaluating compositional understanding, and generating controlled outputs. The Abstract-to-Concrete Translator (ACT) [11] (Chap. 6) demonstrated that abstract and affective terms, often underrepresented in training corpora, can be effectively linked to visual semantics through post-hoc latent adaptation, without need to re-train expensive multi-modal architectures. The Localized VQA-Score (L-VQAScore) [10] (Chap. 7) introduced a new way to evaluate text-to-image generation, localizing the relationship between entities and attributes to assess whether the generated content truly reflects the intended semantics. Finally, the Localized Text and Sketch (LOTS) adapter [9] (Chap. 8) showed how language can be used in combination with visual cues such as sketches to control diffusion-based generation, enabling fine-grained creative design workflows. Together, these contributions assess the second research question,

improving the depth, granularity, and usability of language in multimodal AI systems.

Across both parts, the overarching theme of this thesis is the use of language as a medium for human-machine interaction, from serving as an auxiliary input to becoming a core mechanism of reasoning, communication, and creativity.

9.2 Limitations and Future Directions

Although the contributions presented in this thesis push forward the state of the art in multimodal understanding and generation, several limitations remain. These naturally delineate promising directions for future research.

9.2.1 Data and Generalization

As with most Deep Learning work, the performance and generalizability of the proposed approaches are inherently bounded by the data on which they are trained. While ACT [11], discussed in Chapter 6, was explicitly designed to improve model robustness under novel and abstract linguistic descriptions, the other methods presented in this thesis remain more tightly coupled to the specific data distributions they were trained on.

Beginning with Le-RNR-Map (Chapter 3), one key limitation emerges from the controlled nature of the experimental setup. All human-agent interaction experiments relied on simulated environments with fixed visual conditions: unchanging lighting, no occlusion, and consistent item appearance. Although this may initially seem like a purely visual constraint, the VLM used (CLIP [12]) depends heavily on the alignment between its image and text encoders. Even subtle perturbations in hue, viewpoint, or visibility could disrupt this alignment, leading to failures at inference time when the two modalities cannot communicate. Recent architectures such as QWEN [108], which process vision and language jointly, open a direct avenue for extending this line of research. Future work could explore whether unified VLMs enable more robust multimodal reasoning in realistic, dynamic environments.

A similar limitation affects In&Out [6] and DIAG [8] (Chapter 4, Chapter 5), both of which were developed and evaluated primarily on a single industrial dataset for surface-defect augmentation. While effective in this controlled setting, mask-based inpainting strategies may not be as suitable for more complex forms of anomaly generation, such as logical defects [168] (*e.g.* "4 items expected, but only 3 produced"). In these scenarios, language, not masking, may be the most precise modality for specifying anomalous reasoning. Recent models such as EMU-Edit [169] integrate large language models into generation pipelines, allowing rich editing instructions expressed verbally. Investigating whether such models can support controlled industrial data augmentation, where data scarcity and unfamiliar domains pose major challenges, constitutes a compelling research direction.

Finally, the data used for LOTS [9] (Chapter 8) also presents constraints. Because sketch–text pairs are generated automatically through pre-trained tools, the entire dataset occupies a single stylistic space: sketches share a uniform visual style and descriptions follow a consistent lexical pattern. In contrast, real-world fashion design involves stylistic plurality: people sketch differently (*e.g.*, messy vs. clean sketches), describe garments using diverse vocabularies, and vary widely in level of detail. Bridging this gap will require methods

capable of generalizing across unseen sketch styles and linguistic expressions. Evaluating the robustness of this methodology to these axis of variation would further strengthen the thesis vision of expanding multimodal models beyond the constraints of their training distributions.

9.2.2 Evaluation

A second major limitation in Multimodal Deep Learning, and one that directly intersects with the contributions of this thesis, is the reliable evaluation of generative models. Automatic metrics for image generation typically rely on the internal representations of large, pre-trained networks [16, 74, 170], using their embeddings to approximate aspects of human judgement. While such metrics have accelerated research progress, they inherit the biases and blind spots of the models they build upon.

When evaluating visual quality, ambiguity in human perception poses an inherent problem: two observers may legitimately prefer different outputs, making "image quality" a target that is both subjective and context-dependent. Consequently, works in this thesis, as well as most of the current literature, rely on automatic metrics only in a task-specific sense: FID [74] for distribution similarity, LPIPS [170] for perceptual similarity, and more recently CLIP-based alignment scores for semantic consistency.

However, when language enters the picture, the gap between metrics and human perception becomes more evident. CLIP [12] similarity scores and VQA-based alignment metrics [16] struggle particularly in multi-entity prompts or fine-grained attribute grounding, precisely the scenarios explored throughout this thesis. This limitation motivated the development of the LVQA-Score (Chapter 7), which leverages localized visual question-answering to better assess attribute placement and fidelity. While LVQA-Score improved granularity in evaluating localized alignment, it also highlighted how far the field still is from a robust, reliable, and universally interpretable text-image evaluation framework.

In this sense, automatic evaluation is inseparable from progress in multimodal generation: better models demand better metrics, not just better prompts or architectures. As text-to-image systems continue to grow in complexity [15, 73, 123, 169], developing evaluation methodologies capable of capturing compositionality, attribute localization, and semantic nuance will be essential, not only to measure improvement, but to define it.

9.2.3 Computing Resources

An additional limitation, shared across all works in this thesis, and increasingly central in the broader literature, is the resource constraint imposed by modern Large Vision–Language Models. Every method presented relies, to varying degrees, on large-scale pre-trained models trained on Internet-scale corpora such as LAION [83]: Le-RNR-Map (Chapter 3) builds upon CLIP [12] for vision–language alignment; In&Out [6] and DIAG [8] (Chapter 4, Chapter 5) depend on Stable Diffusion pre-trained models [15, 73]; ACT [11] in Chapter 6 analyzes the behaviour of these pre-trained backbones, and together with both LVQA-Score [10] and LOTS [9] (Chapters 7–8), makes use of large language models for structured refactoring and generation of textual descriptions. This reliance is not unique to this thesis: it reflects the current direction of multimodal AI, where progress is largely driven by accessibility to powerful foundation models.

While resource availability sits slightly beyond the methodological focus of this thesis, focusing on more parameter-efficient approaches, possibly task-specific, in order to no lorger rely on large pre-trained models, is an interesting, and in my opinion necessary, direction for future developments.

The broader lesson that emerges across these contributions is that progress in multimodal learning is increasingly shaped by infrastructure constraints as much as by algorithmic design. Recent parameter-efficient fine-tuning approaches, such as Low-Rank Adaptation (LoRA) [20], are a clear example: they make specialization feasible by lowering the cost of training, and they fit naturally into the current ecosystem of agentic LLMs and large foundation models. At the same time, reducing training cost does not automatically translate into deployability. Inference remains the critical bottleneck whenever the target setting imposes strict limits on latency, memory, or energy.

This tension is particularly evident in scenarios that motivate the first part of the thesis. If, for example, one aims to deploy Chapter 3 on a robotic platform, real-time operation is required under severe on-device constraints, unless one relies on remote computation (with the attendant assumptions about connectivity, privacy, and reliability). From this viewpoint, a natural and high-impact direction is to characterize the accuracy/efficiency frontier of the proposed pipelines, and to study how model compression and quantization affect not only raw performance, but also robustness and calibration in the field.

9.2.4 Focus on Fashion

The second part of this thesis intentionally uses fashion as a primary testbed, because it offers a rich semantic space where language, abstract attributes, localized evaluation, and controllable generation interact in a particularly visible way. Within this setting, the proposed methods improve complementary aspects of multimodal systems: ACT (Chapter 6) targets abstract language alignment, L-VQAScore (Chapter 7) targets localized semantic evaluation, and LOTS (Chapter 8) targets fine-grained conditioning for generation.

Nevertheless, the emphasis on fashion also raises a broader question that is central to the thesis narrative: to what extent do these mechanisms reflect a general *representation shift* principle, rather than properties of a single domain? A meaningful next step is therefore to validate the same ideas in domains with different visual statistics and language conventions. In the case of ACT, this could include studying domain shifts beyond the concrete/abstract axis, and testing whether this training-free adaptation can bridge more structural gaps in how a VLM organizes its joint embedding space. In parallel, demonstrating the effectiveness of L-VQAScore and LOTS outside fashion would strengthen their role as generally useful tools, and would clarify which components transfer directly and which require domain-specific tuning.

9.3 Closing Remarks

I would like to conclude this thesis by reflecting on its overarching narrative, as well as on the broader perspective and future outlook that this work has shaped for me.

This thesis set out to investigate the interplay between visual and textual information, with a particular focus on adding and improving language interaction inside visual models and tasks.

Each chapter represents an incremental step toward this objective. We begin with visually grounded retrieval through Le-RNR-Map, move toward fine-controlled generative editing with In&Out and DIAG, and expand into abstract semantic alignment with ACT. The final works, LVQA-Score and LOTS, bring the trajectory to its natural conclusion, addressing the dual challenge of evaluating alignment and generating images from localized sketch–text conditions. Together, these contributions highlight both the promise and the fragility of multimodal systems: capable of producing remarkable results, yet still sensitive to ambiguity, domain shift, and evaluation uncertainty.

Working through this research has also made me increasingly aware of the resource-centric nature of progress in the field. Scaling is powerful, but not always accessible, and at times the most significant challenges I encountered were not conceptual, but infrastructural: limited by data availability, compute, or feasibility when comparing against foundation-scale models. Rather than discouraging, this has reinforced a conviction: that future advances should not rely exclusively on increasing model size, but also on developing methods that are efficient, adaptable, and capable of specialization without prohibitive cost. Making multimodal intelligence more accessible is, I believe, as meaningful a goal as making it more capable.

Several concrete research directions emerge from this perspective. First, a more systematic analysis of computational efficiency would strengthen the practical value of the proposed tools. Many of the presented pipelines build on large pre-trained models (CLIP, SDXL, DINOv2), which are excellent general-purpose backbones but can be challenging to deploy outside research settings. Future work could therefore report latency and memory profiles for the end-to-end systems, and explore compression and quantization strategies to identify realistic accuracy/efficiency trade-offs for on-site use.

Second, while ACT demonstrates that abstract language can be made more usable in VLMs, it currently focuses primarily on adjectives. Extending the same idea to verbs and richer compositional structures (*e.g.*, short predicate phrases, agent/action/object templates) would broaden applicability to scenarios where what matters is not only *what* is present, but *what is happening* or *what should happen*. This could also expose new failure modes of current VLMs, which often conflate actions with co-occurring objects and struggle with event semantics.

Third, the industrial anomaly detection thread highlights a gap between visually localized defects (scratches, cracks) and logical anomalies such as missing components or incorrect assemblies. Although DIAG is effective for the former, its behavior on the latter is less explored. A natural continuation is to incorporate constraint- or logic-based priors to explicitly model what a *valid* instance should contain, and to diagnose inconsistencies that may not correspond to local texture changes.

Fourth, to support the broader claim that representation shift is a general mechanism rather than a fashion-specific artifact, it will be important to validate ACT fashion. Domains such as medical imaging provide a particularly challenging test because of different visual statistics, different language conventions, and higher importance for semantic correctness; even a small-scale cross-domain study would clarify which parts of the framework transfer

as-is and which require domain adaptation.

Finally, LOTS currently assumes relatively clean sketch/text pairings, whereas real creative workflows can include sparse, messy, or even unpaired sketches. Beyond dataset expansion, an additional analysis could quantify how performance degrades as sketches become noisier or less informative (*e.g.*, stroke dropout, jitter, reduced resolution). In parallel, relaxing the paired-data assumption would present a new scenario for deployment in different design pipelines.

This work has opened me the door to future research opportunities, including the chance to continue exploring text/image representations during my internship at Canva. I see this as a natural continuation of the research journey that shaped this thesis: bringing academic insight closer to real-world application, and contributing to tools that enable creativity rather than constrain it.

Full list of Publications

[1] L. Capogrosso, G. Skenderi, F. Girella, F. Fummi, and M. Cristani. "Toward smart doors: A position paper". In: *International Conference on Pattern Recognition*. Springer. 2022.

[2] L. Capogrosso, A. Mascolini, F. Girella, G. Skenderi, S. Gaiardelli, N. Dall'Ora, F. Ponzio, E. Fraccaroli, S. Di Cataldo, S. Vinco, et al. "Neuro-Symbolic Empowered Denoising Diffusion Probabilistic Models for Real-Time Anomaly Detection in Industry 4.0: Wild-and-Crazy-Idea Paper". In: *Forum on Specification & Design Languages*. 2023.

[3] F. Cunico, M. Emporio, F. Girella, A. Giachetti, A. Avogaro, and M. Cristani. "OO-dMVMT: A Deep Multi-view Multi-task Classification Framework for Real-time 3D Hand Gesture Classification and Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[4] M. Emporio, A. Caputo, D. Pintani, F. Cunico, F. Girella, A. Avogaro, M. Cristani, and A. Giachetti. "gesture based interaction with the Hololens 2". In: *Proceedings of the Biannual Conference of the Italian SIGCHI Chapter*. 2023.

[5] F. Taioli, F. Cunico, F. Girella, R. Bologna, A. Farinelli, and M. Cristani. "Language-enhanced rnr-map: Querying renderable neural radiance field maps with natural language". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[6] L. Capogrosso, F. Girella, F. Taioli, M. Dalla Chiara, M. Aqeel, F. Fummi, F. Setti, and M. Cristani. "Diffusion-Based Image Generation for In-Distribution Data Augmentation in Surface Defect Detection". In: *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. 2024.

[7] F. Cunico, S. Aldegheri, A. Avogaro, M. Boldo, N. Bombieri, L. Capogrosso, A. Caputo, D. Carra, S. Centomo, D. S. Cheng, et al. "Enhancing Safety and Privacy in Industry 4.0: The ICE Laboratory Case Study". In: *IEEE Access* (2024).

[8] F. Girella, Z. Liu, F. Fummi, F. Setti, M. Cristani, and L. Capogrosso. "Leveraging Latent Diffusion Models for Training-Free in-Distribution Data Augmentation for Surface Defect Detection". In: *International Conference on Content-Based Multimedia Indexing (CBMI)*. 2024.

[9] F. Girella, D. Talon, Z. Liu, Z. Ruan, Y. Wang, and M. Cristani. "LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025.

[10] Z. Liu, G. Federico, W. Yiming, and T. Davide. “Evaluating Attribute Confusion in Fashion Text-to-Image Generation”. In: *Proceedings of International Conference on Image Analysis and Processing*. 2025.

[11] D. Talon, F. Girella, Z. Liu, M. Cristani, and Y. Wang. “Seeing the Abstract: Translating the Abstract Language for Vision Language Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025.

References

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning transferable visual models from natural language supervision". In: *Proceedings of International Conference on Machine Learning*. 2021.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[14] A. Brock, J. Donahue, and K. Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *Proceedings of International Conference on Learning Representations*. 2019.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[16] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. "Evaluating text-to-visual generation with image-to-text generation". In: *Proceedings of European Conference on Computer Vision*. 2024.

[17] L. Zhang, A. Rao, and M. Agrawala. "Adding conditional control to text-to-image diffusion models". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[18] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models". In: *Proceedings of the AAAI conference on artificial intelligence*. 2024.

[19] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. "Sigmoid loss for language image pre-training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[21] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra. "Embodied Question Answering in Photorealistic Environments with Point Cloud Perception". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[22] J. Krantz, S. Lee, J. Malik, D. Batra, and D. S. Chaplot. "Instance-Specific Image Goal Navigation: Training Embodied Agents to Find Object Instances". In: *arXiv preprint arXiv:2211.15876* (2022).

[23] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee. "Beyond the nav-graph: Vision-and-language navigation in continuous environments". In: *Proceedings of European Conference on Computer Vision*. 2020.

[24] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra. "ZSON: Zero-Shot Object-Goal Navigation using Multimodal Goal Embeddings". In: *Advances in Neural Information Processing Systems*. 2022.

[25] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song. "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[26] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov. "Object goal navigation using goal-oriented semantic exploration". In: *Advances in Neural Information Processing Systems* (2020).

[27] P. Marza, L. Matignon, O. Simonin, D. Batra, C. Wolf, and D. S. Chaplot. "AutoNeRF: Training Implicit Scene Representations with Autonomous Agents". In: *arXiv preprint arXiv:2304.11241* (2023).

[28] G. Georgios, S. Karl, W. Karan, D. Soham, M. Eleni, R. Dan, and D. Kostas. "Cross-modal Map Learning for Vision and Language Navigation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[29] C. Huang, O. Mees, A. Zeng, and W. Burgard. "Visual Language Maps for Robot Navigation". In: *Proceedings of the IEEE International Conference on Robotics and Automation*. 2023.

[30] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* (2021).

[31] O. Kwon, J. Park, and S. Oh. "Renderable Neural Radiance Map for Visual Navigation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[32] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov. "Learning To Explore Using Active Neural SLAM". In: *Proceedings of International Conference on Learning Representations*. 2020.

[33] T. Chen, S. Gupta, and A. Gupta. "Learning exploration policies for navigation". In: *Proceedings of International Conference on Learning Representations*. 2019.

[34] F. Taioli, F. Giuliari, Y. Wang, R. Berra, A. Castellini, A. Del Bue, A. Farinelli, M. Cristani, and F. Setti. "Unsupervised Active Visual Search with Monte Carlo planning under Uncertain Detections". In: *arXiv preprint arXiv:2303.03155* (2023).

[35] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. "Language-driven Semantic Segmentation". In: *Proceedings of International Conference on Learning Representations*. 2022.

[36] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler. "Open-vocabulary Queryable Scene Representations for Real World Planning". In: *arXiv preprint arXiv:2209.09874*. 2022.

[37] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind. "Unconstrained Scene Generation with Locally Conditioned Radiance Fields". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2021.

[38] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. "LERF: Language Embedded Radiance Fields". In: *arXiv preprint arXiv:2303.09553* (2023).

[39] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei. "Semantic-conditional diffusion networks for image captioning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[40] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems*. 2022.

[41] S. Peng, K. Genova, C. ". Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser. "OpenScene: 3D Scene Understanding with Open Vocabularies". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[42] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. "Open-vocabulary Object Detection via Vision and Language Knowledge Distillation". In: *International Conference on Learning Representations*. 2021.

[43] X. Dong, J. Bao, Y. Zheng, T. Zhang, D. Chen, H. Yang, M. Zeng, W. Zhang, L. Yuan, D. Chen, et al. "Maskclip: Masked self-distillation advances contrastive language-image pretraining". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[44] M. Savva et al. "Habitat: A Platform for Embodied AI Research". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[45] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese. "Gibson env: Real-world perception for embodied agents". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[46] M. Yang, P. Wu, and H. Feng. "MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities". In: *Engineering Applications of Artificial Intelligence* (2023).

[47] T. Wang, Y. Chen, M. Qiao, and H. Snoussi. "A fast and robust convolutional neural network-based defect detection model in product quality control". In: *The International Journal of Advanced Manufacturing Technology* (2018).

[48] C. S. Tsang, H. Y. Ngan, and G. K. Pang. "Fabric inspection based on the Elo rating method". In: *Pattern Recognition* (2016).

[49] S. H. Hanzaei, A. Afshar, and F. Barazandeh. "Automatic detection and classification of the ceramic tiles' surface defects". In: *Pattern Recognition* (2017).

[50] V. Zavrtanik, M. Kristan, and D. Skočaj. "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[51] H. Zhang, Z. Wu, Z. Wang, Z. Chen, and Y.-G. Jiang. "Prototypical Residual Networks for Anomaly Detection and Localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[52] P. Dhariwal and A. Nichol. "Diffusion models beat GANs on image synthesis". In: *Advances in Neural Information Processing Systems* (2021).

[53] J. Božič, D. Tabernik, and D. Skočaj. "Mixed supervision for surface-defect detection: From weakly to fully supervised learning". In: *Computers in Industry* (2021).

[54] J. Zhang, H. Su, W. Zou, X. Gong, Z. Zhang, and F. Shen. "CADN: a weakly supervised learning-based category-aware object detection network for surface defect detection". In: *Pattern Recognition* (2021).

[55] H. Zhang, Z. Wang, Z. Wu, and Y.-G. Jiang. "DiffusionAD: Denoising Diffusion for Anomaly Detection". In: *arXiv preprint arXiv:2303.08730* (2023).

[56] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep unsupervised learning using nonequilibrium thermodynamics". In: *Proceedings of International Conference on Machine Learning*. 2015.

[57] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems*. 2020.

[58] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[59] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. "8-bit Optimizers via Block-wise Quantization". In: *Proceedings of International Conference on Learning Representations* (2022).

[60] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016.

[61] L. Capogrosso, F. Cunico, D. S. Cheng, F. Fummi, and M. Cristani. "A Machine Learning-oriented Survey on Tiny Machine Learning". In: *IEEE Access* (2024).

[62] H. Deng and X. Li. "Anomaly detection via reverse distillation from one-class embedding". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[63] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. "Towards total recall in industrial anomaly detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[64] M. Rudolph, B. Wandt, and B. Rosenhahn. "Same same but Differnet: Semi-supervised defect detection with normalizing flows". In: *Winter Conference on Applications of Computer Vision*. 2021.

[65] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo. "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities". In: *ACM Computing Surveys* (2023).

[66] Y. Chen, Y. Ding, F. Zhao, E. Zhang, Z. Wu, and L. Shao. "Surface defect detection methods for industrial products: A review". In: *Applied Sciences* (2021).

[67] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak. "Unsupervised anomaly detection for surface defects with dual-siamese network". In: *IEEE Transactions on Industrial Informatics* (2022).

[68] C. Luan, R. Cui, L. Sun, and Z. Lin. "A siamese network utilizing image structural differences for cross-category defect detection". In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020.

[69] T. Defard, A. Setkov, A. Loesch, and R. Audigier. "Padim: a patch distribution modeling framework for anomaly detection and localization". In: *International Conference on Pattern Recognition*. 2021.

[70] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *Proceedings of International Conference on Learning Representations*. 2020.

[71] J. Ho and T. Salimans. "Classifier-free diffusion guidance". In: *arXiv preprint* (2022).

[72] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint* (2022).

[73] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". In: *Proceedings of International Conference on Learning Representations*. 2024.

[74] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs trained by a two time-scale update rule converge to a local nash equilibrium". In: *Advances in Neural Information Processing Systems*. 2017.

[75] P. von Platen et al. *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers. 2022.

[76] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint* (2014).

[77] G. Vigliocco, S.-T. Kousta, P. A. D. Rosa, D. P. Vinson, M. Tettamanti, J. T. Devlin, and S. F. Cappa. "The neural representation of abstract words: the role of emotion". In: *Cerebral Cortex* (2014).

[78] M. Brysbaert, A. B. Warriner, and V. Kuperman. "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior research methods* (2014).

[79] X. Yang, H. Zhang, D. Jin, Y. Liu, C.-H. Wu, J. Tan, D. Xie, J. Wang, and X. Wang. "Fashion captioning: Towards generating accurate descriptions with semantic rewards". In: *Proceedings of European Conference on Computer Vision*. 2020.

[80] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2016.

[81] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue. "Contrastive language and vision learning of general fashion concepts". In: *Scientific Reports* (2022).

[82] G. Cartella, A. Baldrati, D. Morelli, M. Cornia, M. Bertini, and R. Cucchiara. "OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data". In: *Proceedings of International Conference on Image Analysis and Processing*. 2023.

[83] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs". In: *arXiv preprint arXiv:2111.02114* (2021).

[84] M. Takagi, E. Simo-Serra, S. Iizuka, and H. Ishikawa. "What Makes a Style: Experimental Analysis of Fashion Prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2017.

[85] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. "Neuroaesthetics in fashion: Modeling the perception of fashionability". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2015.

[86] N. Inoue, E. Simo-Serra, T. Yamasaki, and H. Ishikawa. "Multi-label Fashion Image Classification with Minimal Human Supervision". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2017.

[87] S. Guo, W. Huang, X. Zhang, P. Srikhanta, Y. Cui, Y. Li, H. Adam, M. R. Scott, and S. Belongie. "The iMaterialist Fashion Attribute Dataset". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019.

[88] H. Xiao, K. Rasul, and R. Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[89] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. "Automatic spatially-aware fashion concept discovery". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2017.

[90] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. "Fashion-gen: The generative fashion dataset and challenge". In: *arXiv preprint arXiv:1806.08317* (2018).

[91] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris. "Fashion iq: A new dataset towards retrieving images by natural language feedback". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[92] P. Aggarwal. *Fashion Product Images Dataset (Small)*. 2019. URL: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small.

[93] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. "Where to Buy It: Matching Street Clothing Photos in Online Shops". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2015.

[94] J. Huang, R. S. Feris, Q. Chen, and S. Yan. "Cross-domain image retrieval with a dual attribute-aware ranking network". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2015.

[95] X. Ji, W. Wang, M. Zhang, and Y. Yang. "Cross-domain image retrieval with attention modeling". In: *Proceedings of ACM International Conference on Multimedia*. 2017.

[96] X. Wang and T. Zhang. "Clothes search in consumer photos via color matching and attribute learning". In: *Proceedings of ACM International Conference on Multimedia*. 2011.

[97] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. "Style Finder: Fine-Grained Clothing Style Detection and Retrieval". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2013.

[98] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan. "FashionVLP: Vision Language Transformer for Fashion Retrieval With Feedback". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[99] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao. "Kaleido-bert: Vision-language pre-training on fashion domain". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[100] M. U. Anwaar, E. Labintcev, and M. Kleinsteuber. "Compositional Learning of Image-Text Query for Image Retrieval". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Jan. 2021.

[101] S. Lee, D. Kim, and B. Han. "Cosmo: Content-style modulation for image retrieval with text feedback". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[102] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. "Effective Conditioned and Composed Image Retrieval Combining CLIP-Based Features". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[103] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. "Reproducible Scaling Laws for Contrastive Language-Image Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[104] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. "Eva-clip: Improved training techniques for clip at scale". In: *arXiv preprint arXiv:2303.15389* (2023).

[105] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303.

[106] B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* (1975).

[107] J. Shlens. "A tutorial on principal component analysis". In: *arXiv preprint* (2014).

[108] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution". In: *arXiv preprint arXiv:2409.12191* (2024).

[109] W. Hong, W. Wang, M. Ding, W. Yu, Q. Lv, Y. Wang, Y. Cheng, S. Huang, J. Ji, Z. Xue, et al. "Cogvlm2: Visual language models for image and video understanding". In: *arXiv preprint arXiv:2408.16500* (2024).

[110] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning". In: *Advances in Neural Information Processing Systems* (2022).

[111] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. "CLIPScore: A Reference-free Evaluation Metric for Image Captioning". In: *EMNLP*. 2021.

[112] Y. Kirstain et al. "Pick-a-pic: An open dataset of user preferences for text-to-image generation". In: *Advances in Neural Information Processing Systems* (2023).

[113] X. Wu et al. "Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis". In: *CoRR* (2023).

[114] J. Xu et al. "Imagereward: Learning and evaluating human preferences for text-to-image generation". In: *Advances in Neural Information Processing Systems* (2023).

[115] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. "T2I-compbench: A comprehensive benchmark for open-world compositional text-to-image generation". In: *Advances in Neural Information Processing Systems*. 2023.

[116] M. Yarom et al. "What you see is what you read? improving text-image alignment evaluation". In: *Advances in Neural Information Processing Systems* (2023).

[117] J. Cho et al. "Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation". In: *Proceedings of International Conference on Learning Representations*. 2024.

[118] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou. "When and why vision-language models behave like bag-of-words models, and what to do about it?" In: *arXiv preprint arXiv:2210.01936* (2022).

[119] D. Koishigarina et al. "CLIP Behaves like a Bag-of-Words Model Cross-modally but not Uni-modally". In: *arXiv preprint arXiv:2502.03566* (2025).

[120] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models". In: *Proceedings of International Conference on Machine Learning*. 2022.

[121] P. Esser et al. "Scaling rectified flow transformers for high-resolution image synthesis". In: *Proceedings of International Conference on Machine Learning*. 2024.

[122] A. Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[123] B. F. Labs. *FLUX*. https://github.com/black-forest-labs/flux. 2024.

[124] stability.ai. *SD-3.5-large*. https://huggingface.co/stabilityai. 2024.

[125] vivago.ai. *HiDream-I1-Full*. https://huggingface.co/HiDream-ai. 2025.

[126] M. Ku et al. "VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation". In: *ACL*. 2024.

[127] M. Otani et al. "Toward verifiable and reproducible human evaluation for text-to-image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[128] M. Karpinska et al. "The perils of using Mechanical Turk to evaluate open-ended text generation". In: *arXiv preprint arXiv:2109.06835* (2021).

[129] M. Ding et al. "Cogview2: Faster and better text-to-image generation via hierarchical transformers". In: *Advances in Neural Information Processing Systems* (2022).

[130] H. Chefer et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models". In: *ACM transactions on graphics* (2023).

[131] W. Feng et al. "Training-free structured diffusion guidance for compositional text-to-image synthesis". In: *arXiv preprint arXiv:2212.05032* (2022).

[132] M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie. "Fashionpedia: Ontology, segmentation, and an attribute localization dataset". In: *Proceedings of European Conference on Computer Vision*. 2020.

[133] H. Liu et al. "Improved baselines with visual instruction tuning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[134] W. Dai et al. "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning". In: *Advances in Neural Information Processing Systems*. 2023.

[135] F. Liang et al. "Open-vocabulary semantic segmentation with mask-adapted clip". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[136] T. Ren et al. "Grounded sam: Assembling open-world models for diverse visual tasks". In: *arXiv preprint arXiv:2401.14159* (2024).

[137] N. Ravi et al. "Sam 2: Segment anything in images and videos". In: *arXiv preprint arXiv:2408.00714* (2024).

[138] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models". In: *arXiv preprint* (2023).

[139] Y. Sun, Y. Liu, Y. Tang, W. Pei, and K. Chen. "Anycontrol: create your artwork with versatile control on text-to-image generation". In: *Proceedings of European Conference on Computer Vision*. 2024.

[140] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". In: *Proceedings of International Conference on Learning Representations*. 2021.

[141] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2017.

[142] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang. "Image generation from sketch constraint using contextual gan". In: *Proceedings of European Conference on Computer Vision*. 2018.

[143] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. Torr, and E. Shechtman. "Interactive sketch & fill: Multiclass sketch-to-image translation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[144] S. Koley, A. K. Bhunia, A. Sain, P. N. Chowdhury, T. Xiang, and Y.-Z. Song. "Picture that sketch: Photorealistic image generation from abstract sketches". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[145] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. "Encoding in style: a stylegan encoder for image-to-image translation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[146] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen. "Pretraining is all you need for image-to-image translation". In: *arXiv preprint* (2022).

[147] A. Voynov, K. Aberman, and D. Cohen-Or. "Sketch-guided text-to-image diffusion models". In: *ACM SIGGRAPH*. 2023.

[148] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations". In: *Proceedings of International Conference on Learning Representations*. 2022.

[149] P. Navard, A. K. Monsefi, M. Zhou, W.-L. Chao, A. Yilmaz, and R. Ramnath. "KnobGen: Controlling the Sophistication of Artwork in Sketch-Based Diffusion Models". In: *arXiv preprint* (2024).

[150] S. Koley, A. K. Bhunia, D. Sekhri, A. Sain, P. N. Chowdhury, T. Xiang, and Y.-Z. Song. "It's All About Your Sketch: Democratising Sketch Control in Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[151] X. Wang, H. Li, H. Fang, Y. Peng, H. Xie, X. Yang, and C. Li. "LineArt: A Knowledge-guided Training-free High-quality Appearance Transfer for Design Drawing with Diffusion Model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2025.

[152] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. "Gligen: Open-set grounded text-to-image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[153] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong. "Uni-controlnet: All-in-one control to text-to-image diffusion models". In: *Advances in Neural Information Processing Systems*. 2024.

[154] W. Nie, S. Liu, M. Mardani, C. Liu, B. Eckart, and A. Vahdat. "Compositional Text-to-Image Generation with Dense Blob Representations". In: *Proceedings of International Conference on Machine Learning*. 2024.

[155] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu. "Dense text-to-image generation with attention modulation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[156] V. Goel, E. Peruzzo, Y. Jiang, D. Xu, X. Xu, N. Sebe, T. Darrell, Z. Wang, and H. Shi. "PAIR Diffusion: A Comprehensive Multimodal Object-Level Image Editor". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[157] M. Hu, J. Zheng, D. Liu, C. Zheng, C. Wang, D. Tao, and T.-J. Cham. "Cocktail: Mixing multi-modality control for text-conditional image generation". In: *Advances in Neural Information Processing Systems*. 2023.

[158] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara. "Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[159] Z. Xie, H. Li, H. Ding, M. Li, X. Di, and Y. Cao. "HieraFashDiff: Hierarchical Fashion Design with Multi-stage Diffusion Models". In: *Proceedings of the AAAI conference on artificial intelligence*. 2025.

[160] J. Li, D. Li, S. Savarese, and S. Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *Proceedings of International Conference on Machine Learning*. 2023.

[161] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint* (2023).

[162] M. Li, Z. Lin, R. Mech, E. Yumer, and D. Ramanan. "Photo-sketching: Inferring contour drawings from images". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2019.

[163] D. Lukovnikov and A. Fischer. "Layout-to-Image Generation with Localized Descriptions using ControlNet with Cross-Attention Control". In: *arXiv preprint* (2024).

[164] S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Tseng, and H.-Y. Lee. "Adaptively-realistic image generation from stroke and sketch with diffusion model". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023.

[165] D. Bashkirova, J. Lezama, K. Sohn, K. Saenko, and I. Essa. "Masksketch: Unpaired structure-guided masked image generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[166] M. Oquab et al. "DINOv2: Learning Robust Visual Features without Supervision". In: *TMLR* (2024).

[167] G. Ilharco et al. *OpenCLIP*. 2021.

[168] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger. “Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization”. In: *International Journal of Computer Vision* (2022).

[169] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman. “Emu edit: Precise image editing via recognition and generation tasks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[170] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018.

Appendix A

Author Contributions and Reproducibility

This chapter summarises the publications included in this thesis, specifying my contributions for each work, and provides the resources required to reproduce the reported results.

Chapter 3: Language-enhanced RNR-Map: Querying Renderable Neural Radiance Field maps with natural language

```
@inproceedings{taioli2023language,
  title={Language-enhanced rnr-map: Querying renderable neural radiance field maps with
  author={Taioli, Francesco and Cunico, Federico and Girella, Federico and Bologna, Ricc
  booktitle={Proceedings of the IEEE/CVF International Conference on Computer Vision},
  year={2023}
}
```

Author Contributions

- Experiments with CLIP embedding injection in the RNR-Map.
- Inverse Projection for direction finding.
- Paper writing.

Reproducibility Resources

- Website: https://intelligolabs.github.io/Le-RNR-Map/
- Code: https://github.com/intelligolabs/Le-RNR-Map/tree/main
- Hardware: Experiments were conducted on a single NVIDIA Titan RTX GPU.

Chapter 4: Diffusion-Based Image Generation for In-Distribution Data Augmentation in Surface Defect Detection

```
@conference{capogrosso2024diffusion,
  title={Diffusion-Based Image Generation for In-Distribution Data Augmentation in Surfa
  author={Luigi Capogrosso and Federico Girella and Francesco Taioli and Dalla Chiara, M
  booktitle={International Joint Conference on Computer Vision, Imaging and Computer Gra
  year={2024}
}
```

Author Contributions

- Project direction.
- Research on related works on Anomaly Detection.
- Diffusion augmentation experimentations.
- Paper writing.

Reproducibility Resources

- Website: https://intelligolabs.github.io/in_and_out/
- Code: https://github.com/intelligolabs/in_and_out
- Hardware: Experiments were conducted on a single NVIDIA Titan RTX GPUs.

Chapter 5: Leveraging Latent Diffusion Models for Training-Free in-Distribution Data Augmentation for Surface Defect Detection

```
@InProceedings{girella2024leveraging,
    title     = {{Leveraging Latent Diffusion Models for Training-Free in-Distribution D
    author    = {Girella, Federico and Liu, Ziyue and Fummi, Franco and Setti, Francesco
    booktitle = {International Conference on Content-Based Multimedia Indexing (CBMI)},
    year      = {2024}
}
```

Author Contributions

- State of the Art research on Generative models.

- Experimentations with Stable Diffusion XL model.
- Domain analysis for prompt tuning.
- Paper writing.

Reproducibility Resources

- Website: https://intelligolabs.github.io/DIAG/
- Code: https://github.com/intelligolabs/DIAG
- Hardware: Experiments were conducted on a single NVIDIA RTX 4090 GPU.

Chapter 6: Seeing the Abstract: Translating the Abstract Language for Vision Language Models

```
@inproceedings{talon2025seeing,
  title={Seeing the Abstract: Translating the Abstract Language for Vision Language Mode
  author={Talon, Davide and Girella, Federico and Liu, Ziyue and Cristani, Marco and Wan
  booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
  year={2025}
}
```

Author Contributions

- State of the Art research on Fashion image retrieval.
- Fine-tuning experimentations with contrastive VLMs.
- Paper writing.

Reproducibility Resources

- Website: https://davidetalon.github.io/fashionact-page/
- Code: https://github.com/davidetalon/fashionact
- Hardware: Experiments were conducted on 8 A100 GPUs (CINECA-HPC).

Chapter 7: Seeing the Abstract: Translating the Abstract Language for Vision Language Models

```
@inproceedings{liu2025evaluating,
  title={Evaluating Attribute Confusion in Fashion Text-to-Image Generation},
  author={Liu, Ziyue and Federico, Girella and Yiming, Wang and Davide, Talon},
  booktitle={Proceedings of International Conference on Image Analysis and Processing},
  year={2025}
}
```

Author Contributions

- Idea exploration.
- Paper writing.

Reproducibility Resources

- Website: https://intelligolabs.github.io/L-VQAScore/
- Code: https://github.com/intelligolabs/L-VQAScore
- Hardware: Experiments were conducted on 8 A100 GPUs (CINECA-HPC).

Chapter 8: Seeing the Abstract: Translating the Abstract Language for Vision Language Models

```
@inproceedings{girella2025lots,
  title={LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairin
  author={Girella, Federico and Talon, Davide and Liu, Ziyue and Ruan, Zanxi and Wang, Y
  booktitle={Proceedings of the IEEE/CVF International Conference on Computer Vision},
  year={2025}
}
```

Author Contributions

- Idea conceptualization.
- Project direction.
- Related works research.
- Methodology experimentations.

- Paper writing.
- Paper Oral presentation at ICCV25.

Reproducibility Resources

- Website: https://intelligolabs.github.io/lots/
- Code: https://huggingface.co/federicogirella/lots
- Hardware: Experiments were conducted on 8 A100 GPUs (CINECA-HPC).

Appendix B

Funding and Acknowledgments

All the works I have carried out throughout my Ph.D. have been performed under the NextGenerationEU Italiadomani grant (PNRR, M4C2, Investimento 3.3), funded by NextGeneration EU. Furthermore HUMATICS, a SYS-DAT Group company, partially funded my Ph.D.

The works in this thesis have received additional funding from other different organizations. The following is a list of each project presented in this thesis and their supporting funds.

[6] - Diffusion-based Image Generation for In-distribution Data Augmentation in Surface Defect Detection

Consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043).

[8] - Leveraging Latent Diffusion Models for Training-Free In-Distribution Data Augmentation for Surface Defect Detection

Consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043).

[11] - Seeing the Abstract: Translating the Abstract Language for Vision Language Models

LoCa AI, funded by Fondazione CariVerona (Bando Ricerca e Sviluppo 2022/23)

PNRR FAIR - Future AI Research (PE00000013)

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain.

[10] - Evaluating Attribute Confusion in Fashion Text-to-Image Generation

LoCa AI, funded by Fondazione CariVerona (Bando Ricerca e Sviluppo 2022/23)

PNRR FAIR - Future AI Research (PE00000013)

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain.

[9] - LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing

LoCa AI, funded by Fondazione CariVerona (Bando Ricerca e Sviluppo 2022/23)

PNRR FAIR - Future AI Research (PE00000013)

Consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23062022, ECS_00000043)

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain.