information

MDPI

Article

VAD-CLVA: Integrating CLIP with LLaVA for Voice Activity Detection

Andrea Appiani [1] **and Cigdem Beyan** [2,*]

[1] Department of Management, Information and Production Engineering, University of Bergamo, 24127 Dalmine, Italy
[2] Department of Computer Science, University of Verona, 37134 Verona, Italy
* Correspondence: cigdem.beyan@univr.it

Abstract: Voice activity detection (VAD) is the process of automatically determining whether a person is speaking and identifying the timing of their speech in an audio-visual data. Traditionally, this task has been tackled by processing either audio signals or visual data, or by combining both modalities through fusion or joint learning. In our study, drawing inspiration from recent advancements in visual-language models, we introduce a novel approach leveraging Contrastive Language-Image Pretraining (CLIP) models. The CLIP visual encoder analyzes video segments focusing on the upper body of an individual, while the text encoder processes textual descriptions generated by a Generative Large Multimodal Model, i.e., the Large Language and Vision Assistant (LLaVA). Subsequently, embeddings from these encoders are fused through a deep neural network to perform VAD. Our experimental analysis across three VAD benchmarks showcases the superior performance of our method compared to existing visual VAD approaches. Notably, our approach outperforms several audio-visual methods despite its simplicity and without requiring pretraining on extensive audio-visual datasets.

Keywords: active speaker; voice activity detection; social interactions; vision language models; generative multimodal models; panel discussions

Academic Editors: Yang Gao and Shibo Zhang

Received: 12 January 2025
Revised: 3 March 2025
Accepted: 6 March 2025
Published: 16 March 2025

Citation: Appiani, A.; Beyan, C. VAD-CLVA: Integrating CLIP with LLaVA for Voice Activity Detection. *Information* **2025**, *16*, 233. https://doi.org/10.3390/info16030233

1. Introduction

voice activity detection (VAD) is the process of automatically determining whether a person is speaking or not in a recording, thereby addressing the question of "Who is speaking and When?". This task is crucial in various real-world applications such as human–robot interaction [1], speech diarization [2–4], multiparty dialogues among humans [5], social behavior analysis [6], automatic speech recognition [7], speech enhancement [8], and emotion recognition [9].

VAD has traditionally been approached through audio processing [10–12], which presents challenges, particularly in scenarios like unstructured social gatherings where multiple speakers talk simultaneously or when there is a high number of subjects or in case of a close proximity between speakers. With the advancement of Convolutional Neural Networks (CNNs), the integration of video modality alongside audio has markedly enhanced VAD performance [13,14]. However, audio-visual VAD still faces challenges, particularly in effectively modeling both modalities jointly or requiring speech enhancements, as shown in [15]. Furthermore, audio-visual VAD may not always be applicable due to technical, ethical, or legal considerations regarding audio [16,17]. Consequently, some approaches emphasize the importance of addressing this task solely based on visual data. Indeed, in some

instances, visual VAD has demonstrated superior performance compared to audio-visual VAD [16–18], highlighting the crucial role of the video modality in this domain.

The utilization of head crops, especially the regions encompassing individuals' faces, stands out as the primary visual cues employed in both visual VAD and audio-visual VAD [13–15,19]. While some studies have concentrated on linking lip movements with audio signals [20–23], others have utilized facial landmarks [24,25]. Despite the latest advancements in audio-visual VAD often relying on facial images to achieve state-of-the-art (SOTA) results, some research emphasizes the importance of analyzing upper-body motion [16–18] and gestures [26,27], proving the relevance of these cues in detecting voice activity.

Our proposal follows the findings of [16–18,28], emphasizing the significance of processing **upper-body images** for VAD. It most closely intersects with visual VAD since the input to our model is a video segment that is associated with the VAD label of an individual as *speaking* or *not speaking*. On the other hand, our initial tests with **Generative Large Multimodal Models (LMMs)**, especially when provided with upper-body crop images, along with insights from studies like [29] that focus on emotion recognition, reveal the proficiency of LMMs in describing images by focusing on facial muscles and gestures, which is crucial for capturing essential cues for VAD. Encouraged by these findings, we opt to leverage **text descriptions** automatically generated by Large Language and Vision Assistant (LLaVA) [30,31] using prompt engineering when the input image represents the central frame of a video segment depicting an individual. Additionally, motivated by the enhanced performance of visual-language models (VLMs), particularly CLIP [32] across various visual downstream tasks, e.g., [33–35], we adopt pretrained CLIP models. CLIP [32] employs separate encoders for visual and textual data, which are aligned within a shared embedding space using contrastive loss [32]. We hypothesize that textual embeddings derived from LLaVA-generated descriptions provide complementary information to visual features, capturing, e.g., subtle facial expressions and upper-body movements that are crucial for VAD. The joint learning of visual and textual features within CLIP's contrastive learning framework might enhance the model's ability to distinguish between speaking and non-speaking individuals more effectively than using visual features alone.

This study introduces **VAD-CLVA**, a novel VAD method that leverages CLIP's contrastive learning framework to integrate textual embeddings from LLaVA [30,31] with visual features. This is the first work to incorporate prompt-engineered LLaVA-generated descriptions for VAD. Specifically, our approach processes upper-body video segments through a visual encoder while simultaneously encoding LLaVA-generated textual descriptions, enabling a richer representation of speaking activity. In detail, our visual encoder processes video segments containing upper-body images, while the text encoder handles textual descriptions provided by a LLaVA [30,31]. These encoders produce embeddings, which are then concatenated and fused using a deep neural network to perform the VAD task. While our approach is not inherently multimodal in terms of data since it focuses on video segments comprising only upper-body frames, it is crucial to note the LLaVA's ability to interpret not just body movements like arm gestures but also facial activity. On the other hand, our proposed architecture is multimodal, incorporating both visual and textual cues. To our knowledge, this is the first time that text data are being used for VAD.

VAD-CLVA, tested on three VAD benchmarks, demonstrates superior performance compared to all SOTA visual VAD methods. It also yields promising results, achieving performance levels comparable to or better than several SOTA audio-visual VAD approaches, despite not being pretrained on large audio-visual VAD datasets. This underscores the effectiveness of our model, which is trained and tested directly on visual VAD benchmarks. In detail, the on-par or better performance of VAD-CLVA demonstrates that eliminating the

need for audio data are feasible, making it suitable for scenarios where audio is unavailable or unreliable. By leveraging LLaVA-generated descriptions, VAD-CLVA enhances discriminative power, as its visual-only version performs worse. Moreover, VAD-CLVA does not require domain adaptation to achieve performance comparable to methods that rely on it.

This study seeks to answer the following research questions.

- Can VLMs, particularly CLIP, enhance VAD by incorporating textual descriptions generated by LMMs, particularly LLaVA, alongside visual features?
- Can a model trained on visual data, together with text obtained through prompt engineering, achieve performance comparable to SOTA methods?

The main contributions and findings of our study can be summarized as follows:

- We introduce a novel VAD method that effectively utilizes visual-language pretraining techniques. To the best of our knowledge, this is the first work to adopt CLIP [32] for VAD. Consequently, we demonstrate that our model, VAD-CLVA, surpasses SOTA visual VAD methods, affirming the utility of joint learning of text descriptions and visual features.
- This is also the first attempt to employ LMMs (i.e., LLaVA [30,31]) with prompt engineering to perform VAD and to generate text descriptions corresponding to individuals' speaking activity when their upper-body images are the inputs. While the standalone LMM model may not match the effectiveness of our VAD-CLVA for the VAD task, its text descriptions enhance the utilization of spatio-temporal upper-body features, thereby improving VAD-CLVA's performance.
- Through extensive experimentation, we demonstrate that our approach outperforms all visual methods as well as a standalone LMM. Moreover, our VAD-CLVA consistently achieves results comparable to or surpassing the audio-visual SOTA, even if it employs a simpler pipeline compared to them and without the necessity of pretraining on audio-visual data.

The remainder of this paper is structured as follows. In Section 2, we provide an overview of the existing literature on VAD and previous studies on CLIP, emphasizing the unique aspects of our approach. The proposed method, VAD-CLVA, is described in Section 3 together with its implementation details. Following that, in Section 4, we introduce the datasets, and the evaluation metrics we used in line with the SOTA, and also present a comprehensive ablation study. This section further includes a comparative analysis between VAD-CLVA and SOTA. Finally, Sections 5 and 6 summarize the major findings of this study, highlight the limitations, and present possible future directions.

2. Related Work

In this section, we provide an overview of the current body of literature on VAD and summarize the various applications of CLIP [32]. Additionally, we outline the rationale for employing CLIP [32] and Generative Large Multimodal Models (LMMs) [30,31] in the context of VAD.

2.1. Voice Activity Detection

Earlier works have tackled the task of VAD solely through audio signal processing, as evident from a broad literature, e.g., [10–12]. However, performing audio-based VAD can be challenging, especially in real-life scenarios where sounds may originate from multiple speakers simultaneously and when the speakers are nearby.

Particularly, with the advancement of Convolutional Neural Networks (CNNs), visual information has also been integrated into audio-based VAD, resulting in the development of several multimodal VAD approaches that employ both audio and visual cues. These

studies, which perform audio-visual VAD (also referred to as audio-visual active speaker detection), typically consider the temporal dependency between audio and visual data. This involves the application of Recurrent Neural Network (RNN) [36,37], Gated Recurrent Unit (GRU) [38], Long Short-Term Memory (LSTM) [39,40], and Transformer Layer [13]. Audio-visual VAD, when fully leveraging cross-modal synchronization information, can achieve highly successful performance [15,19]. However, much of the existing work relies on separately encoding the unimodal features of audio and video, limiting the exploitation of cross-modal synchronization information. For instance, in extracting visual features, some studies employ 3D CNNs to capture temporal dependencies from video data [13,14]. Conversely, for audio features, CNNs are often utilized with log-Mel or Short-Time Fourier Transform (STFT) spectrograms as inputs [13], or directly applied to the audio waveform [14]. Furthermore, in many studies, visual information is primarily employed to link the active speaker with speech, reflecting a methodology that does not fully leverage visual data [41]. It is worth noting that audio-visual VAD techniques utilize the head crops of individuals to extract visual features, sometimes with a focus on lip movements [23,42]. However, such approaches are particularly effective only for frontal facial images.

On the other hand, numerous methods exclusively rely on visual data to address VAD. These visual VAD approaches can be categorized into two groups: those that analyze facial cues, such as facial landmarks [5,20,43–46], and those that consider body cues, including hand gestures, head movements, and upper-body motion [16–18,26,27]. Interestingly, some visual VAD studies, despite lacking audio, have demonstrated superior performance compared to a few audio-visual methods [16–18]. Overall, visual VAD studies offer a significant alternative, especially in scenarios where accessing the audio signal is not feasible due to functional, legislative, or moral limitations [16–18,42,45]. On the other hand, visual-only VAD models struggle to match the performance of some of the audio-visual approaches due to the inherent limitations of visual cues in capturing speech activity. While lip movements, facial expressions, and body gestures provide strong signals for VAD, they are often ambiguous, especially in the presence of occlusions, head movements, or speakers with minimal facial motion [19]. Instead, audio provides a direct and reliable indicator of speech presence, making audio-visual models mostly more robust, particularly in challenging scenarios like multispeaker environments or low-resolution videos [17].

In this study, inspired by the improvements in various visual recognition tasks achieved through a VLM, we introduce a method that combines visual and text modalities to tackle the VAD task. To do so, we utilize CLIP [32] while the text input is produced through the LMM: LLaVA [30,31]. This is the first work using video clips and associated generated text within a CLIP framework [32] to perform VAD. The proposed architecture is much simpler for example compared to [15] simultaneously applying speech enhancement, or [13] performing long-term temporal intra-speaker context processing. Besides, we do not require using the pretrained weights on large-scale VAD datasets as applied in [19]. Our method does not explicitly extract detailed facial muscle movements, expressions, or upper-body motion in general. Instead, we rely on image embeddings from CLIP, which inherently capture salient visual cues without specifically modeling facial expressions or fine-grained muscle activity. However, as shown in the following sections, LLaVA-13B [30,31] is used to generate textual descriptions of the visual content, which may include references to facial expressions or gestures. However, this is not under our control and entirely depends on LLaVA's captioning performance. Nevertheless, it provides an additional layer of semantic understanding beyond raw visual embeddings, contributing to the refinement of the VAD process, as evidenced by the ablation study we perform below.

The reason no prior VAD studies have explored multimodal approaches incorporating textual embeddings from LMMs is likely that VAD has traditionally been framed as a

visual, audio, or audio-visual task, with limited exploration of descriptive language representations. Additionally, before the emergence of powerful LMMs, generating semantically rich textual descriptions from images with sufficient accuracy was challenging. The introduction of models like CLIP now enables us to harness textual embeddings that describe subtle visual features relevant to speaking activity, offering a novel and complementary modality for VAD.

2.2. CLIP and LLaVA

A VLM processes images and their corresponding textual descriptions, learning to connect information from both modalities. The visual component of the model usually captures spatial features from images, while the language model encodes information from text. CLIP, which stands for Contrastive Language-Image Pretraining, is one of the most prominent models in this category. The usage of CLIP includes zero-shot classification and fine-tuning for downstream tasks, as demonstrated in various papers, such as [32,47,48]. Unlike frequently applied downstream tasks such as image classification [49,50] and image segmentation [51,52], CLIP is also utilized in diverse applications such as image enhancement [53], monocular depth estimation [33], text-to-shape generation [54], image manipulation [55], medical image processing [52], anomaly detection [56], gaze estimation [35], emotion recognition [34,57,58], and so forth.

Recent studies, such as [34], have emphasized CLIP's ability to extract robust features for representing facial images and expressions, both through fine-tuning and zero-shot learning. Furthermore, our investigations with other LMMs, particularly LLaVA-13B [30,31], have highlighted the potential of these models (see Section 4 for results) in describing a speaking individual. We observed that LLaVA-13B demonstrates adeptness in focusing on hand gestures, facial expressions, and the shape of the mouth. Additionally, the research presented in [29] provides compelling evidence for the efficacy of LLaVA in facial emotion recognition. All these findings have strengthened our confidence in utilizing LLaVA for generating text descriptions through prompt engineering in conjunction with using CLIP to realize VAD.

3. Proposed Method: VAD-CLVA

The structure of VAD-CLVA is depicted in Figure 1. It consists of both a visual and a textual component, predominantly leveraging the CLIP architecture [32]. Our methodology involves short video segments capturing individuals' upper body, accompanied by textual descriptions regarding their speaking status, which are extracted through prompt engineering. The aim is to harness both video embeddings capturing temporal information and text embeddings to strengthen a neural network in determining whether the depicted individual is speaking or not.

3.1. Preliminaries

Contrastive Language-Image Pretraining (CLIP) [32] employs a dual-encoder framework consisting of a visual encoder denoted as $\mathbf{E_v}$ and a text encoder termed as $\mathbf{E_t}$. $\mathbf{E_v}$ processes input images $I \in \mathbb{R}^{H\times W\times 3}$ by dividing them into a sequence of fixed-size patches. These patches, combined with a learnable class token, are transformed to a unified vision-language embedding space, resulting in a final visual feature $f_v = \mathbf{E_v}(I) \in \mathbb{R}^d$, where d represents the dimensionality of the features. On the other hand, $\mathbf{E_t}$ converts textual input (e.g., prompt) shown as Tx into text embeddings, augmenting them with a learnable class token to create an input feature matrix. This matrix is then processed to extract textual features $f_t = \mathbf{E_t}(Tx)$. By aiming to maximize the similarity denoted as `sim` between the matched text-image pairs and to minimize the `sim` for the unmatched pairs, CLIP is

trained using a contrastive loss function. CLIP uses textual prompts to generate specific text features and compute the prediction by calculating the distance to an image feature as:

$$\mathcal{P}(y|v) = \frac{\exp(\mathtt{sim}(f_v, f_t^y)/\tau)}{\sum_{i=1}^{K} \exp(\mathtt{sim}(f_v, f_t^i)/\tau)} \quad (1)$$

where τ is the temperature parameter, controlling the scale or smoothness of the probability distribution.

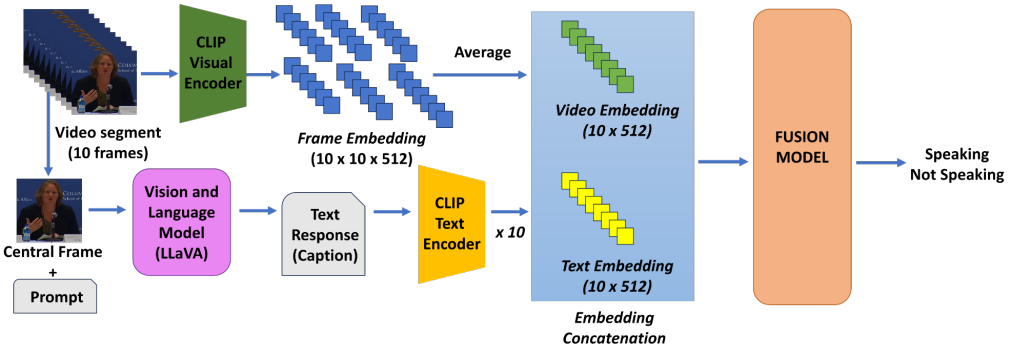

Figure 1. An overview of the proposed approach: VAD-CLVA. Our approach entails short video segments capturing individuals' upper body, along with textual descriptions regarding their speaking status, derived through prompt engineering. The goal is to harness both video and text embeddings to enable a fusion network to determine whether the person depicted is speaking or not. The input video segment consists of 10 frames, each represented as an embedding of size 10×512 after being input to the CLIP visual encoder. These 10-frame embeddings, which are $10 \times 10 \times 512$, are averaged along the temporal channel. The central frame of these 10 frames, together with a prompt, is input to the LLaVa model [30,31] to generate a textual response (caption). The caption is then provided to the CLIP text encoder, resulting in a single text embedding. This text embedding is replicated 10 times and concatenated with the 10×512 video embeddings to be given as an input to a fusion model designed as either an MLP or a Transformer network to predict the VAD label.

3.2. Formal Description

Given a video segment $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_T}\}$ consisting of T frames and the **VAD label** L, we initially preprocess these inputs to enhance the utilization of the pretrained CLIP model. Therefore, the input frames are resized to match the input size expected by **CLIP's visual encoder**, which can be either a Transformer or a Residual Network. These resized frames are then embedded into a set of visual tokens $T \times N \times D$, where N is the number of tokens and D is the dimension of each token. Subsequently, we compute the average of these embeddings to obtain a tensor $F_v \in \mathbb{R}^{N \times D} = \{f_{v_1}, f_{v_2}, \ldots, f_{v_T}\}$ along the temporal channel.

The central frame of the video segment, denoted as $\{v_{T/2}\}$, is utilized to generate text input for **CLIP's text encoder**. Prompt engineering is employed, where $\{v_{T/2}\}$ is paired with a prompt and provided to a **LLaVA** to generate text responses. These responses are then passed to the text encoder of CLIP, resulting in text tokens $F_t \in \mathbb{R}^{N \times D} = \{f_{T/2}, \ldots, f_{T/2}\}$, where $\{f_{T/2}\}$ is replicated to arrive to the size of F_v. Finally, both the visual and textual embeddings are concatenated $[F_v, F_t]$ and fed as input to the **fusion model** FN for classifying the entire video segment as either speaking or not speaking.

The **FN** is designed as either a Multilayer Perceptron (MLP), denoted as $\mathbf{FN_{MLP}}$, or a Transformer, referred to as $\mathbf{FN_T}$, noticing that depending on the size of the training data, one model may outperform the other. Empirical evidence in the next section demonstrates that $\mathbf{FN_T}$ typically requires more training data compared to $\mathbf{FN_{MLP}}$ to achieve better performance. $\mathbf{FN_T}$ includes a normalization layer to process the input, which is transformed

into three branches, i.e., Q, K, and V. Multihead self-attention described in terms of self-attention layers are defined as $MLP(SA(Q,K)V)$, where $SA(Q,K) = \text{Softmax}(\frac{QK^{Tr}}{\sqrt{c}})$ when c denotes feature dimension and Tr is the transpose operation. The outputs are given to the classification head for VAD. On the other hand, $\mathbf{FN_{MLP}}$ consists of multiple dense layers, including an input layer, multiple hidden layers, and an output layer. Each neuron applies an activation function to the weighted sum of its inputs, introducing non-linearity into the network and enabling it to learn complex patterns in the data.

3.3. Implementation Details

The input V comprises frames showing the upper body of an individual for a fixed duration of 10 frames, each sharing the same ground-truth label, as described in [16,18]. During training, if there is no video segment sharing the same ground-truth data for 10 frames, the remaining frames are repeated until a video segment of 10 frames is obtained.

We conducted experiments with pretrained CLIP models, encompassing Residual Networks and Vision Transformers (ViT), which are ResNet101 and ViT-B/16, respectively. These models were selected due to their strong performance across a wide range of vision tasks, including image recognition and feature extraction. We chose them to ensure flexibility in the VAD-CLVA, as both convolutional networks and transformers have proven to be effective in different scenarios. Depending on the dataset size and characteristics, either ResNet101 or ViT-B/16 might be more suitable, allowing us to leverage the strengths of both architectures. Each of these models includes both a visual and a text encoder, accompanied by a preprocessing function and a tokenizer. The preprocessing function is responsible for adapting images to the format accepted by the visual encoder, while the tokenizer divides the text into tokens suitable for input into the text encoder [32].

Each frame within a V undergoes resizing to dimensions of 224 $\times$ 224 to match the input image size expected by the CLIP visual encoder. Subsequently, the frame is partitioned into nine non-overlapping patches. Both these patches (allowing us to capture the local features) and the complete upper-body image (enabling us to extract global features) are then fed into a visual encoder, producing embeddings of size 512 for each. Consequently, the output for each frame yields a vector of dimensions 10 $\times$ 512 (see Figure 2). This process is repeated for every frame in the V, resulting in a total of 100 embeddings with the dimension of 512 (10 $\times$ 10 $\times$ 512). Finally, the average is computed along the temporal dimension to consolidate the embeddings into a final vector of dimensions 10 $\times$ 512, representing the entire V.

As an LMM, we selected LLaVA-13B [30,31]. The frame $\{v_{T/2}\}$, in our case, the 5th frame for a V composed of 10 frames, was chosen and provided as an input to LLaVA-13B, along with a textual prompt, to generate a textual response. Note that the input video segment consists of 10 frames, corresponding to approximately 0.33 s of data, assuming a dataset of 30 frames per second. Based on this, we hypothesize that querying LLaVA-13B with information from all 10 frames would not provide a significant improvement in performance. Instead, querying only the middle frame as we perform must be sufficient for extracting relevant textual information. Additionally, querying just one frame instead of all 10 significantly reduces computational cost, as querying LLaVA-13B is resource-intensive, and this choice results in a considerable reduction in training time.

We experimented with two prompts: *(1) Is the person speaking? Answer with yes or no.* and *(2) Is the person speaking? Explain why in a few words*. The first prompt consistently yields a response of "yes" or "no". For the second prompt, we set the temperature to 0.2 and the maximum number of tokens to 50 (note that this is smaller than CLIP's maximum token limit). The text generated by LLaVA was divided into sentences, which can vary in number across different images. We obtained a CLIP text encoding for each sentence, then averaged

them by summing the embeddings and dividing by the number of sentences generated by LLaVA. By doing so, (i) we aim to ensure that all information is used, (ii) we balance the contribution of each part of the description rather than over-relying on the first few words, and (ii) we naturally avoid exceeding token limits, making it less likely to hit CLIP's input limit.

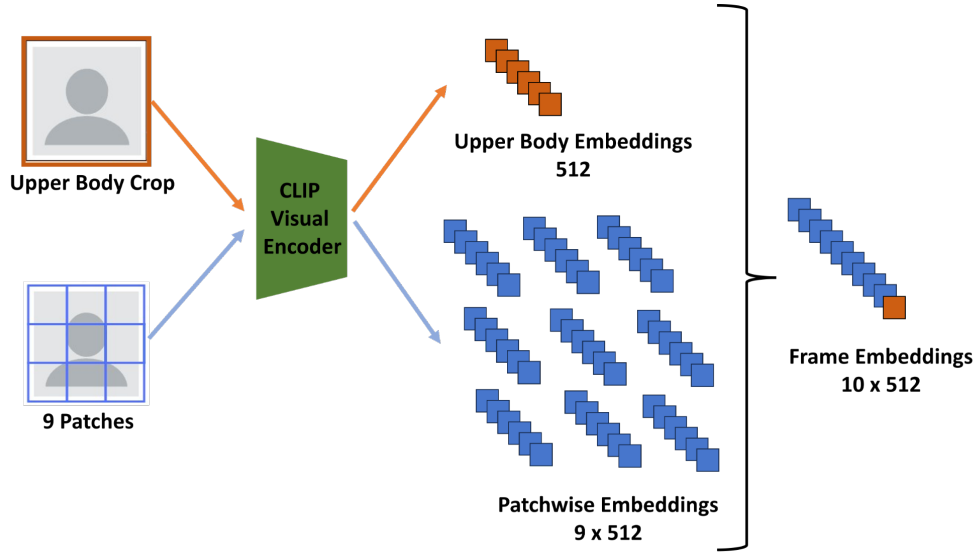

Figure 2. The extraction of visual embeddings involves using the CLIP visual encoder. That encoder takes a frame consisting of an upper-body crop of an individual along with the nine non-overlapping patches obtained from that frame. This process captures both local and global features. From 10 inputs, we obtain a frame embedding size of 10 × 512 (1 × upper-body embeddings + 9 × patchwise embeddings). This procedure is repeated for all frames within a given video segment.

With this second prompt, we aimed to test the expressive capabilities of the model, generating more complex and varied responses. Using the first prompt allowed us to evaluate the standalone performance of LLaVA-13B's VAD capabilities, as explained in the next section. For the second prompt, we computed the similarity score between the captions generated for "no speaking" and "speaking", which resulted in a similarity score on average around 0.5, and we assumed that this score should be sufficient, as 0.5 indicates that there is a moderate level of overlap in the semantic features captured by the prompts. This could mean that the model sees some common characteristics between the classes, such as both potentially involving human behavior or speech-related activities, the same faces, and the same background. However, the similarity is not strong enough to be fully confident that the two classes are the same or even very closely related. Indeed, based on empirical testing, we found that this was sufficient to positively confirming the research questions given in Section 1.

While there is a single set of textual tokens obtained for each video segment, there are visual tokens for every frame within that video segment. To address this imbalance, our solution is to replicate the textual tokens for each frame, aligning the number of textual tokens with the number of visual tokens. This preserves the significance of textual information without any reduction in its importance.

The $\mathbf{FN_T}$ takes inputs consisting of 20 tokens: 10 visual and 10 textual. It includes two attention heads, two linear layers responsible for increasing the dimensionality of the embeddings from 512 to 768, and linear layers responsible for classification, along with two normalization layers processing the classification results to output the final logits. In $\mathbf{FN_{MLP}}$, there are four linear layers with input sizes of 1024, 512, 256, and 1, respectively, with the last layer corresponding to the classification layer. ReLU is used as

the activation function, and each linear layer is followed by batch normalization. Similar to the $\mathbf{FN_T}$, $\mathbf{FN_{MLP}}$ also takes concatenated visual and textual embeddings as input, without considering tokens, resulting in a single embedding of size 1024.

The learning rate for both our $\mathbf{FN_T}$ and $\mathbf{FN_{MLP}}$ models was set to different values: 0.01, 0.001, and 0.0001, with a weight decay of 1×10^{-4}. We trained the models for up to 50 epochs using the Adam optimizer. During training, we utilized the Binary Cross-Entropy Loss with Logits (BCEWithLogLoss) as the loss function. The batch size, following the settings in [17,18], was set to 128, with 64 speaking and 64 not speaking randomly selected V segments used in each batch. We also employed the same data augmentation procedure applied in [17,18].

4. Experimental Analysis and Results

The experimental analysis includes comparisons with the SOTA and an ablation study, where different combinations of the VAD-CLVA were tried along with assessing the LMM's performance for VAD. For evaluation purposes, we utilize Columbia [42], Modified Columbia [16], and RealVAD [17], which are three popular benchmarks for visual VAD, and in particular, they were used by the SOTA presenting upper-body activity-based VAD (see Figure 3).

Figure 3. Example frames from the datasets used in this paper. On the left are the Columbia and Modified Columbia datasets [16,42], and on the right is the RealVAD dataset [17]. Green boxes indicate the active speakers, while red boxes denote other participants with VAD ground-truth who are not speaking at that moment.

The Columbia dataset [42] contains an 87-min video of a panel discussion including several individuals' speaking activity annotations, in which 2–3 speakers are visible at a time. The Modified Columbia dataset [16] is derived from the Columbia dataset [42]. Unlike the Columbia dataset, which contains a significant number of not speaking frames, Modified Columbia has more balanced classes. This more balanced distribution makes the evaluation less skewed, leading to more reliable assessments while the training splits contain fewer samples than Columbia, as noted in [16]. Following SOTA, the evaluations of the Columbia and the Modified Columbia datasets are performed for five panelists: Bell, Bollinger, Lieberman, Long, and Sick. While the Columbia dataset [42] includes bounding box annotations for the head position of each panelist, herein, for both Columbia and Modified Columbia, we use the upper-body crops supplied by [18,28]. On the other hand, the RealVAD dataset [17] comprises an 83-minute panel discussion featuring panelists from various ethnic backgrounds, including British, Dutch, French, German, Italian, American, Mexican, Columbian, and Thai. The video is recorded using a static, mounted camera, capturing the nine panelists in a full shot. The panelists sit in two rows and are engaged in various activities, leading to potential partial occlusions of the upper body.

The standard evaluation settings of visual VAD, including **leave-one-person-out** cross-validation (i.e., in each fold of cross-validation, the testing set comprises data from a single individual, while the training set incorporates data from all other individuals) and **F1-score** as the evaluation metric are used. Results are presented as the F1-score for each person, together with the average and standard deviation across all individuals. The application of leave-one-person-out cross-validation facilitates evaluating the VAD model's ability to generalize to unseen individuals, taking into account the variability in head and body motion patterns among different people, known as the domain-shift problem [17].

4.1. Ablation Study

The results of the ablation study, in which the different combinations of the VAD-CLVA are tested, are given in Tables 1 and 2. Below, we delve into a detailed discussion of our findings.

Table 1. The ablation study evaluates the performance of various combinations of VAD-CLVA on the Modified Columbia dataset [16] in terms of the F1-score (%). The best results are highlighted in bold, while the second-best results are underlined. Vis. stands for visual modality and $\mathbf{E_v}$ represents the visual encoder. Exp. indicates the index of the experiments. See text for the descriptions of *prompt 1*, fixed and variable.

Exp.	Model	Modality	$\mathbf{E_v}$	Text Prompt	Bell	Sick	Long	Boll.	Lie.	Avg.	Std.
1	LLaVA-13B [30,31]	Vis. & Text	X	prompt 1	82.06	54.89	41.41	76.66	76.97	66.40	17.46
2	ViT-B/16 + MLP	Vis.	pretrained	X	66.53	73.88	44.94	83.6	94.96	72.78	18.87
3	ResNet101 + MLP	Vis.	pretrained	X	66.77	84.99	78.76	85.25	90.77	81.31	9.17
4	ResNet101 + MLP	Vis.	fine-tuned	X	71.29	89.98	81.42	83.08	88.88	82.93	7.46
5	$\mathbf{FN_{MLP}}$	Vis. & Text	pretrained	fixed	75.31	91.59	81.61	**86.73**	83.95	83.84	6.04
6	$\mathbf{FN_{T}}$	Vis. & Text	pretrained	fixed	88.00	89.04	81.42	71.23	94.19	84.78	8.83
7	$\mathbf{FN_{T}}$	Vis. & Text	pretrained	variable	83.21	92.29	84.03	83.69	83.56	85.36	**3.89**
8	$\mathbf{FN_{MLP}}$	Vis. & Text	pretrained	variable	76.38	96.1	84.03	76.87	94.85	85.65	9.48
9	$\mathbf{FN_{T}}$	Vis. & Text	fine-tuned	fixed	79.12	85.04	85.25	79.88	**96.84**	85.23	7.08
10	$\mathbf{FN_{MLP}}$	Vis. & Text	fine-tuned	fixed	90.75	96.64	78.76	73.97	97.13	87.45	10.56
11	$\mathbf{FN_{T}}$	Vis. & Text	fine-tuned	variable	85.52	92.09	**86.50**	83.27	96.52	88.78	5.41
12	$\mathbf{FN_{MLP}}$	Vis. & Text	fine-tuned	variable	**96.37**	**98.60**	85.25	78.52	94.02	**90.55**	8.42

Table 2. Average and standard deviation of the performances of various fusion networks (*FN*) on the Columbia [42] and RealVAD [17] datasets in terms of F1-score (%). The best results are in black. Exp. indicates the index of the experiments given in Table 1.

Dataset	Exp.	Avg.	Std.
Columbia [42]	12	93.8	**3.7**
Columbia [42]	11	**95.2**	4.9
RealVAD [17]	12	86.4	6.3
RealVAD [17]	11	**88.2**	**5.3**

As mentioned in Section 3.3, we use the two sets of prompts. In the case of the first prompt, the responses consistently result in either "yes" or "no". In the second prompt, we aim to test the expressive capabilities of the model, obtaining more complex and varied responses such as *The person is not speaking because their mouth is closed and there is no visible movement of the lips.* See Figure 4 for additional examples obtained through the use of LLaVA. Using *prompt 1* allows us to evaluate the standalone LLaVA-13B's [30,31] VAD performance (Exp. 1 of Table 1). Such an analysis leads to overall the worst performance, particularly for subjects Sick and Long, where the F1-scores are below 60%.

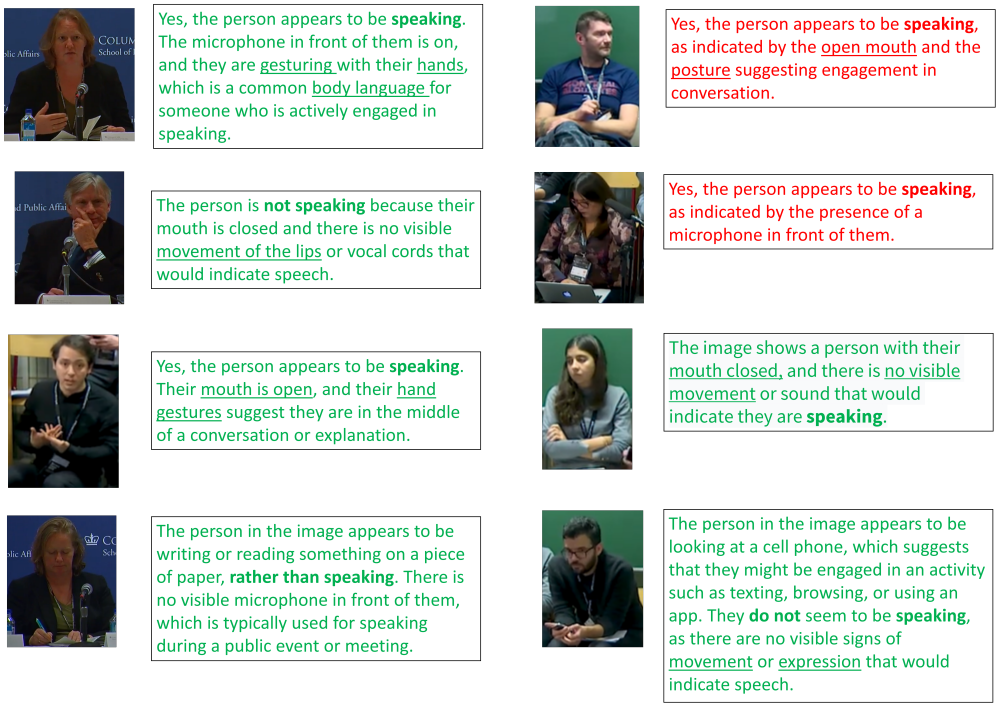

Figure 4. Example text responses obtained upon using LLaVA [30,31] with the second prompt: *Is the person speaking? Explain why in a few words.* In these examples, we did not limit the length of the generated captions, showcasing LLaVA's ability to produce descriptions of varying detail. However, in our implementation, we enforce a maximum sequence length to ensure compatibility with CLIP's text encoder. The green texts are the cases where the VAD class is correctly predicted, while thered texts signify instances of incorrect predictions.

Furthermore, the textual responses produced by LLaVA-13B were converted into textual embeddings utilizing the pretrained CLIP text encoder with ResNet-101. Analysis of the cosine similarity between these embeddings revealed that, on average, the captions generated by the first prompt exhibited a similarity exceeding 0.9. As a result, it can be inferred that classifiers, which would later utilize these embeddings as input, might struggle to discern significant differences between descriptions of a person speaking and not speaking. Therefore, we also tested to convert the results of the first prompt such that instead of "yes", we used *the person is engaged in a conversation* and instead of "no", we used *no one is talking* captions, which are referred as "fixed" in Table 1. As a consequence of this change, the cosine similarity was found to be 0.75, which was assumed to have substantial differences compared to others when used, e.g., for multimodal classification. To sum up, "yes/no" captions were only used to evaluate the classification performance of LLaVA-13B. The highly variable and expressive captions generated by the second prompt were used as textual inputs for the multimodal models referred to as *variable* in Table 1. Additionally, the *fixed* captions: *the person is engaged in a conversation* and *no one is talking* were used as alternative textual inputs for the multimodal models.

To test the performance of the pretrained CLIP visual encoders ResNet101 and ViT-B/16 on the VAD task, we attach an MLP whose design is as described in Section 3.3, since the visual encoders alone lack classification capability. From the observed results (Exp. 2 and 3 of Table 1), it became evident that the performance of ResNet101 not only competes with the more complex ViT-B/16 but is even superior, with an average F1-score greater by 8%. Furthermore, it can be observed how the domain-shift problem has influenced the VAD results, particularly for Bell and Long, for both visual encoders, leading to relatively lower performances with respect to the others.

We further tried to test also the fine-tuned ResNet101 (Exp. 4) for the VAD task considering its better performance than ViT-B/16. Such results indicate a clear improvement over using the pretrained CLIP for almost all individuals. In particular, there is an average improvement in the F1-score of ∼3% compared to the score obtained by classifying embeddings from pretrained ResNet-101. In particular, from these results, it is observed that the domain-shift problem has been effectively mitigated for many panelists compared to the pretrained visual encoder. It is also noticeable that the performance of Bollinger and Lieberman has slightly worsened.

The other experiments involve various combinations of VAD-CLVA, wherein the CLIP encoders remain fixed to ResNet101 while the FN changes across MLP or Transformer, and the visual encoder is utilized as pretrained or fine-tuned, with text embeddings fixed or variable (Exp. 5–12). Despite the FN_{MLP} with variable captions and fine-tuned visual encoder emerging as the best-performing model on average, there are instances where FN_T outperforms FN_{MLP} (e.g., in Exp. 5 and 6, for Lie., Bell, and on average). Overall, fine-tuning and employing variable captions enhance performance across all individuals as well as on average.

We additionally present the results of FN_{MLP} and FN_T, employing variable captions and fine-tuned visual encoders, across both the Columbia [42] and RealVAD [17] datasets in Table 2. Notably, for these datasets where larger training sets per fold are available, one can observe the superior performance of FN_T over FN_{MLP}.

4.2. Comparisons with the SOTA

We compare the effectiveness of VAD-CLVA on Tables 3–6 for the Columbia [42], Modified Columbia [16], and RealVAD [17] datasets, respectively. Overall, VAD-CLVA outperforms all visual VAD approaches as well as some of the audio-visual methods. Considering that the used benchmarks include a single speaker at a time, without overlapping speech, we claim that the audio signal can help significantly to detect whether there is a speaker or not at a certain time. While our focus is not to outperform all audio-visual VAD methods, recognizing the crucial role of audio signals in this task, we find VAD-CLVA's performance in surpassing certain audio-visual models noteworthy. Moreover, such results support the potential to integrate VAD-CLVA with audio signals to enhance overall performance. In detail, for the Columbia dataset [42] (Tables 3 and 4), VAD-CLVA surpasses the performances of all the visual VAD methods as well as audio-visual methods, i.e., SyncNet [23], LWTNet [59,60], and UNICON [19], on average. It is important to notice that the performance of VAD-CLVA on panelist Sick is the best of all methods. Furthermore, when we examine the best performance achieved among the visual VAD methods (indicated as max-V in Tables 3 and 4) and among the audio-visual methods (denoted as max-AV in Tables 3 and 4), it becomes apparent that in four out of five cases, VAD-CLVA outperforms at least one of the two. Consequently, it can be argued that the visual backbones (e.g., ResNet50) utilized in such methods can be substituted with CLIP visual encoder through fine-tuning, and furthermore, the text descriptions encoded with CLIP text encoder can contribute to enhancing performance. We also performed a paired t-test between the performances of VAD-CLVA and the performance of the method in Tables 3 and 4 when $\alpha = 0.05$. The t-test results show a statistical significance between our method and [18,42], LWTNet [59], and UNICON [19], where our average performance is higher. It must be noted that TalkNet [13], despite having a higher average performance than ours, has a performance that is not significant.

Table 3. Comparisons on the Columbia dataset [42]. We report F1-scores (%) for each person, the overall average (AVG), and the overall standard deviation (STD). V, AV, and VT stand for visual, audio-visual, and visual and generated text modalities, respectively. The last two rows show the best of all results excluding VAD-CLVA for audio-visual (denoted as max-AV) and visual (shown as max-V) VAD methods, respectively. Bold results indicate the best of all for each column. The colored results are the ones in which the VAD-CLVA performs better than at least one of the max-AV and max-V.

Method	Venue	Modality	Bell	Boll	Lieb	Long	Sick	AVG	STD
[42]	ECCV 2016	V	82.9	65.8	73.6	86.9	81.8	78.2	8.5
SyncNet [23]	ACCV 2017	AV	93.7	83.4	86.8	97.7	86.1	89.5	5.9
[18]	ICCV 2019	V	89.2	88.8	85.8	81.4	86.0	86.2	3.1
RGB-DI [18]	ICCV 2019	V	86.3	93.8	92.3	76.1	86.3	86.9	7.0
LWTNet [59]	ECCV 2020	AV	92.6	82.4	88.7	94.4	95.9	90.8	5.4
RealVAD [17]	IEEE TMM 2020	V	92.0	**98.9**	94.1	89.1	92.8	93.4	3.6
S-VVAD [16]	WACV 2021	V	92.4	97.2	92.3	95.5	92.5	94.0	**2.2**
[60]	CVPR 2021	AV	95.8	88.5	91.6	96.4	97.2	93.9	3.7
TalkNet [13]	ACM MM 2021	AV	97.1	90.0	**99.1**	96.6	98.1	**96.2**	3.6
UNICON [19]	ACM MM 2021	AV	93.6	81.3	93.8	93.5	92.1	90.9	5.4
ACLNet [15]	IEEE TMM 2022	AV	**97.4**	88.1	97.5	98.5	98.0	95.9	4.4
GSCMIA [61]	IEEE JSTSP 2023	AV	96.3	89.4	**98.7**	98.7	96.8	96.0	3.8
VAD-CLVA (**Ours**)		VT	96.9	86.7	96.0	97.8	**98.8**	95.2	4.9
		max-AV	98.1	89.4	99.1	99.3	98.1		
		max-V	92.4	98.9	94.1	95.5	92.8		

Table 4. Comparisons on the Columbia dataset [42]. We report F1-scores (%) for each person, the overall average (AVG), the overall standard deviation (STD), and the overall median (MED). V, AV, and VT stand for visual, audio-visual, and visual and generated text modalities, respectively. The last two rows show the best of all results excluding VAD-CLVA for audio-visual (denoted as max-AV) and visual (shown as max-V) VAD methods, respectively. Bold results indicate the best of all for each column. The colored results are the ones in which the VAD-CLVA performs better than at least one of the max-AV and max-V.

Method	Venue	Modality	Bell	Boll	Lieb	Long	Sick	AVG	STD	MED
[42]	ECCV 2016	V	82.9	65.8	73.6	86.9	81.8	78.2	8.5	81.8
SyncNet [23]	ACCV 2017	AV	93.7	83.4	86.8	97.7	86.1	89.5	5.9	86.8
[18]	ICCV 2019	V	89.2	88.8	85.8	81.4	86.0	86.2	3.1	86.0
RGB-DI [18]	ICCV 2019	V	86.3	93.8	92.3	76.1	86.3	86.9	7.0	86.3
LWTNet [59]	ECCV 2020	AV	92.6	82.4	88.7	94.4	95.9	90.8	5.4	92.6
RealVAD [17]	IEEE TMM 2020	V	92.0	**98.9**	94.1	89.1	92.8	93.4	3.6	92.8
S-VVAD [16]	WACV 2021	V	92.4	97.2	92.3	95.5	92.5	94.0	**2.2**	92.5
[60]	CVPR 2021	AV	95.8	88.5	91.6	96.4	97.2	93.9	3.7	95.8
TalkNet [13]	ACM MM 2021	AV	97.1	90.0	**99.1**	96.6	98.1	**96.2**	3.6	97.1
UNICON [19]	ACM MM 2021	AV	93.6	81.3	93.8	93.5	92.1	90.9	5.4	93.5
ACLNet [15]	IEEE TMM 2022	AV	**97.4**	88.1	97.5	98.5	98.0	95.9	4.4	**97.5**
GSCMIA [61]	IEEE JSTSP 2023	AV	96.3	89.4	**98.7**	98.7	96.8	96.0	3.8	96.8
VAD-CLVA (**Ours**)		VT	96.9	86.7	96.0	97.8	**98.8**	95.2	4.9	96.9
		max-AV	98.1	89.4	99.1	99.3	98.1			
		max-V	92.4	98.9	94.1	95.5	92.8			

Table 5. Comparisons on the Modified Columbia dataset [16]. Bold indicates the best. We report F1-scores (%) for each person, the overall average (AVG), and the overall standard deviation (STD).

	Bell	**Boll**	**Lieb**	**Long**	**Sick**	**AVG**	**STD**
S-VVAD [16]	86.1	87.7	96.7	84.0	75.1	85.9	**7.8**
VAD-CLVA (**Ours**)	**96.4**	78.5	**94.0**	**85.2**	**98.6**	**90.6**	8.4

Table 6. Comparisons on the RealVAD dataset [17]. Bold indicates the best. We report F1-scores (%) for each panelist, the overall average (AVG), and the overall standard deviation (STD). The same pretraining and training setups are color-coded for clarity. Zero-shot and fine-tuning experiments are separated by a double horizontal line.

Method	**Modality**	**Pretraining Data**	**Training Data**	**Testing Data**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**AVG**	**STD**
[17]	V	Columbia [42]	-	RealVAD [17]	53.6	51.1	41.1	50.2	37.3	50.3	56.8	53.6	69.8	51.5	9.3
UNICON [19]	V	AVA-ActiveSpeaker [38]	-	RealVAD [17]	86.7	78.1	70.5	73.1	68.9	84.9	93.0	80.4	87.0	80.3	7.8
UNICON [19]	AV	AVA-ActiveSpeaker [38]	-	RealVAD [17]	**94.3**	74.0	**89.9**	76.7	**80.6**	**93.6**	**98.8**	83.5	**93.5**	**87.2**	8.3
VAD-CLVA (**Ours**)	VT	Columbia [42]	-	RealVAD [17]	89.0	**81.6**	81.4	**83.4**	79.3	**93.6**	97.2	**85.8**	93.4	**87.2**	**6.4**
[17]	V	-	RealVAD [17]	RealVAD [17]	51.6	53.5	42.9	51.7	44.4	50.5	58.7	67.9	55.8	53.0	7.1
UNICON [19]	V	AVA-ActiveSpeaker [38]	RealVAD [17]	RealVAD [17]	86.9	76.5	81.6	**87.0**	79.6	**88.9**	97.0	84.5	88.9	85.6	5.7
UNICON [19]	AV	AVA-ActiveSpeaker [38]	RealVAD [17]	RealVAD [17]	**96.5**	**81.1**	**86.9**	84.4	**89.9**	85.6	94.9	88.1	**90.9**	**88.7**	**4.7**
VAD-CLVA (**Ours**)	VT	-	RealVAD [17]	RealVAD [17]	91.7	78.8	86.2	87.7	84.4	87.8	**98.5**	**88.9**	89.5	88.2	5.3

For the Modified Columbia dataset [16] (Table 5), VAD-CLVA outperforms S-VVAD [16] on average by 5% in terms of the F1-score. VAD-CLVA only falls short of the S-VVAD method [16] for Boll. It is worth noting that the Modified Columbia dataset is more challenging than the Columbia dataset, as it has fewer training data. Despite this, VAD-CLVA still manages to outperform the SOTA.

For the RealVAD dataset, there is no consistent evaluation approach across the literature. In this study, we adopt the evaluation approach utilized in the original paper [17], which comprises two methods: the first involves traditional training and testing on the same dataset, while the second entails training on the Columbia dataset [42] and subsequently testing the trained model on RealVAD [17]. This latter approach is referred to as zero-shot and/or cross-dataset evaluation. On the other hand, Zhang et al. [19] conducted their evaluation on this dataset by initially pretraining their model on a large dataset known as AVA-ActiveSpeaker [38]. They then either applied the trained model directly to RealVAD (hence, zero-shot) or fine-tuned the model on RealVAD before testing it on the same dataset. Given the scale of AVA-ActiveSpeaker [38] and the training of UNICON [19] on both audio and visual modalities, especially in the fine-tuned scenario, UNICON holds a significant advantage over VAD-CLVA and the approach by Beyan et al. [17], as it has access to a larger and more diverse dataset. Additionally, considering that the RealVAD dataset [17] typically features single speakers at a time (with no multiple speakers or overlapped speech), audio-visual methods like UNICON [19] may have an edge over video-only VAD methods in accurately detecting whether anybody is speaking or not using the audio signal. Nevertheless, as seen in Table 6, VAD-CLVA achieves an on-par performance with the audio-visual UNICON [19] in the zero-shot setting on average (87.2% F1-score), while it outperforms the visual UNICON [19] (87.2% versus 80.3% F1-score) and RealVAD [17] (87.2% versus 51.5% F1-score) in both average performance and across all panelists. On the other hand, in fine-tuning experiments, audio-visual UNICON surpasses VAD-CLVA on average by 0.5% for the F1-score and for several panelists (i.e., P1, P2, P3, and P5), which we attribute to its pretraining on a large multimodal dataset. Conversely, VAD-CLVA outperforms visual-only UNICON [19] with +2.6% and RealVAD [17] with +35.2%, on average. Such results underscore that VAD-CLVA provides a promising alternative to traditional audio-visual VAD methods. The model's performance in the zero-shot setting,

as well as its ability to outperform visual-only and audio-visual models in certain contexts, demonstrates the effectiveness of combining visual and textual cues for VAD.

5. Discussion

This section discusses the proposed method in relation to the research questions outlined in the introduction, as well as various aspects of both the model and the datasets used.

Research Questions. Regarding the two research questions given in Section 1: (a) we have empirically demonstrated that leveraging textual descriptions generated by a separate LMM like LLaVA-13B, combined with the visual features processed by CLIP, enhances VAD. In other words, the fusion of visual and textual embeddings helps improve the accuracy of VAD predictions, highlighting that combining visual and textual cues can provide better context and understanding for classifying speaking and not speaking activities. (b) Furthermore, the experimental analysis shows that VAD-CLVA can achieve performance comparable to or even surpass SOTA audio-visual VAD methods. Despite not using pretrained audio-visual data, the model trained on only visual data combined with text-based descriptions yields strong results. This once again suggests that the text description generated by the model, though simple, can effectively complement the visual features for VAD tasks, challenging the traditional reliance on audio-visual data for such tasks.

Challenges in Capturing Fine-Grained Speech-Related Visual Cues: While CLIP is a powerful model for visual and textual tasks, it was not explicitly trained to solve VAD tasks. As a result, CLIP might not be able to capture very fine-grained visual cues related to speech activities (such as subtle mouth movements or facial expressions) with the same level of effectiveness as models specifically designed for VAD. Specialized VAD models are typically trained on large datasets of speech-related visual data and optimized for detecting small, speech-relevant features that CLIP may not focus on. This discrepancy can lead to performance gaps in certain contexts where highly detailed speech-related cues are necessary for accurate classification. We argue that this might be the reason why some audio-visual model surpasses VAD-CLVA.

The Impact of Textual Description Quality and Their Processing: The performance of VAD-CLVA heavily relies on the relevance of the textual descriptions generated by LLaVA-13B. If the LLaVA-13B fails to capture critical speech-related cues, such as facial expressions or body movements that are indicative of speaking activity, the model might not have enough relevant information to generate accurate embeddings. For instance, if the LLaVA-13B only focuses on irrelevant features in the image, such as background elements (e.g., in our case, the microphone) or non-speech-related body parts, the model's predictions could be less accurate.

The way we process the sentences generated by LLaVA-13B offers several implementation advantages, as described in Section 3.3. However, in our current approach, CLIP processes each sentence separately, which leads to a loss of context between sentences. If a later sentence refers to a subject mentioned earlier, the meaning may be lost. Instead of averaging the sentences, an alternative would be to concatenate the most important sentences, ensuring the total token count does not exceed CLIP's limit. However, there is no automatic mechanism to determine which sentences are the most important. Another option is to weight the sentences differently, but this approach also requires knowing the relevance of each sentence to the task. The best, though more costly, alternative (which will be tested in future work) is to use a Large Language Model (LLM) to summarize or refine LLaVA's raw output before passing it to CLIP, ensuring the text remains informative yet concise enough to fit CLIP's token limit. For instance, using GPT-4o-mini to rephrase or extract key information before feeding it to CLIP seems like a reasonable approach, even

though it would require additional prompt design. Moreover, we have not aimed to test an ensemble of prompts that can potentially improve the results.

Captioner Model Selection. We exclusively assessed LLaVA-13B [30,31] both as a standalone LMM and integrated within the VAD-CLVA framework. While this choice can be justified given LLaVA's demonstrated effectiveness in earlier works, it is worth noting that there are several other models that warrant exploration to provide a more comprehensive evaluation. Furthermore, querying LLaVA to generate text descriptions for each input image adds an additional computational layer to the process, which could introduce latency, especially in real-time applications, even though neither our method nor the SOTA explicitly targets real-time processing. Nevertheless, the inference for our method is completed in milliseconds for a segment of 10 frames, excluding the LLaVA querying, which is performed once and then stored for all datasets. Using an A100 GPU, we were able to obtain LLaVA results within a few hundred milliseconds.

Despite our approach to handling temporal data in the visual domain, we extract textual embeddings from single images. This decision was driven by the need to optimize model efficiency, as obtaining captions for each video frame can be resource-intensive. However, we presume that the use of 10 video frames would not significantly increase the variance in captions. Nonetheless, exploring the processing of spatio-temporal data with LMMs could potentially enhance performance, although this remains an area of investigation beyond VAD specifically. The obtained results suggest the potential benefits of integrating LMM-based pipelines into audio-visual VAD systems, which could enhance performance. Future work will further explore this topic in depth.

Generalization across Different Speakers. The standard deviation results of VAD-CLVA are not satisfactory compared to other methods. This shows that VAD-CLVA is not resilient to the domain-shift problem occurring across different speakers, where with some speakers the head and body activity can be more evident, while some others use less head and body expression when they speak. Indeed, our method neither integrates a domain adaptation procedure such as applied in [18] nor is it pretrained on a large VAD dataset, as applied in [19], which potentially can help to generalize better across different individuals.

Used Datasets. We further discuss the potential limitations of the benchmark datasets used in this study and how they may affect generalizability when applied to real-life scenarios without being fine-tuned. The Columbia [42] and RealVAD [17] datasets are derived from panel discussions. The Columbia dataset [42] consists mainly of controlled environments with similar background conditions across participants, limiting dataset diversity. In contrast, RealVAD [17] provides a more diverse set of scenarios, including multiple subjects in the camera's field of view, and varied ethnic and demographic backgrounds. This diversity potentially can allow models to better handle different speaking and body language styles. RealVAD also includes non-static backgrounds, camera motion, and varying video resolutions, reflecting more dynamic, real-world environments. Additionally, the dataset accounts for natural changes in lighting, shadows, and occlusions, all of which introduce real-world challenges for VAD. However, it is important to note that in these datasets, speakers typically exhibit controlled behaviors and strong verbal communication skills, which may not capture the spontaneous and unpredictable nature of everyday interactions. As a result, VAD models trained on these datasets might perform well in controlled settings like other panel discussions but may face difficulties in generalizing to real-world applications, where speakers show more natural and erratic behaviors.

6. Conclusions

We have introduced a voice activity detection (VAD) method employing Vision-Language Models (VLMs) and Generative Large Multimodal Models (LMMs). It takes as

input a video segment composed of a single person's upper body to predict their VAD label as "speaking" or "not speaking". While this video segment is input to the visual encoder of a certain VLM, called CLIP [32], the central frame of the video segment is used to automatically generate the text descriptions regarding the speaking activity of the person by LMM model LLaVA-13B [30,31] through prompt engineering, which is further given as the input to the text encoder of CLIP [32]. The visual and textual encodings in the latest stage are fused with a simple concatenation and learned with a deep model to perform VAD predictions.

This study is a pioneer in showcasing the contribution of visual-textual models, which improves VAD results, particularly compared to VAD methods relying solely on visual data. It also presents the VAD performance of a standalone LMM model using a single image and a rather simple prompt. Even though the standalone LMM is not sufficient by itself, the text descriptions generated by it contribute to achieving improved VAD when combined with visual features. This highlights the effectiveness of the video–language coupling, as demonstrated across various downstream tasks, suggesting its efficacy over relying solely on visual cues for VAD.

Furthermore, we empirically demonstrate that the proposed method can perform comparably or even better than several audio-visual VAD approaches. This accomplishment is significant, particularly considering that audio is a primary modality for VAD tasks, as indicated by prior studies. Having the audio in the loop is especially effective in scenarios where there are no simultaneous speakers, as observed in the benchmarks utilized in this study. Additionally, the noteworthy performance of our VAD-CLVA is evident, especially considering that some SOTA models have been pretrained on large audio-visual datasets. Such conclusions about the performance of VAD-CLVA are not only applicable to the traditional setting, where the training and testing data share the same distribution (i.e., originating from the same dataset), but also to zero-shot settings in which the model is trained on one dataset and tested on another dataset without fine-tuning.

Author Contributions: Conceptualization, C.B.; Methodology, A.A. and C.B.; Software, A.A.; Validation, A.A.; Formal analysis, A.A.; Writing—original draft, A.A. and C.B.; Writing—review & editing, C.B.; Supervision, C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are openly available in zenodo at https://data.niaid.nih.gov/resources?id=zenodo_3928150 (accessed on 11 January 2025), reference number 3928151.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Skantze, G. Turn-taking in conversational systems and human–robot interaction: A review. *Comput. Speech Lang.* **2021**, *67*, 101178. [CrossRef]
2. Xu, E.Z.; Song, Z.; Tsutsui, S.; Feng, C.; Ye, M.; Shou, M.Z. Ava-avd: Audio-visual speaker diarization in the wild. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3838–3847.
3. Wang, Q.; Downey, C.; Wan, L.; Mansfield, P.A.; Moreno, I.L. Speaker diarization with LSTM. In Proceedings of the 2018 IEEE ICASSP, Calgary, AB, Canada, 15–20 April 2018; pp. 5239–5243.
4. Chung, J.S.; Huh, J.; Nagrani, A.; Afouras, T.; Zisserman, A. Spot the conversation: Speaker diarisation in the wild. *arXiv* **2020**, arXiv:2007.01216.

5. Hung, H.; Ba, S.O. Speech/Non-Speech Detection in Meetings from Automatically Extracted Low Resolution Visual Features. 2009. Available online: https://infoscience.epfl.ch/entities/publication/0659b34f-3f4d-44e6-86a8-898c01b6b857 (accessed on 11 January 2025).
6. Beyan, C.; Katsageorgiou, V.M.; Murino, V. A sequential data analysis approach to detect emergent leaders in small groups. *IEEE Trans. Multimed.* **2019**, *21*, 2107–2116. [CrossRef]
7. Górriz, J.M.; Ramírez, J.; Lang, E.W.; Puntonet, C.G.; Turias, I. Improved likelihood ratio test based voice activity detector applied to speech recognition. *Speech Commun.* **2010**, *52*, 664–677. [CrossRef]
8. Michelsanti, D.; Tan, Z.H.; Zhang, S.X.; Xu, Y.; Yu, M.; Yu, D.; Jensen, J. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1368–1396. [CrossRef]
9. Moine, C.L.; Obin, N.; Roebel, A. Speaker attentive speech emotion recognition. *arXiv* **2021**, arXiv:2104.07288.
10. Moattar, M.H.; Homayounpour, M.M. A simple but efficient real-time voice activity detection algorithm. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, Scotland, UK, 24–28 August 2009; IEEE: New York, NY, USA, 2009; pp. 2549–2553.
11. Minotto, V.P.; Jung, C.R.; Lee, B. Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs. *IEEE Trans. Multimed.* **2014**, *16*, 1032–1044. [CrossRef]
12. Patrona, F.; Iosifidis, A.; Tefas, A.; Nikolaidis, N.; Pitas, I. Visual voice activity detection in the wild. *IEEE Trans. Multimed.* **2016**, *18*, 967–977. [CrossRef]
13. Tao, R.; Pan, Z.; Das, R.K.; Qian, X.; Shou, M.Z.; Li, H. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 3927–3935.
14. Köpüklü, O.; Taseska, M.; Rigoll, G. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1193–1203.
15. Xiong, J.; Zhou, Y.; Zhang, P.; Xie, L.; Huang, W.; Zha, Y. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Trans. Multimed.* **2022**, *25*, 5800–5812.
16. Shahid, M.; Beyan, C.; Murino, V. S-vvad: Visual voice activity detection by motion segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 2332–2341.
17. Beyan, C.; Shahid, M.; Murino, V. RealVAD: A real-world dataset and a method for voice activity detection by body motion analysis. *IEEE Trans. Multimed.* **2020**, *23*, 2071–2085. [CrossRef]
18. Shahid, M.; Beyan, C.; Murino, V. Voice activity detection by upper body motion analysis and unsupervised domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–29 October 2019.
19. Zhang, Y.; Liang, S.; Yang, S.; Liu, X.; Wu, Z.; Shan, S.; Chen, X. Unicon: Unified context network for robust active speaker detection. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 3964–3972.
20. Chung, J.S.; Zisserman, A. Learning to lip read words by watching videos. *Comput. Vis. Image Underst.* **2018**, *173*, 76–85. [CrossRef]
21. Liu, Q.; Wang, W.; Jackson, P. A visual voice activity detection method with adaboosting. In Proceedings of the Sensor Signal Processing for Defence (SSPD 2011), London, UK, 27–29 September 2011.
22. Sodoyer, D.; Rivet, B.; Girin, L.; Savariaux, C.; Schwartz, J.L.; Jutten, C. A study of lip movements during spontaneous dialog and its application to voice activity detection. *J. Acoust. Soc. Am.* **2009**, *125*, 1184–1196. [CrossRef] [PubMed]
23. Chung, J.S.; Zisserman, A. Out of time: Automated lip sync in the wild. In Proceedings of the ACCV 2016 Workshops, Taipei, Taiwan, 20–24 November 2016; Springer: Berlin/Heidelberg, Germany, 2017; pp. 251–263.
24. Geeroms, W.; Allebosch, G.; Kindt, S.; Kadri, L.; Veelaert, P.; Madhu, N. Audio-Visual Active Speaker Identification: A comparison of dense image-based features and sparse facial landmark-based features. In Proceedings of the 2022 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 12-14 October 2022; pp. 1–6. [CrossRef]
25. Huang, C.; Koishida, K. Improved active speaker detection based on optical flow. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 17–18 June 2020; pp. 950–951.
26. Cristani, M.; Pesarin, A.; Vinciarelli, A.; Crocco, M.; Murino, V. Look at who's talking: Voice activity detection by automated gesture analysis. In Proceedings of the Constructing Ambient Intelligence: AmI 2011 Workshops, Amsterdam, The Netherlands, 16–18 November 2011; Springer: Berlin/Heidelberg, Germany, 2012; pp. 72–80.
27. Gebre, B.G.; Wittenburg, P.; Heskes, T. The gesturer is the speaker. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; IEEE: New York, NY, USA 2013; pp. 3751–3755.
28. Shahid, M.; Beyan, C.; Murino, V. Comparisons of Visual Activity Primitives for Voice Activity Detection. In Proceedings of the Image Analysis and Processing—ICIAP 2019: 20th International Conference, Trento, Italy, 9–13 September 2019; Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2019; pp. 48–59. [CrossRef]

29. Xenos, A.; Foteinopoulou, N.M.; Ntinou, I.; Patras, I.e.a. VLLMs Provide Better Context for Emotion Understanding Through Common Sense Reasoning. *arXiv* **2024**, arXiv:2404.07078.
30. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *arXiv* **2023**, arXiv:2304.08485.
31. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv* **2023**, arXiv:2310.03744.
32. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR: Birmingham, UK, 2021; pp. 8748–8763.
33. Auty, D.; Mikolajczyk, K. Learning to Prompt CLIP for Monocular Depth Estimation: Exploring the Limits of Human Language. In Proceedings of the IEEE/CVF ICCV, Paris, France, 1–6 October 2023; pp. 2039–2047.
34. Bondielli, A.; Passaro, L.C. Leveraging CLIP for Image Emotion Recognition. In Proceedings of the CEUR WORKSHOP PROCEEDINGS, Virtual, 4–5 October 2021; Volume 3015.
35. Chen, D.; Gou, G. Unleash the Capabilities of the Vision-Language Pre-training Model in Gaze Object Prediction. In Proceedings of the Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 453–466.
36. Tao, F.; Busso, C. Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection. In Proceedings of the INTERSPEECH, Stockholm, Sweden 20–24 August 2017; pp. 1938–1942.
37. Tao, F.; Busso, C. End-to-end audiovisual speech activity detection with bimodal recurrent neural models. *Speech Commun.* **2019**, *113*, 25–35. [CrossRef]
38. Roth, J.; Chaudhuri, S.; Klejch, O.; Marvin, R.; Gallagher, A.; Kaver, L.; Ramaswamy, S.; Stopczynski, A.; Schmid, C.; Xi, Z.; et al. Ava active speaker: An audio-visual dataset for active speaker detection. In Proceedings of the IEEE ICASSP, Barcelona, Spain, 4–8 May 2020; pp. 4492–4496.
39. Sharma, R.; Somandepalli, K.; Narayanan, S. Crossmodal learning for audio-visual speech event localization. *arXiv* **2020**, arXiv:2003.04358.
40. Shvets, M.; Liu, W.; Berg, A.C. Leveraging long-range temporal relationships between proposals for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9756–9764.
41. Gebru, I.D.; Ba, S.; Li, X.; Horaud, R. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1086–1099. [CrossRef]
42. Chakravarty, P.; Tuytelaars, T. Cross-modal supervision for learning active speaker detection in video. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part V 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 285–301.
43. Joosten, B.; Postma, E.; Krahmer, E. Voice activity detection based on facial movement. *J. Multimodal User Interfaces* **2015**, *9*, 183–193. [CrossRef]
44. Haider, F.; Campbell, N.; Luz, S. Active speaker detection in human machine multiparty dialogue using visual prosody information. In Proceedings of the 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, DC, USA, 7–9 December 2016; IEEE: New York, NY, USA, 2016; pp. 1207–1211.
45. Stefanov, K.; Beskow, J.; Salvi, G. Vision-based active speaker detection in multiparty interaction. In Proceedings of the Grounding Language Understanding (GLU2017), Stockholm, Sweden, 25 August 2017.
46. Stefanov, K.; Beskow, J.; Salvi, G. Self-supervised vision-based detection of the active speaker as support for socially aware language acquisition. *IEEE Trans. Cogn. Dev. Syst.* **2019**, *12*, 250–259. [CrossRef]
47. Wortsman, M.; Ilharco, G.; Kim, J.W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R.G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 7959–7971.
48. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.W.; Yao, Z.; Keutzer, K. How much can clip benefit vision-and-language tasks? *arXiv* **2021**, arXiv:2107.06383.
49. Yuan, M.; Lv, N.; Xie, Y.; Lu, F.; Zhan, K. CLIP-FG:Selecting Discriminative Image Patches by Contrastive Language-Image Pre-Training for Fine-Grained Image Classification. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 560–564. [CrossRef]
50. Srivastava, M.M. RetailKLIP: Finetuning OpenCLIP backbone using metric learning on a single GPU for Zero-shot retail product image classification. *arXiv* **2023**, arXiv:2312.10282.
51. Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. Cris: Clip-driven referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11686–11695.
52. Liu, J.; Zhang, Y.; Chen, J.N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; Zhou, Z. Clip-driven universal model for organ segmentation and tumor detection. In Proceedings of the IEEE/CVF ICCV, Paris, France, 1–6 October 2023; pp. 21152–21164.

53. Liang, Z.; Li, C.; Zhou, S.; Feng, R.; Loy, C.C. Iterative prompt learning for unsupervised backlit image enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8094–8103.
54. Sanghi, A.; Chu, H.; Lambourne, J.G.; Wang, Y.; Cheng, C.Y.; Fumero, M.; Malekshan, K.R. Clip-forge: Towards zero-shot text-to-shape generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18603–18613.
55. Kim, G.; Kwon, T.; Ye, J.C. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2426–2435.
56. Wang, H.; Li, Y.; Yao, H.; Li, X. Clipn for zero-shot ood detection: Teaching clip to say no. In Proceedings of the IEEE/CVF International Conference on Computer Vision, New Orleans, LA, USA, 18–24 June 2023; pp. 1802–1812.
57. Wang, M.; Yang, N. EmoAsst: Emotion recognition assistant via text-guided transfer learning on pre-trained visual and acoustic models. *Front. Comput. Sci.* **2024**, *6*, 1304687. [CrossRef]
58. Garg, B.; Kim, K.; Ranjan, S. From Video to Images: Contrastive Pretraining for Emotion Recognition from Single Image. In Proceedings of the AAAI Conference on Artificial Intelligence, Pomona, CA, USA, 24–28 October 2022; Volume 36, pp. 12951–12952.
59. Afouras, T.; Owens, A.; Chung, J.S.; Zisserman, A. Self-supervised learning of audio-visual objects from video. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 208–224.
60. Truong, T.D.; Duong, C.N.; Pham, H.A.; Raj, B.; Le, N.; Luu, K. The right to talk: An audio-visual transformer approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1105–1114.
61. Sharma, R.; Narayanan, S. Audio-visual activity guided cross-modal identity association for active speaker detection. *IEEE Open J. Signal Process.* **2023**, *4*, 225–232. [CrossRef]