Emotion Recognition from Videos Using Multimodal Large Language Models

future internet

MDPI

Article

Emotion Recognition from Videos Using Multimodal Large Language Models

Lorenzo Vaiani, Luca Cagliero and Paolo Garza *

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy; lorenzo.vaiani@polito.it (L.V.); luca.cagliero@polito.it (L.C.)
* Correspondence: paolo.garza@polito.it; Tel.: +39-011-090-7022

Abstract: The diffusion of Multimodal Large Language Models (MLLMs) has opened new research directions in the context of video content understanding and classification. Emotion recognition from videos aims to automatically detect human emotions such as anxiety and fear. It requires deeply elaborating multiple data modalities, including acoustic and visual streams. State-of-the-art approaches leverage transformer-based architectures to combine multimodal sources. However, the impressive performance of MLLMs in content retrieval and generation offers new opportunities to extend the capabilities of existing emotion recognizers. This paper explores the performance of MLLMs in the emotion recognition task in a zero-shot learning setting. Furthermore, it presents a state-of-the-art architecture extension based on MLLM content reformulation. The performance achieved on the Hume-Reaction benchmark shows that MLLMs are still unable to outperform the state-of-the-art average performance but, notably, are more effective than traditional transformers in recognizing emotions with an intensity that deviates from the average of the samples.

Keywords: video–language large language models; emotion recognition; emotional reaction intensity estimation; multimodal learning

Citation: Vaiani, L.; Cagliero, L.; Garza, P. Emotion Recognition from Videos Using Multimodal Large Language Models. *Future Internet* **2024**, *16*, 247. https://doi.org/10.3390/fi16070247

Academic Editors: Filipe Portela and Athanasios D. Panagopoulos

Received: 1 June 2024
Revised: 9 July 2024
Accepted: 11 July 2024
Published: 13 July 2024

1. Introduction

Thanks to its ever-increasing diffusion, video content is gradually replacing traditional textual and image web sources. Video-sharing platforms like YouTube, TikTok, and Twitch have attracted millions of social users [1]. Every day, they process millions of videos, thus requiring automated solutions for efficient and effective content classification, annotation, and retrieval. Recognizing human emotions in videos is particularly relevant to content-sharing platforms as it enables smart applications and services such as healthcare monitoring [2], AI chatbots [3], and engagement and gaming [4].

This work studies the problem of emotional reaction intensity (ERI) estimation from video sources in which the simple processing of facial expressions is not sufficient to detect the correct emotional reaction, e.g., adoration, amusement, anxiety, disgust, empathic pain, fear, and surprise. Rather than proposing ad hoc image processing techniques for emotion recognition, our purpose is to explore the capabilities of state-of-the-art multimodal learning systems that effectively combine visual and acoustic sources. To this end, this study analyzes a video benchmark collection released by the organizers of the MuSe-Reaction challenge [5]. Videos in the collection show people's reactions captured by a front-facing camera. Human reactions can be detected and evaluated by processing both a subject's face and voice.

State-of-the-art approaches rely on transformer architectures [6] that jointly process the visual and audio streams. They adopt either modality-specific [7,8] or cross-modal [9] fusion techniques to combine the separate inputs and then perform vision–language classification on top of the encoded inputs.

In parallel, the progress of Large Language Models has allowed the evolution of traditional text-only Large Language Models (LLMs) towards the combined processing of

multiple modalities. Recent LLMs such as GPT-4 [10], LLaVA [11], Video-LLaVa [12], and LaViLa [13] support visual content as part of the LLM prompts or responses beyond plain text. However, their performance on emotion recognition tasks is still largely unexplored.

This paper studies the application of Multimodal Large Language Models (MLLMs) to estimate the emotional reactions in videos. It explores three alternative strategies. The first one directly applies recently proposed video–language LLMs (Video-LLaVa [12]) to estimate emotional reactions from video clips. We cast the problem of ERI estimation to a multi-regression task in which the MLLM predicts the corresponding level of intensity of each emotion. The second strategy applies probing [14] on top of MLLM embeddings. Finally, the third strategy integrates MLLM features into ViPER [9], a state-of-the-art multimodal architecture for video emotion recognition.

The results achieved on the Hume-Reaction benchmark [5] give interesting insights into the performance comparison between MLLMs and traditional transformers. Despite their promising performance, MLLMs in a zero-shot setting are still incapable of achieving a higher correlation score on the analyzed video clips. The combination of MLLMs with transformer-based architectures turns out to be marginally beneficial in the average performance metrics. However, a deeper analysis of the per-class results highlights the higher capability of MLLM-based approaches to correctly estimate emotional reaction intensities that deviate from the average of the entire collection. Importantly, their higher effectiveness in predicting atypical emotion intensities could be particularly helpful in situations in which fine-tuning ad hoc models is unfeasible due to a lack of training data or limited computational resources.

The remainder of this paper is organized as follows. Section 2 overviews the existing video–language LLMs and transformer-based architectures for ERI estimation from videos. Section 3 introduces the task and the benchmark dataset. Section 4 describes the LLM-based methods. Section 5 presents the experimental results achieved on benchmark data. Finally, Sections 6 and 7 summarize the main findings, draw conclusions, highlight the main limitations of the proposed method, and discuss the future research extensions of the present work.

2. Related Works

2.1. Emotion Recognition

Emotion recognition encompasses a variety of related tasks that differ in the modality involved in the input data, e.g., facial expression recognition (FER), speech emotion recognition (SER) and textual emotion recognition (TER) [15].

Recently, the interest of the research community has mainly focused on recognizing emotions from multimodal sources such as videos [16–18]. Here, the key challenge is properly extracting and combining features from the input video since the discriminating information conveying the emotion is often cross-modal. However, a unified solution that has proved to outperform all existing approaches in unimodality, bimodality, and multimodality scenarios is still missing [19]. This work focuses on a particular emotion recognition subtask, i.e., emotional reaction intensity estimation [20].

2.2. Emotional Reaction Intensity Estimation

The MuSe 2022 (Multimodal Sentiment Analysis Challenge) research challenge [5] first employs Hume-Reaction, a benchmark for ERI estimation from videos. The task organizers invite researchers to explore the complementary role of multimodal information in the emotion recognition task. The same video corpus has been used in further competitions, such as the ABAW 2023 (Affective Behavior Analysis in the Wild) research challenge [21]. Task participants mainly adopt transformer-based architectures [6] and focus on facial details to address the issue. The main limitation of transformer-based models is the need for large-scale training data that are, unfortunately, not always available in several domains and scenarios. Some research efforts have been devoted to exploiting modality fusion layers [7,8], each one relying on visual [22] and audio [23] encoders. Adopting separate

per-modality encoders limits the potential of attention-based classifiers as they neglect cross-modality interactions.

The authors of [20] propose a dual-branch network that processes both visual and acoustic information, employing spatial (for vision only) and temporal (for both modalities) transformer-based encoders. They also propose a modality dropout fusion layer to combine modalities, proving its effectiveness with respect to simple concatenation. The approach described in [24] is based on the PosterV2-ViT model, a transformer-based architecture designed to extract features from the Hume-Reaction dataset. The authors also combine these visual features with the precomputed DeepSpectrum audio features to further improve the performance. In [9], the authors propose ViPER, a multimodal architecture designed to combine features from an arbitrary number of sources. All the information is extracted at the frame level and concatenated across modalities before feeding a Perceiver model, a transformer-based modality-agnostic architecture [25]. Beyond visual and acoustic features, it includes Facial Action Units and textual features to enhance the results. Particularly, textual features are obtained using the CLIP [26] model to align video frames with pre-defined templates.

Unlike ViPER [9], this work focuses on exploring the capabilities of visual and video LLMs in video emotion recognition. It proposes both a probing network tailoring LLMs to the emotion recognition task and an extension of the state-of-the-art ViPER [9] architecture integrating LLMs to generate textual video- and frame-level textual descriptions.

2.3. Multimodal LLMs

The rapid expansion of online multimodal sources, such as multimedia documents, videos and audio signals, has prompted the evolution of traditional text-only Large Language Models (LLMs) towards the combined processing of multiple modalities. Table 1 summarizes the main characteristics of state-of-the-art Multimodal LLMs. To the best of our knowledge, none of the existing models have already been used to address the task of emotion recognition from videos.

Table 1. Classification of state-of-the-art Multimodal LLMs.

Type	Name	Year	Size	Open	Architectural Details	Downstream Task	Pre-Train
Vision–Language	LLaVA [11]	2023	7B, 13B, 34B	Yes	CLIP [26] extracts visual features, which are projected in the word embedding space before feeding the LLM (decoder only) with both visual and textual tokens	Visual Question Answering (in terms of conversation detail, description, complex reasoning)	QA pairs created using ChatGPT and GPT-4 on top of COCO images [27]
	Open-Flamingo [28]	2023	3B, 4B, 9B	Yes	CLIP [26] extracts visual features; text with interleaved images is passed to the LLM to generate the response	Visual Question Answering	LAION-2B [29], Multimodal C4, ChatGPT-generated data
	GPT-4 [10]	2023	>70B	No	N/A	N/A	N/A
	Mini-GPT-4 [30]	2023	7B, 13B	Yes	ViT [22] backbone plus a Q-Former [31] to extract visual features, used to feed Vicuna model together with textual tokens. Two-stage training: (i) general training to acquire visual knowledge, (ii) high-quality training using a designed conversational template	Visual Question Answering, image captioning, meme interpretation, receipt generation, advertisement creation, and poem composition	SBU [32], LAION [33]
	BLIP-2 [34]	2023	3B, 7B	Yes	Q-Former [31] + LLM trained with image–text contrastive learning, Image-grounded Text Generation and image–text matching	Visual Question Answering, image captioning, image–text retrieval	COCO [27], Visual Genome [35], CC3M [36], CC12M [37], SBU [32], LAION400M [33]
	Fuyu [38]	2023	8B	Yes	Image patches are instead linearly projected into the first layer of the transformer; there is no image encoder	Visual Question Answering, image captioning	N/A

Table 1. *Cont.*

Type	Name	Year	Size	Open	Architectural Details	Downstream Task	Pre-Train
Video–Language	Video-LLaVA [12]	2023	7B	Yes	United visual representation (video + images) before feeding the LLM	Image Question Answering, video understanding	LAION-CC-SBU, Valley [39], WebVid [40]
	Merlin [41]	2023	7B	No	Specifically trained to causally model the trajectories interleaved with multi-frame images	Future reasoning, identity association ability, Visual Question Answering	A lot of datasets (captioning + detection + tracking)
	VTimeLLM [42]	2023	7B, 13B	Yes	CLIP [26] as visual encoder to feed the LLM, which is trained to be aware of temporal boundaries in videos	Temporal Video Grounding, Video Captioning	ActivityNet Captions [43], CharadesSTA [44]
	Video-ChatGPT [45]	2023	7B	Yes	CLIP [26] used to extract frame representations, combined to obtain temporal and spatial video representation, used to feed the LLM	Video-based Generative Performance Benchmarking, Question–Answer Evaluation	Automatically generated data enriched by human annotators
Audio–Visual	Audio–Visual LLM [46]	2023	7B, 13B	Yes	CLIP [26] and CLAP to extract visual and audio features, respectively, projected into the LLM hidden space	Video-QA, Audio–Visual-QA, audio captioning	custom dataset, part of the LLaVA dataset, part of Valley dataset
	Macaw-LLM [47]	2023	7B	Yes	Unimodal feature extraction, alignment module to align each modality feature before feeding the LLM	Image/video understanding, visual-and-audio question answering	COCO [27], Charades [48], AVSD [49]

This work classifies the proposed solutions according to the type of supported inputs as follows:

- Vision–language LLMs, which handle combinations of images and text;
- Video–language LLMs, which are capable of automatically recognizing and interpreting video content as a stream of visual and textual sources;
- Audio–visual LLMs, which combine acoustic and visual information together.

To encode multimodal content, the most established approaches envisage the use of pre-trained vision models to extract textual information from videos and then format them as prompts for LLMs to generate responses, or the combination of LLMs with pre-training or fine-tuning strategies of vision/acoustic/time series models to create a unified representation. Most recent studies mainly focus on the latter approach.

State-of-the-art vision–language LLMs (e.g., [11,28]) leverage constrastive pre-training on image–text pairs to capture cross-modality relations. They are trained to align associated images and text together in a unified embedding space and are then fined-tuned for the Visual Question Answering task. Given an image, the LLM can be instructed in natural language to predict the most relevant text snippets conditioned to both downstream task and visual content.

Video–language LLMs adopt the following pre-training approaches to interpret video content [50]:

- Frame-based methods, which handle each video frame independently using various visual encoders and image resolutions;
- Temporal encoders, which treat videos as cohesive entities, emphasizing the temporal elements of the content [51].

Commonly, video–language models are not fine-tuned for a specific given task but are rather used in a zero-shot setting. Unlike Merlin [41] and VTimeLLM [42], Video-LLaVA [12] handles both videos and images as input, generating a unified video–text–image representation.

Similar to vision–language models, *audio–visual LLMs* align and combine different modalities, including the audio stream, to understand video and answer spoken questions.

3. Task and Dataset Description

The task of this study is the recognition of emotional reactions in videos. For this research, the Hume-Reaction dataset, a large-scale, multimodal dataset designed explicitly for the Emotional Reactions Sub-Challenge (MuSe-Reaction) [5], was employed. The dataset is notable for its extensive collection of naturalistic emotional reactions. The dataset

annotations correspond to the intensity scores (ranging from 0 to 1) of several emotions. Thus, the problem can be formulated as a multi-regression problem, where the goal is to predict the intensity scores of each involved emotion.

These scores are self-annotated by video subjects, and they relate to seven different emotions: *Adoration*, *Amusement*, *Anxiety*, *Disgust*, *Empathic Pain*, *Fear*, and *Surprise*. This set of emotions may differ from previously predefined sets of basic emotions, e.g., the Paul Ekman categorization [52], because they were specifically designed by the dataset authors to better represent the reactions elicited by the video subjects in the dataset. However, the approaches presented in this work can be straightforwardly extended to other emotion types.

The Hume-Reaction dataset is notable for its extensive collection of naturalistic emotional reactions. It comprises recordings from 2222 subjects, amounting to over 70 h of data. It also includes audio and video recordings, capturing the subjects' vocal and facial reactions while reacting to an unknown short video clip. All data samples were gathered in an uncontrolled environment, with subjects recording their responses in diverse at-home settings. These settings introduce a variety of noise conditions, making the dataset robust for real-world applications. After viewing each trigger video clip, each recorded subject reported the emotions they experienced and rated the intensity of each emotion. These self-reported data serve as the ground truth for training and evaluating emotion recognition models. For each selected emotion, subjects rated the intensity on a scale from 0 to 1.

The dataset is divided into three different splits:

- A training set made of 15,806 samples from 1334 different human subjects, for a total of 51 h of video recordings. This split was employed in our experimentation to train our models.
- A development set made of 4657 samples from 444 different human subjects, for a total of almost 15 h of video recordings. This split was employed in our experimentation to evaluate the proposed approaches.
- A private test set made of 4604 samples from 444 different human subjects, for a total of almost 15 h of video recordings. Labels of this split are not publicly available. Thus, this split was not used in this research.

A detailed analysis of the dataset's actual scores reveals that the dataset covers the entire spectrum of intensity for each emotion, from 0 to 1. However, there are substantial differences in the average value of the scores for each emotion. Table 2 reports each emotion's average ground truth score. This variability in average scores reflects the diverse emotional expressions and intensities captured in the dataset, adding an additional layer of complexity to the prediction task.

Additionally, Figure 1 shows the probability density functions of all emotion scores annotated in the dataset generated by a Gaussian Kernel Density Estimation. They all exhibit a bimodal distribution, featuring significant peaks around zero and one, indicating distinct clusters of low and high intensities within the dataset. In detail, all emotions show a higher peak of around 0, except for amusement and surprises, which have a higher density of around 1.

Table 2. Ground truth average scores (± standard deviation) for each emotion.

Emotion	Average Scores ± Std
Adoration	0.3218 ± 0.3612
Amusement	0.6204 ± 0.3669
Anxiety	0.3431 ± 0.3245
Disgust	0.2432 ± 0.2866
Empathic Pain	0.2152 ± 0.3032
Fear	0.2824 ± 0.3363
Surprise	0.5307 ± 0.3219

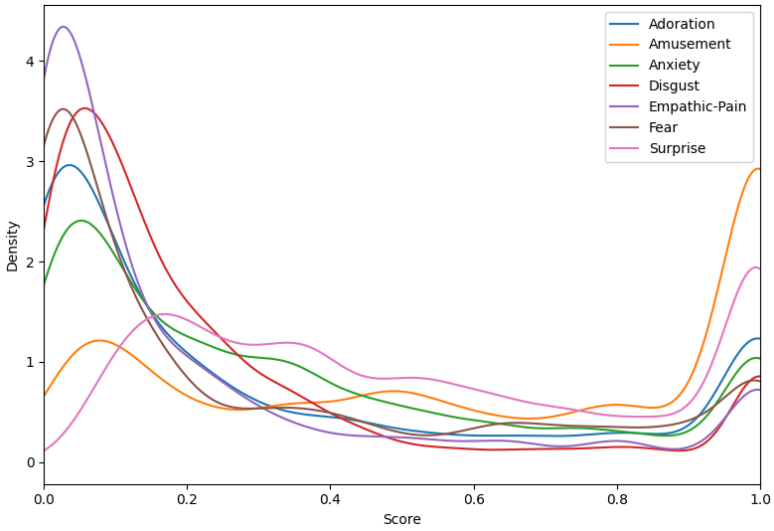

Figure 1. Probability density function of all emotions in the dataset generated by a Gaussian Kernel Density Estimation.

4. Methods

Our methodology involves three main approaches: (1) Video-LLaVA$_{querying}$, i.e., direct querying of Video-LLaVA [12] to obtain the emotional reaction intensities (see Section 4.1); (2) Video-LLaVA$_{prompting}$, i.e., probing with fine-tuning on the embeddings produced by Video-LLaVA [12] (see Section 4.2); and (3) VIPER-VATF [9], i.e., integration of textual features extracted from generated video descriptions into a state-of-the-art transformer-based architecture (see Section 4.3).

4.1. Direct Querying of MLLM

The current study uses Video-LLaVA [12], a state-of-the-art Large Language Model for video understanding, to query the emotion scores directly. The model is prompted with a specific question to assess the intensity of each of the seven emotions in the video. The prompt was carefully designed to elicit detailed responses about the emotional content of the videos:

```
"Can you assign a score between 0 and 1 to each of these emotions based on
  what is expressed by the subject in the video:  adoration, amusement,
          anxiety, disgust, emphatic pain, fear and surprise?"
```

4.2. Probing Network

Multimodal LLMs are complex systems with many parameters, making them challenging to train from scratch. One effective strategy to leverage their capabilities without extensive retraining is the use of probing networks. Probing involves fine-tuning a small set of additional parameters, typically linear layers, to adapt the model for a specific task. This approach is computationally efficient and allows us to extract useful information from the pre-trained embeddings of the LLM.

In our experiment, a probing strategy has been applied to Video-LLaVA [12] by fine-tuning a small regressor on the model's embeddings. Figure 2 schematizes this approach. First, Video-LLaVA is used to process the entire video sequence. The Video-LLaVA [12] encoder was frozen throughout the process to preserve the pre-trained general knowledge while limiting the computational time and resources needed. Moreover, since this MLLM does not employ any special token to represent the input sequence, average pooling is applied to all output tokens to obtain a final video representation. The probing network comprises two linear layers with an activation function between them. The final layer of the probing network ends with a neuron for each emotion, to which a sigmoid function is applied to obtain a score between 0 and 1 for each one. Finally, this probing network is

trained as a regressor to predict the emotion scores based on the embeddings provided by Video-LLaVA [12].

Additionally, we employ two different prompts during the embedding extraction phase to examine their impact on the performance of the probing network. The first prompt is more generic and asks for a simple video description:

```
"Describe the video."
```

The latter prompt is more specific and asks to focus on the emotions:

```
"Describe the reaction of the subject in the video."
```

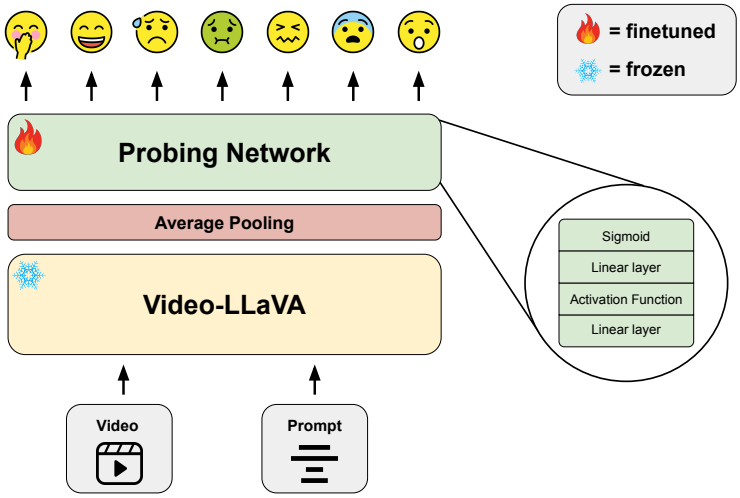

Figure 2. Sketch of the probing pipeline. Video-LLaVA produces meaningful video embeddings through token average pooling, to which a probing network is applied to predict emotion scores. The circle on the right side focuses on the internal structure of the probing network.

4.3. Integrating MLLM-Generated Description Features into a Transformer-Based Architecture

This approach generates textual descriptions of the videos using Video-LLaVA [12] with the same prompts employed in the probing strategy. These descriptions are then used to extract textual features, which were integrated into a state-of-the-art multimodal architecture that combines visual, acoustic, and textual information.

Figure 3 shows how textual features from Video-LLaVa [12] are integrated. First, it generates detailed textual descriptions for each video. These descriptions aim to capture the emotional nuances expressed by the subjects in the videos. Then, a text embedding model, i.e., RoBERTa, extracts meaningful textual features from the generated descriptions. These embeddings capture semantic information relevant to the emotions. Subsequently, the extracted textual features are integrated into an existing multimodal framework, namely ViPER [9], whose architecture is specifically designed to leverage visual, acoustic, and textual features to address the video emotion recognition task. This integration is achieved by replacing the textual embeddings produced by ViPER [9] with those extracted from Video-LLaVA-generated texts. Notably, Video-LLaVA [12] creates a single textual description for the entire video, whereas ViPER [9] exploits frame-level representations. To inject the new textual embeddings into the existing ViPER [9] framework, the new textual embedding is replicated to enrich each ViPER [9] multimodal token.

We also explore an alternative approach where, instead of using Video-LLaVA [12] to describe the entire video, LLaVA [11] is used to describe individual video frames separately. The aim is to enrich each multimodal token with a different textual embedding tailored to the specific frame rather than replicating the same embedding across all tokens. This method provides us with a more granular alignment of textual and visual information, potentially enhancing the model's ability to capture frame-specific emotional nuances. A sketch of this approach is depicted in Figure 4.

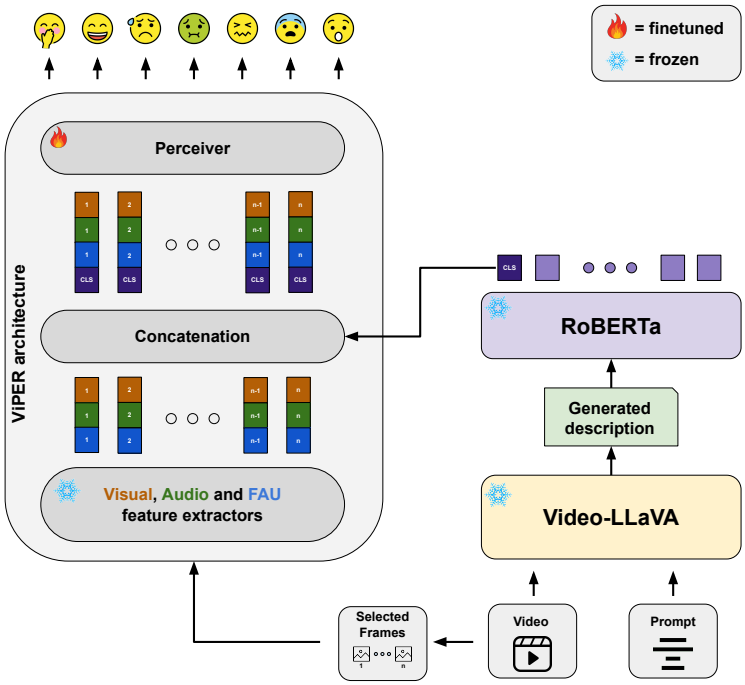

Figure 3. Integration of Video-LLaVA [12] textual features into the state-of-the-art ViPER [9] architecture. First, the entire video is passed as input to the Video-LLaVA [12] model. This MLLM produces a unique description that is encoded and concatenated to each input token of the Perceiver module of ViPER [9].

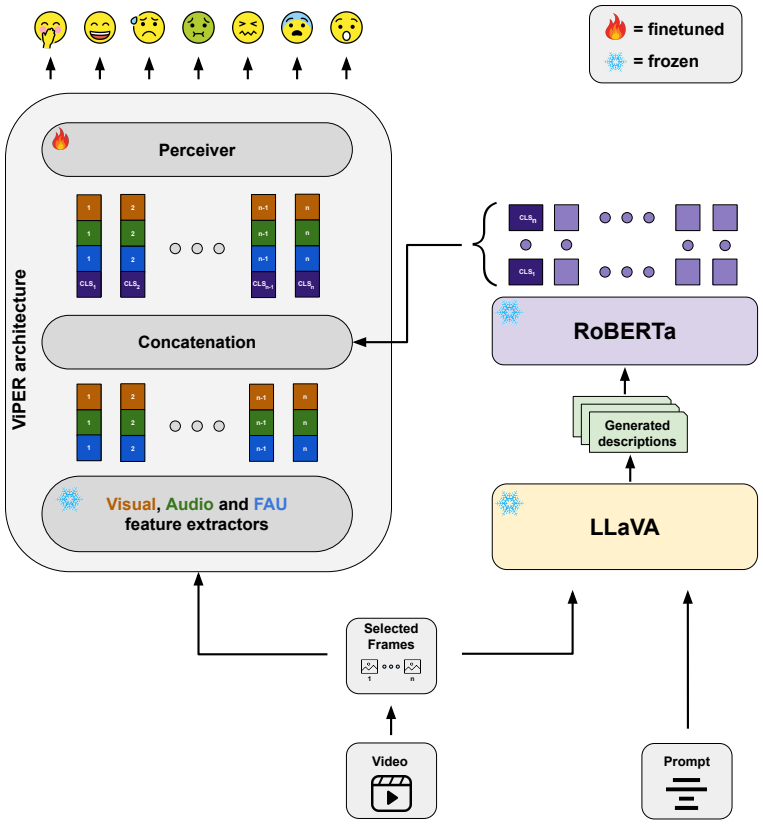

Figure 4. Integration of LLaVA textual features into the state-of-the-art ViPER [9] architecture. The video is sampled to extract 32 equally spaced frames, used to feed the LLaVA model. Then, it produces a different description for each frame. Finally, these descriptions are encoded and concatenated to the corresponding input token of the Perceiver module of ViPER [9].

5. Experimental Results

This section describes the empirical evaluation of the proposed approaches for emotion recognition. We performed quantitative (Section 5.2) and qualitative (Section 5.3) analyses to assess the impact of LLMs.

5.1. Experimental Setup

The experiments were executed on a machine equipped with an 18-core Intel Core i9-10980XE processor, an Nvidia A6000 GPU, and 128 GB of RAM. We chose $n = 32$ equidistant frames from each video, including the first and last frames. Both LLaVA and Video-LLaVA were used with default settings to perform inference. The probing network and the Perceiver module of ViPER [9] were fine-tuned for using the AdamW optimizer and the Mean Squared Error (MSE) as loss function. The probing network was trained for a maximum of 50 epochs using a learning rate equal to 10^{-4}, while the Perceiver module was fine-tuned for a maximum of 20 epochs using a learning rate equal to 10^{-5}.

5.2. Quantitative Results

Table 3 presents the performance of the proposed approaches for video emotion recognition, evaluated using the mean Pearson correlation [53] among all involved emotions. The first half of the table reports the dataset author's baselines and original ViPER [9] results. The second half shows our querying, probing, and integration results exploiting video and visual LLMs.

The proposed methods involving Video-LLaVA [12] and LLaVA [11] show varying levels of effectiveness:

- **Querying Video-LLaVA**: Directly querying Video-LLaVA [12] for emotion scores resulted in a mean Pearson correlation of 0.0937, which is lower than all baselines, indicating limited effectiveness for this approach. Additionally, it was observed that the generated text often used the same exact score value or a limited range of values for some emotions, e.g., the score 0.4 appears 2444 times out of 4657 in the Anxiety predictions. This suggests that text generation in a zero-shot fashion may be unsuitable for regression tasks, as it lacks the precision required for accurate scoring across a continuous range.
- **Probing Video-LLaVA**: Fine-tuning with probing strategies showed improvements, with mean Pearson correlations of 0.2333 for Prompt 1 and 0.2351 for Prompt 2. Although these scores did not surpass the baselines, they highlighted the potential of probing strategies. Additionally, this result indicates that the prompts used, whether general or specific for emotion recognition, do not greatly impact the performance. We also studied the impact of the employed activation function within the probing network. Table 4 reports the results obtained using seven different activation functions while adopting Prompt 2. Noteworthy is that the variation in performance as the activation function varied was very limited, i.e., from 0.2315 to 0.2353. However, the ReLU-based functions achieved a slightly superior result.
- **Integrating Video-LLaVA textual features**: Integrating Video-LLaVA-generated textual features into the ViPER-VATF [9] framework showed competitive performance, with mean Pearson correlations of 0.3004 for the general prompt and 0.3011 for the specific prompt, closely matching the performance of the original ViPER-VATF [9]. If the results are broken down by observing each emotion separately, these approaches surpassed the classical ViPER-VATF [9] for specific emotions. Specifically, using Prompt 1 yielded better performance for Anxiety and Empathic Pain, while Prompt 2 performed better on Adoration, Anxiety, and Surprise. On the other hand, the CLIP-based approach achieved the highest results in Amusement, Disgust, and Fear. Table 5 reports the breakdown results. Furthermore, we compared the impact that textual, acoustic and FAUs features had when combined with visual ones. The results are reported in Table 6. It is important to note that the textual features extracted from Video-LLaVA [12], although we did not have a contribution at the level of the FAU features, always

brought a benefit when injected into the model; this is in contrast to the acoustic ones, which occasionally did not improve or even worsened the performance of the model.

- **Integrating LLaVA textual features**: Using LLaVA [11] to describe video frames separately and integrating these frame-specific textual features into the ViPER-VATF [9] framework resulted in a mean Pearson correlation of 0.2895, indicating the viability of this alternative approach. However, integrating Video-LLaVA [12] textual was still better (up to 0.3011 vs. 0.2895).

Table 3. Mean Pearson correlations achieved by proposed methods. The higher result is highlighted in boldface.

Model	Mean Pearson Correlation
Baseline_{FAU}	0.2840
$\text{Baseline}_{VGGFace2}$	0.2488
ViPER-V	0.2712
ViPER-VATF_{CLIP}	**0.3025**
$\text{Video-LLaVA}_{querying}$	0.0937
$\text{Video-LLaVA}_{probing_1}$	0.2333
$\text{Video-LLaVA}_{probing_2}$	0.2351
$\text{ViPER-VATF}_{Video-LLaVA_1}$	0.3004
$\text{ViPER-VATF}_{Video-LLaVA_2}$	0.3011
$\text{ViPER-VATF}_{LLaVA}$	0.2895

Table 4. Activation function impact in the probing network. Higher result is highlighted in boldface.

Activation Function	Mean Pearson Correlation
Linear	0.2338
ReLU	0.2351
Leaky ReLU	**0.2353**
Tanh	0.2332
Sigmoid	0.2343
GELU	0.2315
ELU	0.2326

Table 5. Single-emotion Pearson correlation depending on the textual features employed in the ViPER-VATF [9] approach. The higher result, separately for each emotion, is highlighted in boldface.

Textual Features	Mean Pearson Correlation							
	Adoration	Amusement	Anxiety	Disgust	Empathic Pain	Fear	Surprise	Average
CLIP	0.2575	**0.3651**	0.3294	**0.2755**	0.2824	**0.3190**	0.2890	**0.3025**
Video-LLaVA_1	0.2575	0.3646	**0.3303**	0.2632	**0.2886**	0.3122	0.2865	0.3004
Video-LLaVA_2	**0.2624**	0.3631	0.3297	0.2708	0.2763	0.3155	**0.2918**	0.3011

Table 6. Ablation study on different modalities.

Image	Audio	Text	FAU	Mean Pearson Correlation
✓	-	-	-	0.2712
✓	✓	-	-	0.2748
✓	-	✓	-	0.2758
✓	-	-	✓	0.2978
✓	✓	✓	-	0.2758
✓	-	✓	✓	**0.3011**
✓	✓	-	✓	0.2924
✓	✓	✓	✓	**0.3011**

5.3. Qualitative Analysis

To better understand the performance differences between the ViPER [9] models based on CLIP and those based on Video-LLaVA [12], we conducted a qualitative analysis by discretizing the predictions and ground truth (GT) into bins of 0.1 range. This was done separately for each emotion. The confusion matrices reveal how often predictions fell within certain ranges of the true emotion scores. Examining these matrices allows us to observe prediction patterns and identify areas where each approach excelled or fell short.

A key observation from the confusion matrices is the range of prediction values:

- ViPER-VATF$_{CLIP}$: The predictions are more concentrated near the GT average value. This indicates that the CLIP-based solution is more focused on predicting scores that are close to the average GT value. It suggests a tendency to overfit on the average value, making it more accurate for samples whose scores are near the mean.
- ViPER-VATF$_{Video-LLaVA}$: The predictions cover a wider range of values. This means that the LLM-based solution is better at predicting scores that can be considered outliers with respect to the average value. These outliers include cases where an emotion is particularly evident or notably missing.

Figures 5 and 6 show the confusion matrices obtained using the original ViPER-VATF [9] architecture for the Adoration and Empathic Pain emotions, respectively. Notably, the predictions focus in the range $[0.2, 0.5]$ for Adoration and $[0.1, 0.4]$ for Empathic Pain, with just a few predictions in the adjacent bins. Figures 7 and 8 show the confusion matrices obtained by integrating the textual features from Video-LLaVA [12] into the ViPER-VATF [9] architecture for the same emotions. It can be observed that the majority of predictions fall in a range of values wider with respect to the previous case, i.e., $[0.1, 0.6]$ for Adoration and $[0.0, 0.5]$ for Empathic Pain.

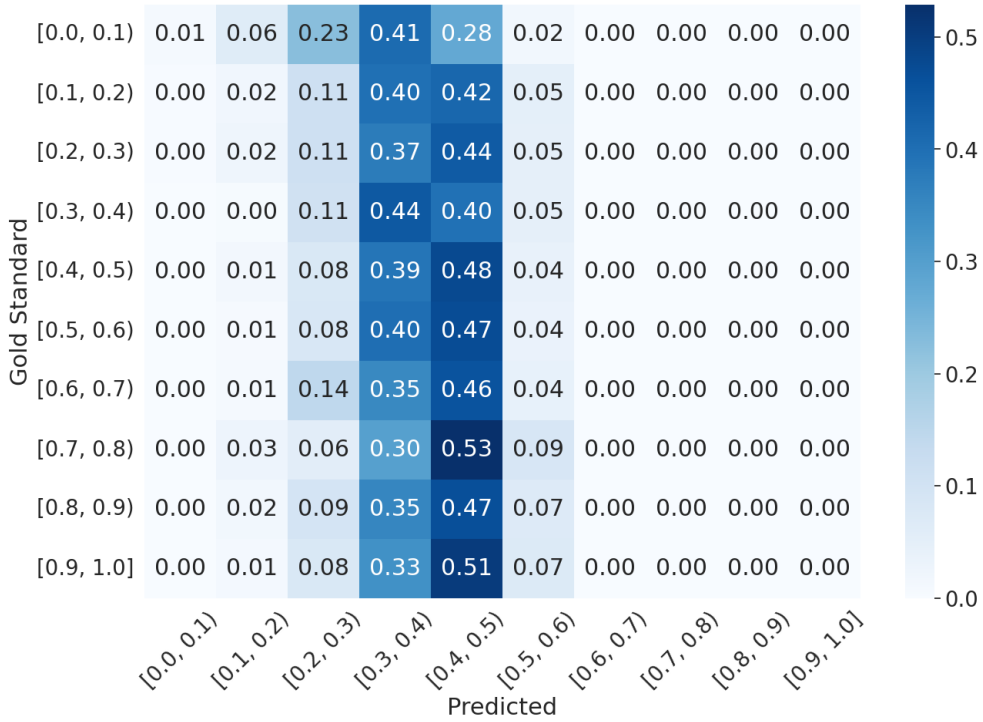

Figure 5. ViPER-VATF [9] based on CLIP [26] textual features prediction for the Adoration emotion. The average prediction is 0.3737 ± 0.0867, with a minimum prediction score of 0 and a maximum of 0.5500.

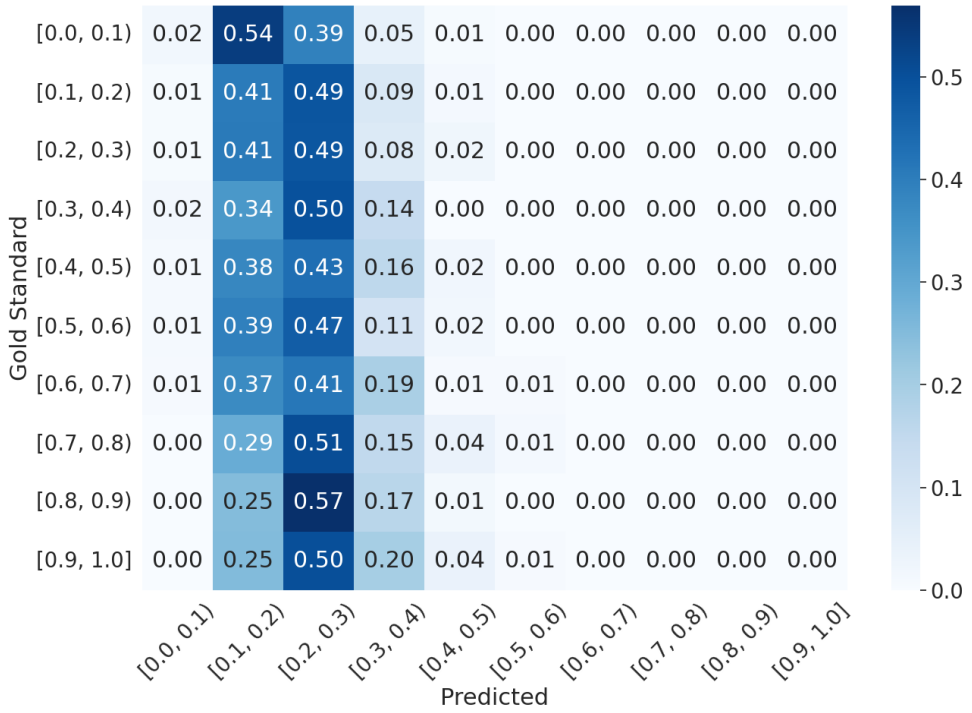

Figure 6. ViPER-VATF [9] based on CLIP [26] textual features prediction for the Empathic Pain emotion. The average prediction is 0.2088 ± 0.0669, with a minimum prediction score of 0.0812 and a maximum of 0.5513.

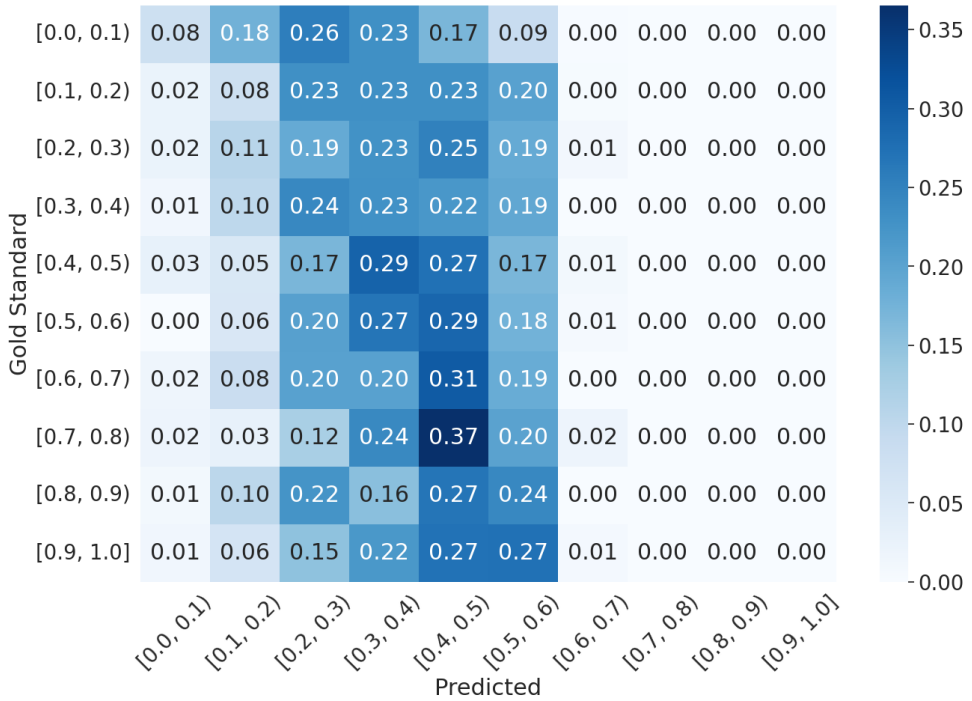

Figure 7. ViPER-VATF [9] based on Video-LLaVA [12] textual features prediction for the Adoration emotion. The average prediction is 0.3465 ± 0.1387, with a minimum score of 0 and maximum score of 0.6289.

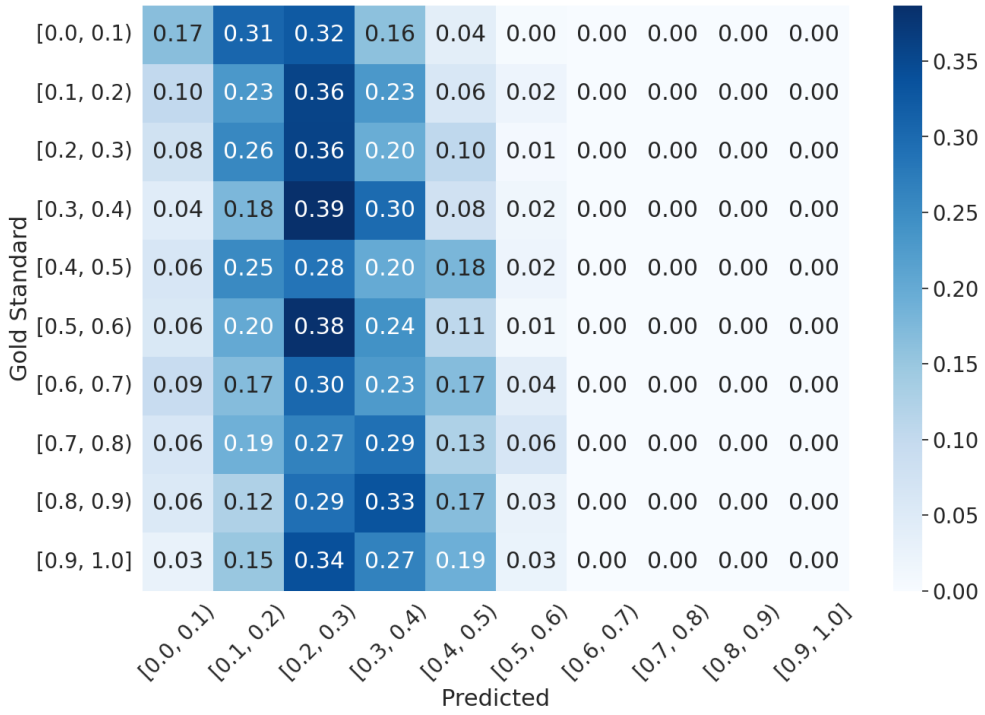

Figure 8. ViPER-VATF [9] based on Video-LLaVA [12] textual features prediction for the Empathic Pain emotion. The average prediction is 0.2372 ± 0.1099, with a minimum score of 0.0369 and a maximum score of 0.5848.

The wider range of predictions in the Video-LlaVA-based approach suggests its superior ability to detect and accurately score emotions that deviate significantly from the mean. This capability is particularly valuable in scenarios where certain emotions are either strongly expressed or barely present, which is often critical for nuanced emotional understanding.

In summary, while the CLIP-based solution shows robustness in predicting common emotional expressions, the LLM-based solution offers a more comprehensive approach capable of recognizing and scoring a wider variety of emotional intensities, including extreme cases. This qualitative analysis underscores the potential for combining both approaches to achieve a more balanced and accurate emotion recognition system.

6. Discussion

Current research in the field of emotion recognition has largely explored the use of transformers. However, training such models requires a large set of high-quality annotated examples and is potentially costly. This work explores the parallel direction of using pre-trained Multimodal Large Language Models. Due to the specificity of the ERI estimation task, the capabilities of Multimodal LLMs in a zero-shot setting are questionable. Our results show that querying Video-LLaVA [12] directly for emotion scores resulted in a mean Pearson correlation of 0.0937, lower than all baselines, indicating limited effectiveness for this approach. However, fine-tuning with probing strategies showed improvements, with mean Pearson correlations of 0.2333 for Prompt 1 and 0.2351 for Prompt 2, highlighting the potential of probing strategies despite not surpassing baselines. Additionally, integrating Video-LLaVA-generated textual features into the ViPER-VATF [9] framework showed competitive performance, with mean Pearson correlations of 0.3004 for the general prompt and 0.3011 for the specific prompt, closely matching the performance of the original ViPER-VATF [9] (0.3011 vs. 0.2895).

6.1. Multimodal LLM vs. Transformers

A strategy based on Multimodal LLMs has several advantages with respect to traditional techniques such as ViPER [9], which typically utilizes the CLIP model to align video

frames with predefined textual templates. Our approach replaces the CLIP model with the Video LLM, obviating the need to pre-define textual templates to match frames. Moreover, this method enhances the system's extensibility, as the inclusion of new emotions in the recognition task is automatically handled by Video-LLaVA [12], eliminating the need for additional template redefinitions. This flexibility simplifies adapting the model to recognize new emotions, improving its scalability and applicability to a broader range of emotional contexts. Additionally, the integration of Video-LLaVA [12] textual features surpassed the classical ViPER-VATF [9] for specific emotions such as Anxiety and Empathic Pain, indicating a robust performance for extreme emotion cases without any ad hoc model fine-tuning.

6.2. Application Scenarios

These preliminary results support the application of video–language LLMs in a variety of real-life application contexts, including item recommendations on multimedia platforms, sentiment analysis in the financial domain, healthcare monitoring, and engagement analysis for learning analytics applications.

6.3. Limitations

The main limitations of the present work are (1) the limited adaptability of existing MLLMs such as Video-LlaVa [12], which, for example, hinder the application of in-context learning strategies; (2) the sensitivity of the proposed approach to the presence of bias and to data overfitting; and (3) the limited accountability and interpretability of the proposed solutions.

7. Conclusions and Future Works

The empirical results shown in this study confirm the potential of MLLMs in addressing complex video understanding. Pre-trained Multimodal LLMs allow us to achieve interesting performance in detecting a broader range of reaction scores despite leaving room for improvements. Specifically, integrating Video-LLaVA [12] textual features into the ViPER-VATF [9] framework resulted in a mean Pearson correlation of up to 0.3011, demonstrating its effectiveness. Additionally, the qualitative analysis highlights a key advantage of the new approach: it predicts a wider intensity range for every emotion compared to the original ViPER [9]. This means that the Video-LLaVA-based approach is better at recognizing and scoring extreme emotion cases, which is critical for nuanced emotional understanding. However, this study also highlights several limitations, such as the need for large ad hoc training datasets for model fine-tuning and challenges in integrating multimodal modalities.

As future work, we plan to extend the scope of our analysis to other emotion recognition scenarios, explore the use of audio–language LLMs, and integrate LLMs into diverse transformer-based architectures.

Author Contributions: Conceptualization, L.V., L.C., P.G.; methodology, L.V., L.C., P.G.; software, L.V.; validation, L.V.; formal analysis, L.V.; investigation, L.V., L.C.; resources, L.V., L.C., P.G.; data curation, L.V.; writing—original draft preparation, L.V., L.C.; writing—review and editing, L.V., L.C., P.G.; visualization, L.V.; supervision, L.C., P.G.; project administration, L.C., P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. The research uses data released by third parties. Data were obtained from Hume AI Inc. and are available for research purposes only by contacting competitions@hume.ai.

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

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
MLLM	Multimodal Large Language Model
ERI	emotional reaction intensity

References

1. Bartolome, A.; Niu, S. A Literature Review of Video-Sharing Platform Research in HCI. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; CHI '23. [CrossRef]
2. Hossain, M.S.; Muhammad, G. Cloud-Assisted Speech and Face Recognition Framework for Health Monitoring. *Mob. Netw. Appl.* **2015**, *20*, 391–399. [CrossRef]
3. Zhang, Z.; Coutinho, E.; Deng, J.; Schuller, B. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Trans. Audio, Speech Lang. Proc.* **2015**, *23*, 115–126. [CrossRef]
4. Szwoch, M. Design Elements of Affect Aware Video Games. In Proceedings of the Mulitimedia, Interaction, Design and Innnovation, Warsaw, Poland, 29–30 June 2015; MIDI '15. [CrossRef]
5. Christ, L.; Amiriparian, S.; Baird, A.; Tzirakis, P.; Kathan, A.; Mueller, N.; Stappen, L.; Messner, E.; König, A.; Cowen, A.; et al. The MuSe 2022 Multimodal Sentiment Analysis Challenge: Humor, Emotional Reactions, and Stress. In Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, Lisboa, Portugal, 10 October 2022.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
7. Sun, L.; Xu, M.; Lian, Z.; Liu, B.; Tao, J.; Wang, M.; Cheng, Y. Multimodal Emotion Recognition and Sentiment Analysis via Attention Enhanced Recurrent Model. In Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, Virtual Event, 24 October 2021; MuSe '21, pp. 15–20. [CrossRef]
8. Sun, L.; Lian, Z.; Tao, J.; Liu, B.; Niu, M. Multi-Modal Continuous Dimensional Emotion Recognition Using Recurrent Neural Network and Self-Attention Mechanism. In Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop, Seattle, WA, USA, 16 October 2020; MuSe'20, pp. 27–34. [CrossRef]
9. Vaiani, L.; La Quatra, M.; Cagliero, L.; Garza, P. ViPER: Video-based Perceiver for Emotion Recognition. In Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, Lisboa, Portugal, 10 October 2022; MuSe' 22, pp. 67–73. [CrossRef]
10. OpenAI. GPT-4 Technical Report. *arXiv* **2023**. [CrossRef]
11. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *arXiv* **2023**. [CrossRef]
12. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment before Projection. *arXiv* **2023**, arXiv:2311.10122. [CrossRef]
13. Zhao, Y.; Misra, I.; Krähenbühl, P.; Girdhar, R. Learning Video Representations From Large Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 6586–6597.
14. Huang, J.; Chang, K.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1049–1065. [CrossRef]
15. Hazmoune, S.; Bougamouza, F. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108339. [CrossRef]
16. Zhou, H.; Meng, D.; Zhang, Y.; Peng, X.; Du, J.; Wang, K.; Qiao, Y. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; ICMI '19, pp. 562–566. [CrossRef]
17. Liu, C.; Jiang, W.; Wang, M.; Tang, T. Group Level Audio-Video Emotion Recognition Using Hybrid Networks. In Proceedings of the 2020 International Conference on Multimodal Interaction, Virtual Event, 25–29 October 2020; ICMI '20, pp. 807–812. [CrossRef]
18. Qi, F.; Yang, X.; Xu, C. Zero-shot Video Emotion Recognition via Multimodal Protagonist-aware Transformer Network. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; MM '21, pp. 1074–1083. [CrossRef]
19. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA, 13–15 October 2004; ICMI '04, pp. 205–211. [CrossRef]
20. Yu, J.; Zhu, J.; Zhu, W.; Cai, Z.; Xie, G.; Li, R.; Zhao, G.; Ling, Q.; Wang, L.; Wang, C.; et al. A dual branch network for emotional reaction intensity estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–23 June 2023; pp. 5810–5817.

21. Kollias, D.; Tzirakis, P.; Baird, A.; Cowen, A.; Zafeiriou, S. Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–23 June 2023; pp. 5888–5897.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929
23. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv* **2021**, arXiv:2110.13900.
24. Li, J.; Chen, Y.; Zhang, X.; Nie, J.; Li, Z.; Yu, Y.; Zhang, Y.; Hong, R.; Wang, M. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–23 June 2023; pp. 5837–5843.
25. Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; Carreira, J. Perceiver: General perception with iterative attention. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual Event, 18–24 July 2021; pp. 4651–4664.
26. 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. PMLR, Virtual Event, 18–24 July 2021; pp. 8748–8763.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision— ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.
28. Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv* **2023**, arXiv:2308.01390. [CrossRef]
29. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 25278–25294.
30. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv* **2023**, arXiv:2304.10592. [CrossRef]
31. Zhang, Q.; Zhang, J.; Xu, Y.; Tao, D. Vision Transformer with Quadrangle Attention. *arXiv* **2023**, arXiv:2303.15105. [CrossRef]
32. Ordonez, V.; Kulkarni, G.; Berg, T. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2011; Volume 24.
33. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv* **2021**, arXiv:2111.02114.
34. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597. [CrossRef]
35. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
36. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2556–2565.
37. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3558–3568.
38. Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; Taşırlar, S. Introducing our Multimodal Models. 2023. Available online: https://www.adept.ai/blog/fuyu-8b (accessed on 1 June 2024).
39. Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.; Lu, P.; Wang, T.; Wei, Z. Valley: Video assistant with large language model enhanced ability. *arXiv* **2023**, arXiv:2306.07207.
40. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 1728–1738.
41. Yu, E.; Zhao, L.; Wei, Y.; Yang, J.; Wu, D.; Kong, L.; Wei, H.; Wang, T.; Ge, Z.; Zhang, X.; et al. Merlin: Empowering Multimodal LLMs with Foresight Minds. *arXiv* **2023**, arXiv:2312.00589.
42. Huang, B.; Wang, X.; Chen, H.; Song, Z.; Zhu, W. VTimeLLM: Empower LLM to Grasp Video Moments. *arXiv* **2023**, arXiv:2311.18445. [CrossRef]
43. Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; Carlos Niebles, J. Dense-captioning events in videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 706–715.
44. Gao, J.; Sun, C.; Yang, Z.; Nevatia, R. Tall: Temporal activity localization via language query. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5267–5275.
45. Maaz, M.; Rasheed, H.; Khan, S.; Khan, F.S. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv* **2023**, arXiv:2306.05424.
46. Shu, F.; Zhang, L.; Jiang, H.; Xie, C. Audio-Visual LLM for Video Understanding. *arXiv* **2023**, arXiv:2312.06720. [CrossRef]

47. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv* **2023**, arXiv:2306.09093. [CrossRef]
48. Sigurdsson, G.; Russakovsky, O.; Farhadi, A.; Laptev, I.; Gupta, A. Much ado about time: Exhaustive annotation of temporal data. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Austin, TX, USA, 30 October–3 November 2016; Volume 4, pp. 219–228.
49. Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T.K.; Hori, C.; Anderson, P.; et al. Audio Visual Scene-Aware Dialog. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
50. Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. Video Understanding with Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.17432. [CrossRef]
51. Tan, Q.; Ng, H.T.; Bing, L. Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 14820–14835. [CrossRef]
52. Ekman, P. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*; Henry Holt and Company: New York, NY, USA, 2004.
53. Tan, P.N.; Steinbach, M.; Karpatne, A.; Kumar, V. *Introduction to Data Mining*, 2nd ed.; Pearson: London, UK, 2018.