This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Face-LLaVA: Facial Expression and Attribute Understanding through Instruction Tuning

Ashutosh Chaubey Xulang Guan Mohammad Soleymani
Institute for Creative Technologies, University of Southern California
{achaubey, xulanggu}@usc.edu, soleymani@ict.usc.edu

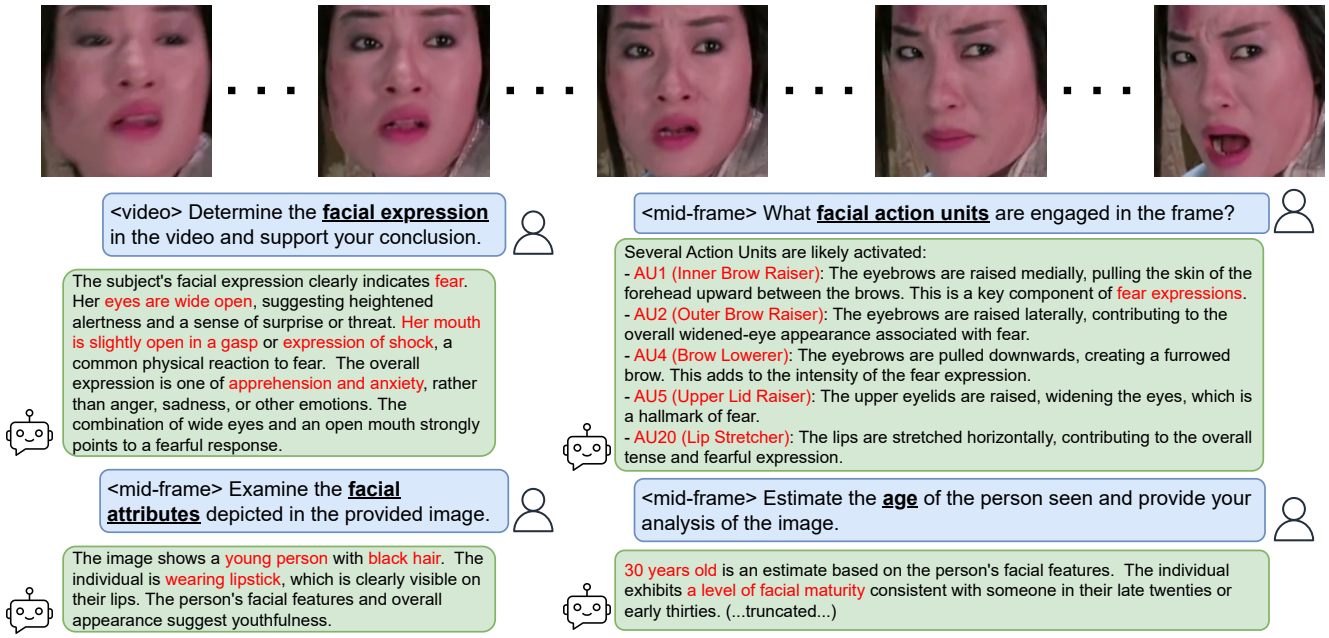


Figure 1. We introduce Face-LLaVA, a multimodal LLM for facial analysis and reasoning which can perform a variety of facial analysis tasks. Instead of only predicting the labels, Face-LLaVA provides a detailed description (or reason) for its predictions.

Abstract

The human face plays a central role in social communication, necessitating the use of performant computer vision tools for human-centered applications. We propose Face-LLaVA, a multimodal large language model for face-centered, in-context learning, including facial expression and attribute recognition. Additionally, Face-LLaVA is able to generate natural language descriptions that can be used for reasoning. Leveraging existing visual databases, we first developed FaceInstruct-1M, a face-centered database for instruction tuning MLLMs for face processing. We then developed a novel face-specific visual encoder powered by Face-Region Guided Cross-Attention that integrates face geometry with local visual features. We evaluated the proposed method across nine different datasets and five different face processing tasks, including facial expression recognition, action unit detection, facial attribute detection, age estimation and deepfake detection. Face-LLaVA achieves superior results compared to existing open-source MLLMs and competitive performance compared to commercial solutions. Our model output also receives a higher reasoning rating by GPT under a zero-shot setting across all the tasks. Both our dataset and model are available at https://github.com/ihp-lab/Face-LLaVA/ to support future advancements in social AI and foundational vision-language research.

1. Introduction

The human face serves as a fundamental conduit for social communication. Hence, face analysis including facial expression recognition [79], action unit (AU) detection [99], facial attribute detection [101], age estimation [2] has received considerable attention. These tasks have a plethora of applications such as human-computer interac-

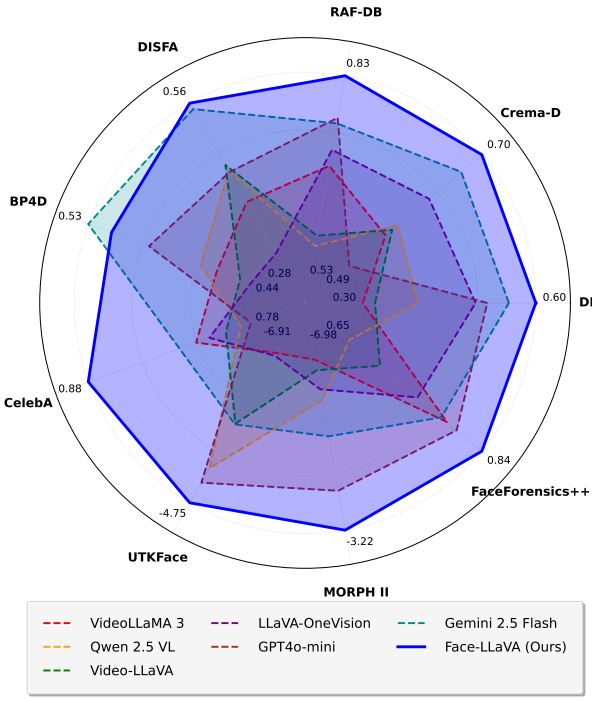


Figure 2. Comparison of *Face-LLaVA* with existing MLLMs for zero-shot performance on different face analysis benchmarks (values normalized and the metric is negative MAE for UTKFace [94] and MorphII [60]).

tion [65], social behavior assessment [55], e-learning [20] and surveillance [54], among others.

Existing face analysis approaches suffer from two major limitations: (i) they are developed for specific tasks such as expression recognition [7, 85] or attribute detection [45, 48], limiting their generalizability; and (ii) they primarily output labels without natural language description for their predictions. The ability to use natural language to describe predictions is crucial for critical applications such as healthcare and surveillance [11, 80].

To address the first limitation, recent work has explored generalist face models capable of handling multiple tasks [6, 58], but these lack a natural language interface and offer no reasoning transparency. The emergence of Multimodal Large Language Models (MLLMs), such as LLaVA [29, 42, 93], has enabled their application to vision tasks, including face analysis. EmoVIT [82] improves emotion understanding via instruction tuning, while Emotion-LLaMA [8] combines audio, video, and text for emotion reasoning. Foteinopoulou et al. [12] study LLMs for deepfake detection. However, these are task-specific and often rely on background or audio cues. Recent models like FABA [37] and VL-FAU [15] begin to bridge vision and language reasoning for faces, but remain limited to expression understanding. Moreover, current MLLM-based efforts in facial attribute understanding [31, 32, 67, 95] lack support for video-based analysis.

In this paper, we propose Face-LLaVA, a generalist MLLM that can perform perception with reasoning on different face analysis tasks. Fig. 1 shows the interaction of a user with our model. Note that given a video or an image, our model can recognize facial expressions, facial action units, age, and facial attributes and identify face forgery (see Fig. 3). This advancement paves the way for developing general-purpose AI-driven social agents with enhanced facial reasoning capabilities, making them more applicable to real-world scenarios.

Training an MLLM for face analysis requires a vast amount of instruction-tuning data covering different tasks. However, existing instruction-tuning datasets [8, 37, 39] for related tasks are relatively small, task-specific, and often limited to either images or videos (see Tab. 1). To address this limitation, we introduce *FaceInstruct-1M*, a large-scale face-specific dataset comprising one million samples, including around 850k images and 120 hours of video data, annotated with task-specific instructions and descriptions across five distinct face analysis tasks. Furthermore, as the dataset is automatically generated, we employ a GPT-4o [72]-assisted evaluation and filtering process to ensure high-quality data for fine-tuning the MLLM.

General-purpose vision encoders used in MLLMs, such as CLIP [59] and LanguageBind [100], are not designed to extract face-specific features such as facial landmarks, which are crucial for fine-grained face analysis. To address this, we propose a novel architecture that is specifically designed for face processing to enhance its performance. Unlike previous approaches [37], our method is more efficient in utilizing the LLM's context window and is generalizable to both images and videos. Specifically, we introduce a *face-region landmark projector*, which maps face expert (landmark) features to face-region tokens and refines visual features via cross-attention with the extracted visual tokens.

We conduct extensive experiments on nine datasets spanning five different face analysis tasks, demonstrating that our method outperforms existing open-source MLLMs across all benchmarks in a zero-shot setting (Fig. 2) and achieves performance competitive with supervised task-specific approaches. To further assess the reasoning capabilities of our model, we conduct automatic GPT-based evaluations comparing our approach against recent MLLMs and showcasing its superior generative and captioning capabilities.

In summary, the key contributions of this work are as follows. (i) We introduce *FaceInstruct-1M*, a large-scale face analysis dataset comprising one million instruction-tuning samples, including both images and videos. The dataset spans five key face-related tasks: expression recognition, action unit (AU) detection, attribute detection, age estimation, and deepfake detection. (ii) We propose *Face-*

LLaVA, a multimodal large language model architecture for face analysis that incorporates face landmark features using a face-region projector and face-region-guided cross-attention, enabling effective instruction tuning for face processing tasks. (iii) Through extensive experiments, we demonstrate that our approach outperforms existing MLLMs in a zero-shot setting on nine datasets across five face analysis tasks, achieving superior performance on both traditional benchmarks and GPT-assisted evaluations of generated responses. Additionally, we show that our method achieves competitive performance with supervised task-specific techniques across all tasks.
FaceInstruct-1M and *Face-LLaVA* model weights are available at https://github.com/ihp-lab/Face-LLaVA/.

2. Related Works

2.1. Traditional face analysis

Extensive research has been conducted on individual face analysis tasks, including expression recognition [7, 68, 76, 81, 85, 86, 91, 92], action unit detection [36, 46, 63, 69], age estimation [2, 13, 66], attribute detection [45, 48, 62], and deepfake detection [1, 9, 56]. Recently, the focus has shifted toward generalist models [51, 57, 58] and learning robust facial representations [6, 64, 98]. Faceptor [58] extends FaRL [98] and SWINFace [57] with a single encoder and dual-decoder transformer for multi-task face analysis. MARLIN [6] and PrefAce [19] use self-supervised learning on videos via masked autoencoders (MAE) [16, 74] to learn robust representations, while PCL [44] employs a pose-disentangled decoder for contrastive learning. Building on these, our work presents a generalist face analysis model that goes beyond perception by incorporating reasoning over predictions, enhancing interpretability and context-awareness.

2.2. Face analysis using instruction tuning

Multimodal LLMs improve visual reasoning by leveraging the broad knowledge of LLMs [40, 42, 73, 88, 93]. However, their application to face analysis is limited due to the lack of large-scale instruction-tuning data. Existing face datasets [22, 45, 49, 50, 60, 61] mainly provide categorical or numeric labels (Tab. 1). Additionally, current instruction-tuning datasets for emotion recognition [8, 39, 43, 82] often depend on context or audio rather than facial cues. FABAInstruct [37] targets face analysis but is limited to still images and affective behavior. In contrast, we introduce *FaceInstruct-1M*, a large-scale instruction-tuning dataset for diverse face-related tasks.

While prior work has explored reasoning in VLMs for face tasks [8, 12, 82], comprehensive face-specific reasoning remains limited. AU-LLAVA [18] predicts AUs and intensities but lacks reasoning. VL-FAU [15] and EmoLA [37] incorporate visual-language frameworks and landmark priors for interpretable affective analysis. EMO-LLaMA [83] enhances facial encoding for reasoning and dialogue.

Face-MLLM [67] uses Gemini to annotate Laion-Face [97] for instruction tuning but focuses on perception without any reasoning and depends on Gemini's capabilities. FaVChat [95] deals with facial analysis for videos but can perform only facial attribute understanding. In contrast, Face-LLaVA supports both images and videos, builds on standard datasets, and can perform multiple face analysis tasks.

Dataset	Face Spec.	Modality Img.	Modality Vid.	Number of Samples Expr.	AU	Attr.	Age	DF.	Tot.	Anno. Type
RAF-DB [34]	✓	✓	✗	30k	-	-	-	-	30k	C
AffectNet [50]	✓	✓	✗	440k	-	-	-	-	440k	C
DFEW [22]	✓	✗	✓	16k	-	-	-	-	16k	C
FERV39K [78]	✓	✗	✓	39k	-	-	-	-	39k	C
DISFA [49]	✓	✓	✗	-	140k	-	-	-	140k	C
BP4D [90]	✓	✓	✗	-	150k	-	-	-	150k	C
AffWild2 [27]	✓	✓	✗	2.6M	2.6M	-	-	-	2.6M	C+N
CelebA [45]	✓	✓	✗	-	-	200k	-	-	200k	C
MORPH II [60]	✓	✓	✗	-	-	-	50k	-	50k	N
UTK Face [94]	✓	✓	✗	-	-	-	20k	-	20k	N
FF++ [61]	✓	✗	✓	-	-	-	-	5k	5k	C
Fake-AVC [25]	✓	✗	✓	-	-	-	-	20k	20k	C
EMER [39]	✗	✗	✓	450	-	-	-	-	450	D
MAFW [43]	✗	✗	✓	10k	-	-	-	-	10k	C+SD
EmoVIT [82]	✗	✗	✓	1.6k	-	-	-	-	1.6k	I+D
MERR [8]	✗	✗	✓	30k	-	-	-	-	30k	I+D
FABA Instr. [37]	✓	✓	✗	20k	20k	-	-	-	20k	I+D
Face-Instruct-1M	✓	✓	✓	**350k**	**290k**	**200k**	**70k**	**95k**	**1M**	**I+D**

Table 1. Comparison of the proposed *FaceInstruct-1M* dataset with various annotated datasets of face-related tasks. C:category, SD:short description, I:instruction, D:description, N:number (regression), DF.: deepfake detection, Face Spec.: Face-Specific., AU: action unit.



Figure 3. *FaceInstruct-1M* dataset samples for different tasks.

3. FaceInstruct-1M

We present *FaceInstruct-1M*, a dataset for instruction-tuning MLLMs on face-centric tasks. As shown in Tab. 1, it includes instructions and descriptions for expression recognition, AU detection, deepfake detection, attribute detection, and age estimation. Unlike prior datasets [22, 45, 49, 61, 78] that target single tasks, *FaceInstruct-1M* spans

Task	Datasets	Mean GPT Rating (1-10)			
		Label Acc.	Desc.-Vid. Consistency	Desc.-Lab. Consistency	Overall
FaceInstruct-1M					
Expression	DFEW [22], MAFW [43], FERV39k [78] Crema-D [24], AffectNet [50], RAF-DB [34]	8.81	8.74	8.79	8.71
AU	DISFA [49], BP4D [90]	8.17	8.23	7.62	7.99
Attribute	CelebA [45]	8.79	8.75	8.52	8.69
Age	MORPH II [60], UTK Face [94]	8.63	7.92	7.98	7.99
Deepfake	FF++ [61], Fake AV Celeb [25]	8.84	8.13	8.60	8.20
FABAInstruct [37]					
Expression	AffectNet [50]	8.79	6.44	6.32	6.22

Table 2. Datasets used to construct the *FaceInstruct-1M* and GPT [72] ratings for different aspects of the data. Ratings for FABAInstruct [37] are calculated as a reference.

multiple domains. Existing instruction-tuning [8, 82] and captioning datasets [39, 43] include background and audio cues, while ours focuses purely on facial information. Although FABA-Instruct [37] is face-specific, it is image-only, smaller in size, and limited in scope. To our knowledge, *FaceInstruct-1M* is the first large-scale, multi-task instruction dataset for face analysis using both images and videos.

3.1. Annotation Pipeline

FaceInstruct-1M is built using images, videos, and annotations from existing task-specific datasets (Tab. 2, Appendix D). As we focus only on faces, we crop face regions and retain videos with a single subject throughout (Appendix E.1).

Although many existing annotations are categorical or minimal, they effectively guide automatic description generation (see Appendix E.5 for analysis). To create video descriptions, we provide the face video, its label, and task-specific prompts to Gemini Flash [70], chosen for its accuracy, speed, and low cost. All the prompts used for annotation are detailed in Appendix E.2.

Inspired by FABA-Instruct [37], we pair each sample with a matching instruction. We curate 100 hand-crafted instructions per task and randomly associate them with generated descriptions. See Appendix E for details.

3.2. Dataset Rating and Filtering

Tab. 2 lists the datasets used in *FaceInstruct-1M*, some of which [22, 43, 45, 78] include background and audio, making their annotations influenced by non-facial cues. Furthermore, since our annotations are generated using an LLM, it is crucial to perform a sanity check and apply filtering to ensure high data quality.

We use a GPT-4o [72]-based rating pipeline that scores each sample based on (i) accuracy of the label w.r.t. the video, (ii) consistency between the description and the video, (iii) consistency between the description and the label, and (iv) overall visual quality. Tab. 2 shows that *FaceInstruct-1M* receives consistently high ratings. For calibration, we also evaluate FABA-Instruct [37]. Samples with an average rating less than 6 are discarded, filtering out about 7% of the dataset (Appendix E.4 for details).

3.3. Qualitative analysis

Fig. 3 presents samples from the *FaceInstruct-1M* dataset across different tasks. Notably, for video inputs, the descriptions capture subtle variations in various facial attributes and justify the responses to the given instructions based on these observations. It is also important to highlight that the dataset encompasses tasks reliant on muscle movements (e.g., expression recognition, AU detection, and deepfake detection), facial attributes (including age), and overall video quality (deepfake detection). More data samples are present in Appendix E.7.

3.4. Evaluation

Since an MLLM trained on our dataset generates natural language, its evaluation must consider both (i) metrics for face analysis, such as average recall or accuracy (Appendix F.3), and (ii) the reasoning capabilities w.r.t. instructions. Similar to FABAInstruct [37], we assess traditional metrics using synonym matching (refer Appendix F.1) for expression recognition, attribute detection, and deepfake detection, while for AU detection and age estimation, we extract numerical values directly from the generated text.

To assess MLLMs' reasoning on face tasks, traditional text metrics like BLEU [53] or ROUGE [41] are limited by shallow n-gram matching. Instead, we use GPT-4o mini [72] to evaluate: (i) alignment of reasoning with the video, (ii) consistency with the ground-truth label, and (iii) overall completeness in supporting the label based on the video, similar to Sec. 3.2.

We also curate a test set of 500 samples per task from DFEW [22], DISFA [49], CelebA [45], UTKFace [94], and FaceForensics++ [61]. Samples are selected from top-GPT-rated data and refined to match class distributions.

4. Face-LLaVA

Fig. 4 illustrates the Face-LLaVA architecture for instruction-tuning on face images and videos. Similar to prior MLLMs [40, 42], it consists of a patch-based vision encoder $\mathbf{E_V}$ to encode input image (or video), a tokenizer $\mathbf{E_T}$ to encode the text instruction and a large language model decoder Φ to generate responses according to the instruction x_i. Inspired by Lin et al. [100], we use LanguageBind [100] vision encoders for video and image and learn a joint vision projector $\mathbf{P_V^\theta}$ from the visual token space to the language token space. Specifically, given a visual input x_v, we get visual tokens $h_v \in \mathbb{R}^{T\times N\times d}$ as follows,

$$h_v = \mathbf{P_V^\theta}(\mathbf{E_V}(x_v)) \tag{1}$$

where T is the number of video frames (1 for images), N is the number of visual tokens from the patch-based vision

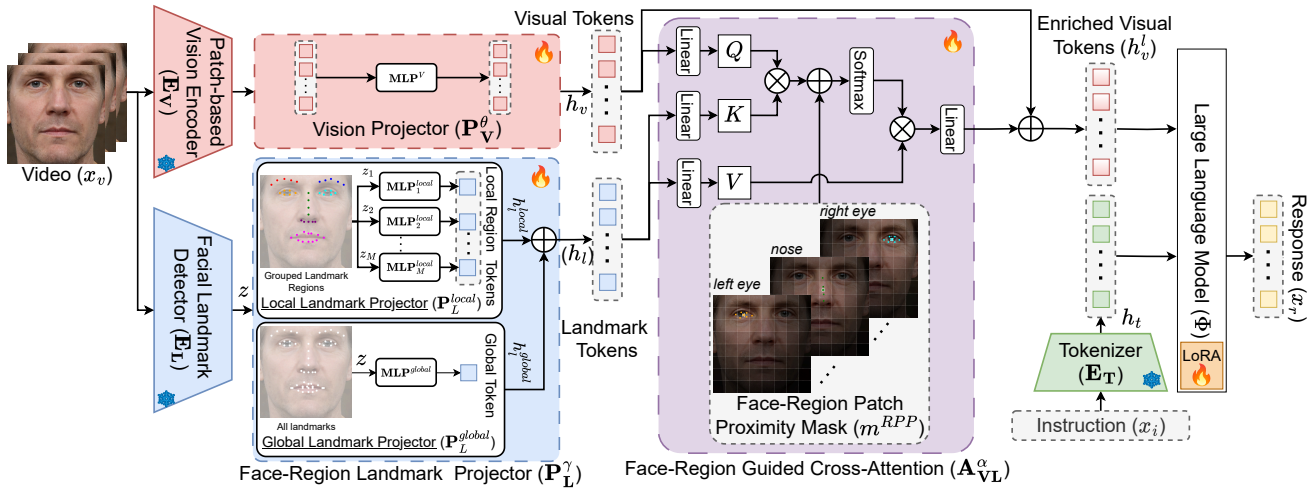


Figure 4. Proposed Face-LLaVA architecture. We group landmark points into different face regions and project them through a face-region landmark projector. The visual tokens are enriched by the landmark tokens through cross-attention and then passed as input to the LLM.

encoder $\mathbf{E_V}$ (256 for LanguageBind), and d is the dimensionality of hidden representation of the LLM.

Since we are dealing with faces, the visual information should also incorporate features specifically important for face processing which can be obtained from a *face-expert* model $\mathbf{E_L}$. In Face-LLaVA, we use a landmark detector as our *face-expert* model to extract face-specific features from the visual input x_v. We also experimented with a face-parsing model as the expert, but that led to extra computing due to pixel-wise information and no returns (see Sec. 5.5). Formally, we obtain the normalized 2-D landmark coordinates, $z = \mathbf{E_L}(x_v)$ where $z \in \mathbb{R}^{T\times L\times 2}$. These expert features are then fed to the proposed novel *Face-Region Landmark Projector* and *Face-Region Guided Cross-Attention* modules.

4.1. Face-Region Landmark Projector (FRLP)

To project the facial landmark features to the token space, one can simply flatten the L facial landmarks into the LLM's d-dimensional token space by projection [37]. However, this maps the entire face landmark features into a single token without preserving information about what individual landmark points mean. Moreover, for a lot of complex face-perception tasks, humans tend to look at a group of facial muscles or regions, such as eyes or lips, to make informed decisions. For example, to detect an expression of disgust, visual features around the nose to capture wrinkling are important. Therefore, we propose a *Face-region Landmark Projector* to group landmarks into different face regions and then project them into individual face-region tokens through separate MLPs. Specifically, we use two sub-modules to tokenize the landmark features.

The *local landmark projector* $\mathbf{P}_L^{local}$ generates M tokens each corresponding to a separate face-region,

$$h_l^{local} = \{\mathbf{MLP}_i^{local}(z_i) \;\; \forall \;\; i \in \{1 \dots M\}\} \tag{2}$$

where $z_i \in \mathbb{R}^{T\times L_i\times 2}$ are the L_i 2-D landmark points corresponding to the i^{th} face region. We use a total of 9 landmark groups corresponding to the following face regions - *face boundary, left/right eye, left/right brow, nose, nostril, lips and teeth*.

Similar to EmoLA [37], we also project all the L landmark points through a *global landmark projector* $\mathbf{P}_L^{global}$ to get a single global landmark token for each frame which can capture the overall facial structure and dynamics beyond individual regions.

$$h_l^{global} = \mathbf{MLP}^{global}(z) \tag{3}$$

and the final landmark tokens h_l are given by,

$$h_l = h_l^{global} + h_l^{local} \tag{4}$$

where the global token h_l^{global} is broadcasted in the token dimension. By combining the local region and the global landmark tokens, our approach ensures that both local and holistic facial features are effectively modeled. To keep the extra compute to a minimum, we use single-layer MLPs for all the landmark projections.

4.2. Face-Region Guided Cross-Attention(FRGCA)

We propose to use the landmark tokens h_l generated by the FRLP module via cross-attention with the visual tokens h_v. Such an architecture poses two benefits - (i) using cross-attention enables weighing visual tokens that are closer to salient face regions and are highly likely to be used by the downstream face processing tasks, and (ii) as opposed to EmoLA [37], the M landmark tokens per frame do not have to be appended to the visual tokens and passed as input to the LLM thus saving the LLM's context window. Formally, we obtain the key ($K \in \mathbb{R}^{M\times d_{attn}}$) and value ($V \in \mathbb{R}^{M\times d_{attn}}$) vectors using h_l, and the query

($Q \in \mathbb{R}^{N \times d_{attn}}$) vector using h_v by passing through separate linear layers. Note that we have dropped the time dimension T for simplicity.

To further enforce the attention weights on face regions, we compute a *face-region patch proximity* (RPP) mask, which is inversely proportional to the pairwise 2-D distances between the centroids of the visual patches used as inputs by $\mathbf{E_V}$ and the centroids of the different face regions as shown in Figure Fig. 4. Mathematically, the elements of the mask m^{RPP} are given by

$$m^{RPP}_{ji} = -||centroid(z_i) - centroid(h_{v,j})||_2 \quad (5)$$

where $centroid(z_i)$ is the centroid of i^{th} face region, $centroid(h_{v,j})$ is the centroid of the j^{th} visual patch associated with the visual token $h_{v,j}$, and $m^{RPP} \in \mathbb{R}^{N \times M}$.

The RPP mask is used to guide the attention weights before the softmax layer. The entire FRGCA module $\mathbf{A^{\alpha}_{VL}}$ can be summarized as,

$$h^l_v = Linear\left(Softmax\left(\frac{QK^T}{\sqrt{d_{attn}}} + m^{RPP}\right)V\right) + h_v \quad (6)$$

4.3. Training

We train the Face-LLaVA model in multiple stages so that the visual and landmark tokens can be aligned with the language tokens. We initialize the vision encoder ($\mathbf{E_V}$), vision projector ($\mathbf{P^{\theta}_V}$), tokenizer ($\mathbf{E_T}$) and the LLM (Φ) from the pretrained weights of Video-LLaVA [40]. For landmark detection, we use pretrained weights from FAN [5].

Face-Region Pretraining. In this stage, we only train the FRLP module $\mathbf{P^{\gamma}_L}$ and the FRGCA module $\mathbf{A^{\alpha}_{VL}}$ and keep all the other weights frozen. This ensures that the newly initialized modules generating landmarks tokens get aligned with the visual and the language tokens generated by $\mathbf{P^{\theta}_V}$ and $\mathbf{E_T}$ respectively.

Finetuning. In this stage, we train the vision projector $\mathbf{P^{\theta}_V}$ and the LLM model Φ along with $\mathbf{P^{\gamma}_L}$ and $\mathbf{A^{\alpha}_{VL}}$ with a lower learning rate. For the LLM, we use a LoRA [17] of rank 8 and scaling factor of 4. This stage jointly finetunes the model to improve the instruction following capabilities.

For both stages, the model is trained in an autoregressive fashion by maximizing the likelihood of the response x_r:

$$P(x_r|h^l_v, h_t) = \prod_{i=1}^{\mathbb{L}} P_{\Theta}(x^i_r|h^l_v, h_t, x^{[1:i-1]}_r) \quad (7)$$

where x^i_r is the i^{th} response token, $\mathbb{L}$ is the length of the response x_r, and $x^{[1:i-1]}_r$ are the response tokens generated before the i^{th} token. Learning rates for the pretraining and finetuning stages are 1e-4 and 2e-5 respectively, and all the models are trained for one epoch (see Appendix G).

5. Experiments

5.1. Datasets and Evaluation Protocol

Perception. We perform extensive evaluations of the proposed approach on nine benchmarks including DFEW [22], Crema-D [24], and RAF-DB [34] for facial expression recognition; DISFA [49] and BP4D [90] for action unit detection; FaceForensics++ [61] for deepfake detection; MORPH II [60] and UTK Face [94] for age estimation and CelebA [45] for facial attribute detection (Appendices D and F.3 for details). Since these datasets are also used to construct the *FaceInstruct-1M* dataset, we report performance on these datasets in the following two settings. In the *zero-shot* setting, we remove the entire task-specific dataset from *FaceInstruct-1M* for tuning our model, and in the *fine-tuned* setting we finetune the *zero-shot* model according to the official splits or training protocol of the benchmark (Appendix F.3).

Reasoning. As mentioned in Sec. 3.4, we perform reasoning evaluation using GPT4o-mini [72] on *FaceInstruct-1M test* set consisting of 500 samples per task. We do not use the existing EMER [39] benchmark for reasoning evaluation because we want to assess the reasoning based on facial expressions and not the background context or audio.

5.2. Evaluation on traditional tasks

We assess Face-LLaVA against multiple baselines under both *zero-shot* and *fine-tuned* settings, following the evaluation protocol detailed in Sec. 5.1. Unless otherwise, we apply the official inference code to face-cropped inputs for *zero-shot* baselines (Appendix G.2) and report the officially published results for *fine-tuned* baselines.

Facial Expression. As shown in Tab. 3, Face-LLaVA outperforms all baselines on all benchmarks in the zero-shot setting. It also performs competitively in the fine-tuned setting, despite lacking access to background or audio cues. It is also important to note that, in contrast to other MLLM baselines [8, 37], our model for finetuned setting is still trained to provide descriptions rather than outputting a single categorical label.

Action Units. Tab. 4 summarizes results on DISFA [49] and BP4D [90]. Since these datasets have disjoint AU annotations, we augment *FaceInstruct-1M* with Gemini [70] annotations on AffectNet [50] for zero-shot training. Face-LLaVA outperforms all baselines in both settings, achieving a 10% relative gain on DISFA over the previous fine-tuned SOTA.

Age, Attribute, and Deepfake. As shown in Tab. 5, Face-LLaVA surpasses all zero-shot baselines on Morph II [60], UTKFace [94], CelebA [45], and FaceForensics++ [61]. Despite being an MLLM, it matches regression-based baselines in age estimation. For attribute detection, due to the lack of large-scale supervised datasets, we annotate FFHQ

DFEW [22]			Crema-D [24]			RAF-DB [34]	
Method	UAR ↑	WAR ↑	Method	UAR ↑	WAR ↑	Method	Acc. ↑
Closed - source models							
GPT4o-mini [72]	0.426	0.518	GPT4o-mini [72]	0.410	0.486	GPT4o-mini [72]	0.758
Gemini-1.5F [70]	0.433	0.481	Gemini-1.5F [70]	0.465	0.635	Gemini-1.5F [70]	0.685
Gemini-2.5F [71]	0.483	0.555	Gemini-2.5F [71]	0.498	0.671	Gemini-2.5F [71]	0.748
Zero-shot							
Vid.LLaMA 3 [88]	0.286	0.305	Vid.LLaMA 3 [88]	0.397	0.546	Vid.LLaMA 3 [88]	0.671
Qwen 2.5 [73]	0.293	0.399	Qwen 2.5 [73]	0.395	0.566	Qwen 2.5 [73]	0.526
Vid.-LLaVA [40]	0.220	0.326	Vid.-LLaVA [40]	0.367	0.557	Vid.-LLaVA [40]	0.545
LLaVA-Vid. [93]	0.375	0.498	LLaVA-Vid. [93]	0.478	0.618	LLaVA-OV [29]	0.700
EmoLA* [37]	0.346	0.449	EmoLA* [37]	0.431	0.618	EmoLA [37]	0.741
Emotion LLaMA [8]	0.456	0.594	Emotion LLaMA [8]	0.225	0.308		
Face-LLaVA (Ours)	**0.467**	**0.601**	**Face-LLaVA (Ours)**	**0.612**	**0.705**	**Face-LLaVA (Ours)**	**0.834**
Fine-tuned							
EC-STFL [22]	0.454	0.565	Lei et al. [28]	0.645	0.648	RUL [91]	0.890
Former-DFER [96]	0.537	0.657	PTH-Net [33]	0.699	0.700	EAC [92]	0.909
GCA+IAL [30]	0.557	0.692	MAE-DFER [68]	0.773	0.774	TransFER [85]	0.909
M3DFEL [76]	0.561	0.693	MTL-ER [89]	0.745	0.756	Xue et al. [86]	0.920
MAE-DFER [68]	0.634	0.744	MT-Former [84]	0.793	0.807	POSTERv2 [47]	**0.922**
S2D [7]	0.618	0.760	MTCAE-DFER [81]	**0.847**	**0.850**	EmoLA [37]	0.921
EMO-LLaMA [83]	0.602	0.659					
Emotion-LLaMA [8]	0.642	**0.771**					
Face-LLaVA (Ours)	**0.645**	0.768	**Face-LLaVA (Ours)**	0.832	0.838	**Face-LLaVA (Ours)**	0.921

Table 3. Comparison of the proposed approach with recent MLLMs and supervised techniques for emotion recognition on the DFEW [22], Crema-D [24] and RAF-DB [34] dataset. *:Only using middle video frame.

Method	Average F1 ↑	
	DISFA [49]	BP4D [90]
Closed-source models		
GPT4o-mini [72]	0.429	0.496
Gemini-1.5F [70]	0.515	0.532
Gemini-2.5F [71]	0.545	0.530
Zero-shot		
VideoLLaMA 3 [88]	0.374	0.458
Qwen 2.5 VL [73]	0.431	0.467
Video-LLaVA [40]	0.442	0.445
LLaVA-OneVision [29]	0.280	0.439
EmoLA [37]	0.418	0.407
Face-LLaVA (Ours)	**0.556**	**0.517**
Fine-tuned		
ATCM [21]	0.615	0.642
ReCoT [36]	0.626	0.648
KS [35]	0.628	
ME-GraphAU [46]	0.631	0.655
$JÂA$-Net [63]	0.635	0.624
PIAP-DF [69]	0.638	0.641
VL-FAU [15]	0.665	**0.658**
AU-LLaVA [18]	0.525	0.603
EmoLA [37]	0.651	0.642
Face-LLaVA (Ours)	**0.729**	**0.658**

Table 4. Comparison of the proposed approach with MLLMs and supervised techniques on DISFA [49] and BP4D [90].

Age Estimation (MAE ↓)			Face Attribute (mAcc. ↑)		DeepFake Det. (Acc. ↑)	
Method	M [60]	U[94]	Method	CA [45]	Method	FF [61]
Closed-source models						
GPT4o-m [72]	4.09	5.04	GPT4o-m [72]	0.780	GPT4o-m [72]	0.807
Gem.-1.5F [70]	4.78	6.13	Gem.-1.5F [70]	0.814	Gem.-1.5F [70]	0.770
Gem.-2.5F [71]	5.29	5.88	Gem.-2.5F [71]	0.823	Gem.-2.5F [71]	0.785
Zero-shot						
V-LLaMA3 [88]	6.98	6.91	V-LLaMA3 [88]	0.813	V-LLaMA3 [88]	0.793
Qwen 2.5 [73]	6.09	5.25	Qwen 2.5 [73]	0.786	Qwen 2.5 [73]	0.653
V-LLaVA [40]	6.75	5.89	V-LLaVA [40]	0.795	V-LLaVA [40]	0.697
LLaVA-OV [29]	6.33	6.87	LLaVA-OV [29]	0.805	LLaVA-V [93]	0.751
Face-LLaVA	**3.22**	**4.75**	**Face-LLaVA**	**0.878**	**Face-LLaVA**	**0.843**
Fine-tuned						
PML [10]	2.15	-	Liu et al. [45]	0.873	MesoNet [1]	0.705
Berg et al. [4]	-	4.55	MOON [62]	0.909	Xception [9]	0.869
DLDL-v2 [14]	1.97	4.42	SwinFace [57]	0.913	MARLIN [6]	0.894
MWR [66]	2.00	4.37	Faceptor [58]	0.914	M2TR [77]	0.929
Faceptor [58]	**1.96**	4.10	DMM-CNN [48]	0.917	F3-Net [56]	**0.930**
Face-LLaVA	**1.96**	**4.06**	**Face-LLaVA**	**0.921**	**Face-LLaVA**	0.901

Table 5. Comparison of the proposed approach with recent MLLMs and supervised techniques on the age estimation, face attribute detection and deepfake detection tasks on different datasets [45, 60, 61, 94]. M: MORPH II [60], U: UTKFace [94], CA: CelebA [45], FF: FaceForensics++ [61].

[23] through Gemini without labels and use that to train Face-LLaVA for zero-shot setting. For deepfake detection, results on low-quality FaceForensics++ videos show Face-LLaVA significantly outperforms zero-shot MLLM baselines, which often hover around random (80%) accuracy. Notably, our model uses only eight frames per video, compared to more frames in fine-tuned baselines.

It is also important to note that even though Face-LLaVA is constructed using an older version of Gemini (1.5 Flash), we outperform Gemini-2.5 Flash on almost all the facial analysis benchmarks under the zero-shot setting.

5.3. Evaluation of reasoning

As mentioned in Sec. 3.4, we compare the reasoning capabilities of baselines with Face-LLaVA using GPT evaluation on the *FaceInstruct-1M test* set. Face-LLaVA achieves superior ratings compared to other MLLM baselines [29, 40, 73, 88]. Across all tasks, Face-LLaVA's mean rating for completeness of reasoning is around 33% higher than the best baseline, with a mean rating of 7.60/10. Moreover, Face-LLaVA responses achieve higher mean ratings than the baselines on response-video consistency and response-GT consistency as well highlighting excellent vision language alignment and high accuracy. Detailed results can be found in Tab. 9 in Appendix C. Fig. 5 presents a comparison of reasoning generated by different MLLMs for expression recognition in a video (Appendix H for more samples).

5.4. Visualizing landmark tokens

To visualize token representations learned by the FRLP module, we plot t-SNE [75] embeddings of the h_l tokens for a DISFA [49] subject in Fig. 6. DISFA is chosen as its action units map directly to facial regions. Frequently co-occurring AUs (e.g., AU25 and AU26) show overlapping representations, while mutually exclusive AUs (e.g., AU2 and AU4) appear well-separated.

5.5. Ablation Study

We perform an ablation study (Tab. 6) on DFEW [22] in a zero-shot setup using model trained on MAFW [43], FERV39k [43], and Crema-D [24]. Our baseline, Video-LLaVA [40], uses only visual tokens h_v. We can observe that using landmark tokens h_l (or h_l^{local}/h_l^{global}) as additional inputs with h_v improves baseline performance but not as much as incorporating the landmark tokens using cross-attention with h_v. Moreover, using FRGCA, which incorporates masked attention through RPP mask m^{RPP},

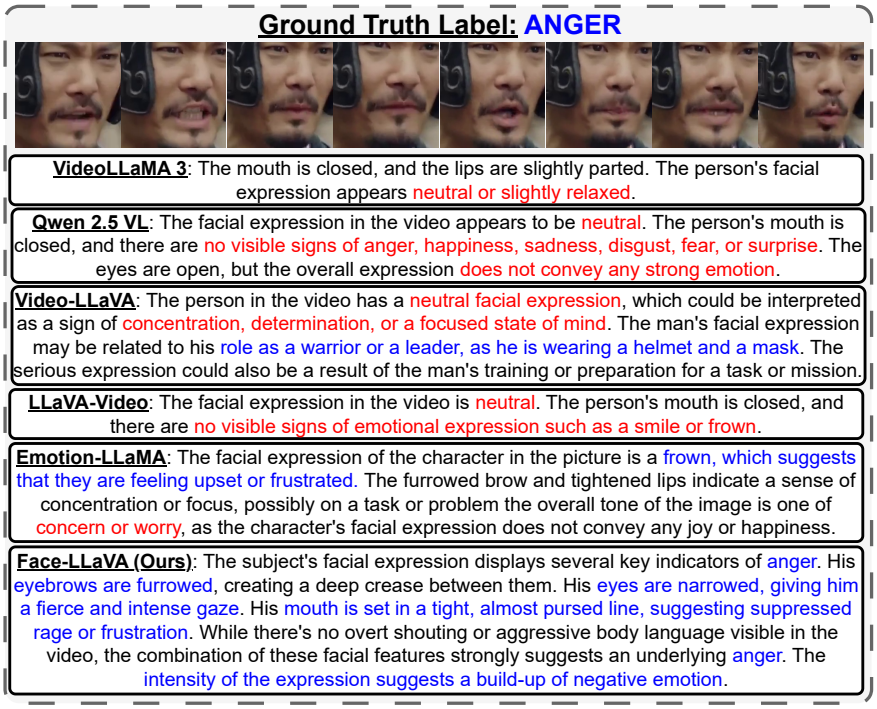


Figure 5. Comparison of Face-LLaVA with other MLLM baselines on the task of facial expression recognition. Red text indicates misaligned text and blue text indicates aligned/correct text according to the video and ground truth label.

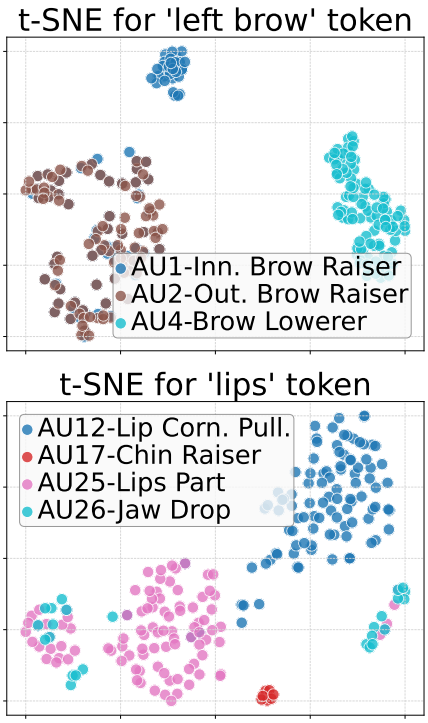


Figure 6. t-SNE plots of landmark region tokens h_l for *left brow* and *lips* token for subject SN032 of DISFA [49] dataset.

Model	Landmark Projector	Attention	Input tokens	DFEW [22] UAR	DFEW [22] WAR
Baseline	-	-	h_v	0.391	0.479
Baseline + Landmarks	only global	-	$h_v + h_l^{global}$	0.402	0.483
	only local	-	$h_v + h_l^{local}$	0.409	0.491
	FRLP	-	$h_v + h_l$	0.410	0.491
	only global	naive cross	$h_v^{l_{global}}$	0.401	0.483
	only local	naive cross	$h_v^{l_{local}}$	0.409	0.494
	FRLP	naive cross	h_v^{l}	0.416	0.512
	-	KP-RPE [26]	h_v^{KP-RPE}	0.422	0.509
	-	Q-former [87]	h_v^{Qform}	0.420	0.504
	only local	FRGCA	$h_v^{l_{local}}$	0.412	0.511
	FRLP	**FRGCA**	h_v^{l}	**0.424**	**0.520**
Baseline + Face Parsing	-	naive cross	h_v^{p}	0.413	0.498

Table 6. Ablations showing the effectiveness of the proposed FRLP and FRGCA modules compared to other design choices for zero-shot performance on DFEW [22].

leads to a larger performance gain than a simple cross-attention. Lastly, substituting landmarks with face-parse heatmaps (encoded via $\mathbf{E_V}$) yields competitive results but increases compute and memory due to their size.

We also compare our architecture with other face-guided attention architectures in Tab. 6. Specifically, we compare with KP-RPE [26] and Q-former based on ConsisID [87]. Since these architectures are not incorporated to MLLMs, we include the implementation details in Appendix G.2.1.

6. Ethics and Limitations

As our approach involves face analysis, we acknowledge ethical concerns related to privacy, bias, and potential misuse. *FaceInstruct-1M* is constructed from publicly available data with appropriate licensing, and we will release only annotations and instructions, ensuring compliance with data protection regulations like GDPR. However, we recognize that facial analysis technologies may impact privacy and encourage responsible use. Since our dataset is derived from existing sources, it may inherit biases, and addressing these remains for future work. Our research is intended for ethical academic use, and we strongly discourage any usecase that infringes on individual rights or promotes discrimination.

7. Conclusions

This work advances MLLMs for facial analysis across diverse tasks, including expression recognition, AU detection, attribute detection, age estimation, and deepfake detection. We introduce *FaceInstruct-1M*, a large-scale dataset with over 1M samples, automatically annotated using Gemini and GPT-4o, and propose Face-LLaVA, an MLLM leveraging novel FRLP and FRGCA modules to enrich visual representations for advanced face analysis. Extensive evaluations demonstrate superiority over open-source MLLMs and competitive performance against task-specific methods, with GPT-4 confirming its advanced reasoning capabilities.

Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 3, 7

[2] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. Age estimation via face images: a survey. *EURASIP J. Image Video Process.*, 2018(1), 2018. 1, 3

[3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *CoRR*, abs/1907.05047, 2019. 5

[4] Axel Berg, Magnus Oskarsson, and Mark O'Connor. Deep ordinal regression with label diversity. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2740–2747. IEEE, 2021. 7

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 6, 5

[6] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1504, 2023. 2, 3, 7

[7] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, pages 1–15, 2024. 2, 3, 7, 16

[8] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3, 4, 6, 7

[9] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 3, 7

[10] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10498–10507, 2021. 7

[11] Mohammad Ennab and Hamid Mcheick. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. *Diagnostics (Basel)*, 12 (7):1557, 2022. 2

[12] Niki M Foteinopoulou, Enjie Ghorbel, and Djamila Aouada. A hitchhiker's guide to fine-grained face forgery detection using common sense reasoning. In *Advances in Neural Information Processing Systems*, pages 2943–2976, 2024. 2, 3

[13] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 712–718. AAAI Press, 2018. 3

[14] Bin-Bin Gao, Xin-Xin Liu, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Learning expectation of label distribution for facial age and attractiveness estimation, 2021. 7

[15] Xuri Ge, Junchen Fu, Fuhai Chen, Shan An, Nicu Sebe, and Joemon M Jose. Towards end-to-end explainable facial action unit recognition via vision-language joint learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8189–8198, New York, NY, USA, 2024. ACM. 2, 3, 7

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 3

[17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[18] Guohong Hu, Xing Lan, Hanyu Jiang, Jiayi Lyu, and Jian Xue. Towards unified facial action unit recognition framework by large language models, 2024. 3, 7, 2

[19] Siyuan Hu, Zheng Wang, Peng Hu, Xi Peng, Jie Wu, Hongyuan Zhu, and Yew Soon Ong. Preface: Face-centric pretraining with self-structure aware distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 12538–12546, 2024. 3

[20] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019. 2

[21] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 7, 2, 3

[22] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2881–2889, 2020. 3, 4, 6, 7, 8, 5, 10, 11, 16, 22

[23] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial

Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228, 2021. 7

[24] Michael K Keutmann, Samantha L Moore, Adam Savitt, and Ruben C Gur. Generating an item pool for translational social cognition research: methodology and initial validation. *Behav. Res. Methods*, 47(1):228–234, 2015. 4, 6, 7, 5, 10, 16

[25] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3, 4, 5

[26] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–255, 2024. 8, 18

[27] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition, 2019. 3

[28] Yuanyuan Lei and Houwei Cao. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, 14(4):2954–2969, 2023. 7

[29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 7, 3

[30] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. *Proc. Conf. AAAI Artif. Intell.*, 37(1): 67–75, 2023. 7

[31] Jingzhi Li, Changjiang Luo, Ruoyu Chen, Hua Zhang, Wenqi Ren, Jianhou Gan, and Xiaochun Cao. Faceinsight: A multimodal large language model for face perception. *arXiv preprint arXiv:2504.15624*, 2025. 2

[32] Junzhe Li, Xuerui Qiu, Linrui Xu, Liya Guo, Delin Qu, Tingting Long, Chun Fan, and Ming Li. Unif2ace: Fine-grained face understanding and generation with unified multimodal models. *arXiv preprint arXiv:2503.08120*, 2025. 2

[33] Min Li, Xiaoqin Zhang, Tangfei Liao, Sheng Lin, and Guobao Xiao. Pth-net: Dynamic facial expression recognition without face detection and alignment. *IEEE Transactions on Image Processing*, 34:30–43, 2025. 7, 16, 17

[34] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 3, 4, 6, 7, 5, 17

[35] Xiaotian Li, Xiang Zhang, Taoyue Wang, and Lijun Yin. Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20922–20932, 2023. 7, 2

[36] Yifan Li, Hu Han, Shiguang Shan, zhilong ji, Jinfeng Bai, and Xilin Chen. Recot: Regularized co-training for facial action unit recognition with noisy labels. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 3, 7, 2

[37] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning. *European Conference on Computer Vision (ECCV) 2024*, 2024. 2, 3, 4, 5, 6, 7, 16, 17

[38] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W Schuller, and Jianhua Tao. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614, New York, NY, USA, 2023. ACM. 10

[39] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*, 2023. 2, 3, 4, 6

[40] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3, 4, 6, 7, 2, 17

[41] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 4

[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 4

[43] Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. *MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild*. ACM, New York, NY, USA, 2022. 3, 4, 7, 5, 10

[44] Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, and Zhe Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9717–9728, 2023. 3

[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 2, 3, 4, 6, 7, 5, 17, 20, 24

[46] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1239–1246, 2022. 3, 7, 2

[47] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster++: A simpler and stronger facial expression recognition network, 2023. 7

[48] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep Multi-Task Multi-Label CNN for Effective Facial Attribute Classification . *IEEE Transactions on Affective Computing*, 13(02):818–828, 2022. 2, 3, 7

[49] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3, 4, 6, 7, 8, 2, 5, 11, 17, 24

[50] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019. 3, 4, 6, 5

[51] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel. Facexformer: A unified transformer for facial analysis. *arXiv preprint arXiv:2403.12960*, 2024. 3

[52] Nostalgebraist. interpreting GPT: the logit lens — AI Alignment Forum — alignmentforum.org. https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. [Accessed 11-07-2025]. 2, 3

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 4

[54] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey, 2024. 2

[55] Tonya S Pixton. Happy to see me, aren't you, sally? signal detection analysis of emotion detection in briefly presented male and female faces. *Scand. J. Psychol.*, 52(4):361–368, 2011. 2

[56] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, pages 86–103, Cham, 2020. Springer International Publishing. 3, 7

[57] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2024. 3, 7

[58] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. In *Computer Vision – ECCV 2024*, pages 240–260, Cham, 2025. Springer Nature Switzerland. 2, 3, 7

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[60] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006. 2, 3, 4, 6, 7, 5, 11, 17

[61] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 4, 6, 7, 5, 11, 17, 21, 23

[62] Ethan M. Rudd, Manuel Günther, and Terrance E. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Computer Vision – ECCV 2016*, pages 19–35, Cham, 2016. Springer International Publishing. 3, 7

[63] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jâa-net: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321 – 340, 2020. 3, 7, 2

[64] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[65] Yan Shi, Zijun Zhang, Kaining Huang, Wudi Ma, and Shanshan Tu. Human-computer interaction based on face feature localization. *Journal of Visual Communication and Image Representation*, 70:102740, 2020. 2

[66] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18739–18748, 2022. 3, 7

[67] Haomiao Sun, Mingjie He, Tianheng Lian, Hu Han, and Shiguang Shan. Face-mllm: A large face perception model, 2024. 2, 3

[68] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Maedfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 6110–6121, 2023. 3, 7

[69] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12879–12888, 2021. 3, 7, 2

[70] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 4, 6, 7, 2, 3

[71] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 7, 2, 3

[72] OpenAI Team. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. [Accessed 25-02-2025]. 2, 4, 6, 7, 3, 17

[73] Qwen Team. Qwen2.5-vl, 2025. 3, 7, 2, 18

[74] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 3

[75] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 7

[76] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17958–17968, 2023. 3, 7, 16

[77] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2TR: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, New York, NY, USA, 2022. ACM. 7

[78] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20922–20931, 2022. 3, 4, 5, 10

[79] Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, and Zhongxue Gan. A survey on facial expression recognition of static and dynamic emotions, 2024. 1

[80] Chongke Wu, Sicong Shao, Cihan Tunc, Pratik Satam, and Salim Hariri. An explainable and efficient deep learning framework for video anomaly detection. *Cluster Comput.*, 25(4):2715–2737, 2022. 2

[81] Peihao Xiang, Kaida Wu, Chaohao Lin, and Ou Bai. Mtcaedfer: Multi-task cascaded autoencoder for dynamic facial expression recognition, 2024. 3, 7, 16, 17

[82] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[83] Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. Emo-llama: Enhancing facial emotion understanding with instruction tuning, 2024. 3, 7

[84] Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *Computer Vision – ECCV 2022*, pages 304–321, Cham, 2022. Springer Nature Switzerland. 7

[85] Fanglei Xue, Qiangchang Wang, and Guodong Guo. TransFER: Learning Relation-aware Facial Expression Representations with Transformers . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3581–3590, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 2, 3, 7, 17

[86] Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 14(4):3244–3256, 2023. 3, 7, 17

[87] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025. 8, 18

[88] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3, 7, 2, 18, 21

[89] Tenggan Zhang, Chuanhe Liu, Xiaolong Liu, Yuchen Liu, Liyu Meng, Lei Sun, Wenqiang Jiang, Fengyuan Zhang, Jinming Zhao, and Qin Jin. Multi-task learning framework for emotion recognition in-the-wild. In *Computer Vision – ECCV 2022 Workshops*, pages 143–156, Cham, 2023. Springer Nature Switzerland. 7

[90] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. Best of Automatic Face and Gesture Recognition 2013. 3, 4, 6, 7, 2, 5, 11, 17

[91] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. In *Advances in Neural Information Processing Systems*, pages 17616–17627. Curran Associates, Inc., 2021. 3, 7

[92] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision – ECCV 2022*, pages 418–434, Cham, 2022. Springer Nature Switzerland. 3, 7

[93] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 2, 3, 7, 18

[94] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2, 3, 4, 6, 7, 5, 17, 25

[95] Fufangchen Zhao, Ming Li, Linrui Xu, Wenhao Jiang, Jian Gao, and Danfeng Yan. Favchat: Unlocking fine-grained facial video understanding with multimodal large language models, 2025. 2, 3

[96] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021. 7

[97] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022. 3

[98] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022. 3

[99] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *Vis. Comput.*, 36(5):1067–1093, 2020. 1

[100] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4

[101] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 1