

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

Supplementary Material

Table of Contents

- **Ethics Statement** Appendix A
- **Limitations and Future Work** Appendix B
- **Detailed Results** Appendix C
 - GPT Evaluation Appendix C.1
 - Effect of FRGCA module Appendix C.2
- **Task-specific Datasets Used** Appendix D
- **Additional Details about FaceInstruct-1M** Appendix E
 - Data preprocessing Appendix E.1
 - Getting descriptions/reasoning from Gemini Appendix E.2
 - Task-specific instructions Appendix E.3
 - GPT-4 Filtering Appendix E.4
 - Ablation for data annotation pipeline Appendix E.5
 - Data statistics Appendix E.6
 - Data samples Appendix E.7
- **Evaluation** Appendix F
 - Synonyms matching Appendix F.1
 - String parsing Appendix F.2
 - Traditional metrics and evaluation protocol Appendix F.3
 - GPT-Evaluation Appendix F.4
- **Implementation Details** Appendix G
 - Face-LLaVA Appendix G.1
 - Baselines Appendix G.2
- **Reasoning Comparison with Baselines** Appendix H
- **Failure Cases** Appendix I

A. Ethics Statement

As the proposed approach involves face analysis, we acknowledge the ethical considerations associated with privacy, bias and potential misuse.

Privacy and data protection. *FaceInstruct-1M* is constructed using existing task-specific datasets with appropriate licensing for research use. We will only release the annotations and instructions for *FaceInstruct-1M* and not the video or images from the individual data sources, and the users of the dataset are recommended to obtain the videos from their original sources with appropriate consent. We do not collect or use private user data, ensuring compliance with data protection regulations such as GDPR. However, we recognize that facial recognition technologies may pose risks to individual privacy, and we encourage responsible use of our dataset and model.

Bias and fairness. Facial analysis models often exhibit biases due to imbalanced training data, leading to disparities in performance across different demographic groups. Since our dataset is based on existing datasets for face analysis, it inherits some biases that are already present in those datasets. Evaluation of model bias is however left as a future work.

Responsible use. Face analysis technologies have applications in various domains, including healthcare and accessibility. However, they also present risks if misused for mass surveillance, or profiling. Our research is intended for academic and ethical use, and we strongly discourage any application that infringes upon individuals’ rights, promotes discrimination, or compromises security.

We believe in transparency and open research. Upon acceptance, we will release our dataset and model to support further advancements in social AI development.

Method	AU1 ↑	AU2 ↑	AU4 ↑	AU6 ↑	AU9 ↑	AU12 ↑	AU25 ↑	AU26 ↑	Avg. F1 ↑
<i>Closed-source models</i>									
GPT-4o-mini [72]	0.292	0.302	0.565	0.416	0.493	0.244	0.582	0.536	0.429
Gemini-1.5F [70]	0.504	0.558	0.421	0.376	0.512	0.405	0.724	0.624	0.515
Gemini-2.5F [71]	0.491	0.521	0.494	0.530	0.517	0.610	0.647	0.617	0.545
<i>Zero-shot</i>									
VideoLLaMA 3 [88]	0.101	0.253	0.394	0.369	0.450	0.395	0.493	0.534	0.374
Qwen 2.5 VL [73]	0.204	0.291	0.664	0.571	0.468	0.301	0.478	0.475	0.431
Video-LLaVA [40]	0.167	0.368	0.459	0.518	0.508	0.532	0.501	0.483	0.442
LLaVA-OV [29]	0.044	0.143	0.201	0.146	0.469	0.182	0.520	0.535	0.280
EmoLA [37]	0.193	0.141	0.406	0.513	0.434	0.515	0.608	0.535	0.418
Face-LLaVA (Ours)	0.531	0.620	0.559	0.578	0.537	0.521	0.582	0.521	0.556
<i>Fine-tuned</i>									
ATCM [21]	0.461	0.486	0.728	0.567	0.500	0.721	0.908	0.554	0.615
ReCoT [36]	0.513	0.362	0.668	0.501	0.524	0.788	0.953	0.697	0.626
KS [35]	0.538	0.599	0.692	0.542	0.508	0.758	0.922	0.468	0.628
ME-GraphAU [46]	0.546	0.471	0.729	0.540	0.557	0.767	0.911	0.530	0.631
J \hat{A} A-Net [63]	0.624	0.607	0.671	0.411	0.451	0.735	0.909	0.674	0.635
PIAP-DF [69]	0.502	0.518	0.719	0.506	0.545	0.797	0.941	0.572	0.638
VL-FAU [15]	0.609	0.564	0.740	0.463	0.608	0.724	0.943	0.665	0.665
AU-LLaVA [18]	0.520	0.592	0.444	0.308	0.223	0.661	0.908	0.546	0.525
EmoLA [37]	0.505	0.569	0.835	0.552	0.431	0.801	0.916	0.600	0.651
Face-LLaVA (Ours)	0.636	0.623	0.790	0.733	0.710	0.832	0.902	0.606	0.729

Table 7. Comparison of the proposed approach with recent MLLMs and supervised techniques on the 8 AUs of the DISFA [49] dataset.

B. Limitations and Future Work

While this work demonstrates pioneering efforts in using MLLMs for general face analysis, there are some limitations that can be addressed by future works. First, Face-LLaVA is trained on single-turn conversations and hence lacks advanced abilities such as conversationing and chain-of-thought reasoning. Such training will require augmenting *FaceInstruct-1M* with conversation and reasoning data. Second, we only explored face-perception tasks and not face recognition or dense prediction tasks. While reasoning makes less sense in some of those tasks, there exists a potential to explore the performance of MLLMs on other facial tasks. Finally, since our dataset is automatically annotated using closed-source MLLMs, Gemini and GPT4, it contains some noise introduced by model hallucinations.

C. Detailed Results

This section contains detailed results from Sec. 5.

Action unit detection. Tabs. 7 and 8 expand Tab. 4 to show the F1 scores for individual action units. Notice that for the finetuning setting on DISFA [49], Face-LLaVA not only outperforms the baselines on average F1 score, but achieves the best F1 score on majority of the possible AUs in the dataset. A similar observation can be made for the analysis on the BP4D dataset [90] in Tab. 8.

C.1. GPT Evaluation

We report the mean GPT-4o-mini scores for all the tasks and scoring criterias in Tab. 9. For all five tasks, the reasoning capabilities of Face-LLaVA generated results are rated higher than the baselines. Moreover, the high consistency of reasoning with ground truth suggests that Face-LLaVA provides a description or reason that aligns with the correct ground truth label. Notice that for AU detection, all the general MLLMs perform poorly compared to Face-LLaVA and manual verification of the results revealed that these models have less knowledge about the Facial Action Unit Coding System (FACS) and hence their outputs include hallucinations. To compare the quality of responses generated by our model in comparison to the baselines, please refer to Appendix H.

C.2. Effect of FRGCA module

To assess the impact of the FRGCA module, we apply the *logit lens* [52] technique to intermediate visual representations h_v and h_v^l —before and after the FRGCA block. Specifically, we use the LM head (*unembedding layer*) to predict the output

Method	AU1 ↑	AU2 ↑	AU4 ↑	AU6 ↑	AU7 ↑	AU10 ↑	AU12 ↑	AU14 ↑	AU15 ↑	AU17 ↑	AU23 ↑	AU24 ↑	Avg. F1 ↑
<i>Closed-source models</i>													
GPT4o-mini [72]	0.458	0.449	0.630	0.618	0.331	0.393	0.660	0.525	0.516	0.440	0.465	0.462	0.496
Gemini-1.5F [70]	0.533	0.556	0.606	0.687	0.487	0.432	0.756	0.395	0.543	0.427	0.466	0.493	0.532
Gemini-2.5F [71]	0.511	0.511	0.602	0.619	0.481	0.492	0.673	0.405	0.471	0.534	0.503	0.559	0.530
<i>Zero-shot</i>													
VideoLLaMA 3 [88]	0.426	0.450	0.444	0.488	0.448	0.362	0.508	0.417	0.502	0.454	0.499	0.498	0.458
Qwen 2.5 VL [73]	0.260	0.422	0.549	0.702	0.328	0.503	0.579	0.441	0.455	0.446	0.459	0.461	0.467
VideoLLaVA [40]	0.477	0.495	0.461	0.396	0.296	0.470	0.357	0.491	0.489	0.459	0.470	0.483	0.445
LLaVA-OV [29]	0.342	0.427	0.373	0.473	0.392	0.335	0.454	0.482	0.462	0.505	0.516	0.507	0.439
EmoLA [37]	0.185	0.143	0.584	0.541	0.304	0.266	0.647	0.375	0.419	0.500	0.450	0.468	0.407
Face-LLaVA (Ours)	0.496	0.503	0.644	0.690	0.346	0.465	0.748	0.398	0.510	0.451	0.462	0.487	0.517
<i>Fine-tuned</i>													
JAA-Net [63]	0.538	0.478	0.582	0.785	0.758	0.827	0.882	0.637	0.433	0.618	0.456	0.499	0.624
PIAP-DF [69]	0.542	0.471	0.540	0.790	0.782	0.863	0.895	0.661	0.497	0.632	0.499	0.520	0.641
ATCM [21]	0.517	0.493	0.610	0.778	0.795	0.829	0.863	0.676	0.519	0.630	0.437	0.563	0.642
ReCoT [36]	0.515	0.478	0.589	0.792	0.802	0.849	0.884	0.616	0.533	0.646	0.518	0.554	0.648
ME-GraphAU (SWIN) [46]	0.527	0.443	0.609	0.799	0.801	0.853	0.892	0.694	0.554	0.644	0.498	0.551	0.655
AU-LLaVA [18]	0.582	0.459	0.619	0.786	0.756	0.878	0.905	0.590	0.324	0.625	0.305	0.403	0.603
EmoLA [37]	0.574	0.524	0.610	0.781	0.778	0.819	0.895	0.605	0.493	0.649	0.460	0.524	0.642
VL-FAU [15]	0.563	0.499	0.626	0.795	0.801	0.826	0.886	0.668	0.513	0.635	0.513	0.571	0.658
Face-LLaVA (Ours)	0.541	0.610	0.642	0.801	0.663	0.718	0.861	0.644	0.566	0.627	0.579	0.649	0.658

Table 8. Comparison of the proposed approach with recent MLLMs and supervised techniques on the 12 AUs of the BP4D [90] dataset.

Method	Reason-Video Consistency						Reason-GT Consistency						Reasoning Completeness					
	Emo.	AU	Attr.	Age	DF.	All	Emo.	AU	Attr.	Age	DF.	All	Emo.	AU	Attr.	Age	DF.	All
GT from <i>FaceInstruct-1M</i>	9.47	8.52	9.80	9.27	8.85	9.18	9.70	8.84	9.88	9.55	9.56	9.51	9.21	8.26	9.75	9.02	8.41	8.93
VideoLLaMA 3 [88]	5.14	2.58	6.90	5.82	7.02	5.49	5.27	2.06	6.27	5.13	7.64	5.27	4.90	2.73	6.50	5.37	6.51	5.20
Qwen 2.5 VL [73]	5.82	3.02	5.48	7.36	5.48	5.43	5.96	2.54	4.86	7.02	5.76	5.23	5.57	3.34	5.21	6.89	5.30	5.26
Video LLaVA [40]	4.31	2.58	5.82	7.79	6.19	5.34	4.47	2.06	5.28	7.10	6.58	5.10	4.20	2.73	5.30	7.00	5.84	5.01
LLaVA-OV [29]	7.11	2.18	6.08	7.97	6.69	6.01	7.30	1.95	5.48	7.44	7.33	5.9	6.67	2.34	5.72	7.52	6.19	5.69
EmoLA [37]	7.33	5.17	-	-	-	-	7.58	5.04	-	-	-	-	6.81	5.32	-	-	-	-
Emotion-LLaMA [8]	6.77	-	-	-	-	-	6.90	-	-	-	-	-	6.50	-	-	-	-	-
Face-LLaVA (Ours)	7.95	6.90	8.34	7.68	8.56	7.89	8.14	6.68	8.13	7.53	9.20	7.94	7.79	6.62	7.89	7.59	8.11	7.60

Table 9. Mean GPT4o ratings (on a scale of 1-10) for different methods in a zero-shot setting on the *FaceInstruct-1M* Test Set.

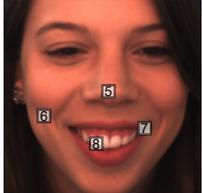
Sample Image	P #	h_v	h_v^l
	1	e, ye	(B)row, (e)ye Lower
	2	(N)ose, tip	nose, dis(gust), wr(inkle)
	3	-	lips, part
	4	(j)aw	(j)aw, dropped
	5	nose, (che)eks	nose, tip
	6	(che)eks, ros(y)	(che)eks, ros(y)
	7	lip	lip, corner
	8	lip, (e)ye	teeth, lip

Table 10. Logit lens [52] for visual tokens before (h_v) and after (h_v^l) FRGCA. Landmark-guided cross attention helps in correcting spatial token representations. Zoom-in to see patch numbers (P#).

language tokens based on h_v and h_v^l at different patch locations in an input image and report the face related tokens present in top-10 output tokens in Tab. 10. An ideal visual patch representation should have relevant language tokens associated with it according to its location in the image. The FRGCA module corrects incorrect spatial tokens (red in Tab. 10) and introduces refined ones (green in Tab. 10) via cross-attention with landmark tokens.

D. Task-specific datasets used

As mentioned in Sec. 3, *FaceInstruct-1M* is constructed using task-specific face analysis datasets for facial expression recognition, facial action unit detection, age estimation, facial attributes detection and deepfake detection. We present a summary of these datasets in this section. We refer the readers to Appendix F for a detailed description of the evaluation protocol on these datasets.

D.1. Facial Expression Recognition

Dynamic Facial Expression in-the-Wild (DFEW) [22] is a large-scale facial expression dataset comprising 16,372 video clips sourced from movies. Each clip is manually annotated by 12 expert annotators, with 10 independent labels per clip. The dataset includes seven basic emotion categories: happiness, sadness, neutral, anger, surprise, fear, and disgust. Although some clips have multiple emotion labels, we observed that the perceived emotion is often ambiguous in these cases. Therefore, we conduct all experiments on the single-labeled subset of 11.7k clips. As a multimodal dataset, DFEW contains audio and background context (e.g., multiple actors, body gestures), meaning the emotion labels may not be solely based on facial expressions.

MAFW [43] is another large-scale dynamic facial expression dataset containing approximately 10k movie clips. Each clip is annotated by 11 professional annotators for 11 emotion categories, including the seven basic emotions plus contempt, anxiety, helplessness, and disappointment. However, since these additional emotion categories have relatively few samples, we exclude them from our experiments. Like DFEW, MAFW is a multimodal dataset containing audio.

FERV39k [78] is a large-scale multi-scene dataset featuring 39k video samples, each categorized into one of seven basic emotions across 22 different scene types. The dataset is annotated through crowd-sourcing and professional annotators, with 30 independent annotations per clip. Unlike DFEW and MAFW, FERV39k does not contain audio, but it still includes background information such as multiple actors, body movements, and hand gestures, which can influence perceived emotions.

Crema-D [24] is an emotional multimodal dataset, containing 7,442 clips from 91 actors, including 48 male and 43 females. They are between 20 and 74, from a variety of races and ethnicities. The actors spoke 12 sentences, which were presented using one of 6 emotion categories, and four different emotion levels. The dataset is annotated by 2,443 participants, promising 95% of the clips have more than 7 ratings.

AffectNet [50] is a large-scale facial expression dataset containing more than 1M facial images. The dataset is collected by querying 1250 emotion related keywords in 6 different languages on the Internet with three major search engines. About half of the retrieved images were manually annotated with seven facial expression categories.

RAF-DB [34] is a large-scale facial expression dataset, containing 29,672 facial images with a variety of age, gender and ethnicity. It is annotated by 40 independent annotators. EM algorithm was applied to filter out unreliable labels.

D.2. Action Unit Detection

DISFA [49] is a non-posed facial expression dataset containing videos of 27 adults with different ethnicities, with high resolution. All video frames are annotated by two human FACS experts for the intensity of AUs (0-5 scale).

BP4D [90] is a 3D video database of facial expressions of 41 young adults between ages 18 to 29, The dataset is manually annotated with 12 action units, and contains automatically tracked head pose and 2D/3D facial landmarks.

D.3. Facial Attribute Detection

CelebA [45] is a large-scale face attributes dataset, containing more than 200K celebrity images, each with 40 attribute annotations. CelebA includes images with large diversity in pose and background.

D.4. Age Estimation

MORPH II [60] is a dataset containing 55134 mugshots, annotated with age estimate, gender, and race classification.

UTKFace [94] is a large-scale face dataset with people aged between 0 to 116 years old, containing over 20K images annotated with age, gender, and ethnicity.

D.5. Deepfake Detection

FaceForensics++ [61] is a dataset consisting of 1000 original video sequence, manipulated with 4 face manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The videos were generated from 977 YouTube source videos.

Fake AV-Celeb [25] is a dataset that includes about 20K manipulated videos generated using various deepfake synthesis methods. The base set consists of 500 real videos of celebrities from YouTube.

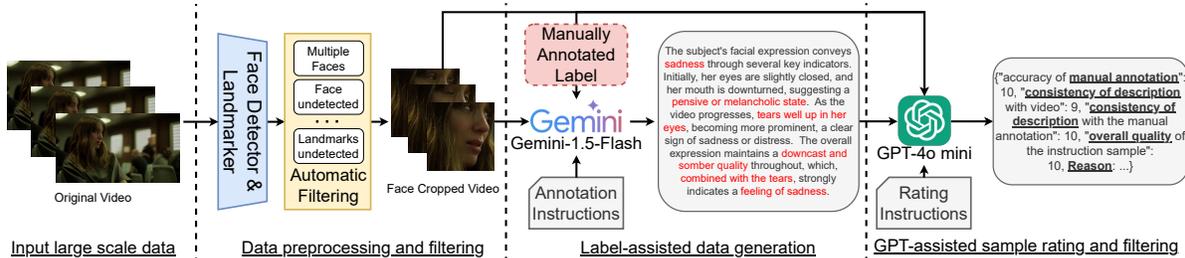


Figure 7. Data annotation pipeline used for creating *FaceInstruct-1M* dataset.

Dataset	Task	Number of Samples		
		Initial	After Preproc.	After GF.
DFEW [22]	Expression	11.7k	6.7k	6.2k
MAFW [43]	Expression	10k	6.9k	6.6k
FERV39k [78]	Expression	39k	30.7k	28.8k
Crema-D [24]	Expression	7.4k	7.4k	6.8k
AffectNet [50]	Expression	287k	280k	260k
RAF-DB [34]	Expression	15k	15k	14.8k
DISFA [49]	AU	131k	130k	123k
BP4D [90]	AU	150k	146k	128k
CelebA [45]	Attributes	203k	201k	196k
UTK Face [94]	Age	24.1k	23.5k	22.8k
MORPH II [60]	Age	50k	49.9k	49k
FaceForensics++ [61]	Deepfake	30k	25.9k	24.7k
Fake AV-Celeb [25]	Deepfake	20k	19.5k	19.3k
Real Faces* [22, 43, 78]	Deepfake	60.7k	44.3k	43.8k
Total	-	1.04M	987k	930k

Table 11. Statistics about data preprocessing and GPT-filtering (GF.) of for each of the constituent datasets of *FaceInstruct-1M*. *: Real faces augmentation for deepfake detection task is created from DFEW, MAFW and FERV39k.

E. Additional details about FaceInstruct-1M

Fig. 7 illustrates the annotation pipeline used to create our dataset from a large-scale, manually annotated dataset.

E.1. Data preprocessing

As mentioned in Sec. 3.1 and shown in Fig. 7, we pre-process the visual inputs, i.e. images and videos in each of the constituent datasets before getting the annotations from Gemini. We use Mediapipe [3] with default parameters to detect face bounding boxes within the video or image. For video datasets, we filter away all the videos in which the face of the subject goes out of the frame and cases where multiple faces are detected. We do not apply any smoothing to the detection bounding boxes for videos as we found that such smoothing fails in cases where the subject’s face has a sudden movement and hence, in such cases, the face moves out of the smoothed bounding box. We do not align the faces after cropping because for some tasks such as facial expression recognition, certain head movements such as nodding and head shakes might contain signals to predict and reason the correct output. After cropping the face bounding box from a video or an image, we resize the image into 256x256 for all downstream processing and training.

After face cropping, we used FAN [5] for detecting 68 2-D facial landmarks for all the samples in constituent datasets. We filter away all the samples in which landmark is not detected even in one of the frames of the video. Such, strict filtering ensures high data quality.

We report the statistics of our data after preprocessing in Tab. 11. Note that for FaceForensics++ [61] dataset, since the number of samples in the training set are quite small (only 4k), we split the dataset into chunks of 3 seconds and use those for all the processing.

E.2. Getting descriptions/reasoning from Gemini

We use Gemini-1.5 Flash [70] for getting all the data annotations for constituent datasets. Note that Gemini-2.0/2.5 were still under an experimental phase during the creation of *FaceInstruct-1M* so we used Gemini-1.5 Flash. Fig. 8 illustrates the prompts that we have used to get annotations from Gemini. Note that we do some prompt tuning for each task and provide negative prompts to control the output format, and to restrict the model from generating disclaimers. Moreover, we explicitly prompt the model to not mention that it has been provided with the ground truth information so that the generated descriptions can directly be used for prompt tuning.

It is also important to note that for the deepfake detection and age estimation tasks, we explicitly ask the model to start its response with the ground truth label so that it is easier to parse the responses, when computing traditional metrics for these datasets.

E.3. Task-specific instructions

To train our model on *FaceInstruct-1M* dataset, we need task-specific instructions for different tasks. To that extent, we carefully collected 100 handcrafted instructions for each of the five tasks. Some example instructions for different tasks are illustrated in Fig. 9. Notice that since the instructions for a particular task are semantically similar, we can use the instructions randomly during training similar to [8] during training. Note that we replace the ‘*video*’/‘*image*’ string in the instruction with appropriate type of data depending on the current sample.

E.4. GPT-4 Filtering

As mentioned in Sec. 3.2, we employ GPT4o-mini to rate the annotations obtained from Gemini to perform additional filtering. Moreover, this automated rating works as a sanity check to understand the quality of annotations generated from Gemini. Fig. 10 shows the rating instructions or prompts that we use for GPT4 assisted rating. We rate the annotations for all the prompts on the same four criteria - (i) accuracy of the manually annotated label w.r.t. the face-cropped video, (ii) consistency of the generated description with the face-cropped video, (iii) consistency of the generated description with the manually annotated label, and (iv) overall quality of the sample based on resolution, visibility of the face, etc. Moreover, we also ask the model to provide a short reason for its ratings to assist us in determining the reason for an unexpectedly high or low rating. Tab. 11 summarizes the number of videos that are filtered after GPT filtering. As mentioned in Sec. 3.2, we filter away all the samples with an overall rating less than or equal to 6, thereby resulting in about 7% of the initial dataset getting filtered away.

The given video has {cur_emotion} emotion. Your task is to reason why the emotion is tagged as "{cur_emotion}". Focus specifically on the facial expression of the subject in the video and describe how it varies.

Give your response in JSON format as following {"emotion":"angry","reason":"reason for your response"}. Strictly follow this format and add details in the reason field of the JSON.

In your provided reason, do not mention that I gave you any prior information about the video.

Facial Expression Recognition

This image is taken from a Facial Action Unit Detection Dataset. According to the dataset, {cur_AU_list} is/are activated in the given image. Please describe why you think {cur_AU_list} is/are activated.

Give your response in JSON format as following {"AU":["AU1", "AU2"],"reason":"reason for your response"}. The field "AU" within the json is a list containing the AUs that are activated on this image depending on your answer. Strictly follow this format and add details in the reason field of the JSON.

In your provided reason, do not mention that I gave you any prior information about the image.

Action Unit Detection

This image is marked to have the following attributes - {cur_attr_list}. Write a paragraph with a few sentences describing the facial attributes of this image. Make sure to use the exact same string attributes as given before in your description. DO NOT Modify the strings in the attributes for each attribute.

Give your response in JSON format as following {"reason":"string describing the facial attributes of the image and reason for any"}. Strictly follow this format and if an attribute is not obvious, then try to reason why that attribute is linked to the image.

In your provided reason, do not mention that I gave you any prior information about the image.

Facial Attribute Detection

The person in this image is {cur_age} years old. Reason why you think the person is {cur_age} years old based on the facial features, hair or anything else.

Give your response in JSON format as following {"age":"10","reason":"reason for your response"}. The field "reason" should start with the age of the person and then following the reason for your answer. Strictly follow this format and add details in the reason field of the JSON.

In your provided reason, do not mention that I gave you any prior information about the image.

Age Estimation

This video has been [manipulated by a deepfake technique](#). Describe the details of the face and video as well that does not seem real in this video as compared to a real video. Please DO NOT describe anything else in the video and avoid any disclaimers.

Give your response in JSON format as following {"fake":"yes","reason":"reason for your response"}. The field "fake" within the json can take values "yes" or "no" depending on whether the given video is fake or not. Start your reason with the word "Real" or "Fake" depending on the case. Strictly follow this format and add details in the reason field of the JSON.

In your provided reason, do not mention that I gave you any prior information about the video.

Deepfake Detection (for fake videos)

This video is [downloaded from youtube](#). Describe the details of face and video as well that make this a real and non-manipulated video. Please DO NOT describe anything else in the video and avoid any disclaimers.

Give your response in JSON format as following {"fake":"yes","reason":"reason for your response"}. The field "fake" within the json can take values "yes" or "no" depending on whether the given video is fake or not. Start your reason with the word "Real" or "Fake" depending on the case. Strictly follow this format and add details in the reason field of the JSON.

In your provided reason, do not mention that I gave you any prior information about the video.

Deepfake Detection (for real videos)

Figure 8. Prompts used for generating descriptions or reasoning from Gemini-1.5 Flash. Notice that we pass the label information about the data through [blue text](#).

- . Analyze the facial expression portrayed in the video.
- . Comment on the expression on the person's face in the video.
- . Give a description of the facial expression shown in the video footage.
- . What does the face in the video look like?
- . What expression is visible on the face within the video recording?
-

Facial Expression Recognition

- . Give the FACS (with explanation) for this image.
- . Specify the activated FACS in the image, with supporting evidence.
- . List the FACS shown in this image, providing your justification.
- . Using FACS, describe the facial expressions in this image and explain your assessment.
- . Analyze this image and indicate which FACS are present and why.
-

Action Unit Detection

- . Document the facial details visible in the image.
- . Describe the countenance presented in the image.
- . Examine the facial features depicted in the provided image.
- . Analyze the person's face in the picture for distinguishing characteristics.
- . Study the image to identify the facial attributes of the individual.
-

Facial Attribute Detection

- . Give your assessment of the person's age from the image, with supporting details.
- . Offer an approximation of the individual's age as shown in the picture and your justification.
- . How old do you think the person in the picture is, and what makes you think so?
- . Give your assessment of the person's age from the image, with supporting details.
- . Assess the age of the individual presented visually and provide your supporting evidence.
-

Age Estimation

- . Use logic to assess whether the video is genuine or fake.
- . Determine if the video is real or fabricated using critical thinking.
- . Reasonably evaluate the video's truthfulness.
- . Assess the video's authenticity using logical reasoning.
- . Consider whether the video is genuine or manipulated through reasoned judgment.
-

Deepfake Detection

Figure 9. Samples of various instructions used for different face analysis tasks. Note that for each task, the instructions are analogous which allows us to use them randomly with any sample belonging to that task during training, thereby augmenting the data size.

You are given 8 frames uniformly sampled from a video along with a description for the facial expression with reason for the video. You are also provided with the ground truth categorical emotional label associated with the video.

Your task is to rate the given video and description pair on a scale of 1(lowest) to 10(highest) based on the following metrics -

- (1) Accuracy of the given ground truth categorical label with respect to the video.
- (2) Consistency of the given description with respect to the video.
- (3) Consistency of the given description with respect to the ground truth label.
- (4) Overall quality of this video-description pair.

Your output should be in JSON format as following {"rating":{"1":"rating for task 1", "2"...}, "reason":"short reason for ratings if any"}.

"Categorical Label": "{cur_emotion}"
"Description": "{desc}"

Facial Expression Recognition

You are given an image along with a description for the facial action units activated with reason for the image. You are also provided with the ground truth action units (FACS-coded) associated with the image.

Your task is to rate the given image and description pair on a scale of 1(lowest) to 10(highest) based on the following metrics -

- (1) Accuracy of the given ground truth action unit labels with respect to the image.
- (2) Consistency of the given description with respect to the image.
- (3) Consistency of the given description with respect to the ground truth label.
- (4) Overall quality of this image-description pair.

Your output should be in JSON format as following {"rating":{"1":"rating for task 1", "2"...}, "reason":"short reason for ratings if any"}.

"Action Unit Label": "{cur_AU_list}"
"Description": "{desc}"

Action Unit Detection

You are given an image along with a description for the face attributes present in the image with reason associated with them. You are also provided with the ground truth face-attributes associated with the image.

Your task is to rate the given image and description pair on a scale of 1(lowest) to 10(highest) based on the following metrics -

- (1) Accuracy of the given ground truth face attribute labels with respect to the image.
- (2) Consistency of the given description with respect to the image.
- (3) Consistency of the given description with respect to the ground truth label.
- (4) Overall quality of this image-description pair.

Your output should be in JSON format as following {"rating":{"1":"rating for task 1", "2"...}, "reason":"short reason for ratings if any"}.

"Face Attributes Label": "{cur_attr_list}"
"Description": "{desc}"

Facial Attribute Detection

You are given an image along with a description for the estimated age of the person in the image with reason associated with them. You are also provided with the ground truth age associated with the person in the image.

Your task is to rate the given image and description pair on a scale of 1(lowest) to 10(highest) based on the following metrics -

- (1) Accuracy of the given ground truth age with respect to the image.
- (2) Consistency of the given description with respect to the image.
- (3) Consistency of the given description with respect to the ground truth age.
- (4) Overall quality of this image-description pair.

Your output should be in JSON format as following {"rating":{"1":"rating for task 1", "2"...}, "reason":"short reason for ratings if any"}.

"Ground Truth Age": "{cur_age}"
"Description": "{desc}"

Age Estimation

You are given 8 frames uniformly sampled from a video along with a description for the authenticity or forgery for the video. You are also provided with the ground truth associated with the video.

Your task is to rate the given video and description pair on a scale of 1(lowest) to 10(highest) based on the following metrics -

- (1) Accuracy of the given ground truth categorical label with respect to the video.
- (2) Consistency of the given description with respect to the video.
- (3) Consistency of the given description with respect to the ground truth label.
- (4) Overall quality of this video-description pair.

Your output should be in JSON format as following {"rating":{"1":"rating for task 1", "2"...}, "reason":"short reason for ratings if any"}.

"Categorical label": "{real/fake}"
"Description": "{desc}"

Deepfake Detection

Figure 10. Rating instructions (Prompts) used for data filtering using GPT4o-mini. Notice that we pass the label information about the data and the description generated from Gemini through [blue text](#).

Constituents	Labels used	GPT-filtering	No. of samples	DFEW [22]	
				UAR	WAR
M, F, C	✗	✗	44.7k	0.298	0.350
M, F, C, MER	✗	✗	89.7k	0.293	0.355
M, F, C	✗	✓	15.4k	0.318	0.375
M, F, C, MER	✗	✓	25.9k	0.327	0.392
M, F, C	✓	✗	44.7k	0.415	0.501
M, F, C	✓	✓	40.2k	0.424	0.520

Table 12. Ablations showing the effectiveness of using annotation labels for data generation and GPT filtering on zero-shot model performance on DFEW [22] dataset. M:MAFW [43], D: DFEW [22], F:FERV39k [78], C: Crema-D [24], MER: MER2023 [38].

E.5. Ablation for data annotation pipeline

To demonstrate the effectiveness of incorporating labels as additional signals and applying GPT-4o-mini filtering during dataset construction, we conduct ablation experiments on a subset of *FaceInstruct-1M* for zero-shot expression recognition on DFEW [22]. As baseline datasets for this study, we use MAFW [43], FERV39k [78], and Crema-D [24]. Additionally, to assess whether increasing the number of unlabeled samples improves performance, we include an unlabeled dataset from MER2023 [38], containing approximately 45k samples.

Tab. 12 summarizes our findings. The results indicate a significant performance improvement when leveraging ground truth labels from constituent datasets to generate annotations using Gemini. Unlike traditional self-supervised learning approaches, where increasing data volume typically enhances performance, we observe that adding more unlabeled data does not lead to better task performance – an important consideration for face analysis. Finally, filtering the dataset with GPT-4o-mini results in additional performance gain, as the data is labeled by one expert model (Gemini) and subsequently rated and filtered by another (GPT), improving overall data quality.

E.6. Data statistics

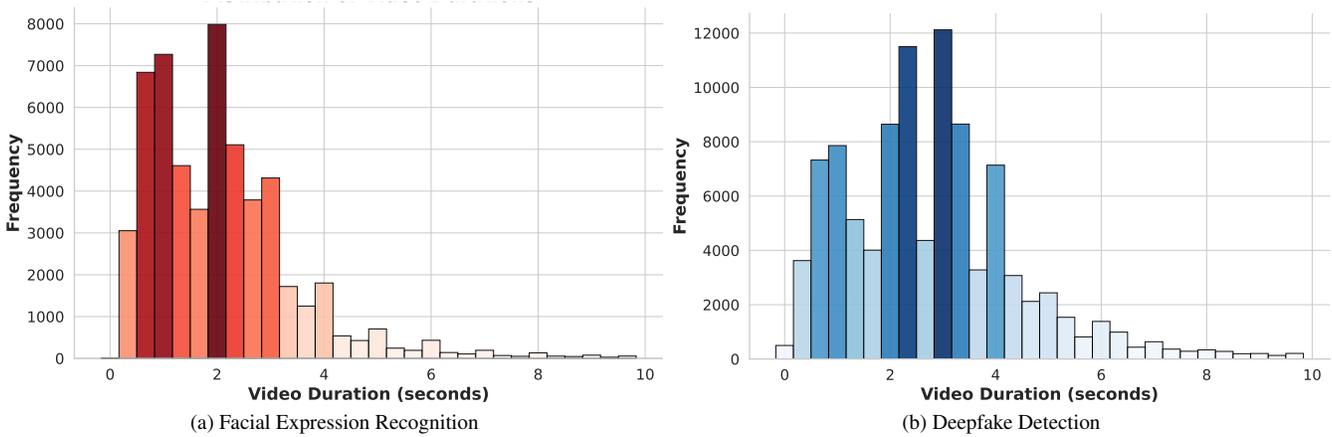


Figure 11. Distribution of video durations for the facial expression recognition and deepfake detection tasks. Note that except these tasks, other tasks are image/frame-based. For deepfake detection, since we cropped the longer videos of FaceForensics++ [61] into chunks of 3 seconds, so we can see a peak around 3 seconds.



Figure 12. Wordcloud for the descriptions or reasons present in *FaceInstruct-1M* belonging to different tasks.

Video durations. Fig. 11 shows the distribution of video durations for video-related tasks in *FaceInstruct-1M*. The total duration of videos for facial expression recognition is around 35 hours, with a mean of 2.28 seconds, and that of deepfake detection is around 84 hours, with a mean of 3.87 seconds. We can observe that the median duration is around 2 seconds for facial expression recognition while it is around 3-4 seconds for deepfake detection. It is also important to mention that since the FaceForensics++ [61] dataset has longer videos, we chunk the videos into smaller ones of around 3 seconds.

Class Distributions. Fig. 13 summarizes the class distribution for different tasks present in *FaceInstruct-1M*. For AU detection, the majority classes are those classes that are present in both BP4D [90] and DISFA [49] datasets. AUs such as AU5 and AU20, being only present in one of the datasets, are highly underrepresented in the dataset. Notice that for facial expression recognition, the classes *surprise*, *fear* and *disgust* are highly under-represented, thus explaining their poor recall on zero-shot performance on DFEW [22] dataset. For deepfake detection, the distribution of *real* and *fake* classes is slightly imbalanced with a bias towards *real* class. For age estimation, since MORPH II [60] only contains face images of people in the age of 16-76, hence the number of samples are more in the 20-40 age group in our dataset. It is important to note that the class imbalances present in *FaceInstruct-1M* is also present in the traditional datasets since *FaceInstruct-1M* is constructed using the images/videos and label information from those datasets.

Word Clouds. Fig. 12 shows the word clouds generated for the descriptions or reasons for the samples belonging to different tasks present in *FaceInstruct-1M*. Word cloud for facial expression recognition further highlights the bias in the dataset for *happiness* class with the word cloud for *happiness* and *smile* being big. Similarly, the word cloud for facial attribute detection correlates with the class distribution for facial attribute detection. It is important to note that all of the word clouds have bigger clouds for face or face regions (such as mouth, eyes, lip, nose, hair) important for reasoning on that task. For example, for deepfake detection, since the texture of skin plays an important role, so *skin* has a larger cloud.

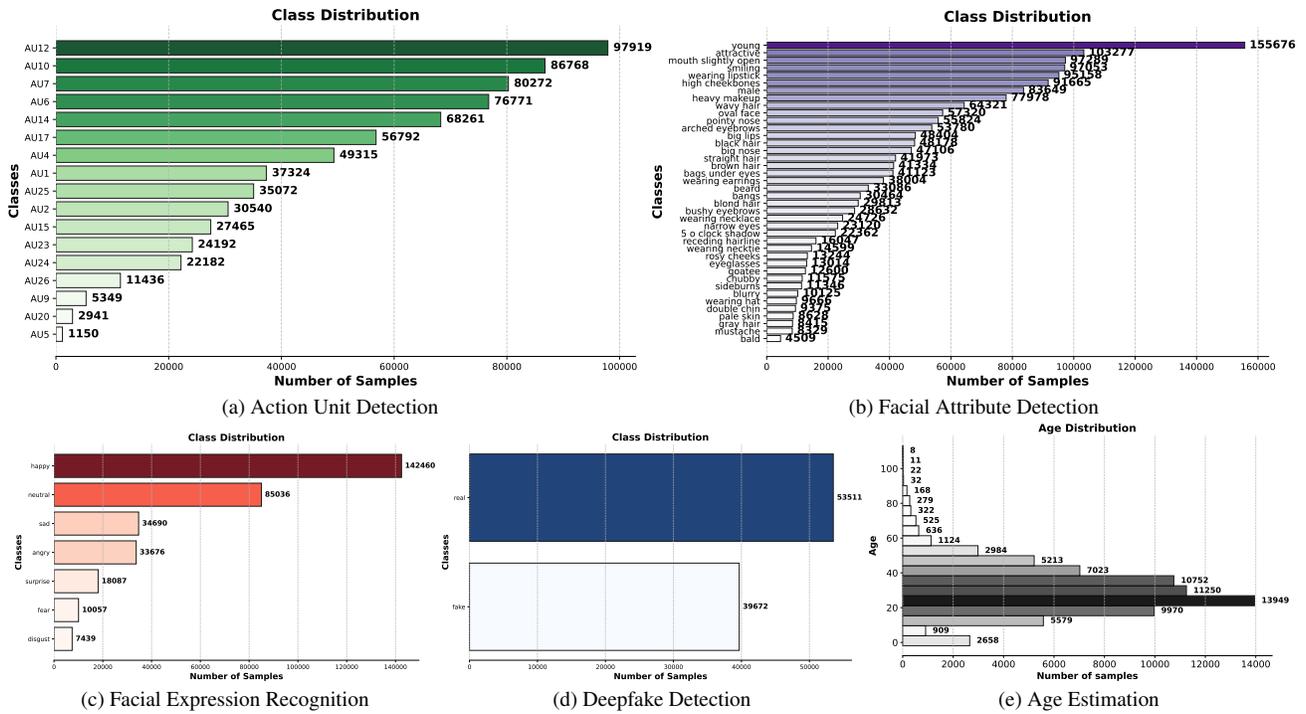


Figure 13. Class distribution for different face tasks present in *FaceInstruct-1M*.

E.7. Data samples

Figs. 14 to 18 show samples of reasoning or descriptions for different tasks present in *FaceInstruct-1M*. Notice in Fig. 14 that the annotations not only capture the overall emotion of the video, but they also contain information about how the facial expression and facial movements varied throughout the video to reason the overall emotion. For facial expression recognition, it is also important to note that in addition to the 7 emotion categories, the annotations also capture sub-emotions or feelings such as revulsion, distaste, etc.

For deepfake detection (Fig. 15), the annotations successfully capture the necessary imperfections or alterations that are visible in the video frames. Moreover, we can also observe that such descriptions incorporate references to different face regions which highlights the importance of different face regions for different face tasks.

For action unit detection (Fig. 16), the annotations capture the overall facial expression in addition to the action units or facial muscle movements responsible for the facial expressions. Thus, such annotations not only contain information about action units activated in the image, but also provide pseudo data for facial expression recognition. This shows how annotations for one task may help improving performance in a related task.

In Fig. 17, we can see that the annotations for facial attribute detection are pretty short and straightforward as they capture features or attributes that are clearly visible in the image without the need to "reason" a particular aspect of the image. Fig. 18 shows the annotations for age estimation task. For this task, notice that since we have prompted Gemini to start descriptions with the ground truth age (see Fig. 8), so all the descriptions start with the ground truth age. This is done to ease string parsing as mentioned later in Appendix F.2.



The baby's facial expression is the primary indicator of happiness. **Initially, a small, gentle smile is present**, with the corners of the mouth slightly turned upwards. The eyes are wide and bright, adding to the overall impression of contentment. As the video progresses, **the baby's mouth opens and closes, appearing as if babbling happily**, and the smile remains consistent though it varies in intensity, sometimes becoming a wider, more pronounced smile. There are no furrowed brows or other features associated with negative emotions. The **overall expression consistently conveys a sense of joy and positivity**.



The subject's facial expression **starts with a neutral look**, then **transitions to a slight grimace**, particularly noticeable around the mouth and nose. Her **lips are slightly pursed, and the corners of her mouth are turned down**. The eyebrows remain relatively unchanged, however, there is a **slight wrinkling around the nose which is a common indicator of disgust**. The overall expression conveys a **sense of revulsion or distaste**, leading to the classification of the **emotion as disgust**.



The subject's facial expression **initially shows a neutral or slightly apprehensive look**. Her eyes are wide but not excessively so, and her mouth is slightly open in a way that doesn't immediately suggest shock or fear. However, **as the video progresses, the widening of her eyes becomes more pronounced, indicating a sudden and unexpected realization**. This intensification of the **eye widening, coupled with the slight parting of her lips, is a classic indicator of surprise**. It's a shift from a **relatively calm expression to one clearly showing astonishment or unexpectedness**. The change in her expression is rapid and significant, highlighting the sudden nature of the surprise.

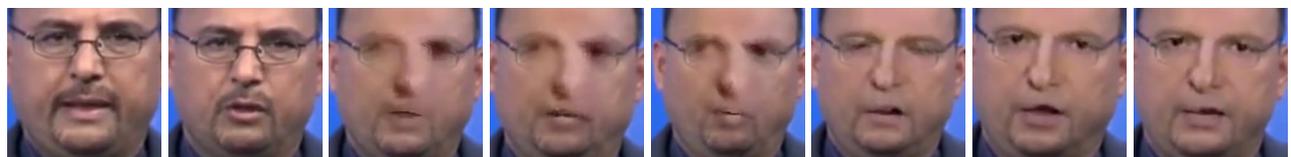
Figure 14. Examples of facial expression recognition task from *FaceInstruct-1M*.



Fake: The face in the video exhibits **unnatural transitions and inconsistencies** in lighting and shading, particularly **around the nose and mouth area**, which is typical of face-swapping techniques. There's a **lack of natural blending between the swapped face and the original video's background**, which makes the face appear pasted onto the video rather than naturally integrated. The **subtle micro-expressions and natural movements of the eyes and mouth seem inconsistent** with the rest of the face, also a common indicator of face-swapping. The overall texture and color of the skin don't fully match the background, adding to the artificial look.

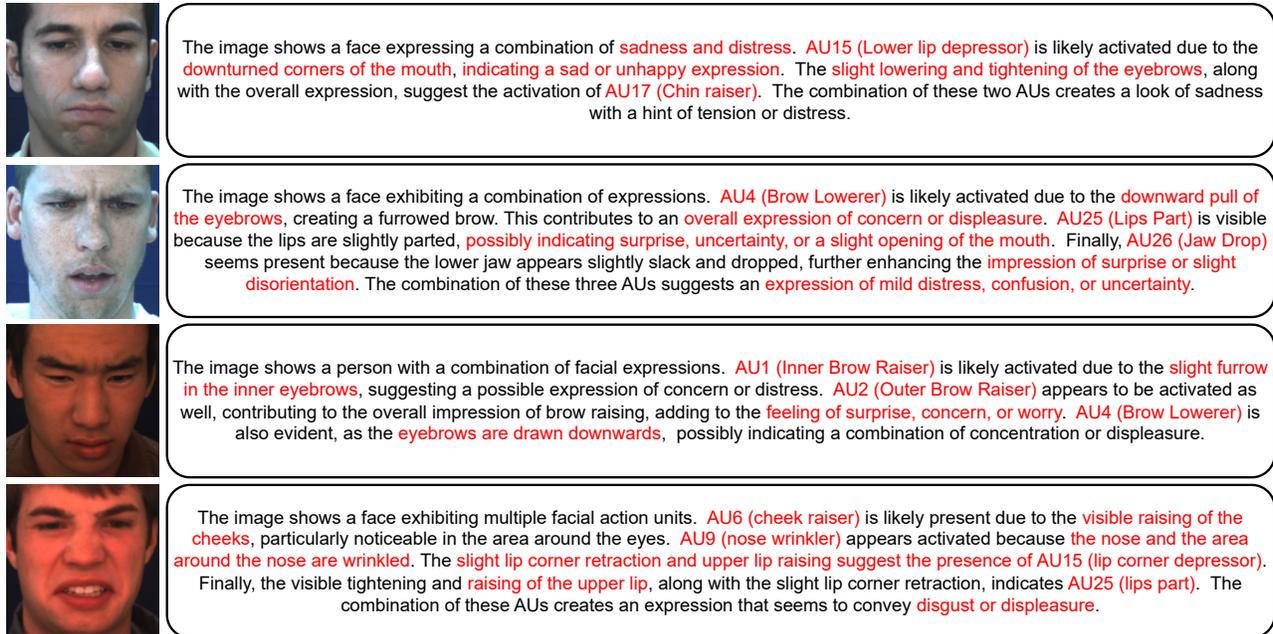


Fake video. **Real skin texture and movement are not consistent**. The face shows **unnatural blurring and smoothing, particularly around the eyes and mouth, indicating artificial manipulation**. The subtle movements of the face lack the **natural nuances of real facial expressions**; transitions between expressions **seem jerky and unrealistic**. There is a **lack of fine detail in the skin; for example, pores and blemishes are missing**, which would be expected in a real video.



Fake: The face in the video is clearly **manipulated**. **Real skin texture and movement are inconsistent** throughout the clip. The **face appears to shift and distort unnaturally, particularly around the eyes and nose, indicating a poor quality face swap**. The **lighting and shadowing on the face also don't match the background consistently**. These inconsistencies are hallmarks of a Faceswap technique gone wrong.

Figure 15. Examples of deepfake detection task from *FaceInstruct-1M*.



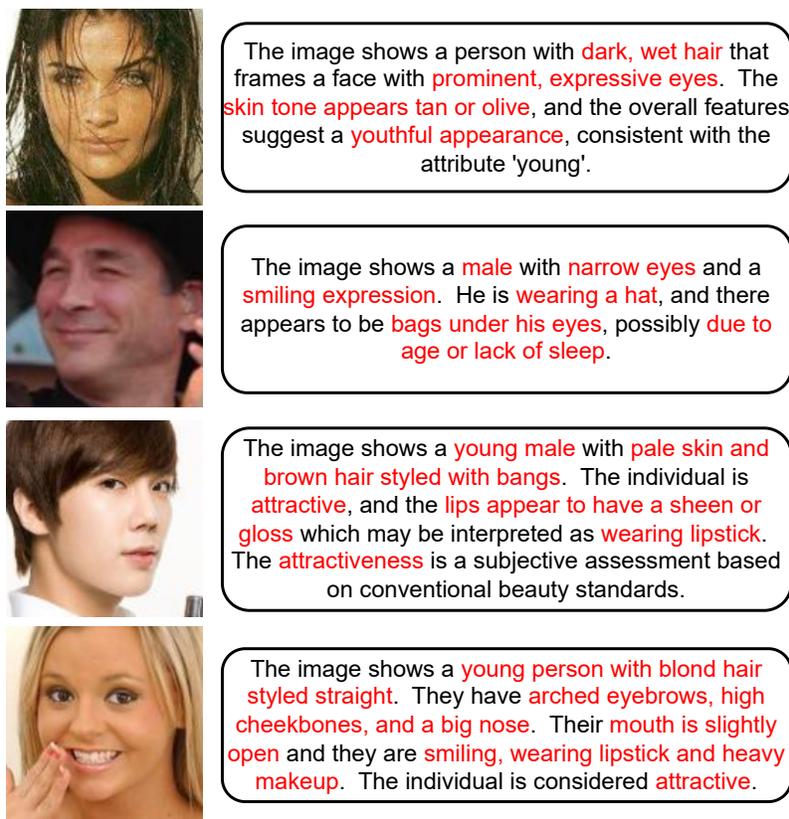
The image shows a face expressing a combination of **sadness and distress**. **AU15 (Lower lip depressor)** is likely activated due to the **downturned corners of the mouth**, indicating a sad or unhappy expression. The slight lowering and tightening of the eyebrows, along with the overall expression, suggest the activation of **AU17 (Chin raiser)**. The combination of these two AUs creates a look of sadness with a hint of tension or distress.

The image shows a face exhibiting a combination of expressions. **AU4 (Brow Lowerer)** is likely activated due to the **downward pull of the eyebrows**, creating a furrowed brow. This contributes to an overall expression of concern or displeasure. **AU25 (Lips Part)** is visible because the lips are slightly parted, possibly indicating surprise, uncertainty, or a slight opening of the mouth. Finally, **AU26 (Jaw Drop)** seems present because the lower jaw appears slightly slack and dropped, further enhancing the impression of surprise or slight disorientation. The combination of these three AUs suggests an expression of mild distress, confusion, or uncertainty.

The image shows a person with a combination of facial expressions. **AU1 (Inner Brow Raiser)** is likely activated due to the **slight furrow in the inner eyebrows**, suggesting a possible expression of concern or distress. **AU2 (Outer Brow Raiser)** appears to be activated as well, contributing to the overall impression of brow raising, adding to the feeling of surprise, concern, or worry. **AU4 (Brow Lowerer)** is also evident, as the **eyebrows are drawn downwards**, possibly indicating a combination of concentration or displeasure.

The image shows a face exhibiting multiple facial action units. **AU6 (cheek raiser)** is likely present due to the **visible raising of the cheeks**, particularly noticeable in the area around the eyes. **AU9 (nose wrinkler)** appears activated because the nose and the area around the nose are wrinkled. The slight lip corner retraction and upper lip raising suggest the presence of **AU15 (lip corner depressor)**. Finally, the visible tightening and raising of the upper lip, along with the slight lip corner retraction, indicates **AU25 (lips part)**. The combination of these AUs creates an expression that seems to convey **disgust or displeasure**.

Figure 16. Examples of action unit detection task from *FaceInstruct-IM*.



The image shows a person with **dark, wet hair** that frames a face with **prominent, expressive eyes**. The **skin tone appears tan or olive**, and the overall features suggest a **youthful appearance**, consistent with the attribute 'young'.

The image shows a **male with narrow eyes** and a **smiling expression**. He is **wearing a hat**, and there appears to be **bags under his eyes**, possibly due to age or lack of sleep.

The image shows a **young male with pale skin and brown hair styled with bangs**. The individual is **attractive**, and the **lips appear to have a sheen or gloss** which may be interpreted as **wearing lipstick**. The **attractiveness** is a subjective assessment based on conventional beauty standards.

The image shows a **young person with blond hair styled straight**. They have **arched eyebrows, high cheekbones, and a big nose**. Their mouth is slightly open and they are **smiling, wearing lipstick and heavy makeup**. The individual is considered **attractive**.

Figure 17. Examples of facial attribute detection task from *FaceInstruct-IM*.



22 years old is a reasonable estimation based on several facial features. The person exhibits a **relatively smooth complexion with minimal visible wrinkles or lines**, suggesting a younger age. The overall shape of the face and the **lack of significant bone structure changes** points towards someone in their early twenties. Additionally, the **person's hairstyle and makeup style** are consistent with those typically associated with young adults in this age range. It's important to note that this is an estimation based on visual cues and could be inaccurate.



85 years old is a reasonable estimate based on the **extensive wrinkling and age spots visible on the skin**. The **deep creases around the eyes, mouth, and forehead** suggest a long history of facial muscle movement and sun exposure. **The thinning of the hair** and the overall texture of the skin also align with the aging process typically seen in someone of this age. While it is difficult to assign a precise age from a photograph, the combination of these features strongly points towards an advanced age within the 80-85 year range.



45 years old is an estimation based on several observable facial features. The individual exhibits **moderate to significant skin texture and some visible lines around the eyes and mouth**, suggesting a person in their **mid-40s**. The **hairline appears to be slightly receding**, which is a common occurrence at this age. The overall facial structure and features show **signs of aging** consistent with a person of this age range. It's important to note that this is an estimation based on visual cues, and individual aging varies greatly.



1 year old: The baby's facial features are consistent with a child of this age. The baby has a **relatively smooth complexion with minimal wrinkles or lines**. The **cheeks are full and round**, which is typical for infants under the age of two. **The size and shape of the head** are also indicative of a baby around one year old. The **presence of only a few teeth** is also consistent with a child who is approximately one year old.

Figure 18. Examples of age estimation task from *FaceInstruct-IM*.

Expression Recognition					Attribute Detection					Deep-Fake Detection	
Happiness	Sadness	Neutral	Anger	...	Attractive	Chubby	Rosy Cheeks	Young	...	Real	Fake
cheerful	crying	calm	annoyed		appealing	plump	blushed cheeks	childish		authentic	fabricated
content	distress	expressionless	enraged		beautiful	puffy face	flushed cheeks	juvenile		genuine	forged
joy	melancholy	unemotional	incensed		good looking	soft cheeks	pinkish cheeks	teenager		legitimate	fraudulent
smiling	sob	unmoving	mad		handsome	round face	red cheeks	youthful		original	manipulated
...

Table 13. Synonyms used for categorizing descriptions into labels for different tasks. We have not shown the complete list of synonyms for all the classes to keep the table succinct.

F. Evaluation

F.1. Synonyms matching

As mentioned in Sec. 3.4, to evaluate the performance of the text generation MLLMs such as Face-LLaVA on traditional face analysis benchmarks, we need to convert the generated text into a prediction label. To that extent we follow synonym matching similar to [37] for facial expression recognition, facial attribute detection and deepfake detection, to categorize the given reason or description to one of the classes. We analyzed the top words from the annotated descriptions of different classes across all tasks and compiled a list of mutually exclusive synonyms for each class, as shown in Tab. 13. For each of the classes within the before-mentioned tasks, we come up with a list of at least 10 synonyms to match to based on the top words occurring in descriptions of our dataset.

Similar to FABAIInstruct [37], to map a given description to a class, we first remove all the negative sentences from the description. Then, we first match for synonyms on the first sentence of the description. If there are matches in the first sentence, then we output the majority voted class to be the dominant class in the given description. If the first sentence did not result in any synonym matches, then we perform majority voting on the entire description. The intuition behind matching the first sentence first is that the response from MLLMs starts with a conclusion or summary sentence and later sentences contain detailed description and reasoning over the first sentence.

While synonyms matching works in an expected way for facial expression recognition and facial attribute detection, manual verification of the descriptions revealed that for deepfake detection the model response might sometimes contain synonyms related to the wrong class more than the predicted class. Since we use majority rating, this would result in the response getting classified to the wrong class even if the response predicts the correct predicted class. To overcome this challenge, we explicitly prompt to start the description of the current sample with the ground truth label of the image (see Fig. 8). While this restricts the description quality in terms of variety in responses, it makes automatic string parsing to extract model prediction from the description easier. Moreover, the goal of this work is not to show that the model can generate varied responses, but rather to exhibit the reasoning capabilities of MLLMs for face-related tasks. Hence, training with such annotations makes Face-LLaVA generate responses that are easier to parse automatically.

F.2. String parsing

For the age estimation and deepfake detection tasks, we simply use regex parsing to convert the given description to categorical or numerical labels (after removing the negative sentences from the description). This works quite well for the AU detection task in fetching the FACS codes from the description, however, for age estimation, sometimes the description contains numerical values other than what the model wants to predict (see Fig. 23). Similar to the previous paragraph for deepfake detection, during data annotation through Gemini, we explicitly prompt to start the description of the current sample with the ground truth age of the person (see Fig. 8).

F.3. Traditional metrics and evaluation protocol

Facial Expression Recognition. For DFEW [22] and Crema-D [24] datasets, similar to previous works [7, 33, 76, 81], we report weighted average recall (WAR) and unweighted average recall (UAR), as the class distribution in these datasets are quite imbalanced. WAR captures the model’s ability to perform well on the majority classes of the dataset. It is calculated as the weighted sum of the recall scores for each class and the weights are determined by the number of samples belonging to a class. Unweighted average recall captures model’s ability to perform well on all the classes including the under-represented classes. It is computed as the average of the recall scores for each class. For DFEW [22], we used the official five-fold cross validation splits to report all the numbers and for Crema-D [24] we perform subject-exclusive 5-fold cross validation to report

```

{...Same instructions as Dataset Rating...}

(1) Consistency or overlap of the given description with respect to the video.
(2) Consistency or overlap of the given description with respect to the ground truth label.
(3) Overall completeness of the description to reason the ground truth label with respect to the video.

{...Same output instructions as Dataset Rating...}

```

Figure 19. Instructions for GPT4o-mini automatic evaluation of the reasoning capabilities of different models. Note that we only change the rating criteria from Fig. 10 to the ones mentioned above to obtain GPT-ratings for the given reasoning outputs for different models.

the results consistent with the baselines [33, 81]. For RAF-DB [34] dataset, similar to the previous baselines [37, 85, 86] we report just the overall accuracy on the official test set consisting of about 3k examples.

Action Unit Detection. In line with the previous works we perform subject exclusive cross validation for reporting our results on BP4D [90] and DISFA [49]. For both the datasets, we report the average F1 score over the set of possible AUs. For DISFA, the available AUs are AU1, AU2, AU4, AU6, AU9, AU12, AU25 and AU26, while for BP4D the AU list contains AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU23 and AU24.

Facial Attribute Detection. We report the mean accuracy over each of the 40 facial attributes in the test set of CelebA [45] dataset containing around 20k images.

Age Estimation. For both MORPH II [60] and UTKFace [94] datasets, we report the mean absolute error between the ground truth and the predicted integer ages on the official test splits for both the datasets. For MORPH II, the test set consists of around 5k images and for UTKFACE the test set consists around 4k images.

Deepfake Detection. We use the official test split of FaceForensics++ [61] dataset to report our numbers. We only use the low quality samples from the dataset for both constructing the *FaceInstruct-1M* dataset (and training) and testing. For all the zero-shot MLLM baselines, we chunk the test videos into a constant duration of three seconds and compute the prediction of a video as the majority voted prediction over its constituent chunks. Since for the MLLM baselines and Face-LLaVA, we can only get a text output for this task, so we only report the accuracy on the test set. Note that, the test set of FF++ is imbalanced and has 80% fake videos and only 20% real videos, hence making 80% accuracy as a baseline.

F.4. GPT-Evaluation

As mentioned in Sec. 5.1, we employ GPT4o-mini [72] for automatically evaluating the reasoning capabilities of baselines in comparison to Face-LLaVA. The instructions (prompts) used for this evaluation are similar to those described in Fig. 10 for dataset rating and filtering. We only change the criteria for evaluating the descriptions to those described in Fig. 19 for this evaluation.

G. Implementation details

G.1. Face-LLaVA

To implement our model, we start with the baseline architecture of Video-LLaVA [40] which has a Vicuna-7B backbone and add our novel FRLP and FRGCA modules (as described in Sec. 4) on top of it. To keep the extra computation introduced by the projection layers minimal, we use single layer MLPs within FRLP, as opposed to a 2-layer MLP with GeLU used in the vision projector [40]. Moreover, we only use a single block of cross-attention inside FRGCA with 8 attention heads. We do not apply layer normalization to the output of FRGCA.

For training, we use deepspeed¹ to parallelize training on a single NVIDIA DGX node with 8*H100 GPUs. We train the model using an AdamW optimizer only for one epoch for both the stages. Face-Region Pretraining takes about 5 hours and Finetuning takes around 12 hours on the entire dataset for the abovementioned environment. Learning rates for the pretraining and finetuning stages are kept as 1e-4 and 2e-5 respectively with cosine learning rate schedule.

G.2. Baselines

We use the official inference code of the baselines. For models which can handle both video and images, we use the corresponding inference code for video and image related tasks. To ensure a fair comparison, for models which have an option to

¹<https://www.deepspeed.ai/>

set the fps [73, 88, 93] for extracting frames for video inputs, we set the fps in such a way that we extract either 8 frames or frames at 1 fps, whichever leads to higher number of frames.

Since we do not finetune the baseline MLLMs and simply use their pretrained weights for inference, we engineer our prompts for each task and model pair to ensure optimum results. For fixed-set prediction tasks such as facial expression recognition, facial attribute detection and deepfake detection, we provide the list of possible classes in the prompt to supervise and restrict the model responses. Moreover, for all the tasks we use a separate prompt for reasoning evaluation and evaluation using traditional metrics. This is done because the responses generated by MLLMs are sometimes hard to parse with the techniques mentioned in Appendices F.1 and F.2. So to report the performance of baselines for Sec. 5.2 we explicitly prompt the baselines to generate its response as a single word or a list of words.

G.2.1. Cross attention baselines

For KP-RPE [26], we use the centroids of the 9 different facial regions as the keypoints to compute relative position embeddings. A single block of self attention (similar to FRGCA) was used to incorporate KP-RPE. We use different relative positional embeddings for different heads of self-attention.

For ConsisID [87], we incorporate their *local facial extractor module* which uses a Q-former architecture for extracting facial features from the visual tokens. We set the number of queries to 256 per frame/image to match the number of tokens input to the LLM as in our proposed approach. We use a lighter landmark detection module instead of face-parsing module used in ConsisID, comparable to our approach and only the last layer features are used.

H. Reasoning comparison with baselines

In this section, we illustrate the reasoning capabilities of Face-LLaVA in comparison to the baseline MLLMs. For each task, we pick diverse samples from the *FaceInstruct-1M* test set to show the comparison. Figs. 20 to 24 show this comparison in detail. Readers are encouraged to refer to the caption of these figures for a detailed description about the samples.

I. Failure cases

This section contains the failure cases for Face-LLaVA for different tasks. Figs. 25 to 29 show examples for each of the tasks where Face-LLaVA gives slightly inconsistent responses. In most of the cases, the inaccuracy of prediction with the ground truth label is due to a corner case (multiple expressions present in the face, no action unit activated, etc.) or when the prediction is difficult to make even for humans (deepfake detection and age estimation). In almost all the cases the model output and descriptions align with the video or the image thereby showing excellent vision language grounding for facial tasks.

Ground Truth Label: FEAR



Video-LLaMA 3: The person in the video is **making a funny face** with their **mouth open and eyes wide**, showing a **playful and expressive** demeanor.

Qwen 2.5 VL: The facial expression in the video appears to be **Disgust**. The person's **eyes are narrowed**, and their **mouth is open** in a way that suggests they are **expressing disgust or aversion** to something.

Video-LLaVA: The person in the video has a **smiling facial expression**, which indicates that they are **happy or content**. This could be due to various reasons, such as receiving good news, enjoying a pleasant conversation, or simply having a positive outlook on life. The smile is a universal symbol of happiness and can convey a sense of warmth and friendliness.

LLaVA-Video: The facial expression in the video is **Disgust**. The person's **mouth is slightly open**, and their **eyebrows are raised**, which are common indicators of disgust.

Emotion-LLaMA: The facial expression of the character in the picture is a **neutral one**, with a **slightly raised eyebrow and an open mouth**, which suggests that they are not particularly emotional or expressive the character's body language, including their posture and hand positioning, is also neutral, **with their arms at their sides** and no visible tension or movement this combination of facial and body language cues implies that the **character is likely feeling a sense of calmness or indifference**, rather than any strong emotions

Face-LLaVA (Ours): The subject's facial expression **strongly indicates fear**. **Initially**, their **eyes are wide** and slightly narrowed, suggesting a sense of apprehension or alarm. The mouth is slightly open, possibly in a gasp or a silent scream. **As the video progresses**, the **expression intensifies**; the eyes remain wide, and the mouth continues to be slightly open, possibly indicating a **heightened state of fear or panic**. The overall expression is one of **distress and anxiety**, consistent with the emotion of fear.

Ground Truth Label: SADNESS



Video-LLaMA 3: The person in the video has a **sad face**, which is evident from their facial expression. The reason for this sad expression is not explicitly stated in the video, but it could be inferred that the person is experiencing **some form of emotional distress or sadness**.

Qwen 2.5 VL: The facial expression in the video appears to be **neutral**. The person's **mouth is closed**, and there are no visible signs of anger, happiness, sadness, disgust, fear, or surprise. The **eyes are slightly narrowed**, which could **indicate a neutral or contemplative state**.

Video-LLaVA: The person in the video has a **frowning facial expression**, which indicates that they are **unhappy or dissatisfied**. This could be due to various reasons, such as a recent event, a personal issue, or simply a momentary mood. The frown on their face suggests that they are not in a **happy or positive state at the moment**.

LLaVA-Video: The facial expression in the video is **neutral**. The person's **eyes are closed**, and there is **no visible smile or frown**.

Emotion-LLaMA: The man in the picture has a **sad expression** on his face, which suggests that **he is feeling emotional or upset** his eyebrows are furrowed, and his eyes appear to be **downcast**, indicating a **sense of melancholy** the overall tone of the image is somber, which further supports the idea that the man is experiencing negative emotions

Face-LLaVA (Ours): The subject's facial expression conveys **sadness** through several cues. **Initially**, his eyebrows are slightly furrowed, and his gaze is downcast, suggesting a **pensive or melancholic state**. His mouth is slightly downturned, further reinforcing the impression of sadness. **As the video progresses**, his eyes appear to well up, **with slight moistness visible**, which is a clear indicator of **impending tears and emotional distress**. The overall expression remains **subdued and somber**, lacking any signs of anger, joy, or surprise, solidifying the interpretation of sadness.

Ground Truth Label: ANGER



Video-LLaMA 3: The person's mouth is slightly open, revealing their teeth, and their eyes are **focused intently on something off-camera**. The expression conveys a **sense of concern or contemplation**, suggesting they might be engaged in a **serious conversation** or deep in thought.

Qwen 2.5 VL: The facial expression in the video appears to be **neutral**. The person's **eyes are open and focused**, but there is no clear indication of anger, happiness, sadness, disgust, fear, or surprise. The **mouth is closed**, and the overall expression does not convey any strong emotion.

Video-LLaVA: The person in the video has a **serious facial expression**, with a **furrowed brow and a look of concentration**. This expression suggests that they are focused on something, possibly a task or a conversation. The person's facial expression is a reflection of their emotional state and the intensity of the situation they are in.

LLaVA-Video: The facial expression in the video is **neutral**. The **person's mouth is closed**, and there are no visible signs of emotional expression such as a smile or frown.

Emotion-LLaMA: In the image, a man with a beard and mustache is **looking directly at the camera**, his **eyes are narrowed**, and his **brows are furrowed**. This facial expression suggests that he is experiencing strong emotions, likely **anger or frustration** the combination of his **intense gaze and furrowed brows** indicates that he is focused on something specific, possibly a problem or issue that is causing him distress. **The context of the scene, which includes a man standing alone in a room, suggests that he may be feeling isolated or disconnected from others, further contributing to his emotional state**

Face-LLaVA (Ours): The subject's facial expression strongly suggests **anger**. His **eyebrows are furrowed**, creating **deep vertical lines between them**. His **eyes are wide and intense**, conveying a **sense of heightened arousal**. His **mouth is slightly open in a grimace**, and his **jaw is tense**, further emphasizing the **tension and frustration**. The overall expression is one of **intense displeasure and potential aggression**, all characteristic of **anger**.

Ground Truth Label: SURPRISE



Video-LLaMA 3: The **mouth is slightly open**, suggesting a **relaxed or neutral expression**. The lips are not tightly closed, indicating a **state of ease or mild engagement**.

Qwen 2.5 VL: The facial expression in the video appears to be **neutral**. The person's mouth is closed, and there are no visible signs of anger, happiness, sadness, disgust, fear, or surprise. **The eyes are looking slightly upwards**, which could **indicate a neutral or contemplative state**.

Video-LLaVA: The person in the video has a **neutral facial expression**, with their **eyes closed and their mouth open**. This expression could indicate that they are either **deep in thought or simply taking a moment to relax and breathe**. The person's facial expression is **not overly expressive**, but it does convey a sense of calmness and introspection.

LLaVA-Video: The facial expression in the video is **neutral**. The **person's mouth is closed**, and there are no visible signs of emotional expression such as a smile or frown.

Emotion-LLaMA: In the image, a young woman is looking directly at the camera with a **neutral expression**, her **eyes are focused on the lens**, and her **mouth is slightly open** the overall facial expression is one of concentration, suggesting that she is engaged in a task or activity the **lack of emotional cues and the focus on the camera's lens suggest that she is not expressing any strong emotions**, which could be interpreted as a sign of indifference or detachment from her surroundings the fact that she is not smiling or showing any other emotional expressions further supports this interpretation, as her facial expression appears to be neutral and unemotional

Face-LLaVA (Ours): The subject's **facial expression changes dramatically, indicating surprise**. **Initially**, her eyes are wide open, and her mouth is slightly parted, a classic sign of surprise. Her **eyebrows are slightly raised**, further emphasizing the **astonishment**. The **change in her expression is sudden and unexpected**, reinforcing the feeling of surprise. The overall effect is a clear visual representation of a **surprised reaction**.

Figure 20. Comparison of descriptions or reasoning obtained by Face-LLaVA with other baselines for facial expression recognition. Blue text indicates alignment with the ground truth and red text indicates wrong reasoning or hallucinations.



Ground Truth AUs:
AU4 (brow lowerer),
AU15 (lip corner depressor)

VideoLLaMA 3: The facial expression in the image suggests the activation of several action units (AUs), which are specific muscle movements that contribute to facial expressions. Based on the description, the following AUs appear to be activated:
- AU11: This is typically associated with the formation of a frown, as it involves the contraction of the muscles between the eyebrows.
- AU23: This is the brow wrinkle, often seen when someone is concentrating or experiencing negative emotions like anger or disgust.
These AUs are commonly used in the Facial Action Coding System (FACS) to describe and analyze facial expressions.

Qwen 2.5 VL: The facial action units (AUs) activated in the given image are:
- AU 12: Brow raiser
- AU 15: Brow lowerer
- AU 20: Upper lip raiser
- AU 45: Lower lip depressor
These AUs contribute to the expression of a raised eyebrow, a furrowed brow, and a slightly lowered upper lip, which together can indicate a state of concern or confusion.

Video-LLaVA: In the image, the facial action units (AU) activated are AU11 (eyebrow raise) and AU53 (lip pucker).

LLaVA OneVision: The facial action units (AUs) activated in the given image are:
- AU 12: Brow Lowering
- AU 4: Cheek Raising
- AU 5: Cheek Depression
- AU 6: Lip Squeezing
- AU 10: Lip Puckering

Face-LLaVA (Ours): The image shows a face expressing a combination of emotions, likely involving displeasure or distress. Let's analyze the AUs individually:
- AU4 (Brow Lowering): The eyebrows are drawn together and lowered, which is a characteristic sign of AU4. This contributes to the overall expression of displeasure or concern.
- AU15 (Lip Corner Depressor): The corners of the mouth are pulled downwards, creating a downturned expression. This is a clear indication of AU15, adding to the feeling of sadness or unhappiness.
- AU17 (Chin Raiser): The chin appears slightly raised, which is consistent with AU17. This subtle movement often accompanies expressions of displeasure or tension, enhancing the overall negative affect.



Ground Truth AUs:
AU6 (cheek raiser),
AU12 (lip corner puller),
AU25 (lips part)

VideoLLaMA 3: The facial action units activated in the image are:
1. AU4: Smiling
2. AU6: Cheek raising
3. AU7: Nose raising
4. AU8: Eyelid closure
These AUs contribute to the overall expression of joy and amusement depicted in the image.

Qwen 2.5 VL: The facial action units (AUs) activated in the given image are:
- AU 6: Cheek Raiser
- AU 12: Lip CornerPuller (Right)
- AU 15: Lip CornerPuller (Left)
These AUs are responsible for the smile and the overall expression of happiness or amusement in the image.

Video-LLaVA: In the image, the man is smiling with his mouth open, which activates the following facial action units (FAU):
1. AU10: Open mouth
2. AU11: Smile
3. AU12: Teeth
4. AU13: Jaw
5. AU14: Tongue
These facial action units contribute to the man's overall happy and joyful expression.

LLaVA OneVision: The facial action units (AUs) activated in the given image are:
- AU 1: The zygomatic major muscle is contracting, which is responsible for smiling.
- AU 4: The zygomatic major muscle is also contracting, which is responsible for the corners of the mouth turning up.
- AU 6: The zygomatic major muscle is contracting, which is responsible for the cheeks puffing out.
- AU 12: The zygomatic major muscle is contracting, which is responsible for the eyes squinting.
- AU 15: The zygomatic major muscle is contracting, which is responsible for the eyebrows rising and
- AU 17: The zygomatic major muscle is contracting, which is responsible for the eyebrows being.

Face-LLaVA (Ours): The image shows a person expressing a smile. AU6 (cheek raiser) is likely activated because the cheeks are raised, creating a visible lift in the area around the eyes. AU12 (lip corner puller) is activated due to the significant upward movement of the corners of the mouth, forming a wide smile. AU25 (lips part) is also present as the lips are visibly parted, revealing the teeth. The combination of these three AUs creates the overall expression of a joyful, open smile.



Ground Truth AUs:
AU 1 (Inner Brow Raiser),
AU 2 (Outer Brow Raiser),
AU 5 (Upper Lid Raiser)

VideoLLaMA 3: The facial action units activated in the given image are:
1. AU42 (Eyebrow raised) - The eyebrows are raised, which is a key indicator of surprise.
2. AU23 (Eye wide open) - The eyes are wide open, contributing to the expression of surprise or shock.
3. AU6 (Nose raised) - The nose is slightly raised, which can be associated with the emotional state being expressed.
4. AU1 (Brow raised and head tilted) - This unit combines the raised eyebrows and tilted head posture, reinforcing the surprised expression.
These AUs collectively create the overall expression of surprise.

Qwen 2.5 VL: The facial action units (AUs) activated in the given image are:
- AU 12: Upper Eyelid Raiser
- AU 25: Lower Eyelid Raiser
- AU 46: Lower Eyelid Depressor
These AUs are responsible for the raised and widened eyes, which is often associated with surprise or shock.

Video-LLaVA: In the image, the man is making a funny face with his eyes wide open and his mouth open. The facial action units (AU) activated in this situation are:
1. AU10: Open mouth
2. AU20: Eyes wide open
3. AU30: Facial expression
The code for these facial action units are:
1. AU10: 0x01 2. AU20: 0x01 3. AU30: 0x01

LLaVA OneVision: The facial action units (AUs) activated in the given image are: AU 01: Brows together AU 02: Forehead wrinkle AU 03: Cheek pucker AU 04: Lip corner puller AU 05: Lip press AU 06: Lip curl AU 07: Lip part AU 08: Lip curl AU 09: Lip curl AU 10: Lip curl AU 11: Lip curl AU 12: Lip curl AU 13: Lip curl AU 14: Lip curl AU 15: Lip curl AU 16: Lip curl AU 17: Lip curl AU 18: Lip curl AU 19: Lip curl AU 20: Lip curl AU 21: Lip curl AU 22: Lip curl AU 23: Lip curl AU 24: Lip curl AU 25: Lip curl AU 26: Lip curl AU 27: Lip curl AU 28: Lip curl AU 29: Lip curl AU 30: Lip curl AU 31: Lip curl AU 32: Lip curl AU 33: Lip curl AU 34: Lip curl AU 35: Lip curl (...truncated output)

Face-LLaVA (Ours): The image shows a face expressing a complex emotion. AU1 (Inner Brow Raiser) is likely activated due to the slight raising of the inner eyebrows, creating a slight furrow between them. This contributes to an expression of concern or worry. AU2 (Outer Brow Raiser) appears activated as well, indicated by the slight elevation of the outer portions of the eyebrows, further enhancing the impression of surprise or concern. Finally, AU5 (Upper Lid Raiser) is evident in the widening of the eyes, with the upper eyelids visibly raised. This contributes to the overall expression of surprise, fear, or wide-eyed attention. The combination of these three AUs suggests a blend of surprise, concern, and perhaps a hint of fear or apprehension.

Figure 21. Comparison of descriptions or reasoning obtained by Face-LLaVA with other baselines for action unit detection. Blue text indicates alignment with the ground truth and red text indicates wrong reasoning or hallucinations. Notice that for AU detection, all the existing baselines hallucinate badly in getting the FACS action unit codes correct. However, our model is able to correctly capture both the FACS encoding and the correct activated AUs. Moreover, the left-most sample had a missing annotation for AU 17 (chin raiser) which is clearly visible in the image and Face-LLaVA correctly identifies the same.



VideoLLaMA 3: The woman's hair is brown and wavy, her facial expression is a smile, and she appears attractive. The other listed attributes do not apply to this image as there is no indication of any of the other features mentioned such as specific makeup, accessories, or additional facial characteristics.

Qwen 2.5 VL: Based on the image provided:
- Arched Eyebrows: The eyebrows appear to be well-groomed and arched.
- Attractive: The person in the image appears attractive, as is often the case with professionally photographed individuals.
- Brown Hair: The hair color appears to be brown.
- Bushy Eyebrows: The eyebrows are bushy, which is a characteristic feature.
- Big Lips: The lips seem to be full, which could be considered "big."
- Brown Hair: As mentioned earlier, the hair color is brown.
- Straight Hair: The hair appears to be straight.

Video-LLaVA: The image features a woman with brown hair, which is a characteristic associated with the term "brown hair".

LLaVA OneVision: The attributes that can be associated with the given image are:
- Brown Hair and Wavy Hair: The hair color is a shade of brown, and it appears to have some waves or curls.
- Smiling: The person in the image is smiling.
- Attractive: The person has a pleasant appearance.
- Big Lips: The lips appear to be full.
- High Cheekbones: The person has prominent cheekbones.
- Straight Hair: The hair is straight.
- Wearing Lipstick: The person is wearing lipstick.
- Young: The person appears to be young.

Face-LLaVA (Ours): The image shows a young woman with attractive features. She has brown hair styled in a wavy manner. Her makeup is heavy, particularly noticeable on her eyes and lips, where she is wearing lipstick. She possesses high cheekbones and an oval face shape. Her mouth is slightly open in a smiling expression.



VideoLLaMA 3: The image is associated with the attributes: 'Attractive', 'Brown Hair', 'Heavy Makeup', 'Wearing Necklace'.

Qwen 2.5 VL: Based on the image provided:
- Arched Eyebrows: The eyebrows appear to be arched.
- Attractive: The person in the image could be considered attractive based on common beauty standards.
- Bags Under Eyes: There seem to be bags under the eyes, indicating tiredness or fatigue.
- Black Hair: The hair appears to be black.
- Big Lips: The lips look relatively large.
- Big Nose: The nose appears to be large.
- Blurred: The image is not blurry; it is clear and sharp.
- Brown Hair: The hair is brown (...output truncated...)

Video-LLaVA: The image features a woman with brown hair, which is a characteristic associated with the image.

LLaVA OneVision: The attributes that can be associated with the given image are: 'Black Hair', 'Bangs', 'Big Lips', 'Brown Hair', 'Bushy Eyebrows', 'Chubby', 'Double Chin', 'Heavy Makeup', 'High Cheekbones', 'Male', 'Mouth Slightly Open', 'Mustache', 'Narrow Eyes', 'Receding Hairline', 'Rosy Cheeks', 'Sideburns', 'Wearing Lipstick', 'Wearing Necklace'. The reason for these associations is based on the visible features of the person in the image.

Face-LLaVA (Ours): The image shows a young woman with attractive features. She has arched eyebrows, big lips, and wavy hair. Her makeup is heavy, including lipstick. The heavy makeup and styling contribute to her overall attractive appearance.



VideoLLaMA 3: The image can be associated with the following attributes: 'Attractive', 'Black Hair', 'Big Lips', 'Male', 'Smiling', 'Wearing Lipstick'. These attributes are inferred from the visible characteristics of the person in the image, such as the long black hair, full lips, and the fact that the individual is smiling. The presence of lipstick suggests a choice to enhance or highlight the lips, which contributes to the attractiveness of the subject.

Qwen 2.5 VL: Based on the image provided:
- Male: The person in the image appears to be male.
- Black Hair: The hair color is black.
- Wavy Hair: The hair seems to have a wavy texture.

Video-LLaVA: The image features a man with a big smile, which is an attractive attribute. The man has a pointy nose, which is another noticeable feature. The image is not blurry, and the man has a bald head, which is a distinctive characteristic. The man is wearing a necklace, which is another noticeable detail.

LLaVA OneVision: The attributes that can be associated with the given image are: 'Black Hair', 'Male', 'Smiling', 'Straight Hair', and 'Wearing Lipstick'. The reason for these associations is that the individual in the image has long, straight black hair, is a male, is smiling, has straight black hair, and is wearing lipstick.

Face-LLaVA (Ours): The image shows a young male with black hair, a beard, and bushy eyebrows. His mouth is slightly open, and he is smiling. He has a big nose and high cheekbones. These attributes are all clearly visible in the provided image.

Figure 22. Comparison of descriptions or reasoning obtained by Face-LLaVA with other baselines for facial attribute detection. Note that for this task, we do not provide the ground truth label in the figure as the ground truth is clearly identifiable from the image. Also, note that for getting the baseline predictions for this task, we provide the list of possible attributes to the input prompt for the baselines to restrict their responses. Moreover, for computing the mean accuracy on CelebA [45] test set for this task for the baselines, we use a separate prompt where we ask the model to just detect the visible attributes and not describe them. This is done because our synonym matching does not work so accurately on formatted text (as shown above) generated by the baselines.

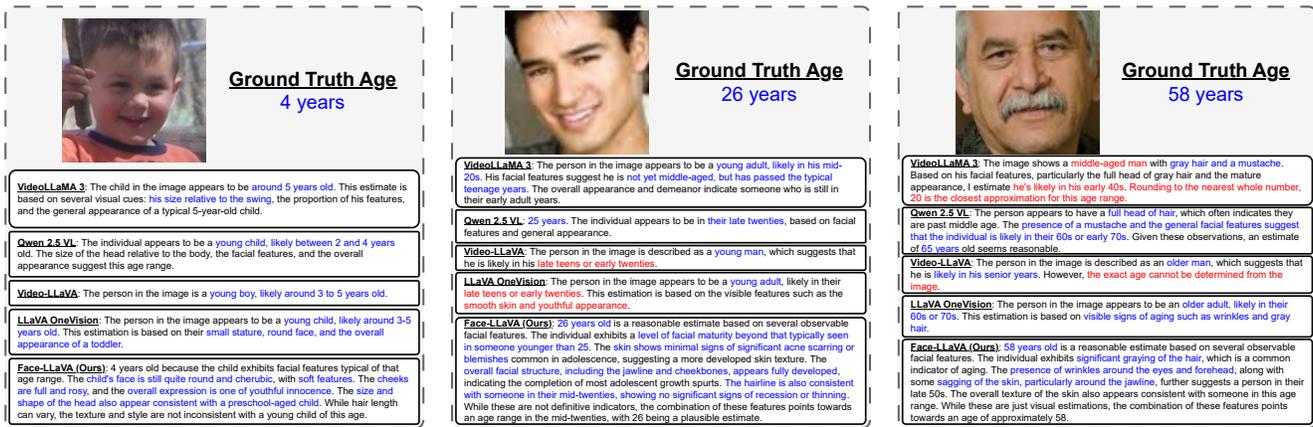


Figure 23. Comparison of descriptions or reasoning obtained by Face-LLaVA with other baselines for age estimation. Blue text indicates alignment with the ground truth and red text indicates wrong reasoning or hallucinations. We can notice that the baselines provide minimal description for reasoning their predictions. Moreover, the reason is not based only on the facial features and has some inconsistencies (e.g. VideoLLaMA 3 [88] makes wrong age prediction even with correct reasoning for the right-most example.) In contrast to the baselines, Face-LLaVA is able to predict the age of the person in the image correctly with a detailed description specifically related to the facial features of the person in the image. Finally, it is important to note that we use a different prompt when we compute the performance of the baselines on traditional metrics (mean absolute error - refer Appendix F.3) as the baselines usually provide an age range in their descriptions and usual string parsing will have some inconsistencies due to that.

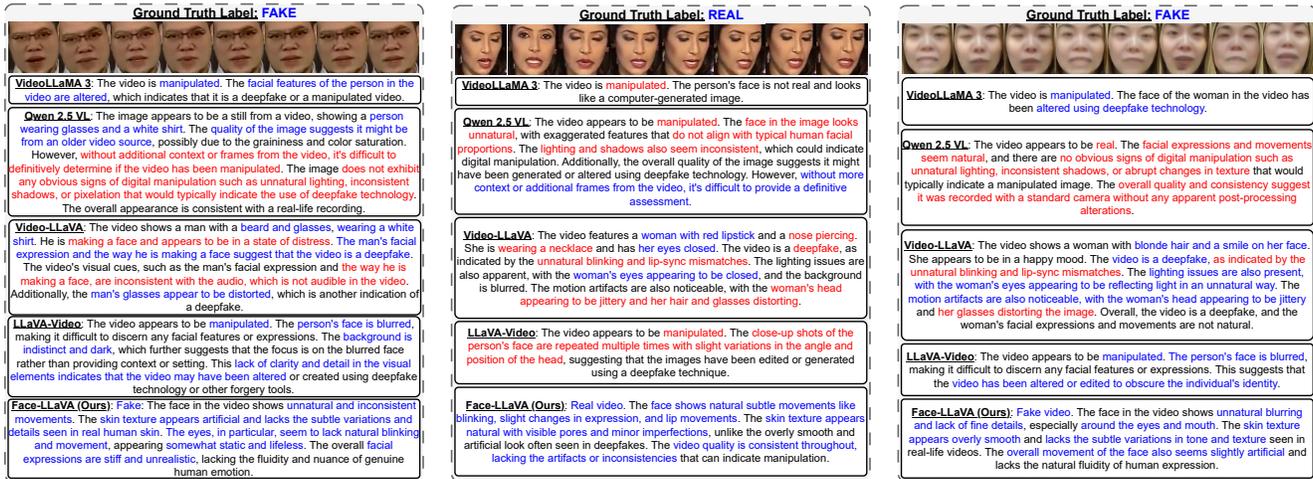


Figure 24. Comparison of descriptions or reasoning obtained by Face-LLaVA with other baselines for deepfake detection. Blue text indicates alignment with the ground truth and red text indicates wrong reasoning or hallucinations. Other MLLMs hallucinate with their reasoning and provide descriptions that are not face-specific. Face-LLaVA on the other hand provides consistent descriptions that are face specific. Similar to attribute detection and age estimation, we use a different prompt for computing accuracy for the baselines on the FaceForensics++ [61] dataset for Tab. 5.

Ground Truth Label: Fear



Face-LLaVA: The subject's facial expression clearly indicates surprise. Initially, her eyes are wide open, and her mouth is slightly parted, suggesting a state of alertness or anticipation. However, as the video progresses, her mouth opens much wider in an 'O' shape, and her eyebrows are slightly raised. This combination of wide eyes, an open mouth, and slightly raised eyebrows is a classic indicator of surprise. The rapid change in her facial expression from a state of mild alertness to a clear expression of astonishment further strengthens the interpretation of the emotion as surprise.

Ground Truth Label: Sadness



Face-LLaVA: The subject's facial expression remains relatively static throughout the video. There are no significant changes in brow furrow, lip corners, or eye widening that would typically indicate emotions like happiness, sadness, anger, or surprise. The mouth is slightly open, which could be interpreted as slight surprise or concern, but it's not intense enough to definitively classify the emotion as anything other than neutral. The overall expression is one of mild contemplation or perhaps slight apprehension, but lacks the intensity or distinct features associated with clearly defined emotions.

Ground Truth Label: Neutral



Face-LLaVA: The subject's facial expression conveys sadness through several cues. His eyebrows are slightly furrowed, indicating a level of distress or worry. His gaze is downcast, avoiding direct eye contact, which is often associated with feelings of sadness or dejection. His mouth is slightly downturned, forming a subtle frown, further reinforcing the overall impression of sadness. The overall expression is one of quiet contemplation and sorrow, rather than an outburst of intense emotion. There is a stillness to his face, lacking the dynamism of other emotions.

Ground Truth Label: Fear



Face-LLaVA: The subject's facial expression conveys sadness primarily through her eyes and mouth. Her eyes appear downcast and slightly watery, lacking their usual brightness. This suggests a feeling of dejection or hopelessness. Her mouth is slightly downturned, forming a subtle frown, further reinforcing the sadness. While the presence of wounds on her face adds to the overall bleakness of the scene, the primary indicator of sadness is the subtle yet noticeable downturn of her mouth and the lack of sparkle in her eyes. The expression is not overtly dramatic, but rather a quiet, subdued sadness.

Figure 25. Failure cases for facial expression recognition task on some samples of DFEW [22]. In the samples above, expressions labeled as ground truth *fear* are often misclassified as *sadness* or *surprise*. Closer inspection of the videos reveals that these instances indeed exhibit mixed emotions. Additionally, ground truth labels are assigned based on the full scene, incorporating background context and audio, whereas Face-LLaVA makes predictions solely from the face-cropped video.

Ground Truth Label: Fake



Face-LLaVA: Real video. The face shows natural skin texture with pores, subtle imperfections, and realistic variations in lighting and shadow across the face. The slight movements of the eyes and mouth are natural and fluid, lacking the stiffness often seen in deepfakes. The video quality is consistent with older recordings, showing some compression artifacts but no obvious signs of digital manipulation or splicing.

Ground Truth Label: Fake



Face-LLaVA: Real video. The video shows a close-up of a woman's face. The face displays natural imperfections, including subtle skin texture, slight asymmetry, and realistic lip and eye movements. The lighting appears consistent and natural, without any obvious signs of digital manipulation or artificial enhancements. The video quality is consistent with older recordings, showing some grain and compression artifacts, which are typical of non-manipulated footage.

Ground Truth Label: Fake



Face-LLaVA: Real video. The face shows natural subtle movements like slight changes in expression and lip movements that are consistent with a real person speaking. The skin texture appears natural with visible pores and minor imperfections, unlike the smooth, flawless skin often seen in manipulated videos. The video quality is consistent throughout, lacking any jarring transitions or inconsistencies that might suggest editing or manipulation.

Ground Truth Label: Real



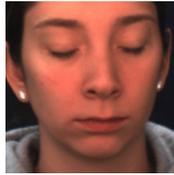
Face-LLaVA: Fake. Real human faces have subtle variations in texture and lighting that are difficult to perfectly replicate. In this video, the face appears overly smooth and lacks the natural imperfections and subtle changes in skin tone and texture that one would expect in a real video. The eyebrows appear unnaturally uniform and sharply defined, lacking the natural variations in hair thickness and direction. The overall appearance of the face is too perfect and lacks the realistic imperfections found in real-life videos.

Figure 26. Failure cases for deepfake detection task on some samples of FaceForensics++ [61]. In the first and third samples above, determining whether the videos are manipulated is challenging even for human beings, as they were altered using a face re-enactment technique (NeuralTextures). In the last example, the model mistakenly classifies a real video as fake, attributing its decision to the smoothness of the skin and makeup around the eyes, interpreting the face as “too perfect” to be real.



Ground Truth Label: No Action Units

Face-LLaVA: The image shows a person with slightly raised eyebrows, specifically the inner parts of the eyebrows. This subtle elevation of the inner eyebrows is characteristic of AU4, brow lowering. While it's not a strong or dramatic display of AU4, the slight downward pull and inward movement of the eyebrows suggest its activation.



Ground Truth Label: No Action Units

Face-LLaVA: The image shows a person with their eyes closed. AU17, which refers to the jaw drop, is not directly visible in this image. However, the closed eyes could be interpreted as a component of a more complex expression, possibly involving a degree of relaxation or calmness, which might be associated with a subtle jaw drop. The lack of other obvious facial muscle movements makes it difficult to definitively rule out a very slight jaw drop, even if it's not strongly expressed.



Ground Truth Label: No Action Units

Face-LLaVA: The image shows a person with their lips slightly pressed together, which is a characteristic of AU26 (Lip Tightening). The corners of the mouth appear slightly downturned, further supporting the presence of AU26. There is no significant lip protrusion or other expressions that would contradict the presence of AU26.

Figure 27. Failure cases for action unit detection task on some samples of DISFA [49]. Our model fails mostly on the cases when the ground truth sample does not have any action units activated.



Ground Truth Label: [Gray hair, Male, High cheekbones, Mouth slightly open, Smiling, Wearing necktie]

Face-LLaVA: The image shows a male with blond hair, a big nose, high cheekbones, and a double chin. His mouth is slightly open, and he is smiling. There are also bags under his eyes.



Ground Truth Label: [Black hair, Male, High cheekbones, Mouth slightly open, Straight hair, Young]

Face-LLaVA: The image shows a young male with brown hair. His mouth is slightly open, and he has narrow eyes. The 'young' attribute is inferred from the overall appearance of the face, which lacks significant signs of aging. The 'male' attribute is determined by the facial structure and features.



Ground Truth Label: [Blond hair, High cheekbones, Mouth slightly open, Oval Face, Smiling, Wearing earrings, Wearing lipstick, Wearing necktie]

Face-LLaVA: The image shows a young woman with attractive features. She has brown hair styled in a short, slightly tousled cut. Her makeup is heavy, particularly noticeable on her eyes and lips, where she is wearing lipstick. She possesses high cheekbones and an oval face shape. Her mouth is slightly open in a smiling expression. She is also wearing earrings.

Figure 28. Failure cases for facial attribute detection task on some samples of CelebA [45]. In the above examples, while the model's predictions slightly differ from the ground truth, they remain consistent with the visual content of the image, suggesting potential imperfections in the dataset annotations of CelebA.



Ground Truth Age: 11 years

Face-LLaVA: 16 years old because the individual exhibits facial features consistent with that age group. The person's facial structure shows a blend of youthful and developing characteristics. The skin appears relatively clear with minimal signs of significant acne scarring or deep wrinkles, common in older individuals. The overall shape of the face and the proportions of the features, such as the nose and jawline, suggest a person in the mid-teens rather than a younger or older age. The hairline and hair texture also appear typical for a 16-year-old.



Ground Truth Age: 12 years

Face-LLaVA: 24 years old is a reasonable estimate based on several observable features. The individual's facial features, particularly the development of the jawline and cheekbones, suggest an age beyond the early twenties. The skin exhibits a texture and tone consistent with someone in their mid-twenties, showing minimal signs of significant aging or acne scarring often associated with younger ages. Furthermore, the overall facial structure and expression convey a maturity level that aligns with the 24-year-old age range. It's important to note that this is an estimation based on visual cues and may not be perfectly accurate.



Ground Truth Age: 80 years

Face-LLaVA: 65 years old is a reasonable estimate based on several observable features. The individual's extensive network of wrinkles around the eyes and forehead, particularly the crow's feet, are consistent with significant sun exposure and the natural aging process over many years. The presence of significant sagging skin on the face, particularly noticeable around the jawline and cheeks, also points towards an older age. The individual's hair is largely white or gray, which is a common indicator of aging. The overall texture of the skin appears thinner and less elastic than that of a younger person. While these are just indicators and not definitive proof, the combination of these features strongly suggests an age in the mid-60s.

Figure 29. Failure cases for age estimation task on some samples of UTKFace [94]. Except for the middle sample, the error in prediction for the other two samples are acceptable and even for humans it is hard to judge the age.