# ContextFace: Generating Facial Expressions from Emotional Contexts

Anonymous ICCV submission

Paper ID *****

## Abstract

*The task of generating 3D facial expressions given various situational contexts is important for applications such as virtual avatars or human-robot interactions. The task is, however, challenging not only because it requires a comprehensive understanding of emotion, expression and contexts, but also there rarely are datasets to support the task. We propose ContextFace, a Multi-modal Large Language Model (MLLM) fine-tuned to generate 3D facial expressions depending on complex situational contexts. To overcome the lack of datasets, we perform a context augmentation to existing emotion recognition datasets; we generate plausible situations and quotes from images and emotions to annotate the dataset. Next, we perform visual instruction tuning of MLLMs on context-augmented datasets to boost its capability of visual synthesis from emotions. Experiments show a superior performance of ContextFace in the zero-shot evaluation of contextual emotion recognition. A qualitative evaluation shows that our method generates expressions consistent with diverse contexts and performs complex emotion reasoning, e.g., speculative generation of expressions of occluded faces through interactive prompting.*

## 1. Introduction

The interpretation of facial expressions inherently depends on the situational context. For instance, a smiling face can convey different emotions – from genuine joy to forced politeness – depending on the situational context. A combined understanding of expression, emotion, and context has recently become important, particularly in the field of human-robot interaction [20, 31, 38, 40]. Prior works on Facial Expression Recognition [30, 34, 39, 43] focused on facial feature analysis, but underexplored the richness of contextual information. Recently, Multimodal Large Language Models (MLLMs) enabled interpretable analysis of expressions by explaining how facial features indicate specific emotional states [21, 24, 45]. MLLMs can perform context-aware emotion recognition [4, 11, 17, 44, 46] to logically infer emotions by sophisticated reasoning, opening new avenues for emotion analysis from multi-modal cues.

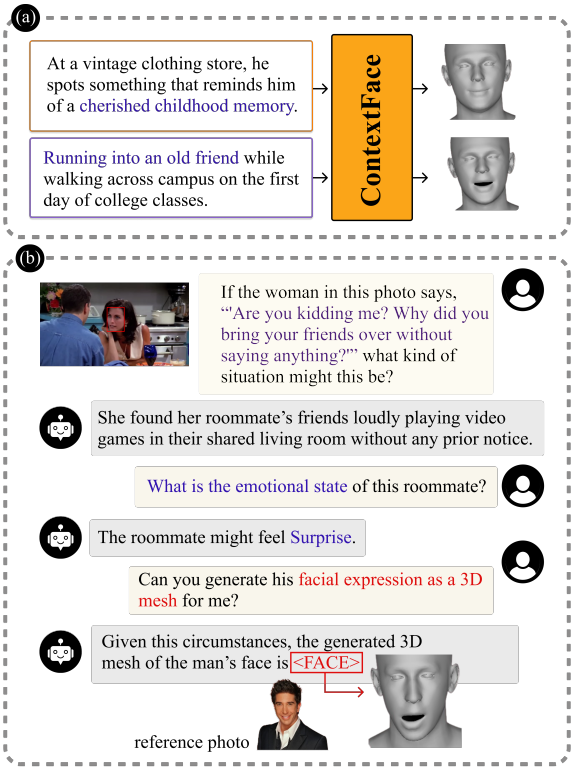


Figure 1. (a) **Context-to-face generation.** ContextFace generates subtly different expressions for the same emotion "Happy". (b) **Speculative generation of expressions.** The emotion of woman with the image alone is ambiguous. But with her quote, ContextFace infers her emotion and generate plausible situations. Moreover, ContextFace speculatively generates the man's expression whose face is not visible. An embarrassed face is generated from special token <FACE> through an interactive prompting. The output is FLAME parameters [19] which is later matched to the reference photo using INFERNO [7] for the final mesh.

While the aforementioned works focus on predicting emotions or textual explanations, we consider a task of *generating 3D facial expressions conditional on complex situational contexts*. The task has a wide range of applications, e.g., interactive visual assistants [25, 27] or virtual

avatars [3, 13], where one needs to aptly synthesize 3D faces that well-reflect emotions arising from various situations. It is challenging to generate expressions capturing the subtlety of situational contexts. Consider two situations: "unexpectedly meeting an old friend on the street" and "cherishing childhood memories". Both will be categorized as "happy" in traditional emotion recognition; however, the faces of the persons in the contexts will subtly differ, although both are expected to be "smiley faces". The task requires reasoning capabilities to derive proper emotion and expressions from situational contexts. However, a lack of datasets that encompass expressions, emotions, and contexts makes it difficult to develop multi-modal models.

In this paper, we propose ContextFace, an MLLM that integrates contextual understanding and emotional reasoning to generate facial expressions. ContextFace is able to generate proper expressions matched to details in the contexts (Fig. 1(a)). Both situations imply emotion "happy", but leading to the generation of different faces. Moreover, ContextFace can perform emotional reasoning in complex contexts. In Fig. 1(b), the model is prompted to 1) generate a plausible situation of the conversation; 2) infer the expression of the man whose face is not visible. Although the emotion of the woman with the image alone is ambiguous, ContextFace can infer her emotion from her quote, create reasonable situations and *speculatively* generate a proper facial expression of the man through an interactive prompting.

The design of ContextFace is based on *visual instruction tuning* [27, 28]. We first create datasets for instruction tuning using *context augmentation*. Specifically, we augment existing emotion-labeled image datasets [9, 16] with two new annotations: situation descriptions and subject quotes. For the annotation, we leverage the capabilities of strong LLMs in grounding emotions in commonly understood social contexts. The datasets will be publicly released for the research in emotional analysis. Next, we construct instruction datasets derived from the newly augmented datasets, fine-tune an MLLM on those instructions. The MLLM uses special token `<FACE>` associated with its hidden state along with the face projection module to estimate FLAME parameters [19]. This enables 3D face reconstruction with expressions based on reference photos using INFERNO [7]. Experiments show the superior performance of ContextFace in contextualized emotion recognition compared to existing MLLMs as well as its capability of emotional reasoning and face generation given various contexts, such as speculative generation of occluded faces through interactive prompting.

Our contribution is summarized as follows. (1) We create and publicly release two emotion datasets augmented with rich contextual information. (2) We propose ContextFace, an MLLM fine-tuned with visual instruction tuning for synthesizing proper expressions from complex situational contexts. (3) Experiments show ContextFace has superior emotional reasoning and recognition capabilities in both quantitative and qualitative aspects.

## 2. Related Work

**MLLMs and Instruction Tuning.** Enabling large language models to process data modalities beyond text has been a foundational area of research [18, 27, 41, 42]. Flamingo [1] proposed general-purpose large vision language models trained from a large-scale web corpus. BLIP [18] proposed an efficient VLM training strategy with Q-Former to bridge the image and text modality using cross-attention. LLaVA [27, 29] proposed a 2-stage visual instruction tuning method, which first aligned vision encoders with LLMs and then fine-tuned the VLM using chat-oriented datasets. The visual instruction tuning was successfully applied to various tasks. LISA [15] leverages the capabilities of LLMs for segmentation tasks that require complex reasoning. GLaMM [37] generates natural conversational texts with integrated pixel-level object segmentation masks in visual interactions. ChatPose [12] enables text and image inputs to generate 3D human body poses while supporting complex reasoning about posture. Inspired by those works, we apply visual instruction tuning for generation of facial expressions given emotional contexts.

**3D Face Reconstruction.** FLAME [19] is a high-quality 3D Face Reconstruction model from a single image, which is a parameterized framework for 3D face representation with separate components for identity, pose, and expression. EMOCA [6] proposed an emotion consistency loss to generate emotional expressions of significantly higher fidelity. Recent works such as EmoTalk [35], EMOTE [8] and EmoFace [26] proposed emotion-aware 3D face reconstruction methods that incorporate emotional expressions into 3D face representations by analyzing speech. However, it is an underexplored task to integrate the emotional expressions and complex situational contexts with ambiguous face images into 3D face representations.

**Multimodal Emotion Understanding.** DialogueLLM [46] proposed a finetuning of LLMs with multimodal emotional dialogues from texts and videos, and released datasets on emotion recognition conversations. There are works [10, 44] which studied reasoning emotions within the situational context, and showed that recognizing accurate visual markers such as bounding boxes could significantly boost the performance. Emotion LLaMA [4] has presented an MLLM that infers human emotions by integrating multiple cues including facial expressions analyzed through action units, utterances, audio tones, and visual context descriptions. Explainable Multimodal Emotion Recognition (MER) [23] was proposed to improve the reliability and accuracy of emotion recognition by generating explanations for their predictions. While those approaches leveraged multimodality to recognize and reason about emotions, our

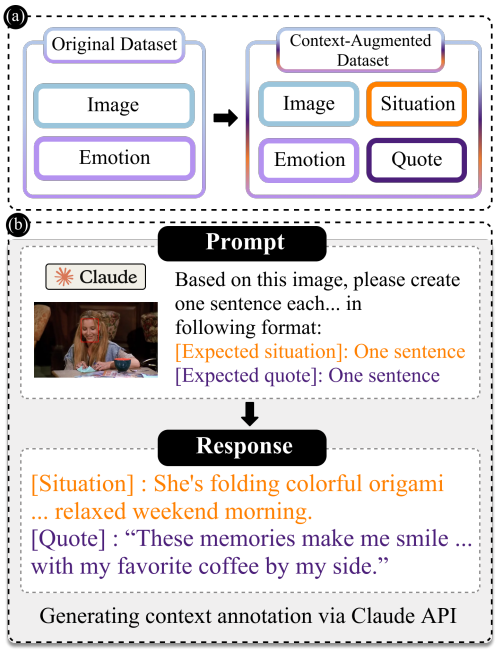


Figure 2. **Contextual Augmentation to Emotion Datasets.** (a) The original dataset is augmented with contexts, plausible situations and quotes, consistent with the image and emotion. (b) The context annotations are generated through prompting the Claude API with images and emotion labels.

research takes a step forward by introducing a novel task of generating facial expressions from emotional reasoning based on complex situational contexts.

# 3. Method

We propose ContextFace, a Multimodal Large Language Model (MLLM) that integrates contextual understanding, emotional reasoning and facial expression generation. We aim to achieve this with *visual instruction tuning* inspired by recent works [12, 15, 37]. However, there rarely exists datasets that integrate context, emotion and facial expressions together. Our strategy to overcome the data scarcity is to perform *contextual augmentation* to existing emotion datasets assisted by LLMs. The augmented datasets enable us to create an instruction-tuning pipeline for an integrated understanding of contexts and emotions to generate appropriate facial expressions. With newly defined instruction tasks and datasets, we propose an efficient MLLM architecture for 3D facial mesh generation. In the following sections, we describe our dataset augmentation approach using LLMs (Sec. 3.1), visual instruction tuning (Sec. 3.2), and the proposed architecture (Sec. 3.3).

## 3.1. LLM-Assisted Context Augmentation to Emotion Dataset

**Datasets.** We construct emotion datasets augmented with contextual annotations. We publicly release two datasets: SFEW-C (Static Facial Expressions in the Wild with Context) and CAER-S-C (Context-Aware Emotion Recognition - Static with Context). Our datasets are based on and enhance SFEW [9] and CAER-S [16], where these original datasets contain samples of image (input) and emotion (label) pairs. We chose SFEW and CAER-S, because they comprise frame captures of human subjects from movies and TV shows, offering a good balance between contextual background information and clear facial expressions of the subjects. We augment each sample of SFEW and CAER-S with contextual information; specifically, situation and quote for the image-emotion pair, e.g., see Fig. 2(a). With the contextual annotations, our datasets will be useful for training MLLMs to understand emotion, context and facial expressions.

**Contextual Annotation.** We generate the contextual annotations assisted by strong LLMs. We leverage the vast visual and textual knowledge of LLMs to infer contexts from image and emotion. Given the original dataset's images and emotion labels as input, we prompted LLMs to generate expected situation and plausible quotes (Fig. 2(b)). The annotations are generated through the Claude-3-5-sonnet-20241022 API [2]. From the original datasets, we excluded several samples that did not meet our criteria, e.g., images without human subjects, those containing adult/inappropriate content rejected by the Claude API, etc. After applying our preprocessing criteria, our final released annotation dataset contains 42,196 training and 20,938 test samples for CAER-S-C, and 890 training and 431 test samples for SFEW-C.

## 3.2. Visual Instruction Tuning

To train ContextFace, we construct instruction tuning datasets derived from the augmented datasets introduced in Sec. 3.1. We define two new tasks: Situation Generation and Expression Generation and build the associated instruction datasets, along with a generic VQA dataset [27] for the instruction tuning of ContextFace.

**Situation Generation Instructions.** Consider the augmented dataset in Sec. 3.1 consisting of image, emotion, situation and quote. From the dataset, we create a task called *Situation Generation* which takes image and quote as input, and predicts situation and emotion as response (see Fig. 3(a)). This task enables a model to learn associations between contextual cues and emotions.

We create a instruction tuning dataset for Situation Generation task as follows. We prepare training pairs $\{\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{txt}}\}$ and their corresponding text output $\mathbf{y}_{\text{txt}}$ from the augmented dataset. Here, $\mathbf{x}_{\text{img}}$ denotes the input image

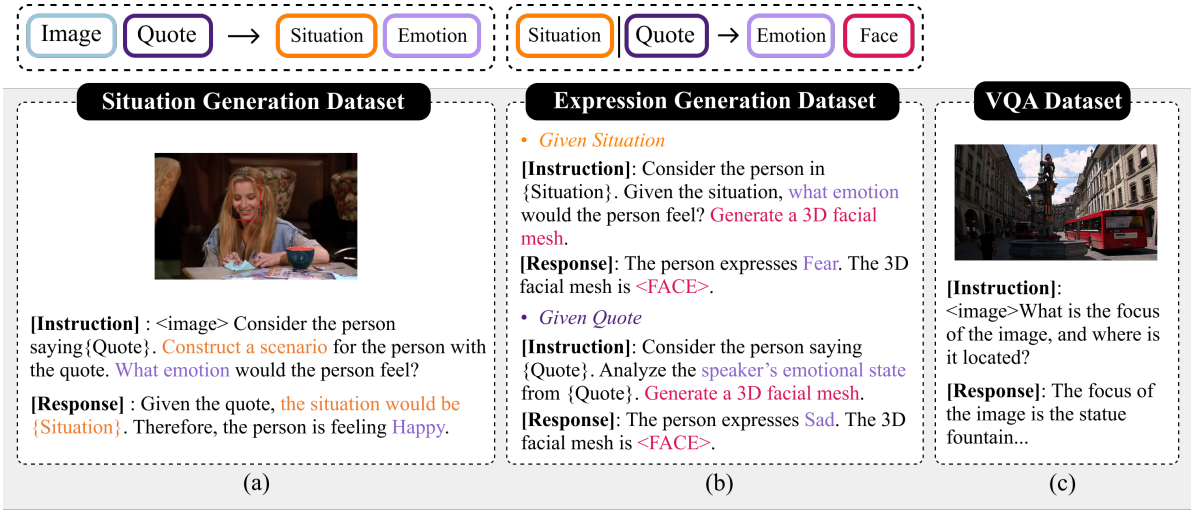


Figure 3. Instruction datasets consisting of three types: (a) Situation Generation Dataset, (b) Expression Generation Dataset, and (c) Visual Question Answering Dataset.

with a red bounding box marking the target person. $\mathbf{x}_{\text{txt}}$ is a text prompt which asks the model to create a scenario from the input quote and predict the emotion from predefined categories. $\mathbf{y}_{\text{txt}}$ denotes the target response describing a plausible situation that fits the quote, followed by the identified emotional state. We organized the data in an instruction-response as follows, with <image> denoting the position of image tokens:

> **INSTRUCTION:** <image> Think about a person who says "I can't believe how therapeutic it is to turn these simple pieces of paper into beautiful little art pieces." - construct the scenario that would result in this statement, and what emotion would they be feeling? Choose from: Angry, Sad, Surprise, Neutral, Fear, Happy, Disgust.
> **RESPONSE**: She's folding colorful origami while enjoying a relaxing afternoon coffee break at home. Threfore, the emotion this person is feeling is Happy.

**Expression Generation Instructions.** We create a task called *Expression Generation* which takes context (situation or quote) as input, and predicts 3D facial expression and emotion as response (see Fig. 3(b)). With this task, the model learns to produce 3D facial expressions that correspond to its emotional analysis of given situations or quotes.

The training data comprises input $\{\mathbf{x}_{\text{txt}}\}$ and target response $\{\mathbf{y}_{\text{txt}}, \beta\}$ obtained from the augmented dataset. $\mathbf{x}_{\text{txt}}$ denotes a text instruction that describes a situation or quote and asks the model to generate an appropriate 3D facial representation. $\mathbf{y}_{\text{txt}}$ is the target response including both emotional inference and the special token called <FACE>. <FACE> is a learnable token from which 3D facial mesh will be predicted. $\beta$ denotes the latent representation of the target 3D facial mesh. Specifically, $\beta$ is a 100-dimensional FLAME expression parameters [19] extracted from the face in the ground truth image using INFERNO [7]. We structure the data in the following instruction-response format:

> **INSTRUCTION:** She's folding colorful origami while enjoying a relaxing afternoon coffee break at home. From the described situation, infer the person's emotional state and generate an appropriate 3D facial mesh that captures the feeling.
> **RESPONSE:** Considering the situation, the person expresses Happy. The 3D facial mesh is <FACE>.

> **INSTRUCTION:** "I can't believe how therapeutic it is to turn these simple pieces of paper into beautiful little art pieces." Please analyze the speaker's emotional state from this quote and generate a corresponding 3D facial mesh.
> **RESPONSE:** Considering the situation, the person expresses Happy. The 3D facial mesh is <FACE>.

**Visual Question Answering Dataset.** To preserve the model's original capabilities in Visual Question Answering during the training on task-specific data, we incorporated VQA dataset (Fig. 3(c)), specifically LLaVA-Instruct-150K [27] which is a visual-language instruction dataset generated by GPT-4, into our training pipeline.

### 3.3. Architecture

**Model Design.** As illustrated in Fig. 4, our model consists of two main components: a multi-modal large language

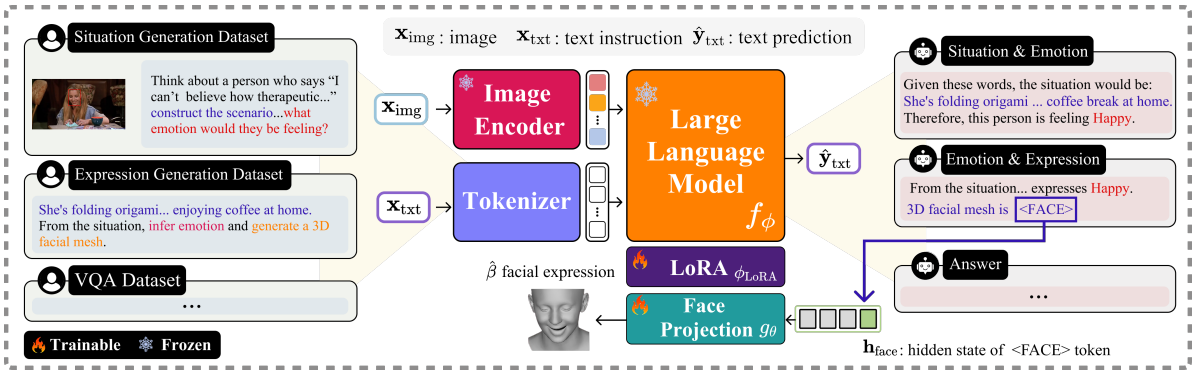


Figure 4. **Pipeline of ContextFace.** The instruction datasets described in Fig. 3 are used as input. Images and text are processed by the image encoder and tokenizer respectively, and are fed into the LLM to produce text output. For the expression generation task, the hidden state of the <FACE> token in the text output is processed through face projection layers to generate the coefficient vector of facial expressions. The coefficient vector is then used to generate the final facial meshes.

model denoted as $f_\phi$ and a face projection model $g_\theta$. Text instructions $\mathbf{x}_{txt}$ serve as inputs to the multi-modal LLM $f_\phi$, resulting in text outputs $\hat{\mathbf{y}}_{txt}$ given by

$$\hat{\mathbf{y}}_{txt} = f_\phi(\mathbf{x}_{txt}). \tag{1}$$

We employed LoRA fine-tuning [14] with trainable parameters $\phi_{LoRA}$ to efficiently adapt the model to our specific task while preserving the generation capabilities of pre-trained vision-language models. If the input prompt requests the generation of 3D facial mesh, the multi-modal LLM $f_\phi$ is trained to include <FACE> token in its text outputs $\hat{\mathbf{y}}_{txt}$. The hidden state of the multi-modal LLM $f_\phi$ corresponding to <FACE> token, denoted by $\mathbf{h}_{face}$, is used to generate the 3D facial mesh. $\mathbf{h}_{face}$ contains prompt-specific contextual information required to generate a facial expression. $\mathbf{h}_{face}$ is subsequently passed through the face projection module $g_\theta$ to obtain the expression coefficient $\hat{\beta}$ given by

$$\hat{\beta} = g_\theta(\mathbf{h}_{face}). \tag{2}$$

We use a two-layer multi-layer perceptron (MLP) for $g_\theta$.

**Loss Functions.** The loss function is designed to optimize both text output $\hat{\mathbf{y}}_{txt}$ and facial expression output $\hat{\beta}$. Firstly, the loss $\mathcal{L}_{txt}$ for the text generation is the cross entropy loss between the predicted response $\hat{\mathbf{y}}_{txt}$ and the ground truth $\mathbf{y}_{txt}$:

$$\mathcal{L}_{txt} = \mathbf{CE}(\mathbf{y}_{txt}, \hat{\mathbf{y}}_{txt}). \tag{3}$$

Secondly, for the facial expression loss denoted by $\mathcal{L}_{face}$, we adopted the L2 loss as the objective function for our facial expression component. The loss compares L2 distance between the predicted $\hat{\beta}$ and ground truth $\beta$ coefficients given by

$$\mathcal{L}_{face} = \|\beta - \hat{\beta}\|_2^2. \tag{4}$$

where $\|\cdot\|_2$ denotes the L2-norm of a vector. The overall loss $\mathcal{L}$ consists of text generation loss $\mathcal{L}_{txt}$ and facial expression $\mathcal{L}_{face}$, which is given by

$$\mathcal{L} = \mathcal{L}_{txt} + \lambda \cdot \mathcal{L}_{face}. \tag{5}$$

where $\lambda$ is a weight that balances the losses.

**Training Process.** ContextFace is jointly trained on three instruction datasets: Situation Generation, Expression Generation, and VQA, in an end-to-end manner. Specifically, the model is jointly trained with batches from different tasks, where each batch contained samples from a single dataset randomly selected according to fixed ratios. Both the token embedding matrix and LLM prediction head are set to be trainable to handle the newly added <FACE> token. Also, the face projection $g_\theta$ is a trainable module to learn the mapping from $\mathbf{h}_{face}$ to facial expression parameters $\hat{\beta}$. Due to the inherently smaller magnitude of the facial expression loss values, we applied a weighting factor $\lambda$ of 10 to this term in the overall loss function during training.

# 4. Experiment

## 4.1. Implementation Details

**Datasets and Baselines.** We use two datasets SFEW [9] and CAER-S [16] which contain captures of movies and TV shows and are suitable for studying contextual emotions. For training and evaluation, we use context-augmented versions: SFEW-C and CAER-S-C as introduced in Sec. 3.1. ContextFace is trained with CAER-S-C. We compare ContextFace with state-of-the-art MLLM models: BLIP-13B [5], LLaVA-v1.5-13B [27], LLaVA-NEXT-13B [29] and Qwen2.5-VL-7B [42]. ContextFace and all the baseline models are evaluated in a zero-shot manner on SFEW-C. In addition, we provide an evaluation on emotion datasets derived from MER2023 [4, 22] datasets: due to space constraints, readers are referred to Sec. 1 of Supplementary Materials.

| Method | Hap | Sad | Neu | Ang | Sur | Dis | Fea | UAR | WAR |
|---|---|---|---|---|---|---|---|---|---|
| *with situation* | | | | | | | | | |
| BLIP-13B [5] | 72.41 | 55.45 | 50.92 | 81.82 | 21.21 | 64.86 | 42.62 | 51.11 | 51.09 |
| llava-1.5-13B [27] | **96.50** | 93.96 | 90.45 | 77.78 | 73.33 | 78.05 | 71.43 | 81.42 | 84.98 |
| LLaVA-NEXT-13B [29] | 93.62 | 93.51 | 88.31 | 91.57 | 69.90 | **85.71** | **86.27** | 86.88 | 89.73 |
| Qwen2.5-VL-7B [42] | 95.10 | 89.44 | 86.96 | 89.04 | 73.79 | 79.17 | 84.00 | 86.08 | 88.33 |
| ContextFace (ours) | 94.20 | **95.89** | **92.77** | **94.94** | **85.45** | **85.71** | 84.31 | **90.13** | **90.68** |
| *with quote* | | | | | | | | | |
| BLIP-13B [5] | 26.51 | 21.95 | 36.44 | 35.79 | 0.00 | 16.00 | 40.00 | 26.35 | 24.52 |
| llava-1.5-13B [27] | 93.53 | 93.15 | 83.22 | 86.36 | 80.73 | 74.42 | 78.00 | 84.12 | 85.66 |
| LLaVA-NEXT-13B [29] | 90.23 | 90.51 | 77.55 | 91.57 | 78.50 | 83.72 | 71.32 | 84.01 | 85.56 |
| Qwen2.5-VL-7B [42] | 90.65 | **97.96** | **91.72** | 92.99 | 80.70 | 76.36 | 90.32 | 90.20 | 91.33 |
| ContextFace (ours) | **96.60** | 92.21 | 84.93 | **94.48** | **91.74** | **95.65** | **92.78** | **93.21** | **94.89** |

Table 1. Zero-shot emotion recognition performance in F1 scores on SFEW-C Dataset. The input data for upper (resp. lower) table is image-situation (resp. image-quote) pair. Emotion categories: Hap (Happy), Sad (Sad), Neu (Neutral), Ang (Angry), Sur (Surprise), Dis (Disgust), Fea (Fear). UAR: Unweighted Average Recall, WAR: Weighted Average Recall.

**Network Architecture.** We utilized LLaVA-v1.5-13B [27] as the backbone multimodal large language model for our implementation. We used the CLIP vision encoder [36] (ViT-L/14 with 336px resolution) of LLaVA-v1.5-13B for visual processing. The face projection module is a two-layer MLP with input, hidden layer and output sizes given by 5120, 5120, and 100 respectively.

### 4.2. Quantitative Results

**Contextual Emotion Recognition.** Tab. 1 shows the performance of contextual emotion recognition. Specifically, the task is to classify images accompanied by one of two context types, situation or quote, into seven categories of emotions. ContextFace achieves superior performance in the majority of emotion categories for both situation and quote contexts, with the highest UAR (Unweighted Average Recall) and WAR (Weighted Average Recall) scores in each contextual setting (90.13%, 90.68% with situation context and 93.21%, 94.89% with quote context). Notably, ContextFace excels at detecting emotions like Disgust and Fear which typical models for emotion recognition struggle to identify. This superior performance suggests that ContextFace has developed a sophisticated understanding of the complex relationship between context and emotion through the proposed framework of visual instruction tuning.

**Facial Expression Prediction.** We assess the performance of predicting facial expression coefficient $\beta$ for given situation or context. The expression-context pairs in SFEW-C are used for evaluation. We measure the error in the prediction in terms of the distance between 100-dimensional coefficient vectors of facial expressions in the L2 and FD (Fréchet Distance) metrics. L2 measures the error in the L2 distance between the coefficient vectors. FD measures the statistical distance between the distributions of coefficient vectors. Since there exist no other methods for contextual expression generation, we perform a self-evaluation adopting the multiple axes-based evaluation framework discussed in previous studies [32, 33] as follows. Tab. 2 shows the comparisons with Random and Mean baselines where we compare the distance between the prediction and a randomly selected sample from test set (Random), or the centroid of the test set (Mean), or the ground truth (Ours). Our model demonstrates a significant improvement over the Random baseline, suggesting that it has successfully learned the relationship between context and facial expressions. Furthermore, our model demonstrates substantially

| Baseline | L2 ↓ | FD ↓ |
|---|---|---|
| *situation* | | |
| Random | 0.34± 0.0102 | 1.98 ± 0.2263 |
| Mean | 0.14± 0.0047 | 14.38 ± 0.4688 |
| **Ours** | **0.07± 0.0055** | **0.81± 0.0711** |
| *quote* | | |
| Random | 0.31± 0.0093 | 2.42 ± 0.1982 |
| Mean | 0.12 ± 0.0042 | 11.76 ± 0.4178 |
| **Ours** | **0.1± 0.0048** | **1.95± 0.1305** |

Table 2. Prediction Performance of Facial Expression. We measure the errors in the predicted facial expression given two context types: quotes and situations. There are two baselines. **Random:** distance between the prediction and a randomly selected sample from the test set (excluding the ground truth). **Mean:** distance between the prediction and the average of samples in the test set. **Ours:** distance between the prediction and the ground truth.

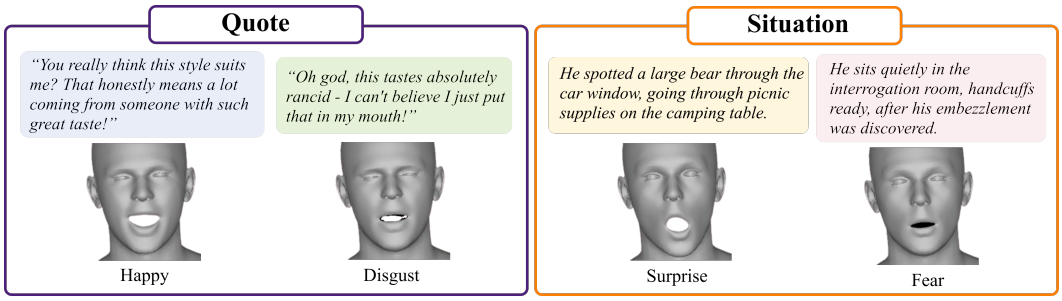


Figure 5. **Contextual Alignment of Emotional Expressions.** Generation results of facial expressions in response to quotes and situations.

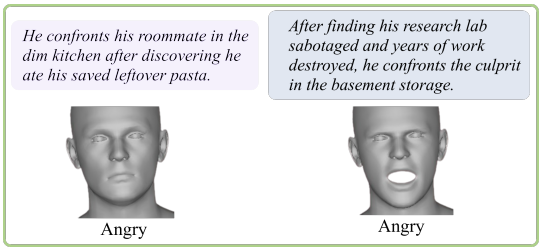


Figure 6. **Contextual Variation within Emotions.** Contextual variations of angry facial expressions in different situations.

| Situation Generation | Expression Generation | UAR | WAR |
|---|---|---|---|
| × | × | 81.42 | 84.98 |
| ✓ | × | 79.34 | 82.48 |
| × | ✓ | 88.33 | 89.40 |
| ✓ | ✓ | **93.21** | **94.89** |

Table 3. **Ablation study.** Performance comparison of emotion inference with different dataset combinations. ✓ (resp. ×) means the model is trained with (resp. without) the dataset.

superior FD scores compared to the Mean baseline, preserving the distinctive elements of specific facial expressions to effectively capture fine facial feature details.

### 4.3. Qualitative Results

**Context-to-Emotional Expression Mapping.** As shown in Fig. 5, ContextFace can generate facial expressions aligned with each emotion by taking textual contexts, such as quotes or situations, as input.

**Intra-Emotion Expressional Diversity.** In Fig. 6, we observe that ContextFace is capable of generating different expressions within the same emotion depending on the situational context. Facial expressions involving anger were generated for different situations. The left image displays a milder expression in response to a minor annoyance, while the right image shows a significantly more intense expression of rage in response to an infuriating situation.

**Speculative Expression Generation through Emotional Reasoning.** We show that our model can generate logically valid situations when given a quote for ambiguous facial expressions. Additionally, based on these generated situations, the model can create appropriate facial expressions. In Fig. 7, we observe that our model correctly infers the emotions of the speaker, even with the same image and different quotes. Moreover, the model can infer the emotional state or facial expression of not only the speaker, but also the other person. As shown in the example, while the quote comes from a woman, the model can *speculatively* generate the facial expression of her boyfriend through an interactive prompting. This suggests that our model has emotional reasoning capabilities to understand interpersonal relationships and infer corresponding emotions or facial expressions.

**Facial expression generation for occluded faces.** As shown in Fig. 8, our model can infer the situation based on contexts (image and quote), and can generate plausible facial expressions, even when the face is intentionally and completely occluded. Our model successfully infers situations and generates appropriate expressions, demonstrating it has comprehensively learned the relationships among context, emotions, and facial expressions.

### 4.4. Ablation study

**Impact of Dataset Combinations.** As shown in Fig. 3, we used three datasets to jointly train ContextFace, i.e., two context-augmented datasets: Situation Generation Dataset, Expression Generation Dataset and one VQA Dataset. We measured the model's ability to infer emotions with or without context-augmented datasets where VQA dataset is used by default. As shown in Tab. 3, the model performed the best when all three datasets are used. The model without Situation Generation dataset performs marginally below the baseline LLaVA-1.5, indicating that balanced training with generalizable datasets rather than task-specific ones is essential for maintaining emotion inference capabilities.

**Impact of Model Scale.** The LLaVA-1.5-13B model demonstrated higher performance in both UAR and WAR compared to LLaVA-1.5-7B, as shown in Tab. 4. This suggests that as model size increases, emotional reasoning and expression generation capabilities improve. However, the

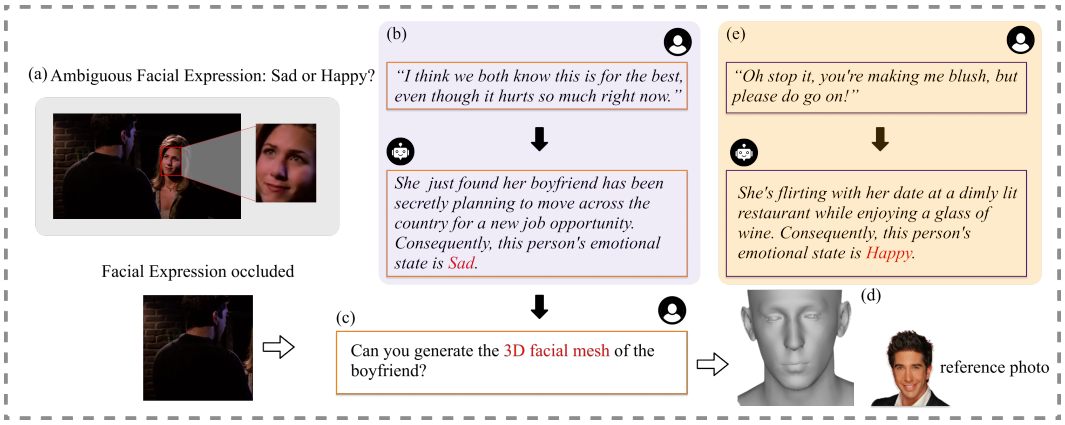


Figure 7. **Speculative Expression Generation through Emotional Reasoning.** (a) The expression in the input image is ambiguous. (b) Given the woman's quote, our model can generate a plausible situation from the image along with the proper emotion. (c) The model is prompted to **speculate** the expression of the man whose face is not visible. (d) The model uses emotional reasoning to generate a sad face which is matched to a reference photo using INFERNO [7]. (e) When prompted with the same input image and a different quote, the model correctly predicts the woman's emotion as "Happy".

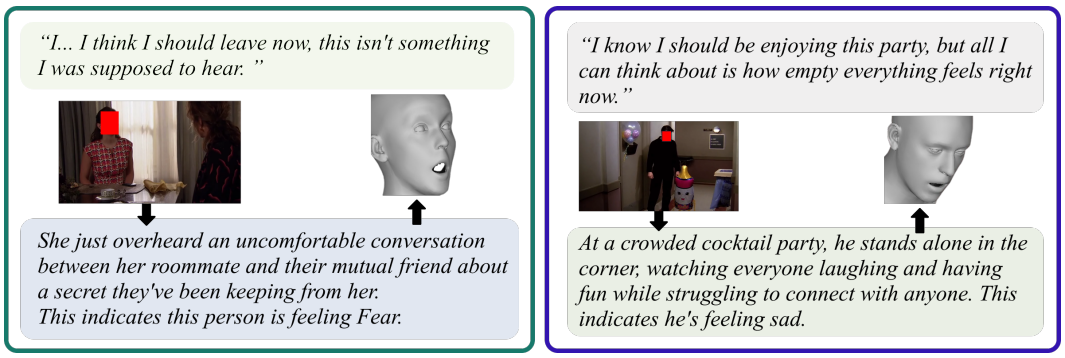


Figure 8. Generating situations and expressions from intentionally occluded faces and quotes.

| Method | UAR (%) | WAR |
|---|---|---|
| LLaVA-1.5-7B | 92.00 | 91.70 |
| LLaVA-1.5-13B | **93.21** | **94.89** |

Table 4. **Ablation study.** Comparing emotion inference performance between LLaVA-1.5-13B and LLaVA-1.5-7B.

7B model maintains performance above 90% of the 13B model, making it a computationally efficient alternative.

## 5. Conclusion

In this paper, we proposed ContextFace, a model that understands the deep association between contextual emotions and facial expressions and generates context-aligned facial expressions. Furthermore, we introduce new CAER-S-C and SFEW-C datasets that enable the proposed contextual learning process. To our knowledge, we are the first to extend MLLMs beyond textual emotion analysis to direct generation of facial expressions, bridging the gap between language-driven emotion interpretation and visual synthesis. The proposed integration will become increasingly valuable as AI services evolve towards human-like interactions through emotionally aware and visually expressive communication.

## References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 2

[2] Anthropic. Claude-3.5-sonnet-20241022. https://www.anthropic.com, 2024. Large Language Model API. 3

[3] Zehranaz Canfes, M Furkan Atasoy, Alara Dirik, and Pinar Yanardag. Text and image guided 3d avatar generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4421–4431, 2023. 2

[4] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning, 2024. 1, 2, 5

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5, 6

[6] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 2

[7] Radek Daněček, Timo Bolkart, and Wojciech Zielonka. Inferno. https://github.com/radekd91/inferno, 2020. 1, 2, 4, 8

[8] Radek Daněček, Kiran Chhatre, Shashank Tripathi, Yandong Wen, Michael Black, and Timo Bolkart. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, page 1–13. ACM, 2023. 2

[9] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011. 2, 3, 5

[10] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4769–4776. IEEE, 2024. 2

[11] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. Contextual emotion recognition using large vision language models, 2024. 1

[12] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. ChatPose: Chatting about 3d human pose. In *CVPR*, 2024. 2, 3

[13] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 2

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 5

[15] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 3

[16] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoonn Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 2, 3, 5

[17] Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. Large vision-language models as emotion recognizers in context awareness, 2024. 1

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2

[19] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans, 2017. 1, 2, 4

[20] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*. 1

[21] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning, 2024. 1

[22] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning, 2023. 5

[23] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *CoRR*, 2023. 2

[24] Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. Affectgpt: Dataset and framework for explainable multimodal emotion recognition, 2024. 1

[25] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 1

[26] Yihong Lin, Liang Peng, Jianqiao Hu, Xiandong Li, Wenxiong Kang, Songju Lei, Xianjia Wu, and Huang Xu. Emoface: Emotion-content disentangled speech-driven 3d talking face with mesh attention, 2024. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 5, 6

[30] Shu Liu, Yan Xu, Tongming Wan, and Xiaoyan Kui. A dual-branch adaptive distribution fusion framework for real-world facial expression recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[31] Zhentao Liu, Min Wu, Weihua Cao, Luefeng Chen, Jianping Xu, Ri Zhang, Mengtian Zhou, and Junwei Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE CAA J. Autom. Sinica*, 4(4):668–676, 2017. 1

[32] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion, 2022. 6

[33] Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen?, 2023. 6

[34] Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. Representation learning and identity adversarial training for facial behavior understanding, 2024. 1

[35] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[38] Niyati Rawal and Ruth Maria Stock-Homburg. Facial emotion expressions in human–robot interaction: A survey. *International Journal of Social Robotics*, 14(7):1583–1604, 2022. 1

[39] Arnab Kumar Roy, Hemant Kumar Kathania, Adhitiya Sharma, Abhishek Dey, and Md. Sarfaraj Alam Ansari. Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition. *IEEE Signal Processing Letters*, pages 1–5, 2024. 1

[40] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. Emotion recognition for human-robot interaction: Recent advances and future perspectives. *Frontiers in Robotics and AI*, 7:532279, 2020. 1

[41] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024. 2

[42] Qwen Team. Qwen2.5-vl, 2025. 2, 5, 6

[43] Chengpeng Wang, Li Chen, Lili Wang, Zhaofan Li, and Xuebin Lv. Qcs: Feature refining from quadruplet cross similarity for facial expression recognition, 2025. 1

[44] Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. Vllms provide better context for emotion understanding through common sense reasoning, 2024. 1, 2

[45] Bohao Xing, Zitong Yu, Xin Liu, Kaishen Yuan, Qilang Ye, Weicheng Xie, Huanjing Yue, Jingyu Yang, and Heikki Kälviäinen. Emo-llama: Enhancing facial emotion understanding with instruction tuning, 2024. 1

[46] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations, 2024. 1, 2