# Anatomy of a Feeling: Narrating Embodied Emotions via Large Vision-Language Models

**Mohammad Saim, Phan Anh Duong, Cat Luong, Aniket Bhanderi, Tianyu Jiang**
University of Cincinnati
saimmd@mail.uc.edu, tianyu.jiang@uc.edu

## Abstract

The embodiment of emotional reactions from body parts contains rich information about our affective experiences. We propose a framework that utilizes state-of-the-art large vision-language models (LVLMs) to generate **E**mbodied **LVLM** **E**motion **Na**rratives **(ELENA)**. These are well-defined, multi-layered text outputs, primarily comprising descriptions that focus on the salient body parts involved in emotional reactions. We also employ attention maps and observe that contemporary models exhibit a persistent bias towards the facial region. Despite this limitation, we observe that our employed framework can effectively recognize embodied emotions in face-masked images, outperforming baselines without any fine-tuning. ELENA opens a new trajectory for embodied emotion analysis across the modality of vision and enriches modeling in an affect-aware setting.

## 1 Introduction

Emotion recognition has been widely studied in both natural language processing and computer vision, with applications ranging from sentiment analysis to human-robot interaction. While facial expressions are often considered the primary channel for conveying emotion, they are not always reliable—for example, when faces are distant, obscured, or deliberately masked. In such cases, the body (torso and limbs) frequently communicates emotion more clearly than the face. For instance, observers often struggle to distinguish between positive and negative feelings from isolated facial expressions, but can do so accurately from bodily cues (Aviezer et al., 2012). These findings highlight that bodily reactions are central to how emotions are expressed and perceived (Fuchs and Koch, 2014); yet, this important dimension remains underexplored in much of the current emotion recognition research.


Figure 1: Illustration of embodied emotion for multi-modal analysis.

The concept of *Embodied Emotion* provides a richer framework for understanding the affective states and making holistic use of all bodily parts for emotion recognition. Rooted in psychology and cognitive science (Niedenthal, 2007; Barsalou, 2008), embodiment theories argue that emotions are not purely mental phenomena but are deeply intertwined with bodily experience and physical response. In parallel, computational work has shown that embodied signals are also crucial for interpreting emotions in text, where bodily actions and sensations often convey affect beyond lexical sentiment (Zhuang et al., 2024; Duong et al., 2025). Together, these insights suggest that posture, gesture, and physiological change are integral to emotional meaning (Barsalou and Wiemer-Hastings, 2005; Niedenthal et al., 2014). Empirical studies further show systematic correlation between specific posture and mood—for example, an upright stance can indicate elevated mood, whereas a slumped shoulder correlates with depression and low arousal (Wilkes et al., 2017).

In this work, we propose to bridge the gap of embodied emotions research with vision modality



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23480–23495
November 4-9, 2025 

by utilizing prevalent Large Vision-Language Models (LVLMs) and their capabilities. While prior approaches to the task of emotion recognition have often relied on models trained and fine-tuned on a variety of features (Tzirakis et al., 2017; Luna-Jiménez et al., 2021), our work utilizes LVLMs in a zero-shot setting to output multi-layered textual descriptions for embodied emotions, as shown in Figure 1. These models do not receive any task-specific training, but are guided by carefully crafted instructions that elicit the underlying emotional manifestations in images. As far as we are concerned, this is the first work in the area of multimodal analysis for embodied emotions. We release the code and outputs of ELENA publicly.[1] We highlight our contributions as below:

1. We introduce ELENA, a novel framework that utilizes structured prompting to guide LVLMs in generating multi-layered narratives of embodied emotion. Our approach moves beyond simple classification to a more explanatory form of affective analysis, and serves as a novel tool for qualitatively probing a model's emotional reasoning.

2. Through systematic face-masking methodology and attention analysis, we provide the empirical evidence that LVLMs exhibit a critical *attention adaptation failure* when recognizing embodied emotion—they do not redirect visual focus to informative bodily regions when facial details are masked, exposing a fundamental vulnerability in visual reasoning.

3. We show that the framework can effectively overcome limitations of LVLMs, achieving credible performance improvements. It suggests that while the model's autonomous visual systems are brittle, their latent knowledge of non-facial emotional signals can be effectively elicited with targeted guidance.

## 2 Related Work

Emotion understanding is a long-standing task in NLP, rooted in the field of affective computing research (Firdaus et al., 2020; Zhuang et al., 2020; Deng and Ren, 2021; Zheng et al., 2023). Psychology and neuroscience theories have long argued that bodily states (e.g., posture and physiological reactions) are integral to emotions (Niedenthal, 2007; Barsalou, 2008; Li et al., 2021).

Prior NLP work has begun to consider the physical expressions of emotion in text. Kim and Klinger (2019) analyzed how authors depict characters' body movements and sensations (e.g., frowning and heart pounding) to imply emotion, and Casel et al. (2021) identified emotion-related components based on a cognitive model. The rise of large language models (LLMs) has prompted a re-examination of emotion understanding in NLP (Liu et al., 2024b; Sabour et al., 2024). However, purely text-based models may overlook nonverbal signals that are critical to emotional expressions (Lian et al., 2025).

Vision-Language Models (VLMs) extend emotion understanding with visual context and commonsense reasoning (Zhang et al.; Xenos et al., 2024), yet evaluations on evoked emotions reveal biases towards specific categories and prompt sensitivity (Bhattacharyya and Wang, 2025). Visual affective computing remains dominated by facial cues (Wang et al., 2024b). In parallel, body-centric approaches model affect via posture and gesture features (Noroozi et al., 2018) and skeleton-based relations among body parts (Shen et al., 2019; Wu et al., 2024; Lu et al., 2025), but progress is limited by scarce, richly annotated datasets. Departing from these facial and skeleton-centric pipelines, we address embodied emotion in images by instructing LVLMs to generate structured, body-part-grounded narratives alongside an Ekman emotion label, and we explicitly disentangle facial versus non-facial evidence via face masking and attention-map diagnostics.

The recently introduced CHEER dataset (Zhuang et al., 2024) represents a first step in this direction—it contains 7,300 instances of body part mentions in narratives, labeled for whether they signal an underlying emotion, e.g., "her hands trembled with fear." This work was carried forward by Duong et al. (2025), who converted the previous binary task to Ekman's six emotion categories, along with improved metrics compared to supervised methods using best-worst scaling. However, the task of embodied recognition remains understudied in vision-language models, as well as the evaluation of the influence of facial parts of the body on emotional expression. Inferring affect purely from bodily posture, movements, and physiological state in images remains an open challenge, which motivates the focus of our study.

[1] https://github.com/cincynlp/ELENA

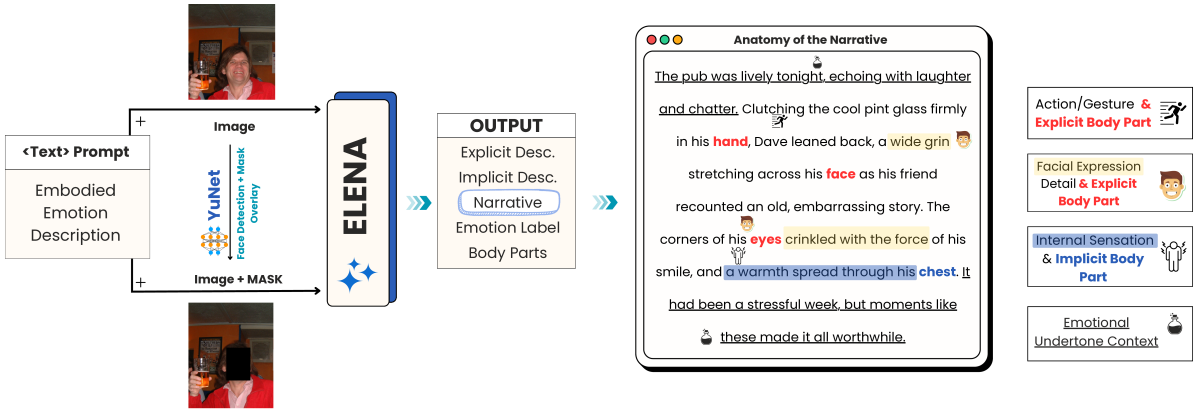


Figure 2: ELENA framework architecture for embodied emotion analysis. The system utilizes LVLMs in zero-shot settings to process structured prompts with images (masked or unmasked), generating all outputs in a single forward pass: emotion labels, explicit descriptions, implicit descriptions, and body parts. Narratives are formed by combining explicit and implicit descriptions, with the right panel illustrating the anatomization of a narrative into its embodied components.

## 3 Methodology

We propose a zero-shot structured prompting approach that utilizes large vision-language models to systematically interpret emotions through bodily expressions. Our methodology includes detailed prompt design, face-masking experiments, attention analysis, and comparative evaluations to robustly assess the models' capability in recognizing embodied emotions.

### 3.1 Datasets and Usage Scheme

We perform our task on three image datasets:

**BESST** (Bochum Emotional Stimulus Set) (Thoma et al., 2013) consists of 1,129 (565 frontal-body + 564 averted-body) high-quality images collected in a lab-controlled environment. An example is shown in Appendix B (Figure 13). Participants for this research were instructed to enact a specific emotion by posing as if they were experiencing it, producing explicit bodily expressions. Each person has their face masked and photographed in both frontal and averted views. BESST includes annotations aligned with Ekman's six basic emotions (Ekman, 1992), along with an additional *Neutral* category.

**HECO** (Human Emotions in Context) (Yang et al., 2022) dataset contains 9,385 images with 19,781 annotated agents depicting people in natural settings. Emotions are labeled using the six Ekman categories, plus two additional positive states: *Peace* and *Excitement*. For consistency, we remap *Peace* to *Neutral* and *Excitement* to *Happiness*, aligning HECO with the seven-category taxonomy used in BESST.

**EMOTIC** (Kosti et al., 2019) dataset includes 23,571 images with 34,320 annotated people in everyday scenarios, annotated with multi-label emotion tags drawn from a fine-grained set of 26 categories. To enable consistent, cross-dataset comparisons, we map EMOTIC's labels to the same seven-category Ekman-based scheme used for BESST and HECO. This normalization ensures a more equitable, "apples-to-apples" evaluation of LVLM performance across datasets. The mapping details and justification are provided in Table 5, Appendix D. For our experiments, we further simplify the evaluation by selecting the dominant emotion per instance, determined by matching gold body bounding boxes and masked face coordinates with high confidence overlap.

### 3.2 Framework Overview

The ELENA framework is illustrated in Figure 2. The pipeline begins with an input image (either masked or unmasked) that an LVLM processes with a text input grounded in the definition of embodied emotion. The input comprises a request for a set of varied features related to the person's emotional state and body parts, for the output (Table 1). In this study, we focus on a salient person—the most notable individual in each image. The model predicts a single label, thereby maintaining uniformity across all images. However, the framework can be modified for multi-label output. The

complete response structure is presented in Table 1. Our framework can be formalized as:

$$\text{ELENA} : I \rightarrow (L, E_e, E_i, N, B) \quad (1)$$

where $I$ represents the input image, $L \in \{$*Happiness, Sadness, Anger, Fear, Disgust, Surprise, Neutral*$\}$ is the predicted emotion label following Ekman's taxonomy, $E_e$ denotes embodied descriptions focusing on visible body parts, $E_i$ represents implicit descriptions capturing internal sensations, and $B$ contains the set of extracted body parts. $N$ is the narrative that we instruct the model to produce, weaving these embodied cues with scene context.

| Output | Description |
|---|---|
| Label $L$ | Single emotion (Ekman's six labels + Neutral). |
| Explicit $E_e$ | Visible body parts and their emotional expression. |
| Implicit $E_i$ | Internal sensations and body parts which are not visible. |
| Narrative $N$ | Description necessarily containing the body part influencing the emotion. Involves an emotional undertone and scene setting. |
| Body Parts $B$ | Identifiable body parts involved in emotion expression. |

Table 1: Components of structured textual output from LVLMs.

We evaluated a suite of proprietary and open-source models, notably Gemini 2.5 Flash (Gemini Team, 2025) and Gemma-3-12B (Gemma Team, 2025) and Llama-3.2-11B/90B (Grattafiori et al., 2024) respectively. Due to computational constraints, we only tested Llama's 90B version on the smaller dataset (BESST, Appendix C). Each narrative follows the definitions in Table 1, with models free to use varied vocabulary within the structural constraints.

For each image, we also query the model to output one of the six corresponding Ekman emotions that match the embodied emotion displayed. If the image does not elicit a clear display of emotion, the model will predict the *Neutral* label. For narratives, more emphasis was placed on describing body-related expressions without explicitly mentioning the emotion to steer the model away from


Figure 3: Example of ELENA compared to generic emotion response. Green highlights correctly predicted labels, while red indicates an incorrect prediction. The red box represents the coordinates of the person's body; the green bounding box is generated using YuNet. The sample is from the EMOTIC dataset.

simply restating the emotion. Figure 3 shows an example response from the ELENA compared to a naive prompt. While the generic prompt captures the broad context of the image, it fails to isolate the fine-grained embodied expressions and lacks specificity in its description.

In the *narrative* response, for the context, we use words like "scene" or "story" to encourage a portrayal-style explanation, allowing the model to come up with the reasoning as to what event and bodily reaction occurred that made the person depict the emotion expressed in the images. Finally, we explicitly request a list of physical features or body parts; this ensures that even if the descriptions and narrative already contain them, the model

isolates them for downstream analysis. Doing this helps us in reliably extracting which body parts the model focuses on for extended statistical comparisons. As a final note, ELENA employs a single, unified prompt that generates emotion labels, descriptions, and narratives simultaneously in one forward pass, ensuring a coherent and jointly reasoned output. The general template for the prompts is available in Appendix B.

### 3.3 Face Masking

It is essential to consider other bodily expressions than the facial region as stimuli for emotion recognition, which is also the central belief of this paper. Furthermore, in many real-world situations, facial expressions may not be fully visible or recognizable, making it challenging to interpret emotions. Therefore, we implement a face masking approach to evaluate whether LVLMs can effectively identify emotions from bodily expressions and to assess the robustness of our pipeline.

The masking procedure is formalized as:

$$I_m(x,y) = \begin{cases} M, & \text{if } (x,y) \in \bigcup_{f \in F} R_f \\ I_o(x,y), & \text{otherwise} \end{cases}$$

The above equation depicts our facial masking procedure, where $I_m(x,y)$ represents the pixel value at coordinates $(x,y)$ in the masked image, $I_o(x,y)$ represents the original pixel value before masking, and $M$ is the mask color. The set $F$ contains all faces detected by YuNet (Wu et al., 2023), while $R_f$ denotes the region of the facial area. For more details on the masking framework, refer to Appendix E. We opted for complete rather than partial face masking to create a more rigorous test of embodied emotion recognition, ensuring models cannot rely on any residual facial cues. Furthermore, partial masking of specific features, such as the eyes or mouth, is often unreliable in datasets like HECO and EMOTIC, which feature many non-frontal views and distant shots.

### 3.4 Attention Visualization

To investigate which regions of input images the LVLMs prioritize when processing emotion-related queries, we conducted an attention analysis using Llama-3.2-11B-Vision. We selected this model due to its open-weight architecture, computational efficiency, and widespread adoption in the research community, which promotes reproducibility. We used a bare-bone prompt ("What is the emotion displayed in the image?") and ran a forward pass through the model to extract attention patterns from the cross-attention layers. We then computed attention heatmaps by averaging attention scores across all tokens within each cross-attention layer. To ensure computational tractability and avoid image tiling artifacts, we resized all images to the encoder's default resolution of 560×560 pixels. From each layer, we extract the first 1,600 visual tokens corresponding to the first image tile, where each token represented a 14×14 pixel patch of the input image. We then generated layer-specific attention heatmaps by computing the average attention score across all attention heads. These scores are finally mapped back to their corresponding spatial locations in the original image to create interpretable visualizations of the model's attention patterns.

## 4 Results and Analysis

With our methodology in place for eliciting embodied emotional indicators from LVLMs, we move to the experimental analysis and results. We evaluate ELENA's performance through two key comparisons: first, we assess how our framework performs against baseline naive prompts across different models and datasets, and second, we examine how face masking affects the model behavior and attention patterns in embodied emotion recognition tasks. We conclude with attention visualization analysis, breakdowns of emotion-specific performance and comparison to a modular baseline.

### 4.1 ELENA vs. Naive Prompt Performance

We compare ELENA's structured prompting approach against a baseline naive prompt to assess whether explicitly incorporating embodied emotion definitions enhances model recognition capabilities. Table 2 presents the performance results across three datasets for multiple LVLMs. ELENA consistently outperforms naive prompts in twelve out of fifteen experimental configurations, with the remaining three showing only narrow performance gaps. Among the evaluated models, Gemini 2.5 Flash achieves the highest F1 scores across most datasets and conditions, demonstrating robust overall performance. The Gemma and Llama models follow closely, yielding comparable results. Further analysis on the BESST dataset (Appendix C) confirms that Gemini 2.5 Flash surpasses even the larger Llama-3.2-90B model.

The performance improvements vary depending

| Dataset | Model | Naive Prompt | | | ELENA | | |
|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 |
| **HECO**$_{Normal}$ | Gemini 2.5 Flash | 38.4 | 30.8 | 30.6 | 45.8 | 31.7 | 34.5$_{\uparrow 3.9}$ |
| | Llama-3.2-11B | 33.2 | 26.6 | 26.2 | 37.1 | 29.3 | 29.5$_{\uparrow 3.3}$ |
| | Gemma-3-12B | 37.0 | 31.7 | 30.6 | 38.4 | 30.1 | 30.1$_{\downarrow 0.5}$ |
| **HECO**$_{Masked}$ | Gemini 2.5 Flash | 27.4 | 19.2 | 16.9 | 33.6 | 34.7 | 31.5$_{\uparrow 14.6}$ |
| | Llama-3.2-11B | 26.5 | 17.7 | 14.4 | 30.9 | 24.9 | 22.5$_{\uparrow 8.1}$ |
| | Gemma-3-12B | 29.2 | 18.2 | 14.3 | 32.1 | 26.2 | 23.6 $_{\uparrow 9.3}$ |
| **EMOTIC**$_{Normal}$ | Gemini 2.5 Flash | 43.9 | 25.1 | 28.6 | 42.8 | 22.6 | 27.7$_{\downarrow 0.9}$ |
| | Llama-3.2-11B | 44.2 | 20.0 | 23.3 | 33.0 | 21.0 | 21.5$_{\downarrow 1.8}$ |
| | Gemma-3-12B | 45.7 | 23.4 | 26.5 | 41.3 | 22.2 | 28.0 $_{\uparrow 1.5}$ |
| **EMOTIC**$_{Masked}$ | Gemini 2.5 Flash | 28.0 | 15.9 | 15.6 | 41.1 | 20.2 | 26.0$_{\uparrow 10.4}$ |
| | Llama-3.2-11B | 35.9 | 16.1 | 17.1 | 32.6 | 17.5 | 20.7$_{\uparrow 3.6}$ |
| | Gemma-3-12B | 39.3 | 17.8 | 17.2 | 39.1 | 19.7 | 24.8$_{\uparrow 7.6}$ |
| **BESST**$_{Masked}$ | Gemini 2.5 Flash | 64.2 | 55.7 | 51.9 | 64.8 | 58.0 | 52.7$_{\uparrow 0.8}$ |
| | Llama-3.2-11B | 36.8 | 30.2 | 22.2 | 48.3 | 44.9 | 37.8$_{\uparrow 15.6}$ |
| | Gemma-3-12B | 63.6 | 39.0 | 32.3 | 55.5 | 46.9 | 41.6$_{\uparrow 9.3}$ |

Table 2: Performance comparison across datasets and prompt types. Values are presented as macro-averaged (%). Subscripts denote image types: $_{Normal}$ for unmasked images and $_{Masked}$ for images with faces masked. $_{\uparrow}$ indicates absolute F1 performance increase with EE prompt compared to Naive prompt, while $_{\downarrow}$ indicates absolute decrease.

on the characteristics of the dataset. On HECO, ELENA shows consistent gains ranging from 3.3 to 14.6 F1 points depending on the model and masking condition. For EMOTIC, results are more mixed, with slight performance drops in some unmasked scenarios, which we attribute to the dataset's rich contextual cues and multiple-person scenarios that can interfere with focused embodied emotion analysis. On BESST, improvements are substantial, with ELENA providing up to 15.6 F1 point gains. These results demonstrate that structured prompting effectively guides LVLMs toward embodied emotion recognition by directing attention to bodily expressions rather than relying solely on facial features or contextual cues. The framework's effectiveness varies with dataset quality and image complexity, but it consistently provides meaningful improvements over generic emotion recognition approaches.

**Takeaway:** ELENA consistently outperforms naive emotion recognition prompts across multiple models and datasets, with performance gains ranging from modest (3-4 F1 points) to substantial (15+ F1 points).

### 4.2 Impact of Face Masking on Embodied Emotion Recognition

To evaluate ELENA's ability to recognize emotions from bodily expressions independent of facial indicators, we implemented a systematic face masking technique using the YuNet detection model, followed by complete occlusion of facial regions. This analysis reveals how models adapt when they rely exclusively on embodied indicators. We essentially duplicate the experiment for each model, with the only change being that the input image now has the face region obscured. We ensure that the prompt remains the same, except that it specifies not to focus on the facial area or the masked region.

We report that performance improvements under masking conditions are most pronounced, with ELENA achieving up to 15.6 F1-point gains over naive prompts (indicated by blue up arrows in the masked sections in Table 2). This suggests that while LVLMs struggle with autonomous attention redirection when faces are masked, structured guidance can effectively compensate for this limitation. From analyzing the performance on masked images, we identify two key insights. First, ELENA succeeds in shifting focus to mention body parts apart from the facial region. Table 3 depicts a substantial shift in highlighting body parts from various anatomical areas of the body. The model's description shifts to focus on the body's limbs, with a primary concentration on the hands, arms, shoulders, and legs. It also sees a percentage increase in the body's torso. Although not in the top ten, narratives still mention facial parts, suggesting that

| Normal Images | % | Face-Masked Images | % |
|---|---|---|---|
| eye | 14.45 | **hand** | **21.82** |
| mouth | 9.66 | **arm** | **18.62** |
| heart | 9.29 | shoulder | 11.76 |
| hand | 9.15 | heart | 5.87 |
| chest | 5.60 | chest | 4.79 |
| arm | 3.95 | torso | 4.49 |
| shoulder | 3.92 | leg | 4.24 |
| face | 3.36 | finger | 4.16 |
| lip | 2.98 | body | 3.60 |
| head | 2.98 | head | 1.80 |
| **Anatomical Regions** | | | |
| Head/Face | 40.63 | **Limbs** | **50.18** |
| Limbs | 17.14 | Torso | 22.98 |
| Internal/Conceptual | 13.88 | Other | 14.07 |
| Torso | 13.21 | Internal/Conceptual | 8.05 |
| Other | 15.13 | **Head/Face** | **4.71** |

Table 3: Top ten body part mentions (%) by ELENA on face masked versus unmasked images. The lower section categorizes all mentions of body parts. The distribution is based on the HECO image dataset. **Bold** highlights dramatic shifts.

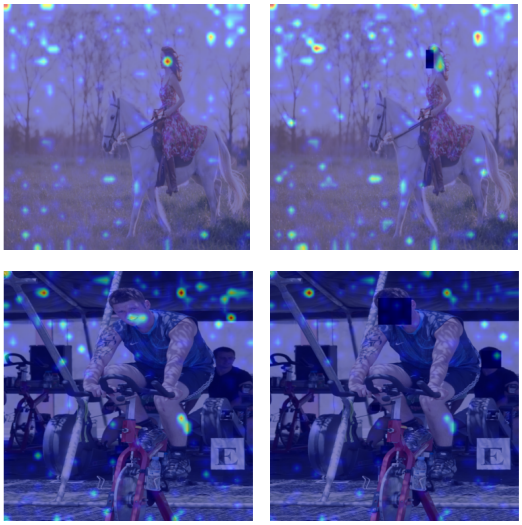

Figure 4: Visualization of Llama-3.2-11B's cross-attention layer 13 for HECO images.

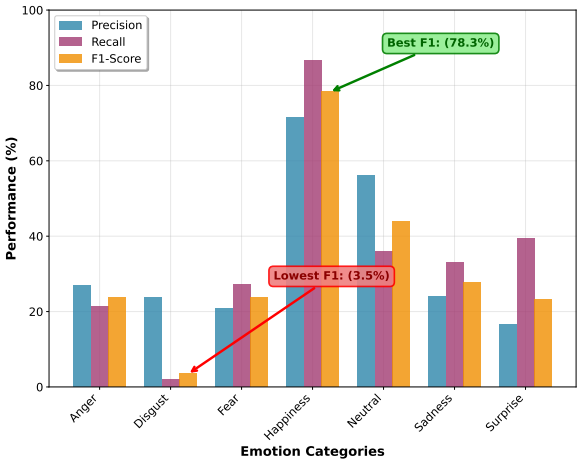


Figure 5: Per-Category Analysis of each Emotion Label. Results are from the Gemini-2.5-Flash Model on the HECO dataset.

models are biased in concentrating on facial features. Other reasons, discovered through manual inspection, included issues with the YuNet model, specifically the failure to mask faces when a scene is filled with a crowd with averted bodies. Another observation is that with the faces masked, models lean towards more fine-grained details surrounding the context.

Second, multiple images are not left with ample context after masking. This can be directly attributed to the images in the dataset, particularly those from the EMOTIC and HECO datasets. Images might be captured in such a way that the facial region is the center of interpretation for the model. Other body parts were rarely visible for the ELENA framework to be effective. Once the face is masked, the image produces very ambiguous and non-coherent narratives, which in turn lead to incorrect label predictions. This highlights a major issue in prevalent emotion-recognition image datasets and further underscores the need for a new, qualitatively filtered dataset. Further arguments on this result are discussed in Appendix D.

**Takeaway:** Face masking reveals that ELENA successfully redirects model attention from facial features to alternative body parts, thus achieving better performance than baselines.

### 4.3 Attention Map Analysis

Through analysis of the attention maps obtained, we identified distinct patterns in Llama-3.2-11B-Vision's behavior during emotion recognition tasks. The model exhibits a pronounced bias toward facial features, particularly the mouth region, in cross-attention layers 13-28 when processing unmasked images. When presented with masked faces, the model demonstrates two problematic response patterns, neither of which constitutes effective adaptation. As illustrated in Figure 4, the model either continues to direct attention toward the masked facial region and its immediate surroundings, essentially attempting to extract information from areas that no longer provide meaningful emotional cues, or it diverts attention away from the face entirely without increasing focus on alternative emotional indicators such as body language. Crucially, the model's attention to non-facial regions remains at


Figure 6: Failure cases of predicting the label *Disgust* in both unmasked and masked conditions. Examples are from the HECO dataset.

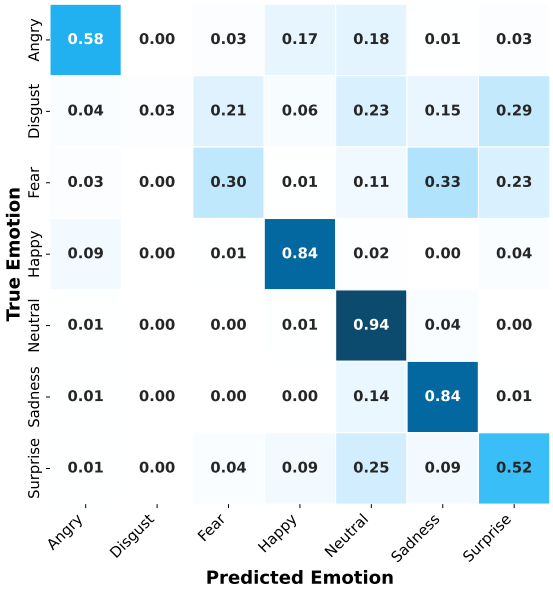


Figure 7: Heatmap visualization of true versus predicted labels of the BESST dataset. Results are from Gemini-2.5-Flash.

baseline levels, comparable to those in unmasked scenarios, indicating a failure to recognize and compensate for the loss of facial information. This behavior could explain the performance degradation observed in masked images, as shown in Table 2.

The findings indicate a fundamental bias in the employed model's attention mechanism: the model does not redirect its focus to non-facial emotional indicators when primary facial indicators are unavailable. Rather than attending to alternative information sources such as body posture or gesture, the model exhibits erratic attention patterns that compromise its effectiveness in masked scenarios. This limitation warrants further investigation, as emotion is fundamentally grounded in the whole body through embodied emotion, extending beyond facial expressions alone. Models that rely predominantly on facial features may demonstrate only a superficial understanding of human emotional expression, suggesting a critical gap in current vision-language model architectures that future research should address to achieve more robust emotion recognition capabilities.

**Takeaway**: The employed LVLM (Llama-3.2-11B) exhibits fundamental attention adaptation failure, maintaining facial bias even when faces are masked, and failing to compensate by increasing focus on informative body regions, revealing an architectural limitation that explains the performance degradation in embodied emotion recognition.

### 4.4 Emotion-Specific Performance Analysis

Figure 5 shows the model's performance for each emotion category and identifies substantial disparities between them. *Happiness* achieves near-ceiling scores, reaching approximately 90% for the best model on the HECO dataset. This high accuracy can be attributed primarily to the distinctive visual cues associated with *Happiness*—particularly smiling faces and highly expressive body language.

In contrast, the recognition of *Disgust* remains notably problematic, with frequent misclassifications. An example is shown in Figure 6. Most of the labels get majorly misclassified into three other labels: *Fear, Sadness, and Surprise*, along with the *Neutral* tag, as evident in Figure 7. The difficulty arises due to definitional ambiguity and subtlety in visual representation, making disgust inherently challenging for models to identify reliably. Another occurrence we came across was the model's refusal to provide an answer to the input query where the gold annotation for the image was the *Disgust* label. We can attribute this partially to the nature of the image; however, it could also be due to the sensitivity of the emotion, which triggers the safety measures of LVLMs. Therefore, this issue is more likely due to the strictness of content filtering rather than a problem with the model's understanding or inappropriate images. Furthermore, this emotion is significantly underrepresented in the datasets, which exacerbates the difficulty in classification by providing insufficient training examples for equitable learning.

Additionally, face masking notably increases the prediction frequency of the *Neutral* category. This

| Method | Precision | Recall | F1 |
|---|---|---|---|
| LLaVA + Fine-tuned BERT | 22.4 | 21.5 | 20.8 |
| **ELENA** | 45.8 | 31.7 | 34.5 |

Table 4: Performance comparison of the image-captioning-plus-classifier (LLaVA + BERT) baseline to Gemini 2.5 Flash on the HECO dataset.

phenomenon occurs when substantial information loss results from the obscuring of informative facial expressions, leaving the models dependent solely on body posture cues. Given that body postures are typically less distinctive and more ambiguous, the models default to predicting *Neutral* in uncertain cases, thus reducing recall for expressive emotional categories. The predicted emotion labels are also scrutinized using a dimensional valence-arousal-dominance (VAD) model. The analysis regarding the VAD scores is shown in Appendix A.

**Takeaway**: Our analysis reveals significant performance disparities across emotions; models excel at identifying specific labels, e.g., *Happiness*, but consistently misclassify others, notably, *Disgust*, due to its visual subtlety in the bodily cues and the nature of definitional ambiguity in images.

### 4.5 Comparison with Modular Baselines

To further contextualize ELENA's performance, we compare it with a supervised modular design. We first apply a widely-used image captioning model, LLaVA-1.5-7b (Liu et al., 2024a), to get the caption of an image, then fine-tune a BERT-base model (Devlin et al., 2019) to predict the emotion label. Here we experiment on the **HECO** dataset. The choice of this dataset stems from our manual inspection, as it provides a better representation of 'emotion-centric' images in a natural setting. This experiment was designed to assess whether standard captioning can retain the subtle, context-rich bodily expressions central to our task.

Our framework shows substantial improvement compared to the LLaVA with a fine-tuned BERT model. The performance gap reveals that embodied emotion information is not sufficiently captured while captioning. For instance, while LLaVA might generate a generic caption like "A person standing alone in a field, lost in thought", ELENA's framework would further capture specific embodied indicators, such as "shoulders slumped with head tilted down, suggesting dejection." This analysis reveals that standard image captions often fail to convey the subtle bodily expressions that are crucial to the embodied emotion task for affective experiences.

## 5 Conclusion

In this work, we introduced ELENA, a framework that employs LVLMs in a zero-shot setting to generate structured narratives that anatomize embodied emotions. Our experiments reveal that ELENA yields credible improvements over naive prompts, demonstrating the models' ability to adapt in both unmasked and masked settings despite information loss. Our experiments showed that the Gemini 2.5 Flash model generally outperformed other models across most datasets. However, labels like *Disgust* proved particularly challenging to identify from bodily expressions. The attention analysis provides a diagnosis by documenting persistent facial bias. It establishes an essential empirical foundation that enables the community to develop targeted solutions, such as fine-tuning with redirected attention mechanisms, for accurate anatomical understanding of embodied emotions.

## Limitations

The proposed work recognizes some of the barriers it encounters in its formation. First, masking the face leaves little to no information in the image, primarily due to the nature of the image. In such cases, the model is forced to output *Neutral* as a response, and this issue extends to the current image datasets, which may not elicit a clear emotional reaction from the person(s) involved, subsequently leading to incorrect analysis, even in human annotations. Second, a direct comparison between ELENA narratives and the CHEER dataset (Zhuang et al., 2024; Duong et al., 2025) would be valuable. However, differences in style make alignment difficult: ELENA outputs are longer and encompass multiple body parts, incorporating a significant amount of context, whereas CHEER sentences are more concise. Therefore, individual descriptions from ELENA need to be broken down into aspect categories to facilitate further analysis of their impact. Third, we believe ELENA-generated narratives could benefit downstream applications such as emotion recognition, conversational agents, or human-robot interaction. We view this as an important direction for future work.

## Acknowledgments

We thank the CincyNLP group for their constructive feedback. We also appreciate the anonymous

EMNLP reviewers for their comments and suggestions, which helped us improve this paper.

## References

Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Sree Bhattacharyya and James Z. Wang. 2025. Evaluating vision-language models for emotion recognition. In *In Findings of the Association for Computational Linguistics (NAACL 2025)*.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *Preprint*, arXiv:2501.17811.

Jiawen Deng and Fuji Ren. 2021. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*.

Phan Anh Duong, Cat Luong, Divyesh Bommana, and Tianyu Jiang. 2025. CHEER-Ekman: Fine-grained embodied emotion classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2025)*.

Thomas Fuchs and Sabine C Koch. 2014. Embodied affectivity: on moving and being moved. *Frontiers in psychology*, 5:508.

Gemini Team. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Evgeny Kim and Roman Klinger. 2019. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics (EMNLP 2021)*.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, and 1 others. 2025. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024)*.

Haifeng Lu, Jiuyi Chen, Feng Liang, Mingkui Tan, Runhao Zeng, and Xiping Hu. 2025. Understanding emotional body expressions via large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*.

Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M Montero, and Fernando Fernández-Martínez. 2021. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22):7665.

Paula Niedenthal, Adrienne Wood, and Magdalena Rychlowska. 2014. Embodied emotion concepts. In *The Routledge handbook of embodied cognition*, pages 240–249. Routledge.

Paula M. Niedenthal. 2007. Embodying emotion. *Science*, 316(5827):1002–1005.

Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, 12(2):505–523.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Zhijuan Shen, Jun Cheng, Xiping Hu, and Qian Dong. 2019. Emotion recognition based on multi-view body gestures. In *2019 IEEE International Conference on Image Processing (ICIP)*.

Patrizia Thoma, Denise Soria Bauser, and Boris Suchan. 2013. Besst (bochum emotional stimulus set)—a pilot validation study of a stimulus set containing emotional bodies and faces from frontal and averted views. *Psychiatry Research*, 209(1):98–109.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yan Wang, Shaoqi Yan, Yang Liu, Wei Song, Jing Liu, Yang Chang, Xinji Mai, Xiping Hu, Wenqiang Zhang, and Zhongxue Gan. 2024b. A survey on facial expression recognition of static and dynamic emotions. *Preprint*, arXiv:2408.15777.

Carissa Wilkes, Rob Kydd, Mark Sagar, and Elizabeth Broadbent. 2017. Upright posture improves affect and fatigue in people with depressive symptoms. *Journal of behavior therapy and experimental psychiatry*, 54:143–149.

Jiehui Wu, Jiansheng Chen, Qifeng Luo, Siqi Liu, Youze Xue, and Huimin Ma. 2024. Upper-body hierarchical graph for skeleton based emotion recognition in assistive driving. In *European Conference on Computer Vision (ECCV 2024)*.

Wei Wu, Hanyang Peng, and Shiqi Yu. 2023. Yunet: A tiny millisecond-level face detector. *Machine Intelligence Research*, 20(5):656–665.

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, and Georgios Tzimiropoulos. 2024. Vllms provide better context for emotion understanding through common sense reasoning. *Preprint*, arXiv:2404.07078.

Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion recognition for multiple context awareness. In *European conference on computer vision (ECCV 2022)*.

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. Visual prompting in LLMs for enhancing emotion recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*.

Wenjie Zheng, Jianfei Yu, Rui Xia, and Shijin Wang. 2023. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2020. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2024. My heart skipped a beat! recognizing expressions of embodied emotion in natural language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

## A VAD Score Interpretation

Apart from narrative description and emotion labels, we also asked the model to provide us with the valence, arousal, and dominance scores. These scores are typically part of many emotion recognition analyses and, therefore, are further explored in detail. Figure 5 reveals distinctive patterns across the seven basic emotions. The *Happiness* label emerges as the highest Valence score emotion (M = 8.51; where M is the mean value of this label), exhibiting substantially higher ratings than all other emotions, which cluster below the midpoint except for surprise (M = 5.22) and neutral states (M = 5.04). Regarding arousal, *Fear*, *Anger*, and *Surprise* demonstrate the highest activation levels (M = 7.38 and 7.32, respectively), whereas neutral expressions predictably show reduced arousal (M = 3.51). This makes sense because the *Neutral* label typically applies to images where there is no body enactment or emotion-intensive activity. In dominance scores, *Happiness* maintains a relatively high dominance (M = 6.28), while *Fear* exhibits markedly low dominance (M = 3.08), reflecting the characteristic vulnerability of the label.

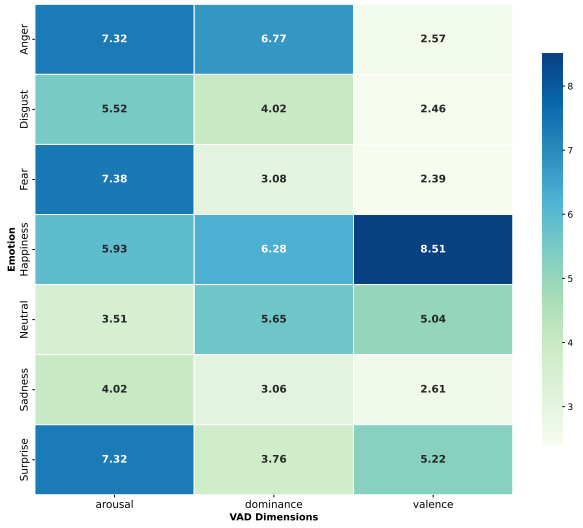


Figure 8: Valence, Arousal, and Dominance score representation associated with each emotion label. Results are from the HECO dataset.

## B Prompts Usage

We employed two prompting strategies in our experiments. First, a naive prompt (Figure 9) asks the models to classify the emotion displayed in the image using a fixed label set, with minimal guidance and no multi-perspective interpretation.

**NAIVE PROMPT**

Your task is to classify emotions in images **apart from masked areas.**
For each image, provide the following in strict JSON format: "emotion": ONLY ONE of: "**Happiness, Sadness, Fear, Anger, Neutral, Surprise and Disgust**". DO NOT OUTPUT ANY OTHER LABEL EXCEPT THE ONES MENTIONED.

DO NOT INCLUDE ANY OTHER EXPLANATORY TEXT OUTSIDE OF JSON.
**IGNORE ANY PERSON IN THE IMAGE WHOSE FACE IS NOT MASKED.**

Figure 9: Naive prompt for masked images.

**SYSTEM PROMPT**

You are an expert in detecting embodied emotions in images. Embodied emotions are physical manifestations of emotional states where:
(1) a physical movement or physiological response is evoked by an emotion, and
(2) the physical movement has no purpose other than emotion expression.

You will analyze images and respond ONLY in valid JSON format with no additional text.

Figure 10: System prompt.

Importantly, it instructed the model to ignore any person whose face was not masked and to output only a single emotion label in JSON format. This setting reflects a constrained classification setup that lacks deeper narrative or reasoning components. In contrast, our primary prompt template employed a structured prompting scheme, combining a system message with a detailed user prompt (see Figures 10 and 11). The system prompt defines the task as embodied emotion recognition and frames the model as an expert in this domain, emphasizing bodily movements as primary indicators of emotion.

The user prompt specifies the expected output format and fields, guiding the model to produce a multi-perspective emotional interpretation grounded in bodily expression. Responses were constrained to strict JSON format, enabling consistent parsing and analysis. All prompts were applied uniformly across images and models without post-hoc adjustment. This ensured comparability between outputs and minimized linguistic variance introduced by prompt phrasing. The prompt design plays a central role in eliciting structured, interpretable outputs from vision-language models in our study.

### B.1 Emotion Mapping and Reasoning Analysis

We adopt the six basic Ekman emotions—Happiness, Sadness, Anger, Fear, Disgust, and Surprise—along with a Neutral category, as our primary label set. This choice strikes a balance between interpretability and coverage, enabling

**USER PROMPT**

Analyze this image and provide the following in strict JSON format:

1. "**explicit_description**": "A short explicit description focusing on visible body parts and their emotional expression",
2. **"implicit_description"**: "A short implicit description capturing internal sensations (body parts which are not visible) (e.g., "My heart raced" rather than "Her eyes widened")",
3. **"narrative"**: "A very concise one line story NECESSARILY CONTAINING THE BODY PART influencing the emotion. It should include the scene and context of the image. Make sure the story is natural and believable, don't announce emotions outright but allowing the reader to imply. ",
4. "**body_parts**": "The specific clear body parts (explicit) involved in expressing the emotion",
5. **"emotion"**: "ONLY ONE of: Happiness, Sadness, Fear, Anger, Surprise, Disgust and Neutral",
6. **"valence"**: 1-10 indicating the positivity/negativity of the emotion,
7. **"arousal"**: 1-10 indicating the intensity of the emotion,
8. **"dominance"**: 1-10 indicating the level of control/power

Figure 11: User prompt.

standardized evaluation while remaining expressive enough to capture a wide range of embodied affect.

This unified 7-category taxonomy was essential for enabling a consistent cross-dataset analysis, as our other datasets were natively annotated with similar categories. Consequently, a direct comparison using mean Average Precision (mAP) against prior EMOTIC benchmarks would be incongruous, as our framework evaluates a fundamentally different task grounded in embodied theory rather than multi-label classification on the 26 categories.

To justify this abstraction, we map each basic emotion to a broader set of fine-grained emotion categories from the EMOTIC dataset (Table 5). For example, the Sadness category includes not only sadness itself but also states such as fatigue, suffering, and yearning. Similarly, Happiness subsumes affective expressions like affection, engagement, and pleasure.

This mapping serves two key purposes:

- It enables aggregation of subtle and diverse affective cues under a shared emotional umbrella, facilitating more robust analysis of model predictions.
- It provides a reasoning structure for validating model outputs against grounded emotional expressions.

While models often use varied language to describe emotion, our mapping enables alignment between expressive counterparts and a more simplified label assignment.

| EMOTIC Emotions | Ekman Emotion |
|---|---|
| Affection<br>Engagement<br>Excitement<br>Happiness<br>Pleasure | Happiness |
| Disconnection<br>Embarrassment<br>Fatigue<br>Pain<br>Sadness<br>Sensitivity<br>Suffering<br>Yearning | Sadness |
| Aversion<br>Disapproval | Disgust |
| Disquietment<br>Fear | Fear |
| Anger<br>Annoyance | Anger |
| Anticipation<br>Confidence<br>Doubt/Confusion<br>Esteem<br>Peace | Neutral |
| Surprise | Surprise |

Table 5: Mapping of EMOTIC emotions to Ekman's six with a *Neutral* category.

### B.2 Dataset Quality Challenges for Embodied Emotion Recognition

While existing emotion recognition datasets, such as EMOTIC, provide bounding box annotations for people in images, many instances present significant challenges for embodied emotion analysis, as illustrated in Figure 12. Our manual analysis reveals that a substantial proportion of images in current emotion recognition datasets are inherently unsuitable for embodied emotion recognition tasks, and sometimes for general emotion recognition too, despite containing valid emotion annotations for general affective computing applications.

The fundamental issue stems from the conceptual difference between contextual emotion recognition and embodied emotion analysis. Embodied emotion recognition specifically requires clear, observable bodily expressions such as posture, gesture, limb positioning, and torso orientation. Many images in datasets contain subjects where these crucial embodied indicators are either absent, ob-

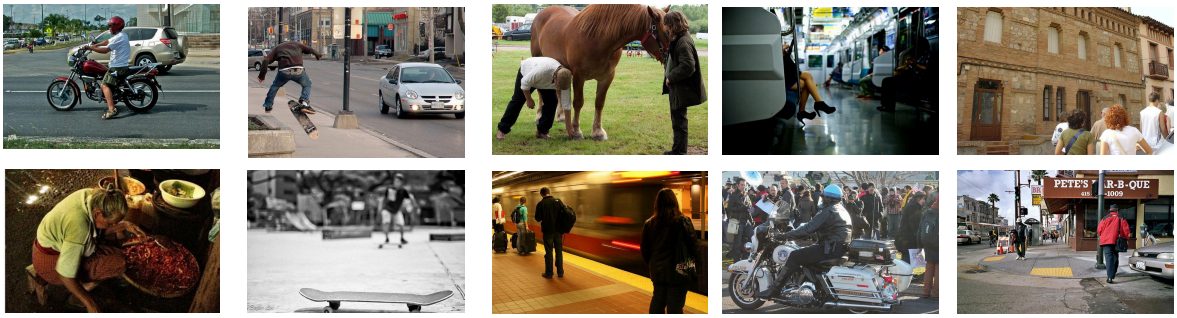


Figure 12: Representative examples from the EMOTIC dataset illustrating challenges for embodied emotion recognition. Many images contain crowded places, distant subjects, ambiguous emotional contexts, or insufficient visibility of bodily expressions, making reliable analysis difficult, even with provided bounding box annotations.

scured, or ambiguous, making reliable analysis difficult even with sophisticated prompting frameworks like ELENA.

Consider the common scenario where multiple individuals appear in a single image, each with varying emotional states. While bounding box annotations attempt to isolate the target subject, the presence of other individuals with potentially conflicting emotional expressions creates visual noise that can mislead the LVLMs. This problem is particularly acute in crowded scenes or social gatherings where the emotional dynamics are complex and the target person's individual embodied expression becomes difficult to distinguish from the collective group behavior. Distance and viewing angle present additional complications. Many images capture subjects at distances where fine-grained bodily details necessary for embodied emotion recognition, such as hand positioning, shoulder tension, or subtle postural shifts, are not visible at sufficient resolution. Similarly, non-frontal views, while valuable for general emotion recognition, often obscure key body parts that serve as primary indicators in embodied emotion analysis. As our experiments demonstrate, these limitations become more pronounced when faces are masked, since the model must rely entirely on these often-inadequate bodily cues.

Perhaps most critically, we observe a systematic bias toward contextually inferred emotions rather than expressions grounded in observable bodily manifestations. For instance, an image might show a person labeled as *Sad* based on the situational context (e.g., standing alone in the rain), yet their actual body posture may not exhibit the characteristic embodied markers of sadness. This disconnect between contextual emotion labels and embodied emotional expressions creates a fundamental training-evaluation mismatch for models designed to recognize emotions through bodily cues.

The implications extend beyond dataset quality to model evaluation. When a significant portion of the evaluation data contains ambiguous or unsuitable examples, performance metrics become unreliable indicators of the actual model's capability. Models may appear to perform poorly on embodied emotion recognition, not due to architectural limitations, but because the evaluation framework includes many instances where even human annotators would struggle to identify clear embodied emotional indicators. This challenge necessitates either more stringent data filtering protocols or the development of specialized datasets explicitly designed for recognizing embodied emotions.

## C Interpreting LVLM Emotion Predictions on BESST

BESST consists of staged emotional expressions in controlled visual contexts, with each emotion category enacted in both frontal and averted views (Figure 13). To analyze how vision-language models interpret embodied emotion, we employ additional models, namely Llama-90B-Vision, Janus-7B (Chen et al., 2025), and Qwen 2.5 (Wang et al., 2024a), to compare their performance with the previous best results by Gemini-2.5-Flash Preview on the BESST dataset. We also tested several other LVLMs on BESST. Despite identical prompting structures, the models diverged in their calibration of embodied emotion. As shown in Table 6, Gemini 2.5 Flash outperforms all other models tested on BESST, achieving the highest F1 score (52.7), followed by Llama-90B (49.3). These differences are consistent with qualitative trends in prediction diversity, suggesting that Gemini's more extensive pretraining and model capacity may support greater


Figure 13: Example from the Bochum Emotional Stimulus Set (BESST), which depicts enactment of defined emotions in full frontal and averted view.

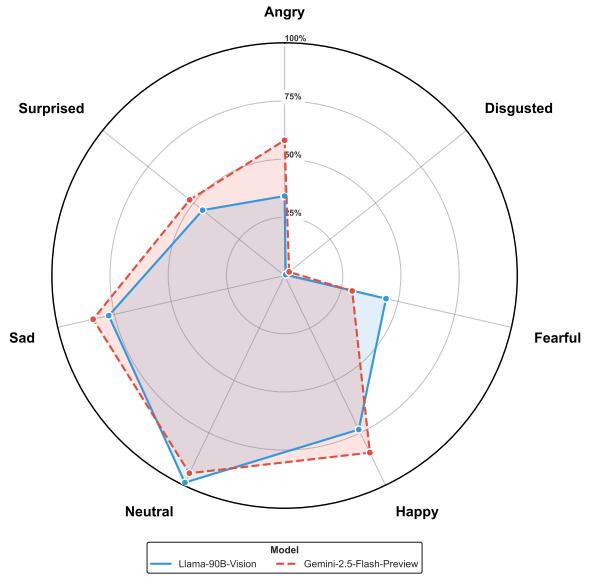


Figure 14: Radar plot of each emotion label predicted by Llama-90B and Gemini 2.5 Flash Preview model on the BESST dataset.

emotion generalization in embodied contexts. Figure 14 visualizes the distribution of predicted emotion labels across the best two models. Both models heavily favored neutral and sadness-related predictions, with notably lower frequencies for anger, disgust, and fear. This may indicate a model-level tendency to downplay high-arousal or negative valence expressions, particularly when bodily cues are ambiguous. Notably, Gemini 2.5 Flash showed a broader emotional spread, selecting anger and surprise more frequently than Llama-90B, suggesting heightened sensitivity to expressive gestures.

| Model | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| Gemma-3-12B | 46.9 | 55.5 | 46.9 | 41.6 |
| Llama-3.2-11B | 44.8 | 48.3 | 44.9 | 37.8 |
| Llama-3.2-90B | 53.4 | 65.81 | 53.5 | 49.3 |
| **Gemini-2.5-Flash** | 57.9 | 64.8 | 58.0 | **52.7** |
| Qwen2.5-VL-7B | 42.7 | 41.2 | 42.7 | 42.4 |
| Janus-Pro-7B | 28.3 | 53.7 | 28.2 | 28.7 |

Table 6: Performance comparison of large vision-language models on the BESST dataset. Scores are macro-averaged.

## D Appendix: Emotion Mapping and Dataset Quality Analysis

We also implement a two-stage pipeline, i.e., description generation followed by emotion parsing. We do so by first generating the narrative descriptions based on the exact definition, without any interpretation of the image involved. In the second step, we feed the narratives as input to perform the emotion classification based solely on the text descriptions. This approach separates the perceptual description task from the emotion understanding task. As BESST contains images of posed emotions in a controlled, lab-based environment with faces already masked, it provides an ideal setting for a clean, controlled test. This allowed us to isolate the core cognitive task of interpreting unambiguous bodily cues, testing the hypothesis of whether unified or sequential processing is more effective without the confounding variables of complex backgrounds or "in-the-wild" ambiguity in static images.

As seen in Table 7, ELENA achieves notably higher prediction metrics than the two-stage pipeline, although the latter suffers less from rare classes being less predicted. However, the results demonstrate that unified structured generation outperforms sequential processing. We believe this is because the single-prompt approach enables the creation of more coherent predictions, as the model jointly considers embodied emotion definitions while generating both narrative descriptions and emotion labels.

## E Appendix: YuNet Face-Masking Experiment Setup

For this task, we utilize the quantized version of YuNet (Wu et al., 2023) to automate face masking in images. YuNet is a fast and accurate

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Two-Step Baseline | 54.2 | 55.1 | 51.0 |
| **ELENA** | 64.8 | 58.0 | 52.7 |

Table 7: Performance comparison with the two-step (description-then-parsing) baseline on the BESST dataset. All metrics are macro-averaged (%).

deep learning-based face detector available for use through OpenCV. We used the block-quantized version in *int-8* precision, as it has nearly the same performance as its standard release. It outputs bounding box coordinates and confidence scores for the detected faces. We then extract the bounding boxes and subsequently replace the pixel data within the region with a uniform mask color, which conceals the detected faces. We set the parameters to detect up to 20 faces, although such images occur rarely. The confidence threshold is taken to be balanced with a value of 0.5.

In short, we duplicate the dataset that masks out every person's face. We conducted this study to investigate the following LVLM characteristic: a model that relies heavily on facial features may exhibit a significant drop in metrics and produce less confident or shorter descriptions when the face is masked. On the other hand, a model that is more *body-aware* might observe a smaller drop in performance.