MASP: Multi-Aspect Guided Emotion Reasoning with Soft Prompt Tuning in Vision-Language Models

SangEun Lee[1], Yubeen Lee[2], Eunil Park[2,*], Wonseok Chae[1,*]

[1]Electronics and Telecommunications Research Institute
[2]Sungkyunkwan University
sange1104@etri.re.kr, lybin1070@gmail.com, eunilpark@skku.edu, wschae@etri.re.kr

Abstract

Understanding human emotions from images is a challenging yet essential task for vision-language models. While recent efforts have fine-tuned vision-language models to enhance emotional awareness, most approaches rely on global visual representations and fail to capture the nuanced and multi-faceted nature of emotional cues. Furthermore, most existing approaches adopt instruction tuning, which requires costly dataset construction and involves training a large number of parameters, thereby limiting their scalability and efficiency. To address these challenges, we propose MASP, a novel framework for Multi-Aspect guided emotion reasoning with Soft Prompt tuning in vision-language models. MASP explicitly separates emotion-relevant visual cues via multi-aspect cross-attention modules and guides the language model using soft prompts, enabling efficient and scalable task adaptation without modifying the base model. Our method achieves state-of-the-art performance on various emotion recognition benchmarks, demonstrating that the explicit modeling of multi-aspect emotional cues with soft prompt tuning leads to more accurate and interpretable emotion reasoning in vision-language models.

Code — https://github.com/sange1104/MASP

Introduction

Vision-language models (VLMs) have shown strong performance across a wide range of multimodal tasks, including visual question answering (Xu et al. 2024; Fang et al. 2025a), image captioning (Yang et al. 2023b), and multimodal dialogue (Gong et al. 2023; Wu et al. 2024). Understanding human emotions from visual inputs is an important task for VLMs, as it plays a vital role in applications such as affective computing (Meudt et al. 2016), human-computer interaction (Cowie et al. 2001), and mental health analysis (Simcock et al. 2020). However, VLMs remain limited in their ability to reason about the emotions from images. Unlike conventional tasks that rely on objective visual perception, emotions are not conveyed through a single visual element but emerge as a result of integrating multiple ambiguous and context-dependent cues, including visual tone, facial expression, and compositional affect (Zhao et al. 2014; Abbas et al. 2025).

Figure 1: Examples showing that emotional perception depends on different visual cues.

Recently, various approaches have been proposed to improve the emotion-recognition capabilities of VLMs. One line of work involves fine-tuning VLMs on emotion datasets to help them better understand and interpret emotional content. For example, several studies (Cheng et al. 2024; Xie et al. 2024) adopted instruction tuning to guide VLMs toward emotion-aware responses.

However, two important challenges remain in the approach of VLMs to emotional recognition. First, while humans interpret emotions by attending to various visual elements, such as facial expressions, scenes, objects, and colors, these cues often interact and jointly contribute to emotional perception. As shown in Figure 1, different emotions are elicited through complementary combinations of visual cues, some of which are shaped by facial expressions and human actions, whereas others are driven by environmental contexts, such as scene composition and visual tone. Nevertheless, most models process images using a single visual encoder and rely on a global visual representation, making it difficult to identify the specific visual factors driving emotional responses, leading to unstructured reasoning and limited interpretability.

Furthermore, most existing approaches rely on instruction tuning to adapt general-purpose VLMs for understanding emotions. This approach requires costly instruction dataset construction, often involving manual annotation or application programming interface-based generation. In addition,

*Corresponding authors

despite the use of parameter-efficient methods such as LoRA or QLoRA, instruction tuning still involves training millions of parameters, limiting scalability and efficiency.

In this paper, we introduce Multi-Aspect guided emotion reasoning with Soft Prompt tuning in VLMs (MASP), a framework that explicitly encodes emotion-relevant visual cues, enabling emotion understanding that mimics human interpretation. To guide the model toward emotion-specific reasoning, MASP employs a soft prompt mechanism that acts as a task-adaptive prior, steering the language model toward emotionally grounded reasoning by contextualizing multi-aspect visual inputs.

Our contributions are summarized as follows:

- We propose MASP, a novel framework that enables emotionally grounded reasoning in VLMs by structurally decomposing emotion-related visual cues. To the best of our knowledge, this is the first study to explicitly separate such cues using aspect-specific queries and cross-attention in VLMs for human-aligned emotion reasoning.
- We introduce a soft prompt mechanism that facilitates parameter-efficient and interpretable adaptation by updating only a small set of task-specific parameters.
- We demonstrate that our method achieves state-of-the-art performance across multiple benchmarks, validating the effectiveness of our framework.

Related Work

Training-free Approaches for Emotion Recognition with VLMs

Recent progress in VLMs highlights their capacity to jointly interpret and reason multimodal inputs, particularly images and text. Models such as BLIP (Li et al. 2022), LLaVA (Liu et al. 2023), and InstructBLIP (Dai et al. 2023) have demonstrated notable capabilities across a range of vision-language tasks, including visual commonsense reasoning, multi-image comparisons, and scene-level description generation. These models typically integrate a pretrained vision encoder with a large language model (LLM) to associate complex visual inputs with the corresponding linguistic representations via multimodal inference. Building on these capabilities, recent research has begun to explore the application of VLMs to affective tasks, such as emotion recognition, suggesting that these models may be extended beyond objective perception to support reasoning about social and cognitive contexts.

These developments indicate that VLMs can be applied to high-level tasks, such as emotion recognition, without task-specific training. In training-free settings, emotion-centric prompts are crafted to enable pretrained VLMs to perform emotion reasoning directly during inference. For instance, Zhang et al. (2024) proposed the set-of-vision prompting (SoV) approach, which leverages visual location information, such as bounding boxes and facial landmarks, as prompts to enhance the emotion recognition performance of VLLMs. Although such approaches offer advantages in terms of scalability and computational efficiency, they often generate superficial and generic responses owing to the absence of explicit modeling of emotional semantics. To address this, Fang et al. (2025b) introduced a training-free and lightweight framework that applies Coarse-to-Fine inference strategies and Focus-on-Emotion Visual Augmentation. This method improves both the precision and efficiency of inference-time emotion recognition by helping models distinguish better between semantically similar emotions and attend to relevant emotional cues while suppressing irrelevant visual information.

Fine-tuned VLMs for Emotion Recognition

To overcome the limitations of training-free approaches, recent studies actively explored the fine-tuning of VLMs specifically for emotion recognition tasks. Xie et al. (2024) employed instruction tuning to guide the generation of emotion-centric responses, while Lee, Lee, and Park (2025) incorporated emotion-specific knowledge into general-purpose VLMs via knowledge distillation.

Etesam et al. (2024) leveraged contextual cues by comparing an indirect approach based on image captioning with direct zero-shot and fine-tuned VLM-based methods. Their experiments on small-scale emotion-labeled datasets demonstrated the effectiveness of fine-tuning, even under limited supervision. Recently, Nonaka and Valles (2024) introduced a fully autoregressive multimodal LLM framework that integrates an instruction-tuned LLaMA3 with a vision encoder. The proposed model exhibited strong performance in conversational emotion recognition, highlighting the potential of context-aware emotional reasoning.

Among these studies, Cheng et al. (2024) was particularly notable for its multi-aspect approach to emotion interpretation. The model utilizes two distinct visual cues that encompass facial expressions and scene context, with each processed by a dedicated modality encoder to facilitate emotion inference. However, because it relies on pretrained facial and scene backbones, emotional reasoning remains implicit, lacking explicit disentanglement and interpretability of the contributing emotional factors.

Building on this insight, we explicitly define six emotional aspects: facial expression, scene, objects, colorfulness, brightness, human action, and leverage them to facilitate the interpretation of visual information in a semantically coherent manner grounded in emotional semantics.

Method

We propose MASP, a two-stage training framework designed to enhance emotion understanding in VLMs. In the first stage, MASP trains six aspect modules, each consisting of a learnable query and a cross-attention mechanism. These modules are optimized to attend to the global visual representation and extract semantically rich embeddings that are specific to each aspect. In the second stage, the extracted aspect embeddings are fused with the global visual representation and used as input to a frozen language model. A learnable soft prompt is prepended to the language input to inject a task-specific inductive bias, enabling the model to better reason about emotions in a parameter-efficient manner.

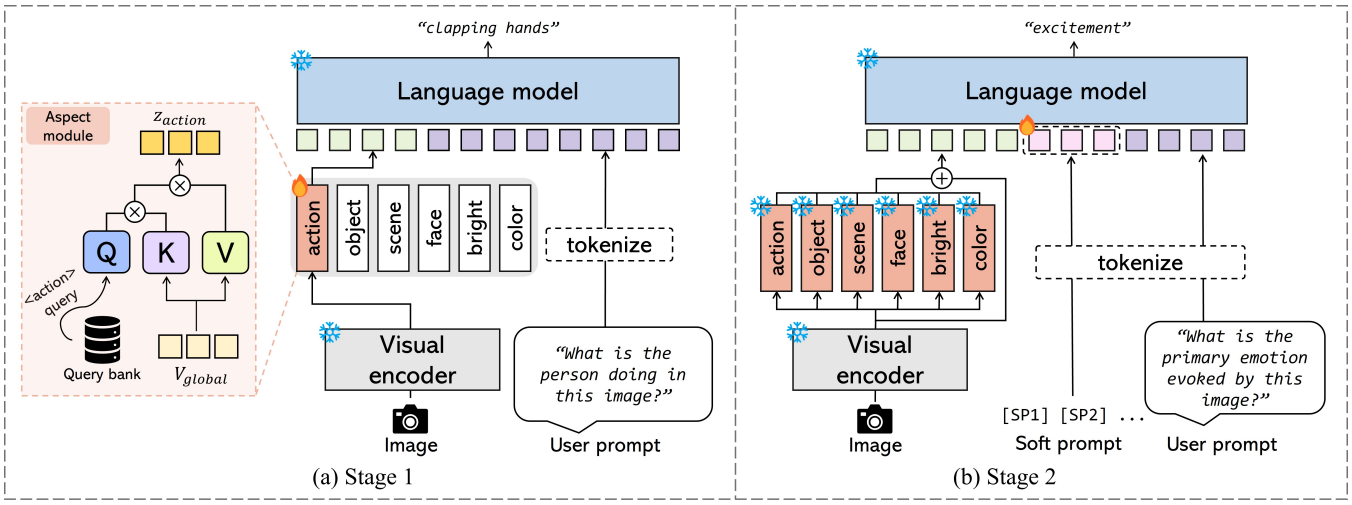

Figure 2: Overview of our MASP framework. In Stage 1, each of the six aspect-specific modules learns a query vector and a cross-attention mechanism to extract aspect-relevant embeddings from the visual representation. Stage 2 freezes these components and trains soft prompts to guide emotion reasoning via a frozen language model.

Preliminary

We consider a VLM that takes an image $\mathbf{I}$ and a natural language query $\mathbf{x}_{emo}$ as input and produces a textual prediction $\hat{y}$ as output. The query prompts the model to choose the most appropriate emotion from a predefined set of categories $\mathcal{Y}$ as follows:

$$\hat{y} = \text{VLM}(\mathbf{I}, \mathbf{x}_{emo}), \hat{y} \in \mathcal{Y}$$

The image $\mathbf{I}$ is first encoded by a visual encoder to obtain a global visual representation of the image. Depending on the model architecture, this representation is either directly fed into the language model or first projected into the language embedding space using an alignment module. The resulting visual feature, whether used directly or after alignment, is then combined with the tokenized query, and the fused sequence is passed to the language model to generate a response.

A prediction is regarded as correct only if the generated response exactly matches the ground-truth label $y^* \in \mathcal{Y}$.

Stage 1: Multi-Aspect Visual Encoding Modules

In Stage 1, our goal is to mimic the human process of emotion recognition, which relies on integrating multiple types of visual cues. Unlike conventional VLMs that encode an image into a single global feature without explicitly focusing on emotion-relevant information, we propose multi-aspect visual encoding modules. These modules are designed to capture six representative visual cues identified as key drivers of emotional perception by Yang et al. (Yang et al. 2023a): *facial expression, scene, objects, colorfulness, brightness*, and *human action*.

For each of these six aspects, we define a learnable query vector $\mathbf{q}^i_{view} \in \mathbb{R}^d$ and an associated cross-attention module $\text{CrossAttn}_i(\cdot)$, where i indicates the corresponding aspect. In this attention mechanism, $\mathbf{q}^i_{view}$ serves as the query, whereas the keys and values are derived from the frozen visual encoder's patch-level image representation $\mathbf{V}_{global} = \mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathbb{R}^{T \times d}$. Each module is independently parameterized and trained to specialize in extracting features $\mathbf{z}_i$ that are relevant to its corresponding aspect, as follows:

$$\mathbf{z}_i = \text{CAttn}_i(\mathbf{q}^i_{view}, \mathbf{V}_{global}) = \sigma\left(\frac{\mathbf{q}^i_{view}\mathbf{K}_i^\top}{\sqrt{d}}\right)\mathbf{V}_i, \quad (1)$$

where CAttn and σ denote cross-attention and softmax, respectively. In addition, $\mathbf{K}_i = \mathbf{W}^i_K\mathbf{V}_{global}$ and $\mathbf{V}_i = \mathbf{W}^i_V\mathbf{V}_{global}$ are the key and value projections for the i-th aspect.

To train the aspect queries $\mathbf{q}^i_{view}$ and attention modules CrossAttn_i, we define a training objective that encourages each aspect embedding $\mathbf{z}_i$ to encode semantically grounded information specific to its corresponding visual cue. During training, a natural language prompt querying a particular aspect of the image (e.g., "What is the background scene?") is tokenized into $\mathbf{t}^i_{aspect}$ and provided with the image. The attention module CrossAttn_i uses the query $\mathbf{q}^i_{view}$ to extract an aspect-specific embedding $\mathbf{z}_i$, which is then passed to the language model together with $\mathbf{t}^i_{aspect}$ to generate a textual response as follows:

$$\hat{y}_i = \text{LM}(\mathbf{z}_i, \mathbf{t}^i_{aspect}) \quad (2)$$

At each training step, a single image–aspect pair is sampled, and only the corresponding aspect module is activated and updated. Although there are six sets of aspect modules and queries in total, only one set receives gradient updates per example, encouraging each to specialize in its designated visual cue without interference.

The model is trained to produce the correct answer based solely on $\mathbf{z}_i$ without access to the global visual representation. This design ensures that each aspect embedding contains sufficient information for reasoning about its corre-

sponding visual aspect. Only the aspect queries and cross-attention modules are updated during training, whereas the visual encoder and language model remain frozen. The model is optimized using a causal language modeling loss, computed as token-level cross-entropy over the ground truth explanation tokens.

Stage 2: Soft Prompting for Emotion Recognition

In Stage 2, we introduce soft prompt tuning to guide the language model toward emotion recognition, leveraging the aspect-specific visual cues extracted in the previous stage. To provide the model with a rich visual context, we first construct a fused visual representation by concatenating the six aspect embeddings $\{\mathbf{z}_{aspect_1}, \ldots, \mathbf{z}_{aspect_6}\}$ with the global image representation $\mathbf{V}_{global}$ from the frozen visual encoder:

$$\mathbf{V}_{fused} = \text{concat}(\mathbf{z}_{aspect_1}, \ldots, \mathbf{z}_{aspect_6}, \mathbf{V}_{global}). \quad (3)$$

The soft prompt $\mathbf{P}_{soft} \in \mathbb{R}^{L_p \times d}$ consists of L_p learnable tokens, where we set $L_p = 32$ in our experiments. These tokens are newly introduced and are not part of the original tokenizer vocabulary. Instead, they are initialized randomly and optimized from scratch during training. Alongside the soft prompt, the emotion query $\mathbf{x}_{emo}$ is a text prompt that asks about the emotion evoked by an image.

The final input to the language model is constructed by sequentially combining the fused visual input $\mathbf{V}_{fused}$, tokenized soft prompt $\mathbf{T}_{soft}$, and tokenized emotion query $\mathbf{T}_{user}$ as follows:

$$\mathbf{x}_{input} = [\mathbf{V}_{fused}; \mathbf{T}_{soft}; \mathbf{T}_{user}], \quad (4)$$

where $[\cdot\,;\,\cdot]$ denotes sequence concatenation. The model is trained to generate the correct emotion label conditioned on $\mathbf{x}_{input}$. Formally, the prediction is obtained as follows:

$$\hat{y} = \text{LM}(\mathbf{x}_{input}) \quad (5)$$

The prediction $\hat{y}$ is supervised using a causal language modeling loss against the ground-truth emotion label. During this stage, only the parameters of the soft prompt $\mathbf{P}_{soft}$ are updated, while all other modules, including the aspect modules and language model, are kept frozen.

While MASP extracts structured visual cues independently, we intentionally avoid using any explicit routing mechanism. Instead, we rely on the attention dynamics of the language model to naturally integrate and interpret the provided aspect information during emotion reasoning. We analyze this behavior through an ablation study of attention patterns across aspects.

Inference Phase

At inference time, the prediction process follows a structured pipeline, as depicted in Algorithm 1. The input image is first encoded using a frozen visual encoder to produce a global visual representation. This representation is passed through six independent aspect modules, each equipped with a query and cross-attention mechanism to extract aspect-specific embeddings. These embeddings are then concatenated with the global visual representation to form the final image features. In parallel, the user prompt and learned soft prompt are tokenized. Finally, the language model takes the tokenized text inputs along with the fused image features as input and predicts an emotion label.

Algorithm 1: MASP Inference Pipeline

Input: Image $\mathcal{I}$, soft prompt $\mathcal{P}_{soft}$, user prompt $\mathcal{P}_{user}$
Output: Predicted emotion label $\hat{y}$

Step 1: Visual Encoding
$\mathcal{V}_{global} \leftarrow \text{VisualEncoder}(\mathcal{I})$

Step 2: Aspect-specific Embedding Extraction
$\mathcal{Z} \leftarrow [\,]$
for *each index* $i \in \{aspect_1, aspect_1, \ldots, aspect_6\}$ **do**
  $z_i \leftarrow \text{CrossAttn}_i(q^i_{view}, \mathcal{V}_{global})$;
  Append z_i to $\mathcal{Z}$

Step 3: Image Feature Aggregation
$\mathcal{V}_{fused} \leftarrow \text{Concat}(z_1, z_2, \ldots, z_6, \mathcal{V}_{global})$

Step 4: Prompt Tokenization
$\mathcal{T}_{soft} \leftarrow \text{Tokenizer}(\mathcal{P}_{soft})$
$\mathcal{T}_{user} \leftarrow \text{Tokenizer}(\mathcal{P}_{user})$

Step 5: Input Construction and Prediction
$\mathbf{x}_{\text{input}} \leftarrow [\mathcal{V}_{fused};\ \mathcal{T}_{soft}; \mathcal{T}_{user}]$
$\hat{y} \leftarrow \text{LanguageModel}(\mathbf{x}_{\text{input}})$
return $\hat{y}$

Experiments

Experimental Setup

Dataset We train and evaluate our model on seven widely used visual emotion datasets that vary in emotion categories, scale, and image sources. We follow the standard train-test splits established in prior studies.

Emoset (Yang et al. 2023a) and FI (You et al. 2016) contain images labeled with one of eight emotion categories: *amusement, anger, awe, contentment, disgust, excitement, fear,* and *sadness*. Emoset, which includes social media images and artistic photographs, comprises 118,102 images. FI includes content from Flickr and Instagram, with a total of 21,824 images. Emotion6 (Peng et al. 2015) provides six emotion labels (i.e., *anger, disgust, fear, joy, sadness,* and *surprise*) and consists of approximately 1,980 images sourced from social media. Flickr and Instagram (Katsurai and Satoh 2016) are binary-labeled datasets (i.e., positive or negative) containing 60,738 and 42,832 images, respectively.

Finally, we include two domain-specific datasets, Abstract and ArtPhoto (Machajdik and Hanbury 2010), which contain 806 and 228 images, respectively. Abstract features emotional abstract paintings, and ArtPhoto consists of artistic photographs shared on online art platforms. Both datasets are annotated using the same eight emotion categories as those used in Emoset and FI.

Model	Emoset	FI	Emotion6	Flickr	Instagram	Abstract	ArtPhoto
Zero-shot							
LLaVA-v1.5 (Liu et al. 2024a)	50.01	53.64	49.83	80.29	85.76	15.21	41.36
LLaVA-Next (Liu et al. 2024b)	55.85	61.59	53.37	51.08	61.69	21.74	41.98
Qwen2-VL-7B (Wang et al. 2024)	54.04	58.16	51.68	81.76	84.96	17.39	43.21
Qwen2.5-VL-7B (Bai et al. 2025)	60.30	67.12	57.74	79.89	83.73	23.91	48.15
InstructBLIP (Dai et al. 2023)	45.98	63.92	54.21	83.12	82.05	21.74	45.68
SEPM (Fang et al. 2025b)	56.04	-	54.21	-	-	-	-
Task-specific							
Emotion-LLaMA (Cheng et al. 2024)	39.62	40.19	44.44	59.95	77.81	15.21	27.77
EmoVIT (Xie et al. 2024)	**83.36**	68.09	57.81	-	-	32.34	44.90
EmoVLM-KD (Lee, Lee, and Park 2025)	79.83	79.51	73.91	88.90	89.59	-	-
MASP	83.21	**80.06**	**74.75**	**89.81**	**90.05**	**45.65**	**53.09**

Table 1: Accuracy comparison between baseline models and MASP on benchmark datasets, with all values reported as percentages (%).

Implementation Details We leverage the pretrained Qwen2-VL-7B model (Wang et al. 2024) as our base VLM. In the first stage, we supervise the learning of the aspect-specific query vectors and cross-attention modules using the aspect annotations provided in the Emoset dataset. These annotations serve as ground-truth labels for each of the six visual cues. Among the six aspects, four are annotated with natural language answers, whereas brightness and colorfulness are given as continuous values between 0 and 1. To support categorical prediction, we discretize these into five ordinal classes (i.e., very low, low, moderate, high, and very high) with 0.2 intervals.

The second stage is trained on the emotion-labeled datasets, where only the soft prompt parameters are updated, while all other modules remain frozen. We optimize each module using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 1.

Baseline Models We compare our proposed model with various baseline models, which fall into two main categories: zero-shot VLMs and instruction-tuned models.

Zero-shot VLMs. We include LLaVA-v1.5 (Liu et al. 2024a), LLaVA-Next (Liu et al. 2024b), Qwen2-VL-7B (Wang et al. 2024), Qwen2.5-VL-7B (Bai et al. 2025) and InstructBLIP (Dai et al. 2023), which represent general-purpose VLMs. In addition, we employ SEPM (Fang et al. 2025b), a recent framework designed to improve emotion perception in VLMs by prompting existing zero-shot models during inference without requiring additional training.

Instruction-tuned models. These models are adapted to the emotion domain via instruction tuning on emotion-labeled data. Emotion-LLaMA (Cheng et al. 2024) extends instruction tuning to a multimodal setting with audio, image, and text inputs; since we focus on visual emotion recognition, only the image–text components are considered. EmoVIT (Xie et al. 2024) aligns visual features with emotion-specific queries through instruction tuning, while EmoVLM-KD (Lee, Lee, and Park 2025) distills an instruction-tuned VLM into a conventional vision backbone.

Evaluation Results

We evaluate MASP on seven emotion recognition datasets, and the results are summarized in Table 1. Compared with zero-shot VLMs and instruction-tuned emotion-specific VLMs, MASP consistently achieves superior performance. Notably, it establishes state-of-the-art results for all datasets except Emoset.

The strongest gains are observed on smaller datasets, such as Abstract and ArtPhoto. In the Abstract, MASP achieves an accuracy of 45.65%, significantly surpassing EmoVIT's performance of 32.34%. On ArtPhoto, MASP records 53.09%, exceeding the previous best record of 45.68%. These results suggest that MASP is particularly effective in low-resource scenarios, where instruction tuning may be difficult because of limited supervision.

MASP also outperforms existing methods on FI, Emotion6, Flickr, and Instagram, surpassing both zero-shot and instruction-tuned baselines. Although MASP falls slightly short of EmoVIT on Emoset, the gap is marginal. Importantly, unlike instruction-tuned models that require extensive instruction-formatted data, MASP relies solely on lightweight soft prompts, offering a more scalable and data-efficient alternative.

Ablation Study

Ablation on Key Components

We conducted an ablation study to assess the contribution of each component of MASP by comparing five configurations: (1) zero-shot VLM, (2) VLM with aspect modules, (3) VLM with pretrained soft prompt, (4) VLM with aspect modules and random soft prompt, and (5) VLM with pretrained soft prompt and aspect modules (i.e., MASP).

As shown in Table 2, adding the aspect modules to the VLM resulted in a noticeable drop in performance, with approximately 13% decrease on the FI dataset and 11% on Emotion6. This suggests that although the aspect modules learn meaningful representations for each aspect during

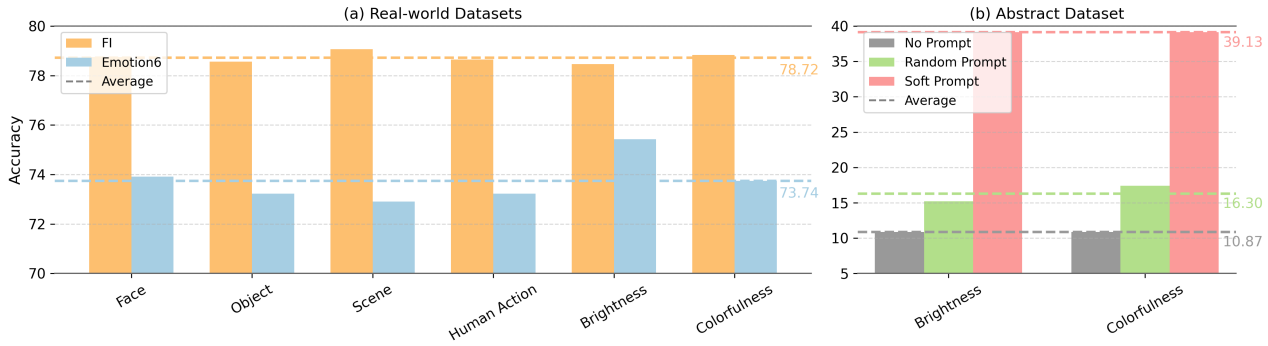

Figure 3: Accuracy comparison of models with single-aspect module for emotion recognition.

	V	A	S	FI	Emotion6
(1) V	✓			58.16	51.68
(2) V + A	✓	✓		44.54	40.40
(3) V + S	✓		✓	76.63	70.88
(4) V + A + rS	✓	✓	✓	41.47	40.07
(5) V + A + S	✓	✓	✓	**80.06**	**74.75**

Table 2: Ablation study of the key components in MASP. V, A, and S denote the zero-shot VLM, aspect modules, and soft prompt respectively.

Stage 1, using them without additional guidance may hinder the VLM's emotion reasoning capabilities. In contrast, adding a pretrained soft prompt to the VLM resulted in clear performance gains (i.e., approximately 18% increase on the FI dataset and 19% on Emotion6). These results suggest that soft prompt tuning alone can effectively adapt the model to an emotion recognition task without fine-tuning.

In contrast, combining the aspect modules with the random soft prompt led to the lowest performance, with the untrained prompt acting as noise rather than providing useful task guidance. Finally, the full MASP model achieved the best results, highlighting that the combination of structured visual representations and guided prompt tuning is crucial for effective emotion understanding.

Ablation on Aspect Module Contribution

To evaluate how much each aspect module contributes to emotion recognition, we conducted an ablation study in which each aspect module was independently attached to the VLM. Specifically, we measured the performance of the VLM with a single aspect module, except for all other modules, to assess the relative influence of each aspect module on the final emotion prediction. This experiment was conducted using FI and Emotion6, which represent general image emotion datasets from social media. In contrast, for abstract images lacking concrete entities, we focused on low-level cues, such as brightness and colorfulness.

As shown in Figure 3 (a), when each aspect module was individually attached to the VLM, the model still achieved a strong performance. On the FI dataset, the average accuracy across all single-aspect settings reached 78.72%, while the MASP model that uses all six aspect modules achieved an accuracy of 80.06%. A similar trend was observed in the Emotion6 dataset, where the single-aspect average was 73.74%, and the MASP model achieved 74.75%. Interestingly, the most effective aspect varied between the two datasets. The scene module led to the highest performance on the FI dataset, whereas the brightness module was the most effective on Emotion6. This observation suggests that the importance of each visual cue depends on the characteristics of the dataset and the type of emotional content it contains. Despite the relatively strong performance of the individual aspect modules, the full MASP model consistently outperformed all single-aspect variants. This indicates that combining multiple aspect modules provides complementary information and results in more accurate and robust emotion recognition results.

As shown in Figure 3 (b), in the abstract image domain, both the brightness and colorfulness modules achieved an accuracy of 39.13%. While these results already represent substantial improvements over the zero-shot baseline accuracy of 17.39%, combining both modules in the MASP model led to a further performance gain, reaching 45.65%. This demonstrates that although each low-level cue is useful on its own, their joint use provides complementary information that significantly enhances emotion recognition performance. We also observe that the *no* and *random* prompt settings resulted in poor performance, even lower than the zero-shot baseline. This finding suggests that untrained or missing prompt signals can introduce noise or ambiguity, ultimately degrading the model's ability to reason about the emotional content.

Ablation on Soft Prompt Length

To investigate the effect of soft prompt length on emotion reasoning in VLMs, we conduct an ablation study by varying the number of learnable prompt tokens. Specifically, we evaluate five different prompt lengths—8, 16, 32, 64, and 128 tokens—on three datasets: FI, Emotion6, and ArtPhoto. The goal is to identify the optimal prompt length that maximizes emotion recognition accuracy while minimizing computational overhead and parameter redundancy.

Dataset	$l = 8$	$l = 16$	$l = 32$	$l = 64$	$l = 128$
FI	77.86	78.64	**80.06**	78.73	79.37
Emotion6	70.20	73.23	**74.75**	73.91	71.88
Artphoto	50.61	49.38	**53.09**	46.30	42.59

Table 3: Accuracy comparison of MASP models with varying soft prompt lengths.

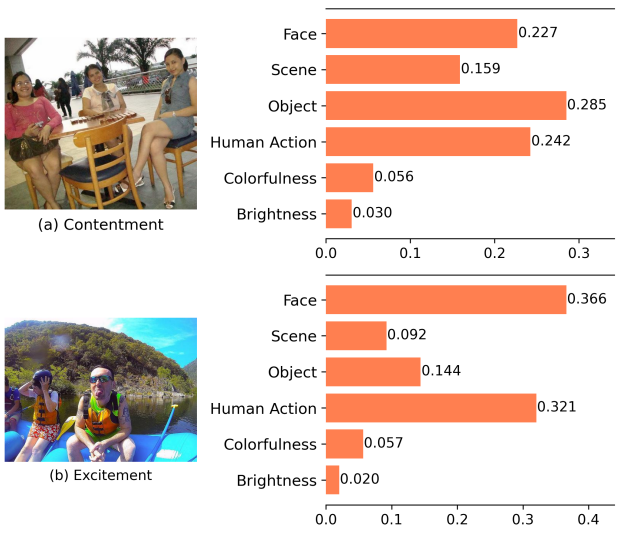

Figure 4: Visualization of attention weights from soft prompt tokens to each aspect embedding.

The results, summarized in Table 3, reveal that the performance is relatively stable across different lengths but with clear trends. In all three datasets, a prompt length of 32 leads to the highest accuracy, suggesting that it provides a favorable balance between representational capacity and efficiency. A similar trend is observed for Emotion6, where the performance peaks at 74.75% with length 32, compared to 70.20% and 71.88% at lengths 8 and 128, respectively.

Notably, the performance drop at longer prompt lengths (i.e., 64 and 128) is especially pronounced on ArtPhoto, where the accuracy drops from 53.09% at length 32 to 42.59% at length 128. This suggests that overly long prompts may introduce noise or redundancy, potentially distracting the model from emotion-relevant cues. Overall, these findings indicate that moderately sized soft prompts are not only sufficient but also optimal for guiding emotion reasoning in VLMs.

Ablation on Prompt-to-Aspect Attention Patterns

To interpret how soft prompts attend to different visual aspects during emotion recognition, we analyzed the attention scores from the final layer of the language model. We extracted the attention matrix and examined the weights from the soft prompt tokens to each of the six aspect embeddings. By averaging these values across all soft prompt tokens, we visualized the attention scores assigned to each aspect, as shown in Figure 4.

In the first image (a), the model primarily attends to object and human action cues, indicating that features such as musical instruments and relaxed postures are critical for interpreting emotion. In contrast, for the second image (b), the model focuses more heavily on facial expressions and human actions, likely because of expressive facial expressions and body postures that suggest active engagement with the environment. These examples imply that the model adaptively shifts its attention across different visual aspects depending on the emotion-relevant cues present in each image, aligning with the intuition that emotional states rely on diverse combinations of visual cues.

	Number of Trainable parameters
Instruction tuning	2,523,136
Soft prompt tuning	114,688

Table 4: Trainable parameters comparison between MASP and QLoRA.

Parameter Efficiency Comparison

In this analysis, we compare the number of trainable parameters required for soft prompt tuning in MASP with those used in instruction tuning based on QLoRA, both applied to the same base model, Qwen2-VL-7B. To ensure a fair comparison, QLoRA is applied to the query and value projection layers across all transformer blocks, which is a commonly adopted and relatively lightweight configuration. In contrast, MASP adopts soft prompt tuning with 32 learnable prompt tokens prepended to the input embeddings.

As summarized in Table 4, the soft prompt tuning approach in MASP introduces only 114 K trainable parameters, in stark contrast to the 2.5M parameters updated by QLoRA. Despite this substantial reduction, MASP consistently achieves comparable or even superior performance across datasets.

Conclusion

We propose MASP, a novel soft prompt-tuning framework for emotion understanding in VLMs. MASP explicitly encodes multi-aspect visual cues using learnable query vectors and cross-attention mechanisms, enabling the model to attend to diverse emotion-relevant features. Unlike previous approaches that rely on global visual representations or require full instruction tuning, MASP updates only a small set of soft prompt parameters. MASP achieves state-of-the-art performance across multiple benchmark datasets while maintaining computational and parameter efficiency. Inspired by the way humans interpret emotions through complementary visual signals, MASP enables structured and cognitively aligned emotional reasoning. Future work will explore cross-aspect interactions and incorporate dynamic routing strategies to enhance both sensitivity and reasoning flexibility.

Acknowledgements

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation) grant funded by the Korea government (Ministry of Science and ICT)(RS-2024-00436936, RS-2025-25440264), and the Technology Innovation Program (RS-2024-00442688) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

Abbas, R.; Ni, B.; Ma, R.; Li, T.; Lu, Y.; and Li, X. 2025. Context-based emotion recognition: A survey. *Neurocomputing*, 618: 129073.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. G. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1): 32–80.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Etesam, Y.; Yalçın, Ö. N.; Zhang, C.; and Lim, A. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4769–4776. IEEE.

Fang, W.; Wu, Q.; Chen, J.; and Xue, Y. 2025a. guided MLLM Reasoning: Enhancing MLLM with Knowledge and Visual Notes for Visual Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19597–19607.

Fang, Y.; Liang, J.; Huang, W.; Li, H.; Su, K.; and Ye, M. 2025b. Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning*.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Katsurai, M.; and Satoh, S. 2016. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2837–2841. IEEE.

Lee, S.; Lee, Y.; and Park, E. 2025. EmoVLM-KD: Fusing Distilled Expertise with Vision-Language Models for Visual Emotion Analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5633–5642.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Machajdik, J.; and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, 83–92.

Meudt, S.; Schmidt-Wack, M.; Honold, F.; Schüssel, F.; Weber, M.; Schwenker, F.; and Palm, G. 2016. Going further in affective computing: how emotion recognition can improve adaptive user interaction. In *Toward Robotic Socially Believable Behaving Systems-Volume I: Modeling Emotions*, 73–103. Springer.

Nonaka, H.; and Valles, D. 2024. Fully Auto-Regressive Multi-modal Large Language Model for Contextual Emotion Recognition. In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0291–0299. IEEE.

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 860–868.

Simcock, G.; McLoughlin, L. T.; De Regt, T.; Broadhouse, K. M.; Beaudequin, D.; Lagopoulos, J.; and Hermens, D. F. 2020. Associations between facial emotion recognition and mental health in early adolescence. *International journal of environmental research and public health*, 17(1): 330.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, J.; Zhong, M.; Xing, S.; Lai, Z.; Liu, Z.; Chen, Z.; Wang, W.; Zhu, X.; Lu, L.; Lu, T.; et al. 2024. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37: 69925–69975.

Xie, H.; Peng, C.-J.; Tseng, Y.-W.; Chen, H.-J.; Hsu, C.-F.; Shuai, H.-H.; and Cheng, W.-H. 2024. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26596–26605.

Xu, D.; Chen, Y.; Wang, J.; Huang, Y.; Wang, H.; Jin, Z.; Wang, H.; Yue, W.; He, J.; Li, H.; et al. 2024. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question an-

swering. In *Findings of the Association for Computational Linguistics ACL 2024*, 4977–4997.

Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; and Huang, H. 2023a. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20383–20394.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2023b. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36: 40924–40943.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Zhang, Q.; Wang, Z.; Zhang, D.; Niu, W.; Caldwell, S.; Gedeon, T.; Liu, Y.; and Qin, Z. 2024. Visual prompting in llms for enhancing emotion recognition. *arXiv preprint arXiv:2410.02244*.

Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.-S.; and Sun, X. 2014. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, 47–56.