Multi-View Emotion Adapter: Structured Internal Emotion Control for Vision–Language Generation

Anonymous ACL submission

Abstract

Despite strong progress in semantic and visual understanding, multimodal image captioning models still rely on prompts or external constraints for emotion control, preventing emotion from acting as a stable internal factor during generation. This leads to emotional expressions that are unstable and difficult to reproduce across layers. We propose the Multi-View Emotion Adapter (MVEA), a lightweight, plug-and-play Transformer module that converts emotion from an external stylistic cue into an internal control signal that propagates across layers. MVEA modulates hidden states from two complementary views—magnitude and direction—allowing emotion to participate stably in multi-layer generation. We further introduce a unified training objective that jointly constrains semantics, visual alignment, and emotion. To support stable training and evaluation, we construct an image–text–emotion dataset of approximately 25K samples covering seven emotion categories. Experiments across multiple mainstream multimodal models show consistent improvements in emotion controllability (Emotion Score +11%–25%, Emotion Accuracy +9%–15%), with significantly higher emotional relevance in both human and GPT-based evaluations. Notably, MVEA enables open-source models to substantially narrow the gap with strong closed-source models such as GPT-4o. Overall, MVEA provides a scalable and interpretable framework for emotion-controllable image captioning.

1 Introduction

Large-scale multimodal models such as GPT-5 and LLaVA have rapidly advanced, and modern generation tasks increasingly require not only accurate content understanding but also the ability to express information through specific emotional perspectives. Stable and controllable emotional expression affects not only output quality but also user experience and downstream task performance, making emotion control a critical component of model alignment capabilities(Zhang et al., 2023). Existing studies explore several strategies for emotional modulation, including steering emotional tendencies during decoding(Liu et al., 2021), dynamically adjusting the generation distribution via attribute-aware future discriminators(Yang and Klein, 2021), and imposing structured condition–based emotional constraints(Wang et al., 2022).

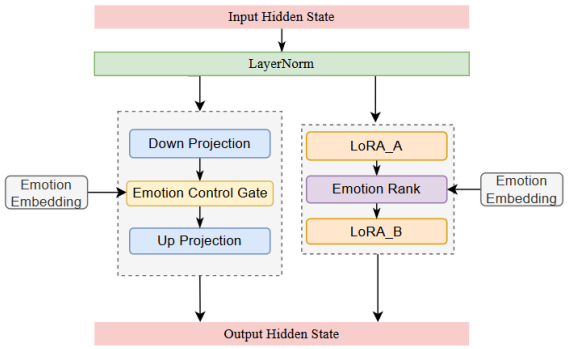

Figure 1: Overview of the MVEA architecture. Through the bottleneck path and the LoRA path, emotion information is injected into the hidden states: the former uses an emotion gate to dynamically modulate feature amplitudes, while the latter adjusts directional offsets through emotion-dependent rank vectors. Together, the two paths update the hidden state and enable emotion modulation during generation.

However, despite substantial progress in visual understanding and natural-language generation, multimodal models still fall short when tasked with producing emotionally controllable outputs. The most striking issue is that emotional signals are neither interpretable nor locatable within current architectures. Although these models can generate text with emotional tendencies, such variation is highly sensitive to prompts and sampling strategies, leading to pronounced instability(Leidinger et al., 2023). At the core of this limitation is the absence

of any explicit structure capable of carrying emotional information: emotion does not participate in generation as an independent computational factor but is instead diffusely encoded as statistical correlations scattered across model parameters(Elhage et al., 2021). Without explicit emotional representations, the model cannot reason about or edit emotional attributes, making stable, controllable, and interpretable emotional expression fundamentally unattainable.

In addition, current multimodal generation pipelines lack an integrated training mechanism that jointly constrains semantic fidelity, visual grounding, and emotional expression. Methods such as FUDGE, which modulate emotion using a future discriminator(Yang and Klein, 2021) improve attribute consistency in text-only settings, yet most optimization objectives focus solely on aligning emotion labels. This leaves no guarantee that the generated text will maintain both content faithfulness and emotional accuracy. More systematic analyses similarly report that even when an image contains clear affective cues, model outputs often remain emotionally inconsistent, revealing that existing objectives fail to establish a unified emotional–semantic space(Jeong et al., 2023). Without joint constraints, emotional information cannot exert a persistent or stable influence throughout the multi-layer generation process. The scarcity of high-quality affective supervision further exacerbates the problem. While datasets such as ArtEmis(Achlioptas et al., 2021) provide image–emotion annotations, the emotional categories are limited and insufficient for learning broad generalization patterns(Huang et al., 2019). Common multimodal corpora such as COCO(Lin et al., 2014) lack systematic emotional expression altogether, causing models to perform poorly on affect understanding tasks(Lu et al., 2024). These limitations underscore the need for a mechanism that builds explicit emotional representations within the model and propagates them through structured pathways, guided by a joint training objective capable of enforcing coherent emotional generation.

In this work, we propose Multi-View Emotion Adapter (MVEA), a lightweight module that builds explicit emotional representations inside the model and integrates them into the generation process through a structured control pathway (shown in Fig 1). The key idea is to introduce an independent emotion control stream alongside the model's original visual content stream, transforming emotion from an external prompt cue into an internal signal that is learnable and capable of exerting persistent, cross-layer influence throughout generation.

To achieve this, MVEA adopts a dual-view design that modulates feature amplitude and direction simultaneously. On one hand, a bottleneck mapping produces emotion-driven dynamic gates that adjust hidden-state magnitude. On the other hand, an emotion-aware low-rank adaptation path (emotion-aware LoRA) alters the directional updates of representations, steering expressive tendencies in an emotion-specific manner.

Beyond architectural integration, we formulate a unified joint training framework that optimizes semantic fidelity, visual grounding, and emotional alignment under a single objective. This allows emotional modulation to shape representations without degrading generation quality. To alleviate the severe scarcity of affect-supervised multimodal data, we further construct a 25,000-sample emotion-annotated image–text dataset, providing reliable visual–affective learning signals.

Experiments across multiple multimodal architectures and parameter scales demonstrate that MVEA brings consistent and significant improvements, confirming its effectiveness and robustness in emotion-controllable image captioning.
Our contributions are as follows:

- We introduce Multi-View Emotion Adapter (MVEA), a dual-path, plug-and-play module designed for emotion-controllable vision–language generation, enabling explicit and interpretable emotional signals to modulate Transformer hidden states at fine granularity.
- We curate and construct a dataset of approximately 25,000 image–emotion–text triplets, providing reliable visual–affective supervision for emotion-controllable captioning.
- We propose a joint learning framework that enforces semantic fidelity, visual grounding, and emotional alignment, allowing vision–language models to incorporate emotional signals into the text generation process. Experiments on four public top VLMs show the effectiveness of our framework.

2 Related Work

Prompt-based / Style-induced This method control emotional expression by adding emotion

prompts or style templates at the input level, without modifying the model's internal representations. Achlioptas et al. (2021) introduce explicit emotion labels and explanatory affective phrases during training, using fixed emotional cues as external prompts to steer caption generation. Li et al. (2021a) extend such prompts from fixed phrases to compositional style signals: they extract style-related phrases and perform emotion-aware rewriting to inject affective style at the input stage. Wang et al. (2023) replace discrete phrases with learnable continuous prompt vectors, modulating emotion through optimized prompt embeddings. Classic learnable-prompt approaches further generalize this idea by introducing trainable soft prefixes or soft prompts, allowing the model to internalize specific tones or emotions solely through optimizing the input prefix—without updating any backbone parameters (Lester et al., 2021; Li and Liang, 2021). In the visual domain, StyleCLIP (Patashnik et al., 2021) maps textual prompts to directions in latent space, enabling cross-modal emotional style editing without any architectural changes. Despite innovations in prompt design and learnable-prompt mechanisms, these approaches share a core limitation: emotional signals remain confined to the input layer, causing emotional control to manifest largely as surface-level lexical shifts rather than deeper representational modulation. This makes it difficult for the model to maintain stable and coherent emotional expression over long sequences.

Post-hoc Emotion Control These methods impose constraints during decoding or directly on the output distribution using external supervision signals, controlling emotion from a post-hoc perspective without constructing any internal emotional representation. InSenti-Cap (Li et al., 2021b) relies on an image-based emotion classifier to determine affective polarity and guides the generator to insert corresponding emotional words during captioning. Shen and Feng (2020) further treats emotion control as a reward-driven bidirectional learning process, using reinforcement learning at the final training stage to strengthen emotional attributes in the output. Among inference-time control methods, Dathathri et al. locally modifies hidden states during sampling through classifier gradients, nudging generation toward the target emotion. Krause et al. (2021) instead uses a small class-conditional language model to reweight token probabilities during decoding. Despite implementation differences, these methods assume that emotion can be controlled at decoding time without being explicitly represented or propagated as an internal signal. As a result, emotional modulation remains local and reactive, failing to persist coherently across layers and often producing unstable or front-loaded emotional expression that weakens over longer sequences, particularly in multimodal settings. Consequently, while post-hoc control can bias emotional tendencies, it lacks the capacity for stable, fine-grained, and interpretable emotional modulation throughout the generation process.

Internal Emotion Modeling Another line of work attempts to inject emotion as an internal signal, distinguishing itself from approaches that rely solely on external prompts or decoding-time constraints. Firdaus et al. (2021) encodes multiple emotions and their intensities into an internal memory module, though its modulation path still centers on a single memory unit. Wang et al. (2022) incorporates character-level emotion embeddings and multitask constraints, providing a representative mechanism for inserting emotional conditions along a single internal pathway. Li et al. (2022) similarly integrates emotion embeddings inside the model, using knowledge-enhanced signals to support emotion understanding. Meanwhile, emotion-trajectory modeling approaches (Xie et al., 2022) introduce internal state planning to control how emotion evolves over time, explicitly tracking affective changes in the latent space as generation unfolds. In more recent work, EmpCRL (Cai et al., 2024) and EmotionVector (Dong et al., 2025) treats emotion as an internally structured causal reasoning chain, learning affective reasoning so that emotional signals influence generation through explicit reasoning pathways in the latent space. Although these methods move beyond external prompting and post-hoc control toward introducing internal emotional mechanisms, the emotional signals they inject typically follow a shallow design. Such mechanisms lack the structured modulation capacity required for emotion to propagate consistently across multiple layers, and they cannot model emotion from multiple complementary perspectives.

In this work, we introduce Multi-View Emotion Adapter (MVEA), a multi-path emotional modulation mechanism that enables emotion to propagate persistently inside the model. MVEA embeds structured, learnable emotional signals throughout multiple Transformer layers, allowing the model to achieve stable, interpretable, and fine-grained emotional control.

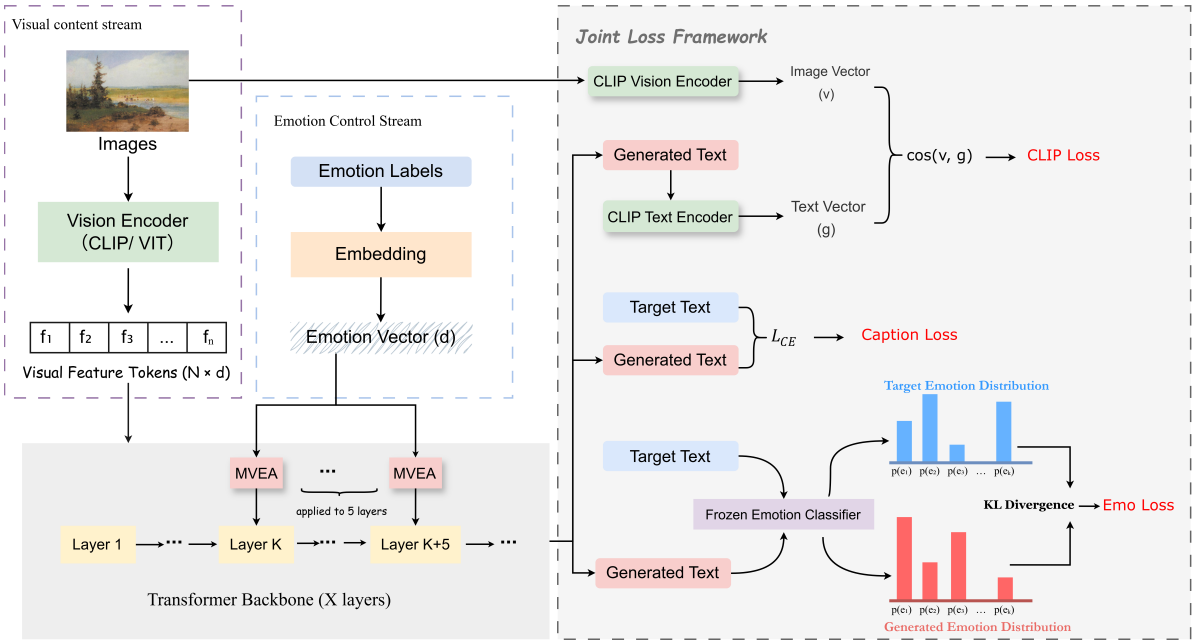

Figure 2: An overview of our model pipeline. The system consists of a visual content stream and an emotion control stream, which are fused within the Transformer backbone through the MVEA modules. The generation process is optimized under a joint loss framework that simultaneously enforces semantic correctness, visual alignment, and controllable emotional expression.

3 Method

As discussed in Section 1, emotion-controllable image captioning faces two key limitations: emotional signals are difficult to control and interpret in multimodal models due to the lack of explicit affect representations, and existing training objectives fail to jointly constrain semantic, visual, and emotional consistency, leaving emotion largely prompt-dependent. To address these issues, we elevate emotion from a prompt-level stylistic cue to a learnable internal control signal and introduce a dual-path emotion adaptor with a joint training framework that integrates emotional modulation directly into the generative process. The proposed module is fully pluggable and applicable to Transformer-based multimodal models.

3.1 Model Architecture

Our system is built upon two information pathways—the visual content stream and the emotion control stream, which are integrated within the Transformer-based text generation backbone. Given an image and a target emotion label, the model first extracts high-dimensional semantic features from its visual tower, while the emotion label is embedded into a continuous emotion representation. These two representations then enter the language backbone jointly and are fused through the MVEA modules inserted into several Transformer layers: the visual stream provides semantic grounding, while the emotion stream modulates hidden-state updates, transforming emotion from a prompt-level artifact into an internal generative factor. The entire process is optimized under a unified joint training objective, and the overall workflow is illustrated in Figure2 .

3.2 Multi-View Emotion Adapter (MVEA)

To allow emotion signals to participate in hidden-state updates in a controllable and interpretable way—while providing a modulation pathway independent of the backbone—we design the Multi-View Emotion Adapter (MVEA). MVEA transforms discrete emotion labels into internal signals that accumulate across multiple layers and continuously influence the hidden representations. Its design focuses on three aspects. First, we introduce an emotion-driven dynamic gating mechanism: the emotion vector generates channel-wise gates within the bottleneck mapping, enabling fine-grained control over activation magnitudes across different emotions and providing an explicit handle on the strength of internal representations:

$$g = \sigma(W_g e), \quad \Delta h_{\text{bott}} = W_u(\phi(W_d h) \odot g) \tag{1}$$

Here, h denotes the hidden state of the current layer, e is the emotion embedding, and W represents learnable linear transformations shared across the formulations. Second, we incorporate an emotion-modulated low-rank LoRA path, where dynamic rank weights derived from the emotion vector rescale the low-rank updates and shift hidden-state directions. This directional modulation encourages consistent emotional style in lexical choice, tonal preference, and other aspects of the generated text.

$$\Delta h_{\text{lora}} = W_B((W_A h) \odot (W_r e)) \tag{2}$$

The two paths operate from complementary perspectives—magnitude modulation and direction modulation—forming a dual-view mechanism that allows emotion signals to influence generation in a richer and more nuanced manner than traditional adapters. Besides, to preserve flexibility and controllability, MVEA is implemented as a pluggable module that explicitly intervenes in representation updates and can be easily fine-tuned for downstream tasks. The structure of MVEA is shown in Figure 1.

3.3 Joint Training Objective

To enable stable coordination among semantic correctness, visual consistency, and emotional controllability, we introduce a joint loss framework tailored for emotion-controllable image captioning. The semantic component of the generated text is supervised using a standard cross-entropy loss. Visual consistency is maintained through a CLIP-score loss based on cosine similarity, which encourages the textual representation to remain close to the corresponding image features in the shared cross-modal space. To ensure that the model expresses the specified target emotion during generation, we further incorporate an emotion-alignment loss. A frozen emotion classifier predicts emotion distributions for both the reference descriptions and the model-generated outputs, yielding the target distribution q and the generated distribution p. The KL divergence between q and p serves as the emotion loss, guiding the emotional distribution of the generated text to converge toward the target emotion.

$$\mathcal{L}_{\text{emo}} = \text{KL}(q \parallel p) = \sum_i q_i \log \frac{q_i}{p_i}, \tag{3}$$

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{cap}}\mathcal{L}_{\text{cap}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{emo}}\mathcal{L}_{\text{emo}}, \tag{4}$$

where q and p are the target and generated emotion distributions, and λ_{cap}, λ_{clip}, λ_{emo} weight the three loss terms.

4 Experiments

4.1 Setup

Datasets To alleviate the limited scale of existing emotion-aware image captioning datasets, we construct a new dataset of approximately 25,000 image–emotion–text triplets spanning seven emotion categories and about 7,600 images. We first filter ArtEmis (Achlioptas et al., 2021) by removing abstract or unsuitable samples, retaining roughly 14,000 instances, and further supplement the data with real-world images from Flickr8k (Hodosh et al., 2013) and COCO (Lin et al., 2014). Captions are obtained via a combination of manual annotation (about 4,500 samples) and LLM-based augmentation (over 7,000 samples). To control noise, all augmented captions are constrained by visually grounded prompts and validated using a frozen RoBERTa-based seven-class emotion classifier (Zhuang et al., 2021), with additional manual inspection applied to remove clear visual–text mismatches. Emotion labels follow the same taxonomy as ArtEmis. As shown in Figure 3, the dataset exhibits mild imbalance without severe long-tail effects. We split the data into training, development (15%), and test (15%) sets, where the development set is used exclusively for hyperparameter tuning. Prompts used for LLM-based augmentation are provided in Appendix B.1.

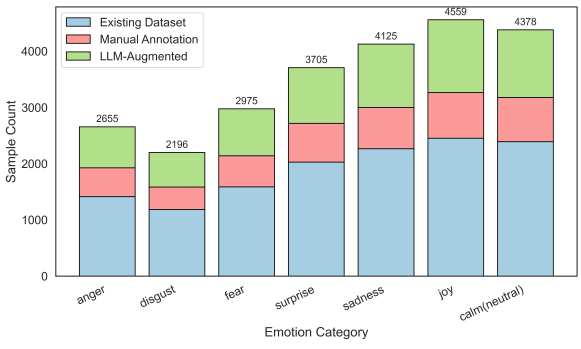

Figure 3: Sample distribution of the seven emotion categories across existing data, manual annotations, and LLM-augmented captions.

Baselines To evaluate the practical gains of MVEA in emotion-controllable generation, we select four open-source multimodal models as comparison baselines: LLaVA-Phi-3-Mini (Liu

et al., 2023), Qwen2.5-VL-7B (Bai et al., 2025), InternVL2-8B (Chen et al., 2024), and LLaVA-v1.6-Vicuna-13B (Liu et al., 2023). They span small to medium-scale Transformer architectures and represent high-performing, widely adopted open-source vision–language models, providing a representative snapshot of current mainstream multimodal modeling and enabling evaluation across different parameter scales and architectural designs. We further include LLaVA-Phi-3-Mini with FUDGE (Yang and Klein, 2021), a decoding-time control baseline that steers emotional expression by reweighting token probabilities during generation without modifying model parameters, as well as Qwen2.5-VL-7B with EmotionVector (Dong et al., 2025), an internal control baseline that injects pre-computed emotion direction vectors into intermediate hidden states during decoding. In addition, we report the zero-shot performance of GPT-4o (Hurst et al., 2024) and Gemini-3-Pro-Preview (Team et al., 2023) as non-controllable references, representing the most widely used GPT model and the strongest currently available Gemini API model, respectively. All open-source models are compared under identical training configurations.

Evaluation Metrics We evaluate emotional controllability and generation quality using a combination of automated and subjective metrics. Emotion-related performance is measured by Emotion Score and Emotion Accuracy. Emotion Score computes cosine similarity between generated text and the target emotion in a seven-class RoBERTa-based affective embedding space (Zhuang et al., 2021), while Emotion Accuracy measures whether the predicted emotion matches the target. We validate this classifier on 5,000 randomly sampled instances, achieving high accuracy ($\approx 95\%$). For text quality, we report CLIP-Score (Radford et al., 2021) for cross-modal consistency, Perplexity for linguistic fluency, and Distinct-n for lexical diversity. Because automated metrics alone cannot fully capture the subjective characteristics of emotional expression, we also incorporate both human and LLM-based subjective evaluations. Human annotators assign 1–5 scores for emotional relevance and image relevance following a unified rubric. In parallel, GPT-4o provides an aggregated score under a fixed evaluation prompt. All evaluation templates and prompts are included in the appendix A. We note that GPT-4o is also used as an automatic evaluator while serving as a zero-shot baseline model, which may introduce potential self-evaluation bias. Therefore, GPT-based scores are reported only as a complementary signal, while all main conclusions rely on automated metrics and blinded human evaluation.

Implementation Details Across all experiments, we adopt a unified training setup for the open-source models: the visual encoder and language backbone are frozen, and only the emotion projection layer and MVEA parameters are updated. Visual features are precomputed under the frozen encoder and fused via each model's native multimodal interface. We first perform lightweight hyperparameter tuning on the development set, after which we fix the joint loss weights to $\lambda_{\text{cap}} = 0.2$, $\lambda_{\text{clip}} = 1.0$, and $\lambda_{\text{emo}} = 3.0$, keeping them consistent across all backbones. All remaining training hyperparameters follow a shared configuration, with full details provided in the Appendix B.2 and B.3.

4.2 Baseline Comparison

We select four mainstream open-source Transformer backbones, ranging from lightweight to large-scale models, to examine the transferability of MVEA. GPT-4o and Gemini-3-Pro-Preview's zero-shot performances are also reported as strong non-controllable references. The full results are shown in Table 1. Figure 4 provides an visualization of how each model changes across the various evaluation metrics.

Overall, MVEA yields consistent emotional enhancement across all model sizes, increasing Emotion Score by 11%–25% and Emotion Accuracy by 9%–15%. The gain is largest on smaller models, while larger models show more saturated improvements, likely because high-capacity backbones already encode stronger implicit affective structure. In human evaluation, emotional relevance (ER) rises by 0.3–1.2 points, whereas image relevance (IR) remains stable or slightly decreases, confirming that MVEA primarily targets the affective dimension rather than visual grounding. Compared with recent related baselines such as FUDGE and EmotionVector, MVEA improves Emotion Score by approximately 10%–20% and Emotion Accuracy by 6%–12%, with consistent gains in both human and GPT-based evaluations. Importantly, GPT-based subjective scores show consistent relative trends with human emotional relevance across models, indicating that the observed gains are not driven by the evaluator itself. Notably, LLaVA-v1.6-Vicuna-13B with MVEA ap-

Model	ES ↑	EA ↑	CS ↑	PPL ↓	D2 ↑	ER ↑	IR ↑	GS ↑
LLaVA-Mini	0.52	0.67	30.99	4.55	0.27	3.32	4.27	3.68
LLaVA-Mini + MVEA	0.65	0.77	26.03	2.74	0.30	4.23	3.79	4.03
LLaVA-Mini + FUDGE	0.61	0.73	28.57	3.44	0.29	3.55	4.13	3.82
Qwen2.5-VL-7B	0.59	0.72	28.23	2.61	0.27	3.67	4.31	3.83
Qwen2.5-VL-7B + EmotionVector	0.64	0.73	23.11	3.91	0.29	3.73	4.09	3.95
Qwen2.5-VL-7B + MVEA	0.68	0.79	23.74	2.41	0.29	4.39	4.02	4.11
InternVL2-8B	0.61	0.74	31.88	2.63	0.27	3.75	4.29	3.76
InternVL2-8B + MVEA	0.71	0.83	26.12	2.51	0.30	4.54	3.87	4.18
LLaVA-v1.6-Vicuna-13B	0.66	0.78	32.34	2.59	0.29	4.28	4.48	4.37
LLaVA-v1.6-Vicuna-13B + MVEA	**0.73**	**0.85**	28.68	2.44	0.29	**4.69**	4.23	4.52
GPT-4o	0.70	0.83	**34.12**	2.35	0.31	4.41	**4.56**	**4.57**
Gemini-3-Pro-Preview	0.72	0.84	32.18	2.39	0.31	4.64	4.42	4.52

Table 1: Comparison across baselines and MVEA-enhanced models. ES = Emotion Score; EA = Emotion Accuracy; CS = CLIP-Score; PPL = Perplexity; D2 = Distinct-2; ER = human-rated emotional relevance; IR = human-rated image relevance; GS = GPT-4o-based subjective score. **Bold** indicates the best result; underline indicates the second best. GPT-4o is used as a fixed evaluator for consistency across models. While its self-evaluation may introduce mild bias, human evaluation results show consistent trends and mitigate this concern.

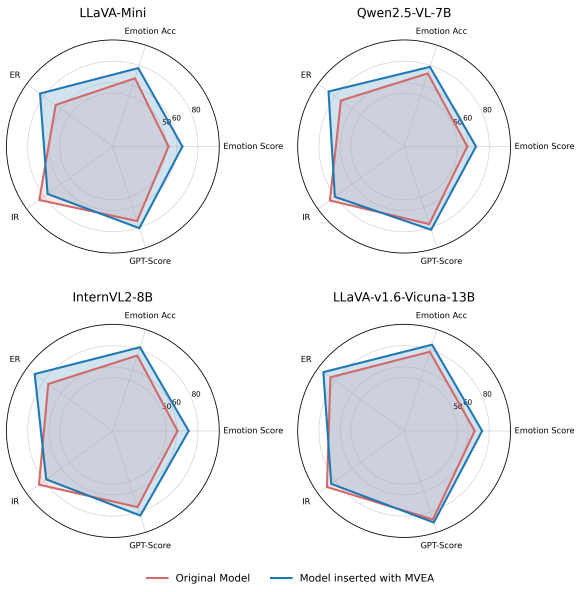

Figure 4: Radar visualization of performance changes before and after inserting MVEA. All metrics are scaled to a 0–100 range, revealing emotional gains alongside the visual trade-off.

proaches the zero-shot emotional performance of top-tier closed-source models. We further provide qualitative examples in Appendix C under fixed images and varying target emotions.

We also observe a clear visual trade-off: CLIP-Score drops by approximately 15%, indicating that the emotion pathway shifts generation away from purely visual semantics toward emotionally aligned descriptions. Nonetheless, Perplexity and Distinct-2 remain nearly unchanged, showing that MVEA preserves the backbone's language-generation quality. Overall, MVEA provides strong, scalable emotional controllability with minimal architectural overhead.

4.3 Ablation Studies

We conduct two ablation studies to analyze how MVEA's design influences controllable generation. The first removes key components—including emotion gating, the LoRA-based modulation path, emotion embeddings, and their loss terms—to isolate each mechanism's contribution. The second varies the insertion depth of MVEA to test how emotional control depends on layer placement. All experiments use LLaVA-Phi-3-Mini as the backbone, with component ablations applied to the final five decoder layers for consistency. The component-wise ablation results are shown in Table 2, and their effects on different emotional metrics are visualized in Figure 5. The overall trend is clear: the components of MVEA contribute to emotional control in a graded manner—removing any single part leads to a 5%–20% performance drop, with the emotion embedding and gating mechanism having the strongest impact. Eliminating either causes a 15%–20% decline in emotion-related

Ablation Setting	ES ↑	EA ↑	CLIP-S ↑	PPL	Distinct-2	Emo Rel ↑	Img Rel ↑	GS ↑
Full MVEA	**0.65**	**0.77**	26.03	2.74	0.30	4.23	3.79	**4.03**
LLaVA-Mini	0.52	0.67	**30.99**	3.55	0.27	3.32	**4.27**	3.68
w/o Gating	0.58	0.71	25.87	2.71	0.26	3.63	3.88	3.78
w/o LoRA Path	0.61	0.74	26.40	**2.55**	0.28	3.71	3.92	3.85
w/o Emotion Embedding	0.55	0.70	28.49	3.13	0.27	3.54	4.11	3.66
w/o Emotion Loss (KL)	0.58	0.72	28.96	2.88	0.28	3.57	4.09	3.70
w/o CLIP Loss	0.64	0.76	24.21	2.75	**0.32**	**4.31**	3.29	3.55

Table 2: Ablation studies of the proposed MVEA module. Best results are shown in **bold**; second-best results are underlined. ES: Emotion Score; EA: Emotion Accuracy; CLIP-S: CLIP-Score; PPL: Perplexity; Distinct-2: lexical diversity; Emo Rel / Img Rel: human relevance ratings (1–5); GS: GPT-4o-based subjective score.

Pos	EA ↑	CS ↑	PPL ↓	ER ↑
First-5 layers	0.72	**5.37**	2.71	3.81
Middle-5 layers	0.74	5.28	**2.66**	4.07
Last-5 layers	**0.77**	5.21	2.73	**4.23**

Table 3: Insertion position ablation for MVEA. Pos = insertion position; EA = Emotion Accuracy; CS = CLIP-Score; PPL = Perplexity; ER = human-rated emotional relevance (1–5). Best results are shown in **bold**.

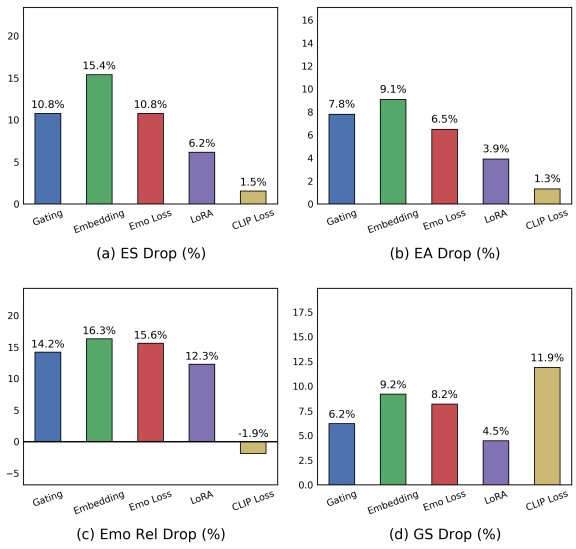

Figure 5: Relative performance drop (%) when removing each component of MVEA. We report the degradation on four emotion-related metrics—Emotion Score (ES), Emotion Accuracy (EA), human-rated Emotion Relevance (Emo Rel), and GPT-based subjective score (GS). For visualization consistency, CLIP-Score is rescaled by a factor of 1/5 before computing.

metrics, while fluency and diversity remain nearly unchanged, indicating that emotional input and amplitude modulation form the core control pathway. By contrast, removing the emotion loss results in a milder degradation, primarily causing emotional distributions to drift while leaving overall caption quality largely stable. The LoRA path and CLIP loss exert weaker effects: LoRA removal introduces about 5%–10% instability in emotional style, while removing CLIP loss reproduces the visual–emotion trade-off discussed in Section 4.2—slightly lower visual consistency but marginally higher emotion scores.

The positional ablation shows that deeper layers are more effective for emotion injection (Table 3). Early-layer insertion yields the weakest control, with emotion accuracy 6–7% below the best setting and a 10% drop in human-rated emotional relevance, indicating signal dilution. Middle-layer insertion improves performance by 3%–4% with negligible visual impact. Placing MVEA in the final five layers achieves the strongest emotional control, with only a modest 2% reduction in visual consistency. Overall, emotion control arises from the synergy of embedding, gating, and joint loss, while deeper insertion maximizes signal strength without compromising generation quality.

5 Conclusion

We propose MVEA, a lightweight modulation module that integrates emotion as a structured internal signal within multimodal generation backbones, supported by a unified joint training objective and a 25K-scale dataset. Experiments across model sizes demonstrate consistent emotional gains (with 9%–25% gains on emotion-related metrics), with performance approaching strong closed-source models. Ablation studies further show that the emotion gate, LoRA path, and emotion loss play complementary roles. This framework advances emotion-controllable image captioning and supports scalable controllable text generation.

Limitations

Emotion–Vision Trade-off Introducing emotion as an explicit internal modulation signal leads to clear gains on emotion-related metrics, while a modest decrease in visual alignment is also observed. This indicates that, under the current modeling and training framework, there remains a structural trade-off between emotional strength and semantic–visual fidelity; the joint loss can balance these objectives but does not fully optimize both simultaneously.

Emotion Representation Setting We adopt discrete emotion labels as control signals to ensure consistency and interpretability of emotion conditions, and to enable stable comparison across models and experimental settings. Under this design choice, the model can reliably switch target emotions, while finer-grained emotional continuity or composite emotional structures are not explicitly modeled. Such extensions would require more complex annotation schemes and evaluation protocols and are beyond the scope of this work.

LLM-based evaluation bias Although we employ GPT-4o to obtain a reproducible subjective evaluation signal, using the same model family for data augmentation and evaluation may introduce self-consistency bias. We mitigate this risk by excluding GPT-4o from all training stages, reporting GPT-based scores only as auxiliary indicators, and conducting blinded human evaluation, which yields trends consistent with automated judgments. Nevertheless, future work may benefit from involving multiple independent LLM evaluators or larger-scale human studies.

Emotion availability in images Our task formulation assumes that a target emotion is provided as an explicit control signal during caption generation. In practice, some images may already contain strong affective cues, raising the question of whether an additional emotion label is necessary. Our goal is not to infer the intrinsic emotion of an image, but to enable controllable emotional expression under a fixed visual scene. Since many images are emotionally ambiguous and can support multiple valid interpretations, the emotion label serves as a controllable conditioning variable rather than redundant supervision. When the image emotion is unambiguous, explicit emotion control may offer limited additional benefit.

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: Affective language for visual art. *CoRR*, abs/2101.07396.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. Empcrl: Controllable empathetic response generation via in-context commonsense reasoning and reinforcement learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5734–5746.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Yurui Dong, Luozhijie Jin, Yao Yang, Bingjie Lu, Jiaxi Yang, and Zhi Liu. 2025. Controllable emotion generation with emotion vectors. *arXiv preprint arXiv:2502.04075*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, and Nicholas Joseph. 2021. A mechanistic interpretation of transformer language models. Transformer Circuits Thread.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2021. More the merrier: Towards multi-emotion and intensity controllable response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12821–12829.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Chenyang Huang, Yong Zhao, Wenya Wang, Bing Liu, Jianbo Chen, Jian Li, Gholamreza Haffari, and Bo Qin. 2019. Emotionx: Multimodal emotion recognition benchmark. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1–6.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1

others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Minseok Jeong, Minchul Kim, Hyesu Mina, and Minjoon Yang. 2023. Is vision-language pretraining enough for affective understanding? In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Anna Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Guodun Li, Yuchen Zhai, Zehao Lin, and Yin Zhang. 2021a. Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5363–5372.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10993–11001.

Tong Li, Yunhui Hu, and Xinxiao Wu. 2021b. Image captioning with inherent sentiment. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zihan Lu, Zirui Zhang, Junqi Han, and Jian Lin. 2024. Emotionbench: Can LLMs and MLLMs understand human emotions? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5999–6021.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Lei Shen and Yang Feng. 2020. Cdl: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2023. Controllable image captioning via prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2617–2625.

Xinpeng Wang, Han Jiang, Zhihua Wei, and Shanlin Zhou. 2022. CHAE: Fine-grained controllable story generation with characters, actions and emotions. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6426–6435, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yuqiang Xie, Yue Hu, Yunpeng Li, Guanqun Bi, Luxi Xing, and Wei Peng. 2022. Psychology-guided controllable story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6480–6492.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Appendix

A Evaluation Details

A.1 Human Evaluation

To complement the limitations of objective metrics in evaluating emotional expression and image–text consistency, we manually curate a set of 500 image–emotion pairs from roughly 2,000 samples across all test sets as our human evaluation subset. Each sample contains an image, a target emotion label, and the outputs generated by different models. Annotators rate the outputs on a five-point Likert scale along two dimensions: emotional relevance (the degree to which the generated text expresses the target emotion through tone, lexical choice, and overall semantics) and image relevance (the consistency between the generated text and the visual content). All samples are evaluated following unified annotation guidelines, and final scores are computed by averaging ratings across annotators. During evaluation, annotators are blinded to the model identities behind each generated text to ensure fairness. The detailed scoring rubric is shown in the Table 4.

A.2 GPT-4o Evaluation Prompt

To obtain a reproducible subjective quality metric aligned with the human evaluation criteria, we further employ GPT-4o to generate a single aggregated score for each caption. Similar to human raters, GPT-4o receives the target emotion, the image, and the model-generated text, and produces an integer score from 1 to 5 according to a unified evaluation guideline. This score reflects an overall judgment of caption quality, jointly considering emotional expression and image relevance rather than separating them into distinct dimensions. To ensure reproducibility, we evaluate using the official GPT-4o API and provide a standardized scoring prompt, allowing the model to assign neutral ratings without knowing the source model of each caption. The full evaluation prompt is shown Table 5.

B Experiment Details

B.1 LLM Data Augmentation Prompts

To expand the emotion-controllable image captioning dataset, we design a data-augmentation pipeline that integrates human filtering, LLM generation, and model-based validation. We first perform manual screening of the raw images, removing abstract artworks and samples with insufficient visual semantics—cases that easily lead to emotional misinterpretation. This ensures that subsequent generation does not produce fake emotions or semantically implausible descriptions. We then use GPT-4o's native multimodal API for augmentation, where each image is provided through the dedicated input image field, and the target emotion label is included directly within the generation instruction. The full prompt is shown in Table 6.

Finally, we apply a frozen seven-class RoBERTa emotion classifier to automatically filter the generated texts, retaining only those whose predicted emotion matches the target label with sufficiently high confidence. Samples with low confidence or mismatched emotional predictions are discarded. To further prevent semantic inconsistencies between the caption and the image, we also conduct manual spot-checking on a subset of approximately 500 samples, improving the overall reliability of the augmentation pipeline.

B.2 Full Training Hyperparameters

To ensure comparability across multimodal models, all experiments employ a unified training configuration: the visual encoder and language backbone remain frozen, and only the emotion projection layer and MVEA parameters are updated. Visual features are extracted once before training using the frozen visual tower and then fed through each model's native multimodal interface. For optimization, we use AdamW and apply cosine learning rate decay to gradually reduce the learning rate over training. The total number of training epochs is set to four; however, based on loss curves and validation performance, we observe that the models generally converge by around the third epoch, with emotion-related metrics stabilizing even earlier—typically within the first two epochs. Subsequent training primarily serves to refine parameter stability and

Score	Emotional Relevance	Image Relevance
1	Emotion is entirely absent.	Severely inconsistent with the image; key visual content is missing.
2	Emotion is expressed weakly; occasional emotional cues appear but overall expression is unstable.	Clear mismatches or major omissions of key visual features.
3	Emotion is somewhat present, but semantic issues or ambiguity make the emotional expression unclear.	Most key features are covered, though some may be incomplete or imprecise.
4	Emotion can be reasonably inferred; the sentence is fluent, though emotional expression may still feel slightly forced.	Description is largely accurate; most key visual attributes are correct with no major deviations.
5	Emotion is expressed naturally and coherently, without being exaggerated or abrupt.	Highly consistent with the image, with both fine-grained details and key features correctly represented.

Table 4: Rubric for human evaluation of emotional relevance and image relevance (1–5).

fine-tune emotional alignment. Training remains stable across hardware configurations ranging from single-GPU RTX 4070 setups to multi-GPU A100 clusters. All hyperparameters are tuned using 5% of the development set and then fixed for all experiments. Closed-source models (GPT-4o) are used solely as uncontrollable reference baselines. We perform inference using the official API with temperature = 0.2 and max tokens = 100, without any parameter updates or post-processing. All input prompts are kept identical across models to ensure reproducibility.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.01
LR scheduler	Cosine decay
Batch size	16 (gradient accumulation = 4)
Epochs	4 (converges by epoch 3)
Max sequence length	256
Insertion layers	Last 5 decoder layers
λ_{cap}	0.2
λ_{clip}	1.0
λ_{emo}	3.0
GPT-4o API	temperature = 0.2, top-p = 0.95

Table 7: Training hyperparameters and configurations for MVEA fine-tuning.

B.3 Joint Loss Weight Selection

To determine suitable weight settings for the joint loss, we conduct both isolated perturbations and

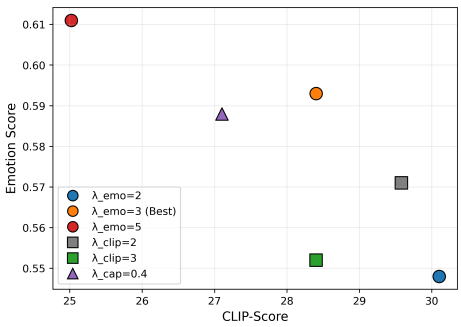

Figure 6: Joint loss weight selection on the development set.

grid-style scans over λ_{cap}, λ_{clip}, and λ_{emo} on the development set. Increasing λ_{emo} strengthens emotional alignment but inevitably weakens image–text consistency, whereas λ_{clip} exhibits the opposite tendency. The caption loss weight λ_{cap} has the smallest impact on overall performance.

Figure X presents six representative configurations selected from the development-set sweep. Circular markers denote varying λ_{emo} while holding $\lambda_{\text{cap}} = 0.2$ and $\lambda_{\text{clip}} = 1$ fixed. Square markers vary λ_{clip} with $\lambda_{\text{cap}} = 0.2$ and $\lambda_{\text{emo}} = 3$ fixed. Triangular markers vary λ_{cap} while keeping $\lambda_{\text{clip}} = 1$ and $\lambda_{\text{emo}} = 3$ fixed. The relative positions in the figure show that $\lambda_{\text{emo}} = 3$ lies in a region that balances emotional strength and visual consistency, so we adopt $\lambda_{\text{cap}} = 0.2$, $\lambda_{\text{clip}} = 1$, and $\lambda_{\text{emo}} = 3$ as the final weight configuration.

Prompt for Emotion-Controlled Caption Generation:

You are now an expert text evaluator. Your task is to assess the quality of a generated caption based on a given image description, the generated text, and a target emotion. You will be provided with an image, a target emotion (a single word), and a model-generated text. Based on the target emotion, you must assign a single overall score following the five-level criterion: assign 1 if the text lacks emotional expression or severely deviates from the target emotion and is inconsistent with the image; assign 2 if emotional expression is very weak and unstable with clear image inaccuracies; assign 3 if a basic emotional tendency is present and generally consistent with the image though some details are insufficient; assign 4 if the text expresses the target emotion naturally and accurately matches the image without major deviations; and assign 5 if the text excels in both emotional expression and image consistency, using a natural tone with complete and coherent details. The generated text does not need to describe all details of the image nor all major features, as long as it corresponds to a meaningful part of the image's primary content. Please provide only one overall score (an integer from 1 to 5), and return only the score with no additional content.

Table 5: The full prompt used to query GPT-4o for subjective aggregated evaluation. This prompt remains fixed across all experiments.

Prompt for Emotion-Controlled Caption Generation

You are a professional image-caption writer. Based on the given image and the target emotion, generate a description that aligns with the specified emotional tone. The target emotion is denoted as `<EMOTION_LABEL>`. The generated caption must strictly adhere to the visual content present in the image and must not introduce any people, objects, actions, or scenes that do not exist. Emotional expression should be conveyed naturally through tone, word choice, and narrative focus, without exaggeration, emotional overloading, or deviation from image semantics. The caption should remain concise and fluent (1–2 sentences), maintaining objectivity and semantic coherence. Emotional cues should be recognizable but should not dominate or overshadow the factual description of the image. The output format requires returning only the final generated caption, without any additional content.

Table 6: Prompt used for emotion-controlled caption generation in the GPT-4o evaluation.

C Additional Qualitative Examples

We present three representative qualitative examples to illustrate how MVEA enables fine-grained emotional control under different target emotions. In each example, the input image is held constant while the emotion label is varied, allowing us to isolate the effect of emotional modulation from visual and semantic factors. This controlled setup makes it possible to examine whether the model can systematically adjust tone, affective intensity, and lexical choice in response to different emotional conditions, without altering or hallucinating core visual content. The selected examples span scenes of varying visual and semantic complexity, including a visually rich environment with multiple salient elements (Fig 7), a real-world scenario centered on a single subject with limited context (Fig 8), and a moderately complex everyday scene (Fig 9). Together, these cases provide a comprehensive view of MVEA's behavior across different levels of visual constraint, demonstrating its ability to produce emotionally distinct yet semantically consistent descriptions for the same image.

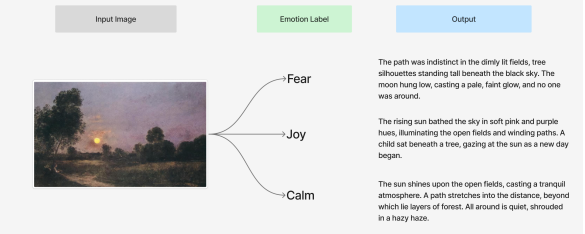

Figure 7: Qualitative examples on a visually rich environmental scene under different target emotions. The captions vary in tone and descriptive focus while preserving the core visual content.

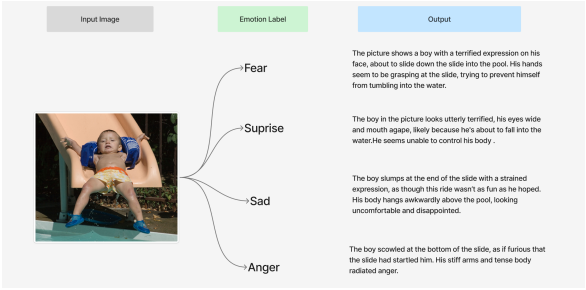

Figure 8: Qualitative results on a real-world single-subject scene with limited visual context. Emotional differences are reflected in narrative focus while remaining grounded in the image.

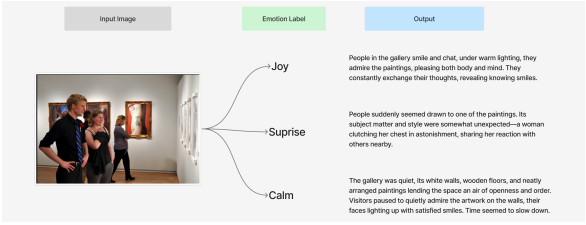

Figure 9: ualitative examples on a moderately complex everyday scene under different target emotions. The generated captions adjust emotional tone without altering the underlying scene semantics.