Semantic-Guided Progressive Multimodal Sentiment Learning with Prompt Interaction

Anonymous ACL submission

Abstract

Multimodal sentiment analysis aims to infer affective states from image–text pairs in social media. Most existing approaches rely on single-step fusion or static representations, treating affective cues as fixed and non-progressive representations. Meanwhile, prompt-based methods typically initialize prompts with sentiment-irrelevant text or random vectors, or inject auxiliary semantics in a one-step manner, failing to explicitly guide semantic evolution. To address these limitations, we propose a semantic-guided progressive framework with stage-wise prompt interaction (SPRO), which organizes multimodal supervision along a cognitively inspired trajectory from Tone to Emotion. Specifically, emotion understanding is decomposed into three successive stages—Tone, Content, and Emotion—corresponding to perceptual appraisal, semantic grounding, and affective reasoning. At each stage, LLM-generated structured captions provide explicit semantic guidance, while learnable multimodal prompts serve as a shared affective interface to progressively align visual and textual representations within a unified semantic space. Furthermore, a dual-path contrastive alignment strategy jointly optimizes image–category and text–category consistency, reinforcing cross-modal consistency. Experiments demonstrate that SPRO achieves superior accuracy and interpretability over state-of-the-art methods. The source code is publicly available.

1 Introduction

With the explosive growth of user-generated multimodal content on social media platforms, multimodal sentiment analysis (MSA) has become an important task for understanding affective meaning from paired text–image posts (Cheema et al., 2021a). Unlike textual sentiment, which is often conveyed explicitly through linguistic expressions, visual affect is typically implicit and semantically indirect, emerging from perceptual cues such as objects, color composition, lighting, and scene context. This intrinsic asymmetry creates a pronounced semantic gap between modalities, making effective sentiment inference dependent on how low-level visual cues are progressively grounded into higher-level affective semantics. Early fusion-based approaches (Wang et al., 2014; Yu et al., 2020) seek to mitigate the semantic gap by directly fusing modality-specific features in a single step, but their reliance on static representations limits their ability to capture implicit visual affect and its interaction with textual sentiment.

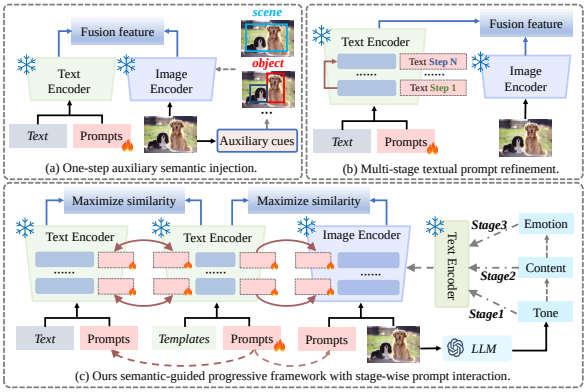

Figure 1: Comparison of multimodal affect modeling paradigms. (a) One-step auxiliary semantic injection without explicit semantic evolution. (b) Multi-stage prompting refines sentiment reasoning within the language branch but remains text-centric. (c) Our framework organizes multimodal supervision in a stage-wise manner from Tone to Content and Emotion via prompt interaction.

More recent studies have introduced prompt learning as a mechanism for injecting auxiliary cues into multimodal representations. Early approaches relied on handcrafted templates or randomized prompt vectors to guide affective prediction (Zhang et al., 2023), while subsequent work incorporated auxiliary visual cues—such as color attributes (An and Wan Zainon, 2023), salient object regions (Yu et al., 2023), and model-generated

descriptions or background knowledge (Wang et al., 2024)—to inject affective signals into fusion-based models, as illustrated in Fig. 1(a). Despite their effectiveness, these methods typically treat prompts as static, single-level signals, injecting semantic information in a one-step manner without explicitly modeling how affective semantics should evolve from perceptual cues to higher-level emotional interpretations.

In parallel, multi-stage prompting strategies have been explored on the text side to enable more structured sentiment reasoning. As shown in Fig. 1(b), these methods refine textual representations through layered reasoning processes and capture richer linguistic semantics (Zhou et al., 2024; Yang et al., 2023b). However, such approaches remain predominantly text-centric and rely heavily on explicit emotional expressions; when textual sentiment cues are weak or implicit, which is common in social media, affective reasoning cannot be reliably grounded in visual semantics, leaving cross-modal affect construction insufficiently modeled.

These limitations suggest that effective multimodal sentiment understanding requires more than static fusion or isolated reasoning, but a semantic-guided modeling of how affective meaning is progressively constructed. From psychological and neuroscientific perspectives, emotional understanding is often described as a progressive process rather than instantaneous recognition. Prior work suggests that affective meaning emerges through interactions between perceptual signals and higher-level semantic interpretations (Scherer et al., 2001; Zhang et al., 2014; Zhu et al., 2015; Sabatinelli et al., 2013), motivating the modeling of sentiment learning as a structured and progressive process that unfolds across different levels of abstraction.

Motivated by these cognitive insights and the aforementioned limitations of existing methods, we propose a **s**emantic-guided **pro**gressive framework with stage-wise prompt interaction (SPRO) for multimodal sentiment analysis, as illustrated in Fig. 1(c). The framework structures visual affect learning into three cognitively meaningful stages and integrates these stages with modality-shared prompts and a contrastive alignment objective. Shared prompt templates enable text and vision to communicate through a common latent space, while LLM-generated structured captions guide the progressive transition from perceptual cues to semantic and affective representations. A dual-path alignment component further organizes the learned representations by jointly optimizing image-category and text-category coherence. Together, these elements form an integrated system that operationalizes the progressive nature of human affect perception. The main contributions of this study are summarized as follows:

- We propose a semantic-guided multimodal prompt interaction mechanism that enables bidirectional exchange of affective cues between text and vision through shared and projected prompt tokens.

- Inspired by cognitive theories of emotion formation, we design a three-stage framework (*Tone* $\rightarrow$ *Content* $\rightarrow$ *Emotion*) that explicitly models progressive semantic abstraction for visual affect construction.

- A dual-path contrastive alignment strategy is designed to unify visual, textual, and categorical affect representations, achieving state-of-the-art performance on multiple social media sentiment benchmarks.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) aims to jointly exploit textual and visual information to enhance affective understanding in social media content. For example, Wang et al. (2014) combined image and text features for Weibo sentiment prediction, while Yu and Jiang (2019) introduced transformer-based architectures to capture sentiment-relevant visual cues. Subsequent work injected auxiliary visual signals such as color attributes (An and Wan Zainon, 2023), salient object regions (Yu et al., 2023), and external knowledge or image descriptions (Wang et al., 2024). More recent approaches, such as MAMSA (Wang et al., 2025), further explored hierarchical attention mechanisms to balance modality contributions.

Despite steady progress, most existing MSA methods still rely on single-step fusion or static attention, treating affective cues as fixed representations. Such designs overlook the progressive and constructive nature of emotion perception, where affective meaning emerges through gradual semantic abstraction from low-level perceptual impressions to higher-level emotional reasoning.

2.2 Vision-Language Models for MSA

Vision-language pre-trained (VLP) models, such as CLIP (Radford et al., 2021), provide strong cross-modal alignment through large-scale contrastive learning and have become popular backbones for affective computing. In multimodal sentiment analysis, CLIP is typically employed as a static dual encoder, whose image and text embeddings are fed into downstream fusion or attention modules. Representative works include CTMWA (Zhang et al., 2024) and MAMSA (Wang et al., 2025), which build task-specific architectures on top of CLIP features to enhance multimodal sentiment prediction. Other approaches similarly adopt CLIP as a visual backbone while relying on lightweight heads or auxiliary objectives for adaptation (Lv et al., 2025).

While effective, these methods largely treat CLIP as a feature extractor and do not explicitly adapt its intrinsic cross-modal alignment for affective consistency. Consequently, emotion representations are typically obtained in a one-shot manner, leaving the hierarchical and progressive construction of visual affect under-explored.

2.3 Prompt Learning for MSA

Prompt learning has emerged as an efficient alternative to full fine-tuning for adapting vision-language models to downstream multimodal tasks. By inserting learnable textual or visual tokens, prompt-based methods steer pretrained representations toward task-relevant semantics with minimal parameter updates. In MSA, prior work has explored mapping object labels or captions into prompts (Wang et al., 2022), applying prompt tuning for few-shot sentiment prediction (Yu et al., 2022), and incorporating topic guidance or LLM-generated affective knowledge through prompts (Liu et al., 2024; Li et al., 2023).

Despite these advances, existing prompt-based MSA methods typically rely on static prompts optimized in isolation or within a single stage. Cross-modal coordination among prompts and the progressive refinement of affective cues are largely absent. As a result, emotion is often treated as a one-step injection rather than a gradually constructed representation, motivating our semantic-guided progressive prompt interaction framework.

3 Method

The overall architecture is illustrated in Fig. 2. We propose a semantic-guided progressive framework with stage-wise prompt interaction (SPRO). SPRO formulates emotion understanding as a progressive visual–semantic construction process, in which affective meaning is gradually abstracted from low-level perceptual cues to high-level emotional reasoning. To this end, SPRO integrates prompt-mediated cross-modal interaction with structured semantic supervision, enabling text and vision to communicate through a shared affective space. On top of this representation space, a dual-path contrastive objective aligns image, text, and sentiment category prototypes to enforce cross-modal affective consistency.

3.1 Prompt-Based Cross-Modal Representation Learning

SPRO employs a prompt-mediated mechanism to construct a unified affective semantic space across post texts, visual features, and sentiment category prototypes. Learnable prompts serve as shared conditioning vectors that inject affective priors into both the text encoder E_T and the visual encoder E_V, enabling heterogeneous sentiment expressions to be aligned within a common embedding space. Unlike static prompt tuning, these prompts are explicitly shared across text, image, and sentiment prototypes, serving as a unified interface for cross-modal affect alignment.

Affective anchors from class templates. For each sentiment class cls, we define a textual template $temp_{cls}$ (e.g., "a sentiment of [positive/neutral/negative]"). A set of b learnable prompt tokens $[\theta_1, \ldots, \theta_b] \in \mathbb{R}^{b \times d_t}$ is prepended to the template and encoded by the text encoder E_T to obtain the class-level affective anchor, denoted as $Z^t_{A_{cls}}$:

$$Z^t_{A_{cls}} = E_T([\,temp_{cls}\,;\,\theta_1, \ldots, \theta_b]), \qquad (1)$$

where $[\ ;\ ]$ denotes sequence concatenation. This embedding serves as the class-level affective prototype, while instance-level text embeddings $Z^t_{x^{(i)}_{text}}$ and image embeddings $Z^v_{x^{(i)}_{img}}$ produced by the corresponding encoders are aligned to these prototypes through prompt conditioning and contrastive objectives.

Prompt-based text–anchor interaction. To capture implicit and ambiguous sentiment cues in social media posts, we encode post text $x^{(i)}_{text}$ and class templates using two textual pathways, denoted as E'_T and E_T, respectively. The two pathways share identical architectures and parameters,

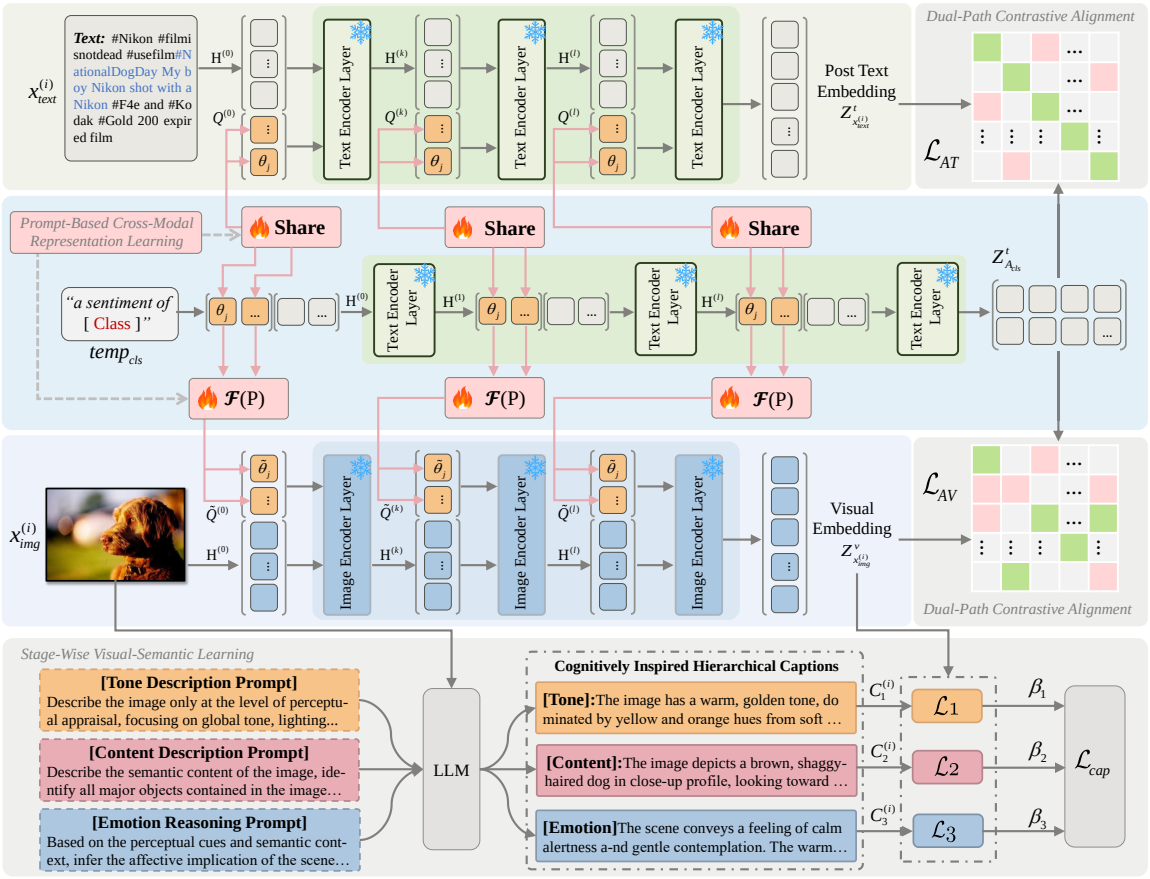

Figure 2: Overall architecture of the proposed SPRO framework. SPRO formulates multimodal sentiment analysis as a semantic-guided stage-wise learning process, where prompt-mediated cross-modal interaction constructs a shared affective space, structured captions provide stage-wise supervision (Tone, Content, and Emotion), and a dual-path contrastive objective enforces affective consistency across images, texts, and sentiment categories.

differing only in their roles for encoding instance-level texts and class-level templates. Both pathways are augmented with the same set of learnable prompt tokens. The resulting post-text representation is computed as

$$Z^t_{x^{(i)}_{text}} = E'_T([\,x^{(i)}_{text}\,;\,\theta_1,\ldots,\theta_b]), \qquad (2)$$

allowing post-level semantics to be directly aligned with affective anchors through shared prompts.

Cross-modal prompt projection. To extend affective prompting into the visual modality, each textual prompt token is projected into the visual feature space via a vision–language coupling function $\mathcal{F}(\cdot)$:

$$\tilde{\theta}_j = \mathcal{F}(\theta_j), \qquad j = 1,\ldots,b. \qquad (3)$$

This produces a set of b visual prompt tokens $[\tilde{\theta}_1,\ldots,\tilde{\theta}_b] \in \mathbb{R}^{b\times d_v}$. Given an image $x^{(i)}_{img}$, these visual prompts are prepended to the image patch embeddings and processed by the visual encoder:

$$Z^v_{x^{(i)}_{img}} = E_V([\,x^{(i)}_{img}\,;\,\tilde{\theta}_1,\ldots,\tilde{\theta}_b\,]), \qquad (4)$$

enabling affective priors derived from text to guide visual feature extraction.

Hierarchical prompting. Input-level prompt tokens act as global shallow prompts that propagate through all Transformer layers, providing consistent affective conditioning. In addition, SPRO introduces layer-specific deep prompts

$$Q^{(l)} \in \mathbb{R}^{b\times d_l}, \qquad l = 1,\ldots,D-1, \qquad (5)$$

which explicitly modulate intermediate representations:

$$H^{(l+1)} = \mathrm{TrBlock}^{(l)}\left([\,Q^{(l)}\,;\,H^{(l)}\,]\right). \qquad (6)$$

Together, shared shallow prompts and layer-wise deep prompts form a hierarchical prompting mechanism that progressively injects affective priors throughout the encoding process, resulting in a unified cross-modal affective representation space. Shallow prompts provide global affective conditioning, while deep prompts modulate intermediate representations to enable more expressive cross-modal adaptation.

3.2 Stage-Wise Visual-Semantic Learning

Built upon the prompt-mediated affective representation space, SPRO formulates emotion understanding as a progressive visual–semantic construction process. Inspired by cognitive theories of emotion perception, visual affect learning is decomposed into three stages: *Tone*, *Content*, and *Emotion*, corresponding to perceptual impressions, semantic recognition, and affective attribution.

For each image–text pair $(x_{text}^{(i)}, x_{img}^{(i)})$, we construct a three-level caption set

$$C_k^{(i)} \in \mathcal{C}^{(i)} = \{C_1^{(i)}, C_2^{(i)}, C_3^{(i)}\}, \tag{7}$$

where each caption level corresponds to a specific abstraction stage, yielding a structured instance $(x_{text}^{(i)}, x_{img}^{(i)}, C_1^{(i)}, C_2^{(i)}, C_3^{(i)})$. Here, $C_k^{(i)}$ provides stage-specific descriptions corresponding to Tone, Content, and Emotion, generated using an instruction-guided LLM with stage-specific prompts.

To enable progressive refinement during training, we introduce a linear ramp scheduling mechanism. Let e denote the current training epoch and T_{ramp} the ramp duration. The activation factor β_k for stage k is defined as

$$\beta_k = w_k \cdot \min\left(1, \max\left(0, \frac{e - s_k + 1}{T_{\text{ramp}}}\right)\right), \tag{8}$$

where s_k denotes the starting epoch at which the supervision of stage k is activated, and $s_1 < s_2 < s_3$ enforces the progressive order from *Tone* to *Content* and *Emotion*.

For each caption level, we compute a caption alignment loss

$$\mathcal{L}_{cap}^{(k)} = -\log \frac{\exp(\text{sim}(Z^v_{x_{img}^{(i)}}, Z^t_{C_k^{(i)}})/\tau)}{\sum_{C'} \exp(\text{sim}(Z^v_{x_{img}^{(i)}}, Z^t_{C'})/\tau)}. \tag{9}$$

The overall progressive caption objective is then defined as

$$\mathcal{L}_{cap} = \sum_{k=1}^{3} \beta_k \, \mathcal{L}_{cap}^{(k)}. \tag{10}$$

3.3 Dual-Path Contrastive Alignment

After obtaining sentiment-aware representations from textual and visual modalities, we introduce a dual-path contrastive alignment strategy to unify emotional understanding across images, texts, and class-level affective prototypes. This design enforces that multimodal representations are organized within a shared affective semantic space.

Specifically, two symmetric contrastive paths are defined: (i) image–prototype alignment and (ii) text–prototype alignment. For an image $x_{img}^{(i)}$ with visual representation $Z^v_{x_{img}^{(i)}}$ and its corresponding affective prototype $Z^t_{A_{cls}}$, the image–prototype contrastive loss is formulated as

$$\mathcal{L}_{AV} = -\log \frac{\exp(\text{sim}(Z^v_{x_{img}^{(i)}}, Z^t_{A_{cls}})/\tau)}{\sum_{cls'} \exp(\text{sim}(Z^v_{x_{img}^{(i)}}, Z^t_{A_{cls'}})/\tau)}. \tag{11}$$

Similarly, for the post-text representation $Z^t_{x_{text}^{(i)}}$, the text–prototype alignment loss is defined as

$$\mathcal{L}_{AT} = -\log \frac{\exp(\text{sim}(Z^t_{x_{text}^{(i)}}, Z^t_{A_{cls}})/\tau)}{\sum_{cls'} \exp(\text{sim}(Z^t_{x_{text}^{(i)}}, Z^t_{A_{cls'}})/\tau)}. \tag{12}$$

These two contrastive objectives jointly encourage both visual and textual representations to converge toward the correct affective prototype, thereby promoting cross-modal semantic consistency.

The final training objective integrates dual-path contrastive alignment with progressive caption supervision:

$$\mathcal{L}_{total} = \lambda_{AV}\,\mathcal{L}_{AV} + \lambda_{AT}\,\mathcal{L}_{AT} + \lambda_{cap}\,\mathcal{L}_{cap}, \tag{13}$$

where λ_{AV}, λ_{AT}, and λ_{cap} control the relative contributions of image–prototype alignment, text–prototype alignment, and stage-wise caption supervision, respectively.

4 Experiments

4.1 Datasets and Implementation Details

We evaluate our framework on two Twitter-based multimodal sentiment datasets, **MVSA-Single** and **MVSA-Multiple** (Niu et al., 2016), where each sample is an image–text pair labeled as *positive*, *neutral*, or *negative*. We follow the preprocessing and data splits of Li et al. (Li et al., 2022) (80%:10%:10% for train:val:test) and report Accuracy (ACC) and Weighted-F1 (F1). Dataset statistics are shown in Table 1.

Models are trained using AdamW with cosine learning-rate decay. We use dataset-specific batch sizes and training epochs. Unless specified otherwise, we set $\lambda_{\text{cap}} = 0.25$, $T_{\text{ramp}} = 15$, temperature $\tau = 0.07$, and the prompt length $b = 16$.

Table 1: Statistics of the MVSA datasets.

Dataset	Label	Train	Val	Test	Total
MVSA-Single	Positive	2147	268	268	2683
	Neutral	376	47	47	470
	Negative	1088	135	135	1358
	All	**3611**	**450**	**450**	**4511**
MVSA-Multiple	Positive	9056	1131	1131	11318
	Neutral	3528	440	440	4408
	Negative	1040	129	129	1298
	All	**13624**	**1700**	**1700**	**17024**

4.2 Main Results

Table 2 reports the performance of unimodal and multimodal baselines on MVSA-Single and MVSA-Multiple. Text-only models consistently outperform image-only models, reflecting that sentiment in social media posts is often expressed explicitly in text, while visual affect is more implicit and challenging to model. This observation highlights the modality imbalance commonly observed in multimodal sentiment analysis.

Multimodal approaches achieve clear improvements over unimodal baselines, demonstrating the benefit of cross-modal integration. However, most existing methods rely on single-step or static fusion strategies, which limits their ability to capture affective cues at different levels of abstraction.

Our proposed SPRO achieves the best overall performance on both datasets. Compared with the strongest baselines, SPRO yields consistent gains in both Accuracy and Weighted-F1, indicating that explicitly modeling emotion construction as a progressive process and strengthening visual representations through guided cross-modal interaction are effective for multimodal sentiment understanding.

4.3 Ablation Studies

We analyze the contribution of each component in SPRO through ablation studies on MVSA-Single and MVSA-Multiple. Results are reported in Table 3. Removing caption-based supervision leads to clear performance degradation, indicating that structured image descriptions provide effective guidance for bridging perceptual and affective representations. Replacing the three-stage progressive supervision with a single-stage variant also results in noticeable drops, confirming that modeling emotion construction as a hierarchical and progressive process is more effective than one-step supervision.

We further observe that removing visual-side prompts causes a substantial decline, especially on MVSA-Single, highlighting the importance of guiding the visual pathway with affective priors. Similarly, removing text-side prompts degrades performance, demonstrating that bidirectional prompt interaction is essential for coherent cross-modal alignment. Overall, the full three-stage progressive model achieves the best performance.

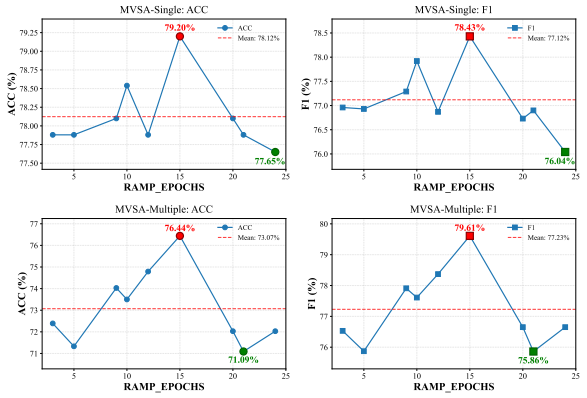

Figure 3: Parameter sensitivity analysis of the ramp-up duration T_{ramp} on the MVSA datasets.

Sensitivity Analysis. We first examine the ramp-up duration T_{ramp}, which controls the progressive activation of the three-stage supervision. As shown in Fig. 3, performance initially improves as T_{ramp} increases and then declines when the ramp duration becomes too large. Optimal performance is achieved around T_{ramp}=15, indicating a balanced trade-off between early training stability and late-stage semantic guidance.

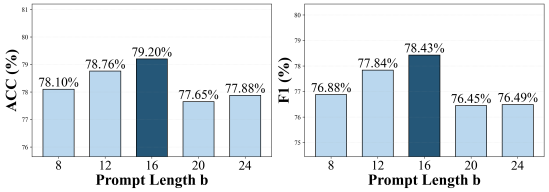

Figure 4: Prompt length sensitivity analysis on MVSA-Single.

We also evaluate the effect of prompt length b. Fig. 4 shows that moderate prompt length yields the best performance, while excessively long prompts lead to performance degradation. This suggests that concise yet expressive prompt contexts are sufficient for effective multimodal affect alignment.

4.4 Visualization Analysis

Representation-Level Visualization. To illustrate how SPRO constructs a coherent affective space,

Table 2: Comparison with unimodal and multimodal baselines on MVSA-Single and MVSA-Multiple. The strongest baseline results are underlined, and improvements of **SPRO** over the strongest baseline are shown in parentheses.

Modality	Model	MVSA-Single		MVSA-Multiple	
		ACC	F1	Accuracy	F1
Text	CNN (Kim, 2014)	68.19	55.90	65.64	57.66
	BiLSTM (Wang and Yang, 2020)	70.12	65.06	67.90	67.90
	BERT (Devlin et al., 2019)	71.11	69.70	67.59	66.24
Image	ResNet-50 (He et al., 2016)	64.67	61.55	61.88	60.98
	ViT (Dosovitskiy et al., 2021)	63.78	62.26	61.94	61.19
Multimodal	CLIP (Radford et al., 2021)	72.20	71.40	69.10	63.40
	MultiSentiNet (Xu and Mao, 2017)	69.84	69.63	68.16	68.11
	Se-MLNN (Cheema et al., 2021b)	75.33	73.76	66.35	61.89
	CLMLF (Li et al., 2022)	75.33	73.46	72.00	69.83
	M2CL (Yang et al., 2023a)	75.50	74.20	73.20	70.50
	SHSL (Han et al., 2024)	75.72	75.20	71.20	70.39
	CTMWA (Zhang et al., 2024)	75.91	75.74	**74.02**	**73.84**
	DRF (Wu et al., 2024)	74.50	74.40	71.00	68.20
	SCDR (Xia et al., 2025)	66.60	–	67.76	–
	MAMSA (Wang et al., 2025)	**77.61**	76.84	–	–
	CoDe (Wu et al., 2026)	76.98	**76.47**	72.74	70.71
	SPRO (ours)	**79.20** (+1.59)	**78.43** (+1.96)	**76.44** (+2.42)	**79.61** (+5.77)

Table 3: Ablation study on the MVSA datasets. All numbers in parentheses denote changes relative to the *w/o Caption Supervision* baseline.

Model Variant	MVSA-Single		MVSA-Multiple	
	ACC	F1	ACC	F1
w/o Caption Supervision(baseline)	74.78	73.47	73.56	70.24
Single-Stage Caption	76.77 (+1.99)	76.12 (+2.65)	74.08 (+0.52)	73.95 (+3.71)
w/o Visual Prompts	72.79 (-1.99)	70.60 (-2.87)	71.62 (-1.94)	73.78 (+3.54)
w/o Text Prompts	75.00 (+0.22)	74.11 (+0.64)	72.21 (-1.35)	76.58 (+6.34)
Three-Stage Progressive	**79.20** (+4.42)	**78.43** (+4.96)	**76.44** (+2.88)	**79.61** (+9.37)

we visualize the learned sentiment representations on MVSA-Single using t-SNE (Fig. 5).

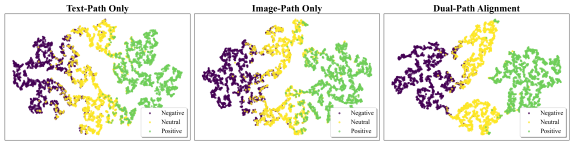

Figure 5: t-SNE visualization of sentiment representations on MVSA-Single. Compared to the unimodal streams, our dual-stream fusion achieves clearer and more compact sentiment clustering.

Compared with unimodal text-only and image-only representations, the proposed dual-path alignment produces more compact and better separated sentiment clusters. While text representations exhibit clearer polarity boundaries, image representations are relatively more diffuse due to implicit visual affect; jointly aligning both modalities toward shared affective anchors yields more discriminative multimodal sentiment representations.

Progressive Visual Attention. Fig. 6 visualizes the attention maps across the three progressive stages and the final aggregation. At Stage-1, the model mainly attends to global tonal regions, while Stage-2 shifts focus toward sentiment-relevant objects and scene elements, indicating enhanced semantic grounding. At Stage-3, attention further concentrates on fine-grained emotional cues such as facial expressions or localized visual details. The final attention aggregates evidence from all stages, resulting in more stable and discriminative patterns.

Overall, these visualizations show that SPRO

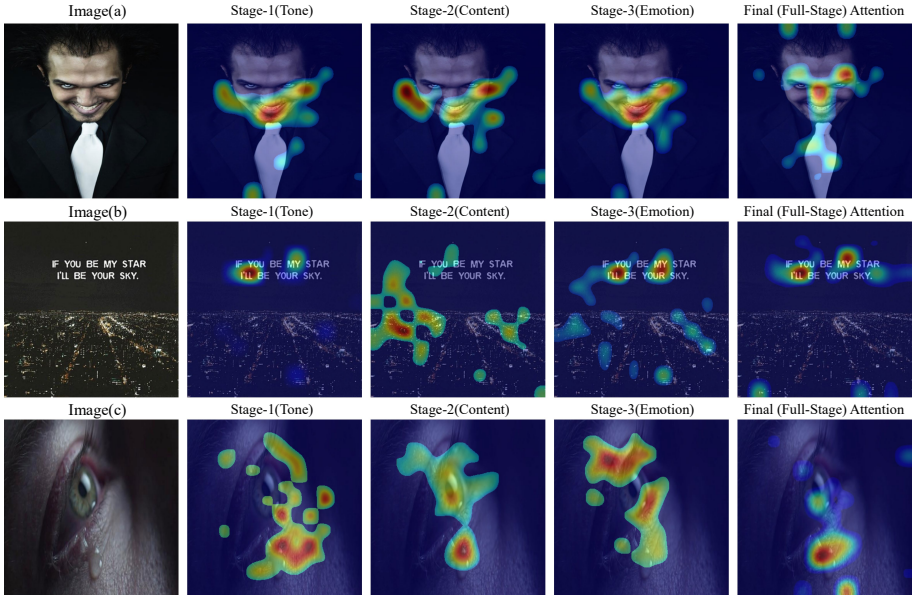

Figure 6: Visualization of progressive visual attention across three stages. Stage-1 attends to tonal regions; Stage-2 focuses on semantic content; Stage-3 captures fine-grained emotional cues; the final attention aggregates all stages.

progressively refines visual affective evidence from coarse perceptual cues to higher-level affective representations, rather than relying on a single-shot fusion process.

4.5 Case Studies

Fig. 7 shows representative examples illustrating the effect of progressive caption supervision. Cases (a) and (b) demonstrate **semantic enhancement**, where caption guidance compensates for misleading or missing textual cues by introducing visual semantics. Cases (c) and (d) illustrate **conflict adjustment**, where the proposed three-stage framework progressively reconciles conflicting textual and visual signals. These examples indicate that SPRO enables more robust multimodal sentiment inference beyond single-step fusion.

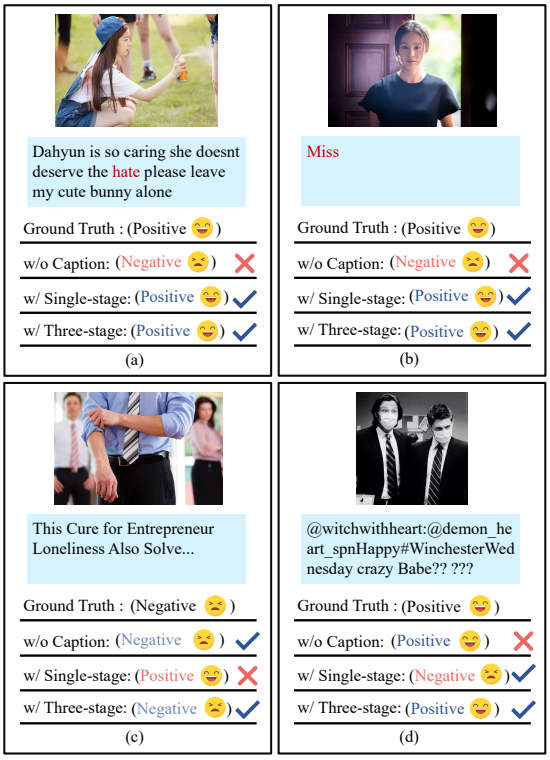

Figure 7: Qualitative case analysis on MVSA dataset. Subfigures (a) and (b) demonstrate **semantic enhancement**, where progressive captions compensate for implicit or missing textual cues. Subfigures (c) and (d) illustrate **conflict adjustment**, where the proposed three-stage supervision reconciles opposing textual and visual polarities to achieve coherent sentiment prediction.

5 Conclusion

We propose SPRO, a cognitively inspired and *semantic-guided* framework that models multimodal sentiment learning as a progressive construction process from low-level tonal cues to high-level semantic and affective representations. By leveraging LLM-generated stage-wise captions as explicit semantic guidance, together with shared multimodal prompts and dual-path contrastive alignment, SPRO effectively integrates visual and textual affective information within a unified affective semantic space. Experiments on MVSA-Single and MVSA-Multiple demonstrate consistent improvements over strong baselines, validating both the effectiveness and interpretability of the proposed semantic-guided progressive learning paradigm.

Limitations

Our work has several potential limitations. While the proposed stage-wise prompted learning framework demonstrates improved performance on multimodal sentiment analysis benchmarks, its effectiveness relies on the quality of the stage-wise caption supervision and prompt design. The generalization of this framework to other domains or more complex affective settings requires further investigation. In addition, the current study focuses on static image–text pairs, and extending the approach to additional modalities or dynamic scenarios remains an open direction for future work.

Ethical Considerations

This work studies multimodal sentiment analysis on publicly available social media datasets. The proposed framework does not introduce new data collection or annotation procedures and does not involve human subjects. While the use of automatically generated captions may reflect biases present in the underlying language models or datasets, this work does not make claims beyond the analyzed benchmarks.

References

Jieyu An and Wan Mohd Nazmee Wan Zainon. 2023. Integrating color cues to improve multimodal sentiment analysis in social media. *Engineering Applications of Artificial Intelligence*, 126:106874.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021a. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding*, MMPT '21, page 37–45, New York, NY, USA. Association for Computing Machinery.

Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. 2021b. A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*, pages 37–45. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xue Han, Honlin Cheng, Jike Ding, and Suqin Yan. 2024. Semisupervised hierarchical subspace learning model for multimodal social media sentiment analysis. *IEEE Trans. Consumer Electron.*, 70(1):3446–3454.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023. Prompting chatgpt in mner: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802. Association for Computational Linguistics.

Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2282–2294, Seattle, United States. Association for Computational Linguistics.

Wuchao Liu, Wengen Li, Yu-Ping Ruan, Yulou Shu, Juntao Chen, Yina Li, Caili Yu, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. 2024. Weakly correlated multimodal sentiment analysis: New dataset and topic-oriented model. *IEEE Trans. Affect. Comput.*, 15(4):2070–2082.

Jialun Lv, Qimeng Yang, Shengwei Tian, Bo Liu, and Long Yu. 2025. Clip-driven attention network for multimodal sentiment analysis. *J. Supercomput.*, 81(8):902.

Teng Niu, Shuhui Zhu, Liangliang Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multiview social data. In *MultiMedia Modeling (MMM 2016)*, volume 9517 of *Lecture Notes in Computer Science*, pages 15–27. Springer, Cham.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Dean Sabatinelli, Andreas Keil, David W. Frank, and Peter J. Lang. 2013. Emotional perception: Correspondence of early and late event-related potentials with cortical and subcortical functional mri. *Biological Psychology*, 92(3):513–519. Specificity, Methodology and Psychopathology of Emotional Attention.

Klaus R. Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York, USA.

Hongbin Wang, Chun Ren, and Zhengtao Yu. 2025. Multimodal sentiment analysis based on multiple attention. *Eng. Appl. Artif. Intell.*, 140:109731.

Min Wang, Dong Cao, Lin Li, Shaozi Li, and Rongrong Ji. 2014. Microblog sentiment analysis based on cross-media bag-of-words model. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 76–80. ACM.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, New York, NY, USA. Association for Computing Machinery.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Jiabo Ye, Ming Yan, and Yanghua Xiao. 2022. Promptmner: Prompt-based entity-related visual clue extraction and integration for multimodal named entity recognition. In *International Conference on Database Systems for Advanced Applications*, pages 297–305. Springer.

Zihan Wang and Bo Yang. 2020. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT. In *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020, Calgary, AB, Canada, August 17-22, 2020*, pages 562–568. IEEE.

Daiqing Wu, Dongbao Yang, Huawen Shen, Can Ma, and Yu Zhou. 2026. Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and decomposition. *Pattern Recognition*, 172:112719.

Daiqing Wu, Dongbao Yang, Yu Zhou, and Can Ma. 2024. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 5780–5789. ACM.

Wuyou Xia, Guoli Jia, Sicheng Zhao, and Jufeng Yang. 2025. Seek common ground while reserving differences: Semi-supervised image-text sentiment recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 29601–29611. Computer Vision Foundation / IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 2399–2402. ACM.

Xiaocui Yang, Shi Feng, Daling Wang, Pengfei Hong, and Soujanya Poria. 2023a. Multiple contrastive learning for multimodal sentiment analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. 2023b. Few-shot multimodal sentiment analysis based on multimodal probabilistic fusion prompts. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 6045–6053. ACM.

Jianfei Yu, Kai Chen, and Rui Xia. 2023. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1966–1978.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 5408–5414, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Jianfei Yu, Jing Jiang, and Rui Xia. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:429–439.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 189–198. ACM.

Baozheng Zhang, Ziqi Yuan, Hua Xu, and Kai Gao. 2024. Crossmodal translation based meta weight adaption for robust image-text sentiment analysis. *IEEE Trans. Multim.*, 26:9949–9961.

Dandan Zhang, Weiqi He, Ting Wang, Wenbo Luo, Xiangru Zhu, Ruolei Gu, Hong Li, and Yue-jia Luo. 2014. Three stages of emotional word processing: An erp study with rapid serial visual presentation. *Social Cognitive and Affective Neuroscience*, 9(12):1897–1903.

Sitao Zhang, Yimu Pan, and James Z. Wang. 2023. Learning emotion representations from verbal and nonverbal communication. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18993–19004. IEEE.

Jingyi Zhou, Jie Zhou, Jiabao Zhao, Siyin Wang, Haijun Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. A soft contrastive learning-based prompt model for few-shot sentiment analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10016–10020. IEEE.

Chuanlin Zhu, Weiqi He, Zhengyang Qi, Lili Wang, Dongqing Song, Lei Zhan, Shengnan Yi, Yuejia Luo, and Wenbo Luo. 2015. The time course of emotional picture processing: An event-related potential study using a rapid serial visual presentation paradigm. *Frontiers in Psychology*, 6:954.