

GSKA: Generative Semantic Anchoring and Bi-directional Knowledge Alignment for Multimodal Sentiment Analysis

Chunling Dong*, Jiao Qian and Qixuan Wang

School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China

* Correspondence: dongchunling@cuc.edu.cn (Dong C)

Abstract

The rapid proliferation of social media has generated an abundance of multimodal data characterized by complex interactions between visual content and user-generated text. Despite recent progress, Multimodal Sentiment Analysis (MSA) remains hindered by the affective gap and the inherent semantic misalignment in unconstrained social media posts. These challenges often lead to severe representation entanglement and modality collapse. In this paper, we propose Generative Semantic Anchoring and Bi-directional Knowledge Alignment (GSKA), a novel framework that formalizes MSA as a knowledge-guided cross-modal alignment task within a learned affective latent manifold. The core of GSKA lies in the introduction of generative semantic anchors, where descriptive captions are utilized to mitigate cross-modal semantic misalignment by distilling model-derived visual context into structured linguistic representations. These model-generated descriptions serve as soft semantic references rather than external factual ontologies, and are integrated through a learnable modulation mechanism that adaptively suppresses potential caption errors. To facilitate deep interplay, we design a bi-directional knowledge alignment module that leverages these anchors as a semantic filter to prioritize emotionally salient regions while suppressing sensory noise. Furthermore, we introduce Emo-Loss, a margin-aware objective function that enforces structured separability in the affective manifold. By leveraging cosine margin optimization, Emo-Loss rigorously widens inter-class boundaries and forces the model to disentangle conflicting affective states into well-separated decision zones, thereby resolving representation overlap. Extensive experiments on the MVSA-Single and MVSA-Multiple datasets demonstrate that GSKA significantly outperforms contemporary state-of-the-art methods. Our results confirm the framework's robustness in resolving semantic ambiguity, achieving superior discriminative power through enforced structural disentanglement.

Keywords: Multimodal sentiment analysis, Knowledge alignment, Generative semantic anchors, Emo-Loss, Affective latent manifold

Citation: Dong C, Qian J, Wang Q. 2026. GSKA: Generative Semantic Anchoring and Bi-directional Knowledge Alignment for Multimodal Sentiment Analysis. *The Knowledge Engineering Review* 41: e013 <https://doi.org/10.48130/ker-0026-0015>

1 Introduction

Emotions represent a fundamental component of human cognition, exerting a profound influence on learning processes, social interactions, and perceptual attitudes. While early investigations into sentiment analysis were predominantly centered on textual content, the rapid proliferation of social media platforms has catalyzed a paradigm shift toward multimodal expression. In modern domains such as healthcare and finance, interpreting human behavior necessitates the robust integration of intertwined visual and linguistic cues^[1,2]. Consequently, the automated detection of sentiment from heterogeneous data streams has emerged as a pivotal research frontier^[3].

A primary obstacle in this field is the affective gap, which is defined as the inconsistency between low-level visual features and high-level emotional perceptions^[4]. Unlike objective classification tasks, emotional perception is inherently subjective and context-dependent. As illustrated in Fig. 1, images with nearly identical semantic content may evoke disparate emotional responses, while semantically divergent objects may elicit the same sentiment. This complexity is further exacerbated in unconstrained social media environments, where cross-modal semantic divergence and linguistic noise are pervasive. In particular, user-generated text frequently provides weakly correlated or ambiguous signals, making reliable sentiment inference even more challenging. As further exemplified in Fig. 2, fine-grained correspondence between emotionally salient image regions and specific textual tokens demands more nuanced interaction. Although previous studies by Xu & Mao^[5] and Wang et al.^[6] attempted to bridge this gap through deep semantic extraction, many existing methodologies model

image-text relationships at a coarse level, neglecting the fine-grained correspondence between emotionally salient visual regions and specific textual tokens.

Furthermore, the inherent subjectivity of emotions and severe linguistic noise in social media often exacerbate this problem. Relying on simple feature concatenation for fusion is insufficient for resolving the semantic gap, frequently leading to modality collapse and overlapping decision boundaries, in which case the model fails to extract discriminative features and disentangle conflicting affective states from noisy inputs^[7,8].

To circumvent these limitations, this paper introduces the GSKA framework. The core philosophy of GSKA is to treat Multimodal Sentiment Analysis (MSA) as a knowledge-guided cross-modal alignment task. We leverage a generative model to distill high-level semantic descriptions from visual inputs, which serve as generative semantic anchors. These anchors function as a stable, visually grounded semantic reference that grounds both raw visual patches and noisy textual tokens in a unified affective latent manifold. To optimize this manifold, we propose a novel Emo-Loss function. Unlike conventional classification objectives that suffer from overlapping decision boundaries, Emo-Loss enforces structured separability by leveraging cosine margin optimization. This rigorously widens inter-class boundaries and forces the model to disentangle conflicting affective states into well-separated decision zones.

The main contributions of this work are summarized as follows:

(1) We propose GSKA, a knowledge-augmented framework that transforms visual understanding into a semantic alignment task. By distilling generative textual descriptions to serve as knowledge anchors,

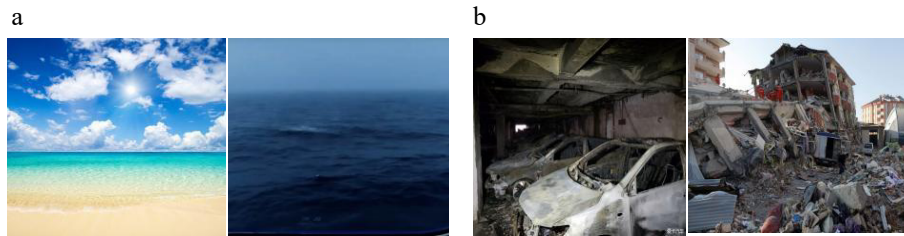
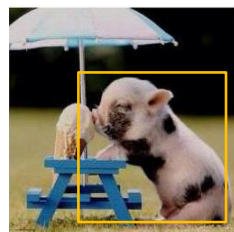


Fig. 1 Examples of the affective gap.



grattis min griskulting!!!!??? va bara tvungen oki s? sch ?
@ingenkommerattrodig #pig #happybday #wow #lovely
#cut...



This **wee** guy didn't pee on my porch. AFAIK. #peegate
#cdnpoli #elxn42

Fig. 2 Examples of image-text pairs collected from Twitter.

the model effectively mitigates cross-modal semantic misalignment and provides a robust reference for noisy multimodal inputs.

(2) We propose an Emotion-oriented Loss (Emo-Loss), a margin-aware objective function that enforces rigorous structural disentanglement and inter-class separability through cosine margin optimization in the affective manifold.

(3) We introduce a bi-directional knowledge alignment module that utilizes generated captions as a semantic filter. This mechanism prioritizes emotionally salient features while suppressing irrelevant noise and preventing modality collapse during the fusion process.

(4) Extensive experiments on MVSA datasets demonstrate that GSKA outperforms contemporary state-of-the-art methods, particularly in resolving representation overlap and achieving superior discriminative power.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed method. Section 4 presents the experimental results and ablation studies. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Related work

2.1 Multimodal sentiment analysis

Multimodal sentiment analysis has transitioned from early feature-centric fusion to deep learning paradigms that recognize affective states as complementary signals. Probabilistic graphical models were established to capture semantic correlations^[9], while structured forests were utilized to bridge the representational gap between low-level visual features and high-level descriptors^[10]. With the advent of Transformers, the research focus shifted toward fine-grained interaction modeling. Consistency-driven frameworks were proposed to integrate hierarchical visual representations^[11], and the relationships between affectively salient image regions and textual elements were investigated^[12]. More recently, Broad Attentive Graph Fusion Networks have been introduced^[13] to capture high-order feature interactions through graph-based reasoning.

In the current landscape, research focus has increasingly shifted toward Parameter-Efficient Fine-Tuning (PEFT)^[14,15] and Contrastive Alignment to adapt large-scale foundation models like Contrastive Language-Image Pretraining (CLIP)^[16] for specialized sentiment tasks^[17–19]. Despite these advances, a fundamental challenge persists: modality incongruence. As depicted in Fig. 3, user-generated text in social media is frequently dominated by subjective expressions, colloquialisms, or linguistic noise, rendering unimodal text-based inference unreliable. Without the structured visual context to ground the semantics, such subjective cues can be highly ambiguous or misleading. Consequently, relying solely on a single modality or employing coarse-grained fusion is insufficient. GSKA addresses this by introducing a generative semantic anchor that distills structured visual descriptions from the visual input. This anchor acts as a stable, third-party reference to ground the noisy textual tokens, thereby stabilizing the cross-modal interaction and preventing the model from being misled by subjective linguistic noise.

2.2 Knowledge-driven image understanding

Image captioning bridges the cognitive gap by transforming visual patterns into semantically meaningful textual expressions^[20]. The integration of deep neural networks in both computer vision and natural language processing has provided a powerful foundation for cross-modal comprehension. The field was significantly advanced by the introduction of BLIP-2^[21], which utilizes pre-trained Querying Transformers to achieve state-of-the-art captioning fidelity. In the context of sentiment analysis, leveraging generative models provides an exogenous source of common-sense knowledge. Generated captions can function as a semantic bottleneck^[22], filtering affectively irrelevant noise (e.g., trivial background objects) before performing sentiment inference. While very recent studies, including Achiam et al.^[23] have begun exploring Large Language Models (LLMs) like GPT-4 to provide reasoning-based knowledge for MSA, the heavy computational overhead of such LLM-based approaches remains a significant bottleneck for real-time applications. GSKA adopts a more streamlined approach by using specialized generative models to produce

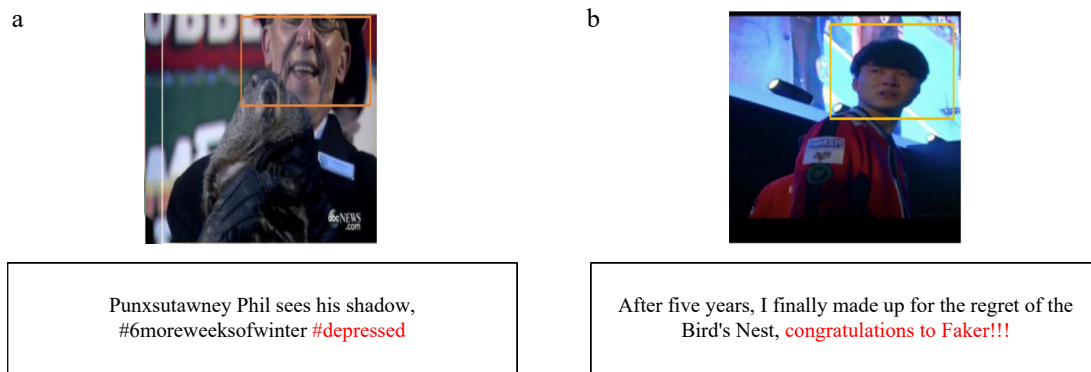


Fig. 3 Examples illustrating the unreliability of unimodal inference when text is dominated by subjective expressions or linguistic noise. The structured visual context provides the necessary semantic grounding for accurate sentiment prediction.

stable semantic coordinates, providing a balanced trade-off between semantic richness and computational efficiency.

While several prior works have incorporated generated captions as additional input features via simple concatenation, this paradigm conflates semantic enrichment with mere feature expansion, leaving the cross-modal alignment process unguided. For instance, CLMLF^[24] appends caption-derived features to the textual stream via contrastive learning, and MSFN^[25] fuses cross-modal representations through semantic alignment networks. While effective, both approaches treat captions as supplementary input rather than as a structured semantic regulator, and neither enforces geometric constraints on the affective representation space. The key limitation is that appending captions increases feature dimensionality without providing structured constraints on how visual and textual signals should interact. Separately, margin-based objectives have demonstrated strong discriminative power in visual recognition. ArcFace^[26], for example, enforces additive angular margins for face identification—a setting characterized by discrete, balanced categories and unimodal input. Directly transplanting such objectives to MSA is non-trivial, as sentiment categories are inherently ambiguous, class-imbalanced, and defined over heterogeneous multimodal signals.

Together, these observations reveal a fundamental gap: while caption-augmented methods enrich feature representations and margin-based methods enforce geometric separability, no existing work integrates both within a unified affective reasoning framework. To our knowledge, no prior work has simultaneously addressed cross-modal semantic misalignment and affective representation overlap within a unified framework. The gap that motivates our work therefore lies not in any single component, but in the absence of a principled integration paradigm that coordinates semantic grounding, cross-modal alignment, and affective space disentanglement toward a common objective.

2.3 Affective space and loss optimization

Traditional sentiment classification typically relies on discrete labels, which often fail to enforce clear inter-class boundaries, leading to severe representation overlap and entanglement in noisy multimodal scenarios^[4]. Affective states can be modeled within a latent manifold where geometric distance reflects emotional separability. Advancing this perspective, prototypical learning has been utilized to align emotional representations across domains, demonstrating that learned class prototypes can effectively anchor heterogeneous features in a shared subspace^[27]. Recent advancements in affective computing have increasingly leveraged contrastive learning objectives and prototypical networks to establish fine-grained correspondences between local visual regions and emotionally salient linguistic tokens^[28,24].

While these strategies enhance the cohesion of multimodal representations, enforcing rigorous decision margins to disentangle conflicting affective states remains a formidable hurdle in latent space optimization. This difficulty arises because ambiguous and noisy inputs often cause representation overlap, making it challenging to maintain distinct class separability and structured geometric layouts through conventional alignment alone.

To address these challenges, our proposed Emo-Loss draws inspiration from margin-based metric learning to enforce inter-class separability. However, unlike the hard additive angular margins established for discrete visual recognition^[26], Emo-Loss models sentiment as a structured cosine manifold. By parameterizing the neutral prototype as the cosine bisector of the polar extremes, we establish a rigorous geometric inductive bias. This design does not aim to mix states transitionally, but rather to anchor the neutral concept at the structural center and maximize the cosine margin against polar sentiments. This ensures that even inherently ambiguous samples are projected into well-separated, unambiguous decision zones.

3 Proposed method

In this section, we detail the architecture of GSKA, a framework designed to alleviate semantic misalignment by standardizing raw multimodal perceptions into semantically grounded linguistic anchors. Unlike traditional models that rely on coarse-grained feature concatenation, GSKA formalizes MSA as a hierarchical alignment process within a learned affective latent manifold. As illustrated in Fig. 4, the framework operates through three synergistic stages: (1) Knowledge-Augmented Encoding, where raw modalities are distilled into regional visual features and generative semantic anchors; (2) Bi-directional Knowledge Alignment, which utilizes a 3-layer co-attention mechanism to model the tri-stream interaction between modalities; and (3) Affective Space Projection, which maps aligned features onto an emotional manifold for similarity-based inference using the proposed Emo-Loss.

3.1 Knowledge-augmented representation learning

To mitigate the inherent semantic divergence between unconstrained visual signals and noisy user-generated text, GSKA casts MSA as a knowledge-guided cross-modal alignment task. By introducing generative semantic anchors as a 'latent reference', we project heterogeneous inputs into a unified Affective Latent Manifold, ensuring that the cross-modal semantic gap is bridged via structured semantic references derived from visual signals. We use the term 'knowledge' to refer to these model-generated semantic priors, acknowledging that

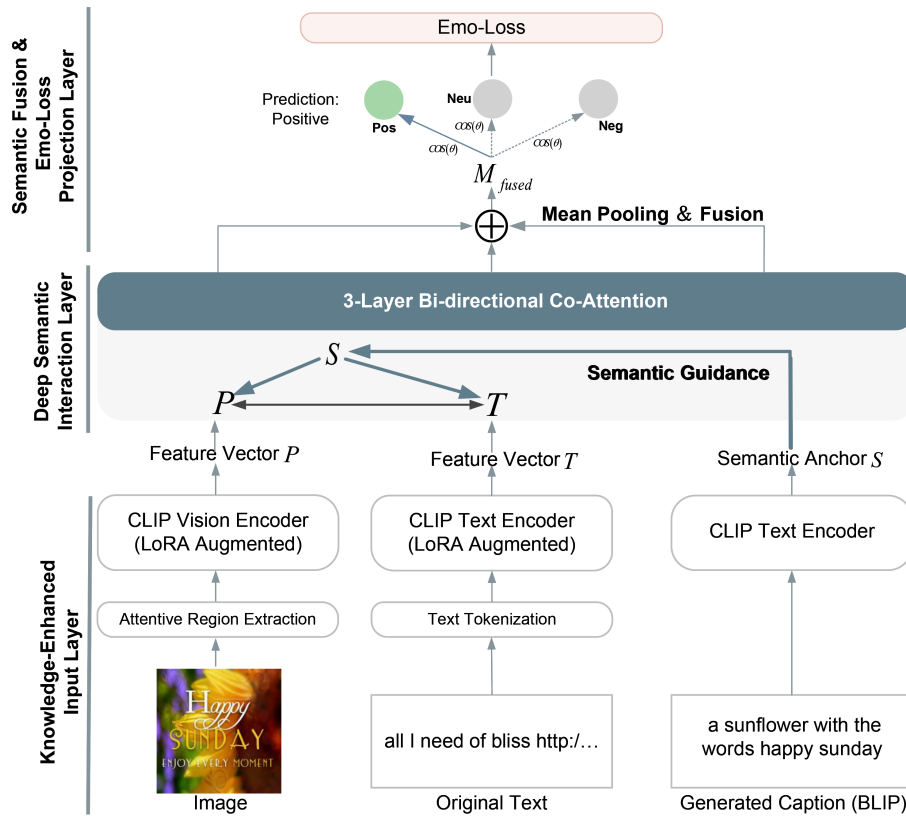


Fig. 4 The overall architecture of GSKA. The pipeline transitions from sensory encoding to knowledge-driven reasoning, culminating in an affective manifold mapping for sentiment prediction.

they represent learned visual descriptions rather than external factual ontologies.

3.1.1 Generative Semantic Anchoring via BLIP

Social media content is inherently characterized by high linguistic noise and sentiment ambiguity. To provide a stable reference, we leverage the BLIP generative model to distill the high-level semantic 'gist' of the scene. Formally, given an image I , the model produces a descriptive sentence S that translates subjective visual patterns into semi-standardized linguistic concepts. As illustrated in Fig. 5, these captions provide explicit affective cues (e.g., distilling 'roller coaster' for Amusement or 'blood' for Fear). These Generative Semantic Anchors (S) function as a semantic bottleneck, grounding the model in structured visual descriptions and filtering out affective-irrelevant noise before the interaction phase.

3.1.2 Representation learning with LoRA-augmented CLIP

To extract high-quality features while maintaining domain adaptability, we utilize the CLIP (ViT-L/14) backbone. We implement Low-Rank Adaptation (LoRA) to fine-tune the encoders efficiently, preventing the catastrophic forgetting of pre-trained multimodal knowledge. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the adapted forward pass is $h = W_0 x + B A x$, where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are learnable low-rank matrices. This results in three distinct feature sets: Visual Patches (P), Textual Tokens (T), and Semantic Anchors (S), all projected into a shared d -dimensional embedding space.

3.2 Hierarchical Bi-directional Knowledge Alignment

The core of GSKA is the 3-layer Bi-directional Co-Attention module. Moving beyond the limitations of coarse-grained

concatenation, GSKA models multimodal fusion as a dynamic reasoning process controlled by a semantic reference.

3.2.1 Bi-directional Knowledge Alignment mechanism

We propose the Bi-directional Knowledge Alignment (BKA) mechanism to facilitate deep interplay between modalities. As depicted in Fig. 6, visual patches P and textual tokens T serve as interactive sequences, while the Semantic Anchor S acts as an exogenous knowledge regularizer.

(1) Cross-modal feature projection

In each alignment layer, input features are projected into independent Query (Q), Key (K), and Value (V) spaces. For the visual stream:

$$Q_p = P W_Q^{(P)}, \quad K_p = P W_K^{(P)}, \quad V_p = P W_V^{(P)} \quad (1)$$

Symmetrically, Query, Key, and Value matrices are derived for the textual stream (Q_t, K_t, V_t) using $W^{(T)}$.

(2) Knowledge-guided cross-attention ($T \leftrightarrow P$)

To disambiguate unconstrained visual signals, we implement a dual-path interaction mechanism. As illustrated in Fig. 6, the Semantic Anchor S provides a 'Knowledge Alignment' signal to calibrate the cross-modal flow. Specifically, we define two symmetric paths: *Text-Guided Visual Alignment* ($T \rightarrow P$) and *Visual-Guided Textual Alignment* ($P \rightarrow T$). These interactions are regulated by a learnable semantic modulation layer $M(\cdot)$ to enforce semantic consistency. The cross-verified representations $\tilde{P}$ and $\tilde{T}$ are synthesized as follows:

$$\tilde{P} = \left(\text{softmax} \left(\frac{Q_p (K_t)^T}{\sqrt{d_k}} \right) \odot M(S) \right) V_t \quad (2)$$

$$\tilde{T} = \left(\text{softmax} \left(\frac{Q_t (K_p)^T}{\sqrt{d_k}} \right) \odot M(S) \right) V_p \quad (3)$$

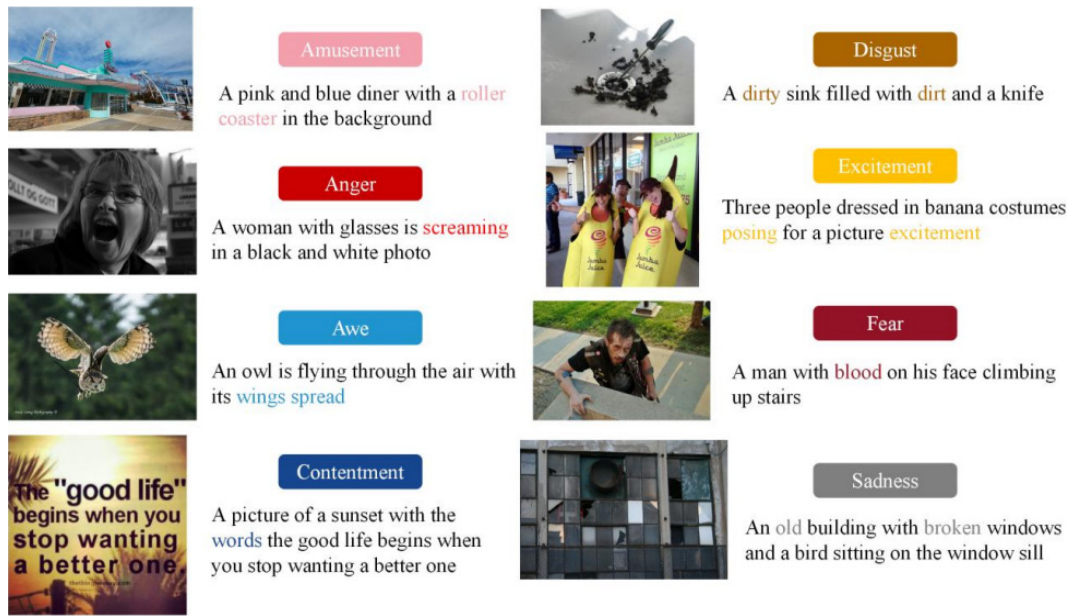


Fig. 5 Visualization of the generative semantic anchoring process. High-level affective indicators are extracted to serve as stable semantic coordinates (**S**).

where $\odot$ denotes the element-wise multiplication and $\sqrt{d_k}$ is the scaling factor. $\mathcal{M}(\cdot)$ represents a learnable transformation mechanism. To ensure full-rank expressive capacity for fine-grained filtering, $\mathcal{M}(\cdot)$ incorporates a two-layer Multilayer Perceptron (MLP) with a ReLU activation to project the Semantic Anchor **S** into a modulation mask of dimension $\mathbb{R}^{1 \times L}$:

$$\mathcal{M}(\mathbf{S}) = \sigma(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{S} + \mathbf{b}_1) + \mathbf{b}_2), \quad \mathcal{M}(\mathbf{S}) \in \mathbb{R}^{1 \times L} \quad (4)$$

where L denotes the sequence length of the attended stream ($L=M$ when modulating $\tilde{\mathbf{P}}$, and $L=N$ when modulating $\tilde{\mathbf{T}}$), with broadcasting applied along the query dimension. Crucially, this modulation layer also serves as an uncertainty management component to mitigate potential 'hallucinations' or semantic drift inherent in generative models. Instead of treating the generated captions as immutable ground-truth facts, GSKA leverages them to synthesize a soft attention mask. This mechanism ensures that if a generated anchor diverges significantly from the primary sensory signals of the visual or textual streams, its modulation weight is adaptively suppressed. Consequently, the anchor functions as a cross-modal semantic filter (represented by the grey 'Knowledge Alignment'

arrows in Fig. 6), effectively re-scaling attention scores to prioritize emotionally salient regions while maintaining high semantic fidelity even in the presence of imperfectly generated knowledge. This dual-path refinement ensures that the cross-modal interaction remains anchored in structured visual descriptions while being robust to unconstrained multimodal noise.

The architectural rationale behind encoding the generative caption via CLIP's text encoder—as opposed to direct visual-to-text alignment—is rooted in mitigating modality asymmetry. Raw visual embeddings (**P**) inherently capture dense, continuous, low-level sensory patterns (e.g., textures, illumination), which exhibit a severe distribution shift from the discrete, highly abstract linguistic features of noisy text (**T**). Direct cross-modal alignment without mediation frequently precipitates representation entanglement or modality collapse. By leveraging BLIP as a semantic bottleneck, the continuous visual manifold is discretized into a structured, human-interpretable linguistic summary. Consequently, the resulting semantic anchor **S** resides within the identical linguistic embedding space as **T**. This homogeneous representation enables **S** to function as an effective exogenous

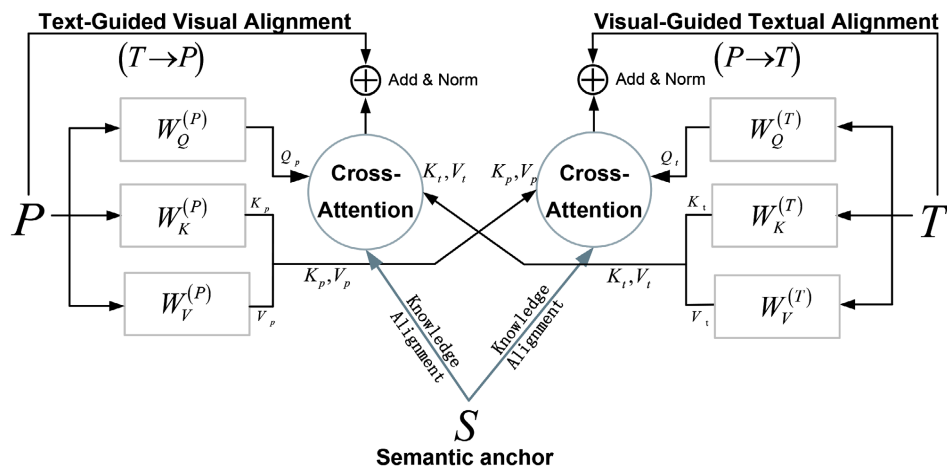


Fig. 6 Detailed mechanism of the Bi-directional Knowledge Alignment (BKA) module. The Semantic Anchor (**S**) modulates the cross-attention flow between vision (**P**) and text (**T**).

modulation signal, filtering cross-attention weights in a semantically coherent manner while explicitly bypassing the sensory noise of the raw pixels.

3.2.2 Residual connection and triple-stream fusion

Following the interaction, we apply a residual connection followed by Layer Normalization (Add & Norm). To consolidate the refined tokens into fixed-length vectors, global mean pooling is performed across each stream to obtain $\tilde{P}_{\text{pool}}$, $\tilde{T}_{\text{pool}}$, and S_{pool} . To dynamically prioritize reliable modalities and mitigate sensory noise, the final multimodal representation M_{fused} is synthesized via a learnable gated fusion mechanism:

$$M_{\text{fused}} = \omega_P \tilde{P}_{\text{pool}} + \omega_T \tilde{T}_{\text{pool}} + \omega_S S_{\text{pool}} \quad (5)$$

where the gating weights $[\omega_P, \omega_T, \omega_S] = \text{Softmax}(\mathbf{W}_f [\tilde{P}_{\text{pool}}; \tilde{T}_{\text{pool}}; S_{\text{pool}}])$ are computed from the concatenated pooled representations. Here, $\mathbf{W}_f$ is a learnable projection matrix. This gated fusion allows the model to assign higher weights to the model-generated semantic anchor (ω_S) and visual patches (ω_P) when the textual modality is dominated by linguistic noise, thereby effectively filtering out irrelevant information and preventing modality collapse. This gated fusion ensures that the final representation encapsulates both latent affective nuances and grounded semantic denotations. To elucidate the information flow within GSKA, Fig. 4 presents a representative case study. As shown in the Knowledge-Enhanced Input Layer, the user-generated text—'all I need of bliss...'—exemplifies the challenge of linguistic abstraction; it conveys high-level sentiment but lacks explicit visual grounding, creating a significant affective gap. Conversely, the generative semantic anchor—'a sunflower with the words happy sunday'—functions as a structured semantic reference, translating raw sensory signals into stable linguistic concepts. Within the Deep Semantic Interaction Layer, this anchor acts as a semantic pivot, enabling the model to synchronize the visual features of the sunflower with the sentiment-rich tokens in the original text. By leveraging the anchor as exogenous guidance, GSKA effectively filters out modality-specific noise and resolves semantic ambiguity. Finally, in the Semantic Fusion and Emo-Loss Projection Layer, the consolidated representation M_{fused} is mapped onto the affective latent manifold, where its cosine proximity to the 'Positive' prototype is maximized. This workflow underscores GSKA's capacity to rectify semantic misalignment by anchoring subjective, noisy expressions within a structured semantic reference system.

3.3 Affective Space Projection and Emo-Loss optimization

To address the inherent representation overlap and subjectivity of emotions, GSKA utilizes an Affective Space Projection mechanism. By structuring the affective manifold with rigorous geometric constraints, our model achieves superior discriminative power for disentangling ambiguous samples.

3.3.1 Affective Projection and Class Prototypes

We define Class Prototypes $\mathbf{C} = \{\mathbf{c}_{\text{pos}}, \mathbf{c}_{\text{neu}}, \mathbf{c}_{\text{neg}}\}$ in the latent space. Among these, $\mathbf{c}_{\text{pos}}$ and $\mathbf{c}_{\text{neg}}$ are learnable embeddings optimized via gradient descent, while $\mathbf{c}_{\text{neu}}$ is not a free parameter but is deterministically constrained as the normalized angular bisector of the two polar prototypes:

$$\mathbf{c}_{\text{neu}} = \frac{\mathbf{c}_{\text{pos}} + \mathbf{c}_{\text{neg}}}{\|\mathbf{c}_{\text{pos}} + \mathbf{c}_{\text{neg}}\|_2} \quad (6)$$

The placement of $\mathbf{c}_{\text{neu}}$ is grounded in two complementary considerations. From a cognitive perspective, neutral sentiment is conventionally understood as the affective midpoint between positive and nega-

tive poles—neither valence dominates, and the neutral category is defined precisely by its equidistance from both extremes^[29]. This motivates anchoring $\mathbf{c}_{\text{neu}}$ at the symmetric center of the polar prototypes as a natural geometric prior. From an optimization perspective, this deterministic placement serves two additional purposes: (i) it enforces equal cosine margins between neutral and each polar class by construction, preventing the neutral region from drifting toward either extreme under gradient pressure; and (ii) it reduces the parameter search space, avoiding degenerate configurations where $\mathbf{c}_{\text{neu}}$ collapses onto $\mathbf{c}_{\text{pos}}$ or $\mathbf{c}_{\text{neg}}$. We therefore treat the geometric parameterization as jointly justified by affective theory and structural convenience, rather than as a claim about the full phenomenological complexity of neutral affect.

Furthermore, we impose a unit norm constraint on all prototypes to ensure they reside on a unified unit hypersphere:

$$\|\mathbf{c}_j\|_2 = 1, \quad j \in \{\text{pos}, \text{neu}, \text{neg}\} \quad (7)$$

This geometric framework ensures a highly robust latent manifold for disentangling affective states, where sentiment polarity is determined purely by the Cosine Proximity (Cosine Similarity) between the fused representation M_{fused} and each prototype:

$$s_j = \cos(\theta_j) = \frac{M_{\text{fused}} \cdot \mathbf{c}_j^\top}{\|M_{\text{fused}}\| \|\mathbf{c}_j\|}, \quad j \in \{\text{pos}, \text{neu}, \text{neg}\} \quad (8)$$

3.3.2 Emo-Loss formulation

Unlike the standard Cross-Entropy loss, which lacks explicit constraints on decision margins, Emo-Loss optimizes the cosine separation between classes. Our proposed Emo-Loss draws inspiration from margin-based metric learning to enforce inter-class separability. However, building on the additive angular margin concept originally proposed for discrete visual recognition^[26], Emo-Loss introduces a Margin-aware Soft Contrastive Loss for structured affective disentanglement. It penalizes samples near the decision boundaries by maximizing the discrepancy Δ between the target similarity s_{target} and its most proximal confounding competitor $s_{\text{other}} = \max_{j \neq y}(s_j)$. This non-linear mapping via a scaled sigmoid function amplifies the gradient signal for ambiguous samples, effectively fostering a highly discriminative affective space.

The full computational process is detailed in Algorithm 1. The final Emo-Loss $\mathcal{L}_{\text{Emo}}$ is calculated by mapping this margin through a scaled sigmoid function to maximize normalized discrepancy:

$$\Delta = s_{\text{target}} - s_{\text{other}}, \quad \mathcal{L}_{\text{Emo}} = 1 - (2 \cdot \sigma(\beta \cdot \Delta) - 1) \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid activation and β is a fixed scaling hyperparameter. This contrastive-style supervision forces the model to cluster similar emotional states while enforcing a rigorous cosine margin against conflicting ones.

Algorithm 1. Emo-Loss computation.

Input: Fused multi-modal feature M_{fused} , target label y , Prototype set $\mathbf{C}$, scaling factor β

Output: Emotional discrepancy loss $\mathcal{L}_{\text{Emo}}$

1. Similarity Calculation: Compute cosine similarities for all classes: $s_j = \text{cosine_sim}(M_{\text{fused}}, \mathbf{c}_j)$ for $j \in \{\text{pos}, \text{neu}, \text{neg}\}$;
2. Target Identification: Retrieve target similarity: $s_{\text{target}} = s_y$;
3. Competitor Mining: Find maximum confounding similarity: $s_{\text{other}} = \max_{j \neq y}(s_j)$;
4. Margin Optimization: Calculate similarity margin: $\Delta = s_{\text{target}} - s_{\text{other}}$;
5. Non-linear Mapping: Normalize margin via scaled sigmoid: $\text{diff} = 2 \cdot \sigma(\beta \cdot \Delta) - 1$;
6. **return** $\mathcal{L}_{\text{Emo}} = 1 - \text{diff}$;

4 Experiments

In this section, we conduct a series of rigorous empirical evaluations to validate the efficacy of the proposed GSKA framework.

4.1 Experimental setup and data curation

We evaluate GSKA on two widely used benchmarks: MVSA-single and MVSA-multiple^[30]. To ensure a fair and reproducible comparison, we strictly adhere to the standardized data curation protocols^[8] as adopted by contemporary SOTA models such as MSFN^[25].

The curation process involves: (1) Consensus-based Quality Filtering: following the standard quality control protocol^[8], we retain only samples that achieved majority annotator agreement, ensuring reliable ground-truth labels for training. Samples are excluded solely on the basis of irreconcilable annotator disagreement—that is, cases where no consensus label can be established due to inconsistent human judgments, not because of cross-modal semantic complexity. For neutral-polar mixed pairs where a majority label exists, the majority label is adopted. (2) Text Preprocessing: textual artifacts (e.g., URLs, user mentions) are removed to improve the signal-to-noise ratio of the linguistic stream.

It is important to emphasize that this standard curation procedure does not remove semantically challenging samples. Cases exhibiting cross-modal incongruity—such as sarcastic text paired with a visually positive image, or subjective colloquial expressions accompanying ambiguous visual content—are fully preserved, as these constitute the core challenging scenarios that motivate GSKA. The resulting datasets therefore remain highly challenging, characterized by the affective gap and cross-modal misalignment that our framework is designed to address.

The final quantitative distribution of samples is summarized in Table 1.

4.2 Implementation details

GSKA is implemented using the PyTorch framework and trained on a single NVIDIA RTX 4090 GPU. We utilize CLIP (ViT-L/14) as the foundational backbone, augmented with LoRA adapters ($r = 8, \alpha = 16$) for parameter-efficient fine-tuning. The BLIP generative model is employed to produce the semantic anchors. We utilize the AdamW optimizer with a decoupled learning rate strategy: 1.0×10^{-5} for the pre-trained foundational backbone and 5.0×10^{-5} for the bi-directional alignment modules. To prevent overfitting, we set the dropout rate to 0.4 and the similarity scaling factor $\beta = 15$. Extensive empirical evaluations indicate that the model performance remains stable within the range of $\beta \in [12, 18]$, with $\beta = 15$ yielding the optimal balance between gradient propagation and class separation. All experiments were conducted over 20 independent runs to ensure statistical significance, with the mean accuracy and weighted F1-score reported. Regarding computational efficiency, although GSKA incorporates a generative component, the semantic anchors can be pre-computed offline for static datasets or processed in parallel in streaming scenarios, imposing negligible additional latency during inference compared to end-to-end LLM-based approaches. Furthermore, by leveraging Low-Rank Adaptation (LoRA), we only fine-tune less than 5% of the

Table 1. Statistics of the datasets.

Dataset	Positive	Neutral	Negative	Total
MVSA-single	2,683	470	1,358	4,511
MVSA-multiple	11,318	4,408	1,298	17,024

total parameters in the foundational encoders. This parameter-efficient design ensures that GSKA maintains high inference throughput on consumer-grade hardware like the NVIDIA RTX 4090, offering a scalable solution for real-time social media sentiment monitoring. In our latency tests, BLIP caption generation takes approximately 10 ms, which can run in parallel with CLIP visual encoding. The end-to-end inference latency of GSKA remains under 25 ms, ensuring practical applicability for near-real-time social media monitoring.

4.3 Results and analysis

To demonstrate the superiority of GSKA, we compare it against three categories of representative baselines: (1) Text-based models: CNN^[31] and BERT^[32];

(2) Image-based models: ResNet50^[33], ViT^[34], RepViT^[35], and TransNeXt^[36];

(3) Multimodal models: this includes earlier benchmarks like HSAN^[8], MultiSentiNet^[5], Co-MN-Hop6^[7] and MVAN-M^[37], as well as contemporary SOTA models including MGNNS^[38], CLMLF^[24], MVCN^[39], ICCI^[40], and MSFN^[25].

The quantitative results on MVSA-single and MVSA-multiple are presented in Table 2.

As observed in Table 2, GSKA consistently outperforms all baseline models across both datasets. Several key observations can be derived from the results. First, multimodal models significantly exceed the performance of single-modality baselines. For instance, on MVSA-single, our framework achieves an accuracy of 80.93% (± 0.2), which is a 9.82 percentage-point improvement over the strongest text-only baseline (BERT) and a 15.22 percentage-point improvement over the strongest image-only baseline (TransNeXt). This underscores the necessity of cross-modal interaction in resolving sentiment ambiguity.

Second, GSKA demonstrates superior performance compared to the most recent multimodal SOTA models. On the MVSA-single dataset, GSKA surpasses MSFN by 1.95%. We attribute this gain to the introduction of generative semantic anchors, which provide structured semantic references that stabilize the alignment process, unlike the coarse-grained concatenation or simple attention mechanisms used in earlier models.

Third, the results on MVSA-multiple further validate the robustness of our framework in handling annotator variability and linguistic noise. GSKA maintains a competitive accuracy of 75.05% (± 0.18),

Table 2. Performance comparison on MVSA datasets. Accuracy and F1-score are reported as a percentage (%).

Modality	Methods	MVSA-single		MVSA-multiple	
		Accuracy	F1	Accuracy	F1
Text	CNN	68.19	55.90	65.64	57.66
	BERT	71.11	69.70	67.59	66.24
Image	ResNet50	64.67	61.55	61.88	60.98
	ViT	63.78	62.26	61.94	61.19
	RepViT	64.82	62.81	62.01	61.35
	TransNeXt	65.71	64.81	62.48	61.52
Multimodal	HSAN	69.88	66.90	67.96	67.76
	MultiSentiNet	69.84	69.63	68.86	68.11
	Co-MN-Hop6	70.51	70.01	68.92	68.83
	MVAN-M	72.98	72.98	72.36	72.30
	MGNNS	73.77	72.70	72.49	69.34
	CLMLF	75.33	73.46	72.00	69.83
	MVCN	76.06	74.55	72.07	70.01
	ICCI	79.33	77.51	73.29	70.06
	MSFN	78.98	78.48	74.75	72.62
	GSKA (ours)	80.93	80.10	75.05	74.57

outperforming established models like CLMLF and MSFN. While the accuracy improvement is 0.3%, GSKA achieves a more substantial 1.95% gain in the weighted F1-score (74.57% vs 72.62%), demonstrating its robustness in handling class imbalance and minority categories. This indicates that the bi-directional knowledge alignment module effectively filters out irrelevant sensory noise, allowing the model to focus on emotionally salient semantic regions even in complex social media environments.

4.4 Ablation study

To systematically evaluate the individual contribution of each component within GSKA, we conduct a series of ablation experiments on the MVSA-single dataset. By removing or replacing specific modules, we demonstrate the necessity of our architectural choices. The results are summarized in Table 3.

The experimental results provide several key insights into the architectural and optimization choices of GSKA:

First, the optimization strategy via Emo-Loss is the most significant factor in achieving superior discriminative power. As shown in Table 3, the variant utilizing standard Contrastive Loss (w/o Emo-Loss w/ Contrastive Loss) yields the most severe degradation, with accuracy dropping by 4.16% and F1-score by 4.05% relative to the full model.

This degradation occurs because standard Contrastive Loss indiscriminately repels non-target classes, which can disrupt the global topology of the affective manifold and cause representation fragmentation. While replacing Emo-Loss with standard Cross-Entropy (w/o Emo-Loss w/ Cross-Entropy) reduces the accuracy drop to 2.61% and the F1 drop to 2.31%, it still lacks explicit margin constraints, leading to residual representation overlap. The consistent superiority of Emo-Loss across both metrics validates that enforcing margin-aware cosine constraints is essential for resolving overlapping decision boundaries and disentangling affective representations.

To further investigate the superiority of Emo-Loss in shaping the representation space, we provide a qualitative comparison via t-SNE

Table 3. Ablation results on the MVSA-single dataset.

Variants	Accuracy (%)	ΔAcc	F1 (%)	ΔF1
GSKA (full model)	80.93	—	80.10	—
w/o Co-Attention	77.42	↓3.51	76.88	↓3.22
w/o Captioning	78.08	↓2.85	78.27	↓1.83
w/o text mediation (Direct Alignment)	77.92	↓3.01	77.45	↓2.65
w/o Emo-Loss (w/ Contrastive Loss)	76.77	↓4.16	76.05	↓4.05
w/o Emo-Loss (w/ Cross-Entropy)	78.32	↓2.61	77.79	↓2.31

↓ denotes the performance drop relative to the full model.

visualization in Fig. 7. As illustrated, (a) the Standard Cross-Entropy variant suffers from significant inter-class overlaps, particularly where neutral samples are widely scattered among positive and negative clusters. (b) The Contrastive Loss variant improves the global layout but fails to achieve intra-class compactness. In contrast, (c) the full GSKA model (with Emo-Loss) produces notably more compact and better-separated clusters. In real-world scenarios, emotional inputs are inherently ambiguous; enforcing a strict margin in the latent space acts as a robust denoising mechanism, compelling the model to make decisive discriminations rather than leaving samples in entangled grey areas. The clear separation of clusters suggests that Emo-Loss successfully prevents the model from collapsing ambiguous signals into overlapping bins, thereby resolving the decision uncertainty.

This visual evidence directly corroborates the quantitative results in Table 3. The clear cluster separation produced by Emo-Loss stems from its joint action on two coupled components: the cosine margin objective explicitly widens inter-class boundaries, while the learnable class prototypes $\mathbf{c}_{pos}$ and $\mathbf{c}_{neg}$ provide stable geometric anchors that the margin is enforced against. Without this prototype-margin coupling—as seen in the Cross-Entropy and Contrastive Loss variants—the decision boundaries remain underconstrained, causing ambiguous samples to scatter across class regions. The resulting improvement in cluster compactness directly translates to the gains in weighted F1-score observed in Table 3, particularly for the minority neutral class where representation overlap is most severe.

Second, the bi-directional co-attention mechanism serves as the architectural foundation for effective cross-modal interaction. In the variant (w/o Co-Attention), the semantic anchor $\mathbf{S}$ is retained but incorporated via simple concatenation rather than through the knowledge-guided cross-attention mechanism, ensuring that the observed performance drop is attributable solely to the removal of the structured alignment process and not to the absence of anchor information. Removing the co-attention module under this controlled setting leads to the second-largest performance degradation, with accuracy declining by 3.51% and F1-score by 3.22% relative to the full model. This confirms that coarse-grained feature concatenation fails to resolve the semantic misalignment often present in noisy social media data. Specifically, the bi-directional alignment in GSKA facilitates a mutual refinement process where visual and textual streams are cross-verified against the generative semantic anchors. This structural design ensures that the model prioritizes emotionally salient regions by maintaining semantic consistency between modalities. Such an adaptive interaction mechanism is more effective at filtering out affective-irrelevant sensory noise than simplistic fusion strategies, providing a robust foundation for subsequent sentiment inference.

Third, the generative semantic anchors are vital for alleviating semantic misalignment. The ablated version without captions (w/o

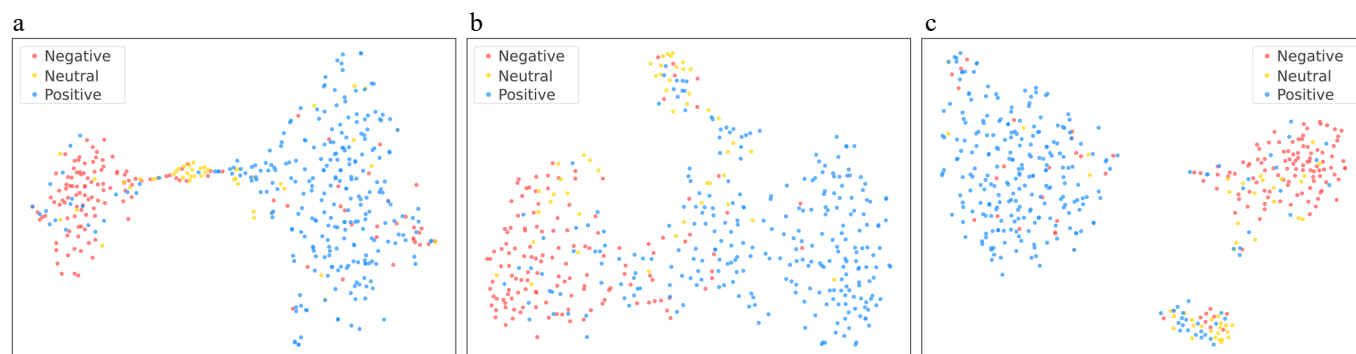


Fig. 7 t-SNE visualization of multimodal embeddings for different variants: (a) Standard Cross-Entropy; (b) Contrastive Loss; and (c) Emo-Loss (full GSKA).

Captioning) shows a 2.85% decrease in accuracy and a 1.83% decrease in F1-score compared to the full model. This justifies our design of using BLIP-generated descriptions as structured semantic references. These anchors provide a stable reference that grounds both raw visual patches and noisy textual tokens, ensuring that the model remains grounded in model-derived scene descriptions before sentiment inference. Notably, the relatively modest drop in F1-score (1.83%) compared to the accuracy drop (2.85%) may suggest that the semantic anchors contribute disproportionately to overall accuracy, particularly on majority classes, while the BKA module's modulation mechanism provides some compensatory robustness even in the absence of captions. From a knowledge engineering perspective, the performance gain of GSKA demonstrates that distilling visual perceptions into explicit linguistic concepts can significantly reduce the entropy of multimodal fusion compared to purely data-driven feature concatenation.

Finally, to verify the necessity of generative semantic anchors, we examine the variant w/o Text Mediation (Direct BLIP + CLIP alignment). In this configuration, the descriptive caption generation process is bypassed, and raw multimodal features are directly integrated into the co-attention modulation layer, while all other architectural components remain unchanged. The omission of this textual intermediary leads to a noticeable performance degradation, dropping to 77.92% in accuracy and 77.45% in F1-score. Intriguingly, this configuration yields even lower metrics than the w/o Captioning counterpart. This phenomenon indicates that introducing unconstrained multimodal

features without explicit linguistic scaffolding inevitably injects substantial alignment noise into the joint latent space. Consequently, these empirical comparisons firmly demonstrate that our proposed generative semantic anchors offer far greater efficacy than direct, unconstrained feature-level alignment.

In summary, the transition from conventional losses and heuristic fusion methods to the integrated GSKA framework enables robust affective reasoning, effectively addressing the inherent subjectivity and noise in MSA.

4.5 Visualization and interpretability of GSKA

To provide qualitative insights into how GSKA establishes fine-grained semantic-spatial grounding, we visualize the learned attention maps in Fig. 8. The results reveal GSKA's remarkable precision in localizing affective hotspots that correspond to specific linguistic cues, highlighting the sophisticated collaborative alignment between user-generated text and generative semantic anchors.

A critical observation is the complementary nature of the dual text streams: while the original text often carries high-level affective connotations (e.g., 'excited', 'mad'), the generative anchors provide explicit semantic denotations (e.g., 'hugging', 'creepy look'). Specifically, in the bottom-left case of Fig. 8, the anchor 'hugging' serves as a crucial semantic bridge, grounding the abstract sentiment 'excited' into concrete visual interactions. Similarly, in the bottom-right sample, the model successfully aligns subjective descriptors such as 'mad' and 'evil' with the subject's intense facial expression by referencing the

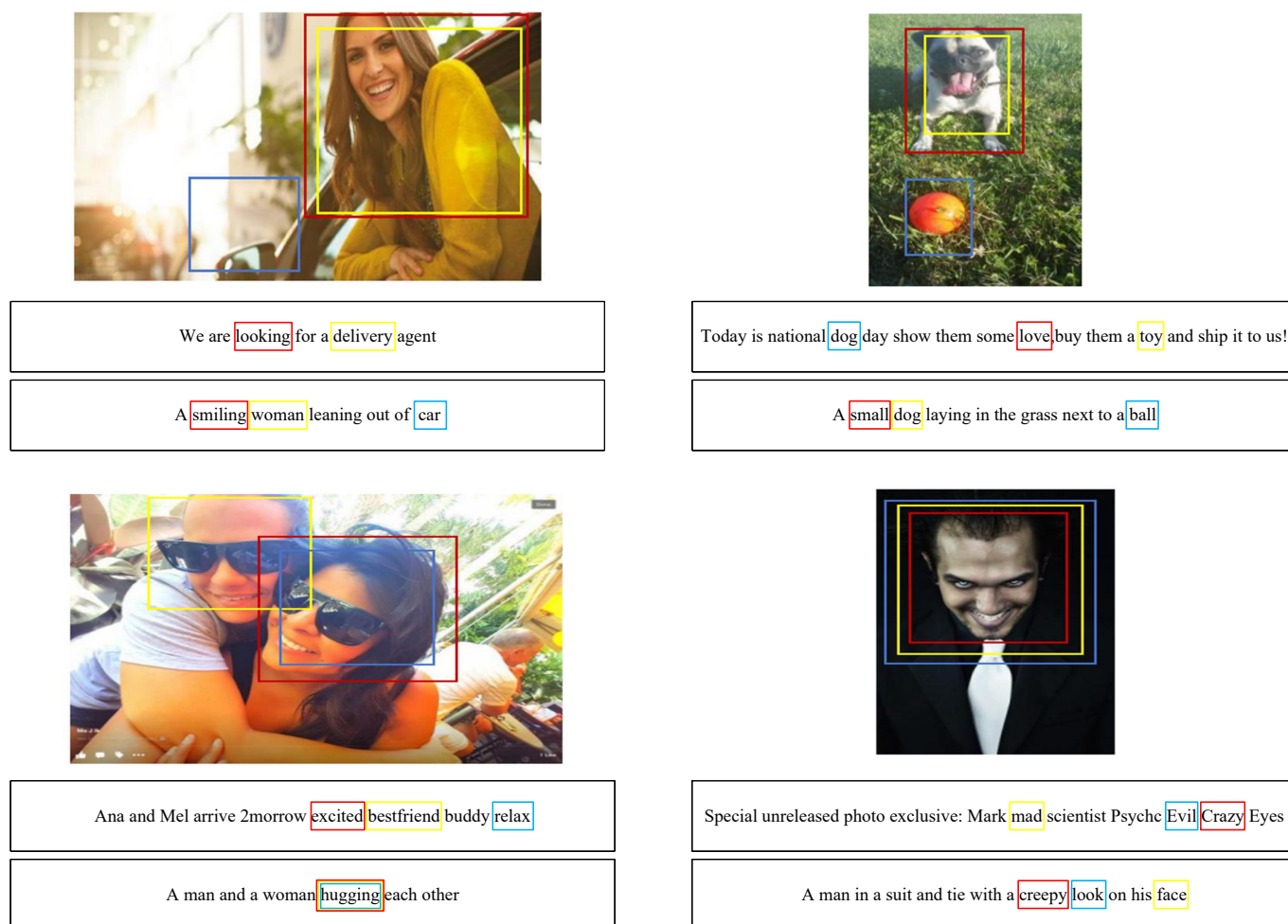


Fig. 8 Visualization of the attention mechanism, highlighting aligned image regions and textual tokens.

anchor 'creepy look'. This synergy confirms that generative anchors function as a semantic refiner, distilling stable semantic coordinates from noisy and unconstrained social media inputs.

From a knowledge acquisition perspective, Fig. 8 reveals that these anchors consistently capture 'action-object' dyads (e.g., 'smiling-woman') that are often omitted in fragmented user-generated text. These dyads provide the structural common-sense semantic context required to resolve visual ambiguity, re-centering the model's attention on shared emotional actions rather than ambiguous background entities. Furthermore, the hierarchical attention intensity—indicated by red, yellow, and blue boxes—confirms GSKA's capacity to effectively suppress background clutter and prioritize sentiment-critical regions. Collectively, this evidence suggests that GSKA successfully transforms unconstrained visual perceptions into structured, task-relevant representations, providing a transparent reasoning path that addresses the 'black-box' nature of multimodal fusion.

5 Conclusions and future work

In this paper, we introduce Generative Semantic Anchoring and Bi-directional Knowledge Alignment (GSKA), a novel framework designed to mitigate cross-modal semantic misalignment and facilitate affective reasoning in MSA. By leveraging generative models to produce semantic anchors, we transform the visual perception task into a knowledge-guided cross-modal alignment task. To capture the complex interplay between modalities, we design a bi-directional alignment mechanism that prioritizes emotionally salient regions while suppressing irrelevant noise. Furthermore, we propose Emo-Loss, a margin-aware objective function that enforces structured separability and disentangles conflicting affective states within a structured affective manifold. Overall, the experimental findings on the MVSA datasets confirm GSKA's clear advantage over current methods in resolving representation overlap and enhancing discriminative capability.

Despite the demonstrated performance gains, we acknowledge certain limitations in the current framework. The reliance on generative anchors introduces additional computational overhead during the pre-processing stage, which may impact deployment in extreme real-time scenarios. Moreover, while GSKA effectively aligns explicit semantic cues, resolving implicit semantic incongruity and abstract metaphors that persist in highly nuanced contexts remains a formidable challenge.

Future research will focus on two primary directions to address these hurdles. First, we aim to enhance the emotional granularity of the generative process by integrating large-scale affective ontologies and leveraging the reasoning capabilities of large language models to better interpret sarcasm and conflicting signals. Second, we will explore parameter-efficient distillation techniques to develop more lightweight interaction modules. This will ensure the scalability of knowledge-grounded affective reasoning for real-time social media monitoring and edge-computing applications, optimizing the trade-off between semantic richness and inference efficiency.

Ethical statements

During the preparation of this work, the authors used ChatGPT (Version: GPT-4o) for language refinement and grammatical polishing. The authors reviewed and edited all content produced with the assistance of this tool, verified its accuracy, and take full responsibility for the integrity and originality of the final manuscript. This work represents the authors' own intellectual contribution, and no AI tool is credited as an author.

Author contributions

The authors confirm their contribution to the paper as follows: study conception and design, draft manuscript preparation: Dong C, Qian J; methodology and algorithm development: Qian J; data collection: Qian J, Wang Q; analysis and interpretation of results: Qian J, Wang Q, Dong C. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. Code and models are available at: <https://github.com/QJ0413/Multimodal-Sentiment>.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos 62572446, 62176240), the Beijing Education Science 14th Five Year Plan Project (Grant No. CDDB25231), the Fundamental Research Funds for the Central Universities (Grant No. CUC26XZ02), and the China Scholarship Council (Grant No. 202507050074).

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Dates

Received 4 March 2026; Revised 7 July 2026; Accepted 26 July 2026; Published online 10 September 2026

References

- [1] Yadav A, Vishwakarma DK. 2020. A deep learning architecture of RA-DLNet for visual sentiment analysis. *Multimedia Systems* 26:431–451
- [2] Das R, Singh TD. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys* 55:1–38
- [3] Wang L, Niu J, Yu S. 2020. SentiDiff: Combining textual information and sentiment diffusion patterns for Twitter sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 44:2026–2039
- [4] Zhao S, Yao X, Yang J, Jia G, Ding G, et al. 2022. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44:6729–6751
- [5] Xu N, Mao W. 2017. MultiSentiNet: a deep semantic network for multimodal sentiment analysis. *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, November 6–10, 2017, Singapore, USA*: ACM. pp. 2399–2402 doi: [10.1145/3132847.3133142](https://doi.org/10.1145/3132847.3133142)
- [6] Wang M, Meng M, Liu J, Wu J. 2023. Adequate alignment and interaction for cross-modal retrieval. *Virtual Reality & Intelligent Hardware* 5:509–522
- [7] Xu N, Mao W, Chen G. 2018. A co-memory network for multimodal sentiment analysis. *SIGIR '18: The 41st international ACM SIGIR conference on research & development in information retrieval, July 8–12, 2018, Ann Arbor, MI, USA*. pp. 929–932 doi: [10.1145/3209978.3210093](https://doi.org/10.1145/3209978.3210093)
- [8] Xu N. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. *2017 IEEE international conference on intelligence and security informatics (ISI), July 22–24, 2017, Beijing, China, USA*: IEEE. pp. 152–154 doi: [10.1109/ISI.2017.8004895](https://doi.org/10.1109/ISI.2017.8004895)
- [9] Chen F, Ji R, Su J, Cao D, Gao Y. 2018. Predicting microblog sentiments via weakly supervised multimodal deep learning. *IEEE Transactions on Multimedia* 20:997–1007

- [10] Dai S, Man H. 2018. Integrating visual and textual affective descriptors for sentiment analysis of social media posts. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), April 10–12, 2018, Miami, FL, USA*. USA: IEEE. pp. 13–18 doi: [10.1109/MIPR.2018.00011](https://doi.org/10.1109/MIPR.2018.00011)
- [11] Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, et al. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management* 56:102097
- [12] Zhu T, Li L, Yang J, Zhao S, Liu H, et al. 2023. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia* 25:3375–3385
- [13] Xie Z, Zhang W, Sheng B, Li P, Chen CLP. 2021. BaGFN: broad attentive graph fusion network for high-order feature interactions. *IEEE Transactions on Neural Networks and Learning Systems* 34:4499–4513
- [14] Houshy N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, et al. 2019. Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning (ICML), June 9–15, 2019, Long Beach, CA, USA*. USA: PMLR. pp. 2790–2799 <https://proceedings.mlr.press/v97/houshy19a.html>
- [15] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, et al. 2022. LoRA: low-rank adaptation of large language models. *The Tenth International Conference on Learning Representations (ICLR), April 25–29, 2022, Virtual Event*. OpenReview.net. <https://openreview.net/forum?id=nZvKeeFYf9>
- [16] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML), July 18–24, 2021, Virtual Event*. USA: PMLR. pp. 8748–8763 <https://proceedings.mlr.press/v139/radford21a.html>
- [17] Chen Z, Xu L, Zheng H, Chen L, Tolba A, et al. 2024. Evolution and prospects of foundation models: from large language models to large multimodal models. *Computers, Materials & Continua* 80(2):1753–1808
- [18] Zhao X, Poria S, Li X, Chen Y, Tang B. 2025. Toward robust multimodal sentiment analysis using multimodal foundational models. *Expert Systems with Applications* 276:126974
- [19] Mu J, Wang W, Liu W, Yan T, Wang G. 2025. Multimodal large language model with lora fine-tuning for multimodal sentiment analysis. *ACM Transactions on Intelligent Systems and Technology* 16:1–23
- [20] Vinyals O, Toshev A, Bengio S, Erhan D. 2015. Show and tell: a neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7–12, 2015, Boston, MA, USA*. USA: IEEE. pp. 3156–3164 doi: [10.1109/CVPR.2015.7298935](https://doi.org/10.1109/CVPR.2015.7298935)
- [21] Li J, Li D, Savarese S, Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML'23: Proceedings of the 40th International Conference on Machine Learning (ICML), July 23–29, 2023, Honolulu, HI, USA*. USA: PMLR. pp. 19730–19742 doi: [10.5555/3618408.3619222](https://doi.org/10.5555/3618408.3619222)
- [22] Hu X, Gan Z, Wang J, Yang Z, Liu Z, et al. 2022. Scaling up vision-language pretraining for image captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18–24 June 2022, New Orleans, LA, USA*. USA: IEEE. pp. 17980–17989 doi: [10.1109/CVPR52688.2022.01745](https://doi.org/10.1109/CVPR52688.2022.01745)
- [23] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. 2023. GPT-4 technical report. *arXiv Preprint:2303.08774*
- [24] Li Z, Xu B, Zhu C, Zhao T. 2022. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. *Findings of the association for computational linguistics: NAACL 2022, Seattle, United States*. USA: ACM. pp. 2282–2294 doi: [10.18653/v1/2022.findings-naacl.175](https://doi.org/10.18653/v1/2022.findings-naacl.175)
- [25] Zhang S, Liu J, Jiao Y, Zhang Y, Chen L, et al. 2025. A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 21:1–22
- [26] Deng J, Guo J, Yang J, Xue N, Kotsia I, Zafeiriou S. 2022. ArcFace: additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44:5962–5979
- [27] Wang Q, Wen Z, Ding K, Liang B, Xu R. 2025. Cross-domain sentiment analysis via disentangled representation and prototypical learning. *IEEE Transactions on Affective Computing* 16:264–276
- [28] Zhuang Y, Bai W, Zhang Y, Deng J, Hu Z, et al. 2025. Multi-level contrastive learning for multimodal sentiment analysis. *IEEE Transactions on Multimedia*. 27:9044–9058
- [29] Russell JA. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39:1161–1178
- [30] Niu T, Zhu S, Pang L, El Saddik A. 2016. Sentiment analysis on multi-view social data. *MultiMedia Modeling. MMM 2016. Lecture Notes in Computer Science*. Cham: Springer. pp. 15–27 doi: [10.1007/978-3-319-27674-8_2](https://doi.org/10.1007/978-3-319-27674-8_2)
- [31] Kim Y. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, Doha, Qatar*. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 1746–1751 doi: [10.3115/v1/d14-1181](https://doi.org/10.3115/v1/d14-1181)
- [32] Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *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), 2019, Minneapolis, Minnesota*. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 4171–4186 doi: [10.18653/v1/n19-1423](https://doi.org/10.18653/v1/n19-1423)
- [33] He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27–30 June 2016, Las Vegas, NV, USA*. USA: IEEE. pp. 770–778 doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90)
- [34] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. 2021. An image is worth 16×16 words: transformers for image recognition at scale. *The Ninth International Conference on Learning Representations (ICLR), May 3–7, 2021, Virtual Event*. OpenReview.net. <https://openreview.net/forum?id=YicbFdNTTy>
- [35] Wang A, Chen H, Lin Z, Han J, Ding G. 2024. Rep Vit: revisiting mobile cnn from vit perspective. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16–22 June 2024, Seattle, WA, USA*. USA: IEEE. pp. 15909–15920 doi: [10.1109/CVPR52733.2024.01506](https://doi.org/10.1109/CVPR52733.2024.01506)
- [36] Shi D. 2024. TransNeXt: robust foveal visual perception for vision transformers. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16–22 June 2024, Seattle, WA, USA*. USA: IEEE. pp. 17773–17783 doi: [10.1109/CVPR52733.2024.01683](https://doi.org/10.1109/CVPR52733.2024.01683)
- [37] Yang X, Feng S, Wang D, Zhang Y. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia* 23:4014–4026
- [38] Yang X, Feng S, Zhang Y, Wang D. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. *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), 2021, Online*. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 328–339 doi: [10.18653/v1/2021.acl-long.28](https://doi.org/10.18653/v1/2021.acl-long.28)
- [39] Wei Y, Yuan S, Yang R, Shen L, Li Z, et al. 2023. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers), 2023, Toronto, Canada*. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 5240–5252 doi: [10.18653/v1/2023.acl-long.287](https://doi.org/10.18653/v1/2023.acl-long.287)
- [40] An J, Wan Zainon WMN. 2023. Integrating color cues to improve multimodal sentiment analysis in social media. *Engineering Applications of Artificial Intelligence* 126:106874



Copyright: © 2026 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit <https://creativecommons.org/licenses/by/4.0/>.