Structures Meet Semantics: Multimodal Fusion via Graph Contrastive Learning

Jiangfeng Sun[1], **Sihao He**[1], **Zhonghong Ou**[1*], **Meina Song**[1,2]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]China University of Petroleum, Beijing at Karamay, China
sun2017@bupt.edu.cn, sihaohe@bupt.edu.cn, zhonghong.ou@bupt.edu.cn, mnsong@bupt.edu.cn

Abstract

Multimodal sentiment analysis (MSA) aims to infer emotional states by effectively integrating textual, acoustic, and visual modalities. Despite notable progress, existing multimodal fusion methods often neglect modality-specific structural dependencies and semantic misalignment, limiting their quality, interpretability, and robustness. To address these challenges, we propose a novel framework called the Structural-Semantic Unifier (SSU), which systematically integrates modality-specific structural information and cross-modal semantic grounding for enhanced multimodal representations. Specifically, SSU dynamically constructs modality-specific graphs by leveraging linguistic syntax for text and a lightweight, text-guided attention mechanism for acoustic and visual modalities, thus capturing detailed intra-modal relationships and semantic interactions. We further introduce a semantic anchor, derived from global textual semantics, that serves as a cross-modal alignment hub, effectively harmonizing heterogeneous semantic spaces across modalities. Additionally, we develop a multi-view contrastive learning objective that promotes discriminability, semantic consistency, and structural coherence across intra- and inter-modal views. Extensive evaluations on two widely-used benchmark datasets, CMU-MOSI and CMU-MOSEI, demonstrate that SSU consistently achieves state-of-the-art performance while significantly reducing computational overhead compared to prior methods. Comprehensive qualitative analyses further validate SSU's interpretability and its ability to capture nuanced emotional patterns through semantically-grounded interactions.

Code — https://github.com/sun2017bupt/SSU

Introduction

Multimodal sentiment analysis (MSA) aims to automatically infer human emotional states by jointly analyzing complementary signals from textual, acoustic, and visual modalities. With the rapid proliferation of multimedia content and the growing need for emotionally intelligent systems in areas such as affective computing, human-computer interaction, and social media analytics, effective multimodal fusion strategies have become increasingly critical.

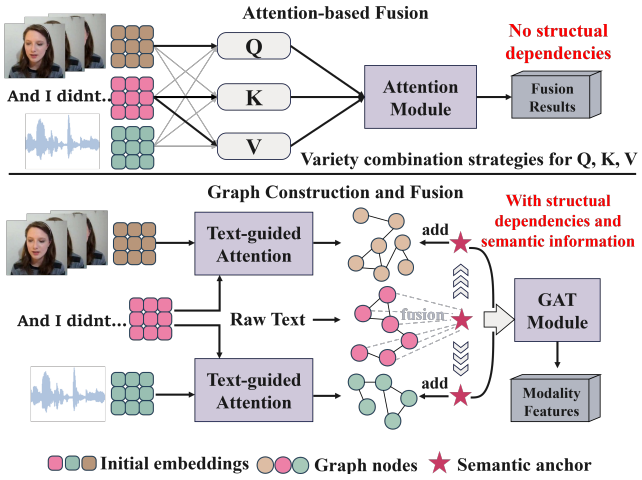


Figure 1: Comparison between attention-based fusion and our proposed SSU-based graph fusion framework.

Attention-based methods have recently emerged as predominant solutions for multimodal fusion, primarily due to their flexibility in modeling dynamic interactions across modalities. However, most existing fusion techniques treat each modality as a simple sequence of features, applying feature-level interactions through standard query-key-value mechanisms. Such approaches frequently disregard inherent modality-specific structural dependencies, including linguistic syntax in textual data and temporal coherence in audio and visual streams. Additionally, current methods implicitly assume that semantic content across modalities is naturally aligned—a simplification that often breaks down under conditions of ambiguous or nuanced sentiment expression.

These fundamental limitations hinder both the representational robustness and interpretability of multimodal sentiment models. Without explicit modeling of structural dependencies or targeted mechanisms to address semantic misalignment across modalities, learned representations tend to be ambiguous, less grounded, and less interpretable. Fig. 1 illustrates these limitations by comparing traditional attention-based fusion with our proposed approach. As highlighted, conventional methods inadequately capture critical intra-modal structural cues and cross-modal semantic coherence, whereas

*Corresponding author.

our proposed framework explicitly addresses these gaps by simultaneously modeling modality-specific structural information and cross-modal semantic alignment within a unified, interpretable representation learning framework.

We propose the **Structural-Semantic Unifier (SSU)**, a novel graph-based framework explicitly designed to unify modality-specific structural dependencies and cross-modal semantic alignment for multimodal sentiment analysis. SSU is built on the fundamental insight that effective multimodal fusion must simultaneously preserve the internal structures inherent in each modality and resolve semantic disparities across modalities—particularly crucial for capturing nuanced affective signals that may be asynchronous or ambiguous.

SSU constructs modality-specific graphs to preserve fine-grained structural dependencies—using syntactic parsing for text and text-guided attention for audio and visual modalities—while introducing semantic anchor as shared reference nodes grounded in linguistic context. These anchors serve to unify heterogeneous modality graphs, enabling coherent and interpretable cross-modal alignment. To further enhance representation quality, SSU incorporates a multi-view contrastive learning objective that jointly enforces task-specific discrimination, structural consistency, and semantic coherence. Together, these components yield robust, discriminative, and semantically aligned multimodal representations.

The main contributions are as follows:

- SSU is a unified multimodal fusion framework that integrates modality-specific structural dependencies with cross-modal semantic alignment.
- A modality-specific graph construction method is proposed, leveraging syntactic parsing for text and text-guided attention for audio and visual streams, and strengthened by a semantic anchor to ensure coherent cross-modal alignment.
- A multi-view contrastive learning objective is formulated that jointly enforces structural consistency, semantic coherence, and task-level discriminability across heterogeneous representations.
- Extensive evaluations on CMU-MOSI and CMU-MOSEI demonstrate state-of-the-art performance and improved computational efficiency.

Related Work

Multimodal Sentiment Analysis (MSA) aims to understand human affective states by jointly analyzing textual, visual, and acoustic signals. Early works typically employ unimodal encoders coupled with straightforward multimodal fusion strategies, such as concatenation (Zadeh et al. 2017; Wu et al. 2023) and attention-based integration (Tsai et al. 2019; Hazarika, Zimmermann, and Poria 2020; Yu et al. 2021). Although these methods achieve reasonable results, they often treat each modality independently, neglecting explicit structural relationships among modalities. Consequently, these approaches may fail to adequately capture nuanced multimodal interactions, especially in complex linguistic scenarios involving sarcasm, irony, or negation.

To address these limitations, recent studies have explored advanced encoders and refined learning objectives. UniMSE (Hu et al. 2022) leverages a pre-trained T5 model for textual representations, while audio and visual modalities are encoded separately through LSTMs, combined with cross-modal and inter-sample contrastive learning to enhance feature interaction. SeMuL-PCD (Anand et al. 2023) integrates modality-specific knowledge via cross-modal peer distillation and contrastive objectives, facilitating richer multimodal representations. MMIM (Han, Chen, and Poria 2021) preserves discriminative multimodal information through mutual information maximization between unimodal and joint representations. Other related efforts include handling modality incompleteness (Peng, Hong, and Zhao 2021; Lin and Hu 2023; Li et al. 2024), dynamic attention modulation strategies (Su et al. 2020; Chen, Huang, and Wang 2022), recurrent modeling of conversational sentiment dynamics (Huddar, Sannakki, and Rajpurohit 2021), and language-guided multimodal interactions (Mai, Xing, and Hu 2021).

In parallel, graph-based methods have demonstrated potential in explicitly modeling structural dependencies among modalities. MMGCN (Hu et al. 2021) introduces multimodal graph convolution networks to jointly capture intra- and inter-modal dependencies, while MGNNS (Yang et al. 2021) constructs sentiment-aware graphs per modality to extract global sentiment characteristics. AMGIN (Gong et al. 2024) employs adaptive gating mechanisms to dynamically integrate modality-specific graphs, promoting robustness and modality interaction. GraphMFT (Li et al. 2023) formulates conversation-based emotion recognition as a heterogeneous graph modeling problem, effectively capturing fine-grained contextual interactions within and across modalities. A multimodal graph contrastive framework (Liang et al. 2024) constructs a single cross-modal text–image graph via external detectors and optimizes supervised contrastive objectives.

Despite strong results, most graph-based methods use static or modality-isolated graphs and lack unified semantic alignment. We instead build dynamic, modality-specific graphs—syntax-induced for text and text-guided for audio/visual—and introduce a global semantic anchor to topologically align modalities. A multi-view contrastive objective further enforces structural coherence and cross-modal consistency, yielding a unified structure–semantics fusion for text–audio–video MSA.

Methodology

We propose the **Structural-Semantic Unifier (SSU)**, a unified framework that integrates intra-modal structure and cross-modal semantics for multimodal sentiment analysis. As illustrated in Fig. 2, SSU comprises three components: (1) modality-specific graph construction to capture structural dependencies; (2) semantic anchor to align modality graphs via language-guided reference nodes; and (3) a multi-view contrastive objective that enforces structural consistency, semantic coherence, and task-level discrimination.

Modality-Specific Graph Construction

Our approach constructs modality-specific graphs to explicitly model inherent structural dependencies and semantic alignments. This process is a crucial departure from conventional fusion techniques, laying the foundation for a richer,

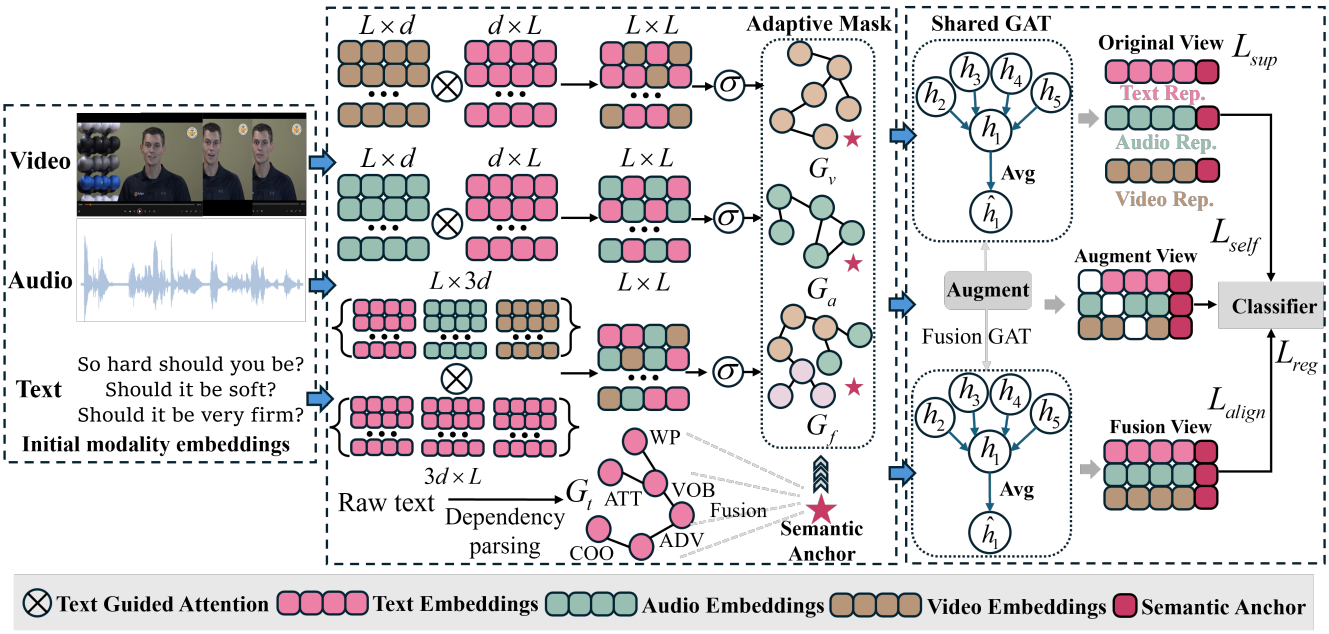


Figure 2: Overview of the SSU framework. Modality-specific graphs are constructed via syntactic parsing for text and text-guided attention for audio and video, capturing intra-modal structural dependencies. Semantic anchor derived from global text semantics is injected into all graphs to unify modalities. A shared GAT encodes structure-aware representations, while a fusion GAT integrates anchor-enhanced graphs. Multi-view contrastive learning is applied across original, augmented, and anchor-guided views to enforce structural consistency and semantic alignment.

more expressive representation space. The methodology is executed in two primary stages: initial feature enhancement and dynamic graph generation.

The textual modality provides an explicit semantic structure crucial for sentiment analysis. We construct the text graph, $\mathcal{G}^t$, based on the syntactic dependency parse tree of the textual transcript. Each word corresponds to a node in the graph, and directed edges represent grammatical relationships (e.g., subject, object, or modifiers). This graph serves as a core structural element, preserving the fundamental linguistic organization that guides the processing of other modalities.

We first enhance the unimodal representations by incorporating semantic information from the text. This is achieved via a cross-attention mechanism, where the textual representations serve as the queries to enrich the non-textual representations. Given the initial feature sequences $X^t \in \mathbb{R}^{T_t \times d}$ and $X^m \in \mathbb{R}^{T_m \times d}$ for text and modality $m \in \{a, v\}$ respectively, the augmented feature sequence $\hat{H}^m$ is obtained by Eq. 1.

$$\hat{H}^m = \text{Attention}(X^t, X^m). \tag{1}$$

The final enhanced representation, H'^m, is then derived by adding this augmented feature back to the original feature sequence via a residual connection (Eq. 2).

$$H'^m = X^m + \hat{H}^m. \tag{2}$$

For the audio and visual modalities, we dynamically construct a sparse and robust graph based on the cross-modal semantic relevance. We compute a raw score matrix $S^m \in \mathbb{R}^{T_m \times T_t}$ that quantifies the semantic similarity between segments of modality m and the text by Eq. 3.

$$S^m = \frac{(X^m \cdot W_Q)(X^t \cdot W_K)^\top}{\sqrt{d'}} + B_{pos}, \tag{3}$$

where $W_Q, W_K \in \mathbb{R}^{d \times d'}$ are learnable projection matrices, and B_{pos} is a learnable temporal position bias matrix that encourages connections between temporally proximate segments. The raw scores are then normalized via a softmax function to obtain the base adjacency matrix A'^m, which is subsequently symmetrized to ensure bidirectional relationships (Eq. 4).

$$A'^m = \frac{1}{2}(\text{softmax}(S^m) + \text{softmax}(S^m)^\top). \tag{4}$$

To enhance robustness and computational efficiency, we introduce an adaptive sparsification strategy that prunes noisy connections. For each sample, we compute a dynamic threshold τ_m based on the mean (μ_m) and deviation (σ_m) of the base matrix A'^m, $\tau_m = \mu_m + \lambda\sigma_m$, where λ is a learnable scaling factor. We then generate a binary mask matrix M^m that only retains edges with scores exceeding this threshold, $M^{m_{ij}} = \mathbb{I}(A'^{m_{ij}}) \geq \tau_m$, where $\mathbb{I}(\cdot)$ is the indicator function. The final, dynamically pruned adjacency matrix A'^m is obtained by applying this mask to the base matrix, as Eq. 5.

$$A^m = A'^m \odot M^m. \tag{5}$$

To better illustrate this process, Fig. 3 provides a visual explanation of how our text-guided attention mechanism enables semantic graph construction. Specifically, textual tokens are first used to query the audio and visual segments

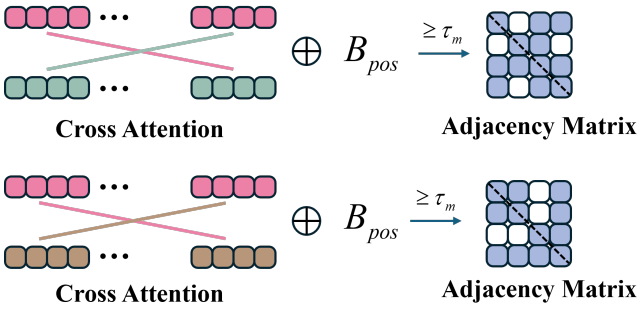


Figure 3: Illustration of text-guided graph construction for audio and visual modalities.

through a cross-modal attention operation. Each non-text segment obtains a text-informed attention vector. Next, pairwise similarity between attention vectors within each modality is computed to capture the semantic relatedness of non-textual segments under textual guidance. The resulting affinity matrix forms the basis of the graph structure, which is then sparsified by thresholding, as described above.

Semantic Anchor Integration

While modality-specific graphs capture fine-grained *intra*-modal structural dependencies, their embeddings still reside in heterogeneous spaces, which may induce cross-modal semantic misalignment. To alleviate this, we introduce a **semantic anchor**—shared latent nodes derived from global textual semantics—serving as a bridge across heterogeneous modality graphs to enforce cross-modal consistency.

Let $X^t \in \mathbb{R}^{N_t \times d}$ denote the contextualized textual sequence. We instantiate the anchor as a global representation vector:

$$\mathbf{z}_a = \text{AvgPool}\big(X^t\big) \in \mathbb{R}^d. \tag{6}$$

This anchor summarizes the utterance-level semantic intent conveyed by the textual modality and, unlike individual tokens, provides a holistic abstraction that is less sensitive to token-level noise.

To unify modalities structurally and semantically, we inject the semantic anchor into each modality-specific graph $\mathcal{G}^m = (\mathcal{V}^m, \mathcal{E}^m)$ by:

- Adding the anchor node to the node set: $\mathcal{V}^m \leftarrow \mathcal{V}^m \cup \{z_a\}$;
- Connecting the anchor to all modality nodes via attention-based edge weights:

$$\beta_i^m = \text{softmax}(z_a^\top W_a H_i'^m), \tag{7}$$

where $W_a \in \mathbb{R}^{d \times d}$ is a learnable projection matrix, and β_i^m indicates the strength of semantic relevance between z_a and node i in modality m.

The resulting augmented graph contains a shared semantic hub that aligns structurally disjoint modality graphs into a common space.

We further construct a **fusion graph** $\mathcal{G}^f$ that combines all modality nodes and the semantic anchor. This graph allows interactions not only within each modality but also across modalities via anchor-mediated connections. Specifically, edges are retained between:

- Modality-specific nodes (from $\mathcal{G}^a$, $\mathcal{G}^v$, and $\mathcal{G}^t$),
- Each node and the anchor node,
- Select cross-modality pairs with high semantic affinity to the anchor.

This design allows the anchor to serve as both a semantic aggregator and a structural bridge across modalities, effectively promoting coherent and interpretable fusion.

Multi-View Contrastive Learning Objective

To further enhance the discriminability, robustness, and semantic coherence of the learned representations, we incorporate a multi-view contrastive learning objective into the SSU framework. This design encourages alignment across heterogeneous modalities, resilience to structural perturbations, and preservation of task-relevant semantics.

We formulate three distinct views of each input:

- **Original View** (*structure-aware*): modality-specific graphs $\mathcal{G}^m$ are encoded via a shared GAT to capture intra-modal structure and anchor-enhanced semantics.
- **Fusion View** (*anchor-guided*): modality graphs are unified through the semantic anchor into a fusion graph $\mathcal{G}^f$, which is encoded using a separate GAT to model cross-modal alignment.
- **Augmented View** (*noisy variant*): random perturbations (e.g., edge addition/deletion) are applied to $\mathcal{G}^m$ to simulate noisy conditions, and representations are re-encoded using the same shared GAT.

Let z_{ori}, z_{aug}, and z_{fuse} denote the representations derived from the original, augmented, and fusion views, respectively. Our contrastive objective consists of three components:

Supervised Discrimination Loss. Given labeled training samples, we enforce supervised separation via a standard cross-entropy loss:

$$\mathcal{L}_{\text{sup}} = \text{CE}(\text{Classifier}(z_{\text{ori}}), y). \tag{8}$$

Structural Consistency Loss. To ensure robustness to structural perturbations, we apply a self-supervised contrastive loss between z_{ori} and z_{aug}:

$$\mathcal{L}_{\text{self}} = -\log \frac{\exp(\text{sim}(z_{\text{ori}}, z_{\text{aug}})/\eta)}{\sum_{z^- \in \mathcal{N}} \exp(\text{sim}(z_{\text{ori}}, z^-)/\eta)}, \tag{9}$$

where η is a temperature parameter and $\mathcal{N}$ denotes the set of negative samples in the mini-batch.

Semantic Alignment Loss. To unify multimodal semantics, we encourage the anchor-guided fusion representation to align with the original:

$$\mathcal{L}_{\text{align}} = \|z_{\text{fuse}} - z_{\text{ori}}\|_2^2. \tag{10}$$

Method	CMU-MOSI				CMU-MOSEI			
	ACC2↑	F1↑	ACC7↑	MAE↓	ACC2↑	F1↑	ACC7↑	MAE↓
Attention-based Methods								
CIA (Chauhan et al. 2019)	79.88%	79.54%	38.92%	0.9147	80.37%	78.23%	50.14%	0.6835
MAT (Delbrouck, Tits, and Dupont 2020)	80.00%	80.00%	35.41%	0.9230	82.00%	82.00%	47.32%	0.7067
TBJE (Delbrouck et al. 2020)	81.00%	78.00%	40.00%	0.8950	82.48%	65.54%	45.52%	0.7441
GATE (Kumar and Vepa 2020)	83.91%	81.17%	42.85%	0.8600	85.27%	84.08%	53.26%	0.6200
MPT (Cheng et al. 2021)	82.80%	82.90%	43.20%	0.7900	82.60%	82.80%	50.60%	0.5800
UniMSE (Hu et al. 2022)	86.90%	86.42%	48.68%	0.6914	87.50%	87.46%	54.39%	0.5238
SPECTRA (Yu et al. 2023)	87.50%	87.20%	49.20%	0.6600	87.34%	87.10%	53.95%	0.5350
MMML+FusionNet (Wu et al. 2023)	88.16%	88.15%	48.25%	0.6429	86.73%	86.49%	51.54%	0.5154
CMPT (Reza et al. 2025)	88.51%	88.34%	49.13%	0.6223	87.03%	86.92%	53.64%	0.5271
Graph-based Methods								
MMGraph (Mai et al. 2020)	81.40%	81.70%	49.70%	0.6082	80.60%	80.50%	32.10%	0.9331
GraphCAGE (Wu, Mai, and Hu 2021)	82.10%	82.10%	35.40%	0.9333	81.70%	81.80%	48.90%	0.6092
CJTF-BERT (Lu et al. 2024)	86.50%	86.40%	47.00%	0.7046	86.10%	86.04%	52.90%	0.5137
MoSARe (Moradinasab et al. 2025)	88.37%	88.10%	49.21%	0.6346	87.03%	86.91%	53.90%	0.5223
SSU (Ours)	**89.32%**	**89.28%**	**51.89%**	**0.5666**	**87.93%**	**87.72%**	**55.29%**	**0.5090**

Table 1: Comparison with state-of-the-art models on CMU-MOSI and CMU-MOSEI datasets. Metrics include binary accuracy (ACC2↑), F1-score (F1↑), 7-class accuracy (ACC7↑), and mean absolute error (MAE↓), where ↑ indicates higher is better and ↓ indicates lower is better.

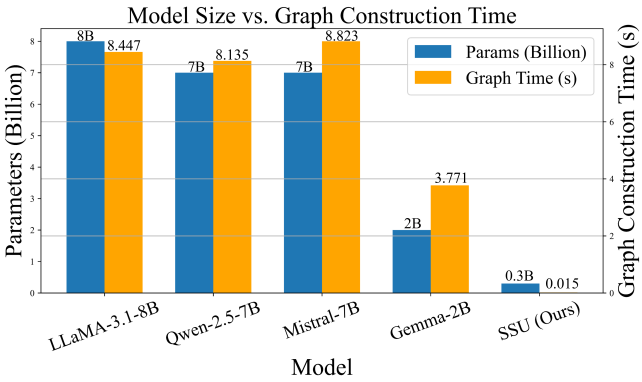


Figure 4: Comparison of model size and graph construction time. SSU achieves the best efficiency–performance trade-off, requiring only 0.3B parameters and approximately 0.015s per batch, significantly outperforming LLM-based graph constructors with substantially lower computational overhead.

Final Objective. Here, $\mathcal{L}_{\text{reg}}$ denotes the regression loss for continuous sentiment prediction. The total training loss is a weighted sum of the objectives:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \qquad (11)$$

where λ_{sup}, λ_{self}, and λ_{align} are hyperparameters that control the balance among different learning signals. This multi-view contrastive formulation encourages the model to learn structure-aware, semantically aligned, and perturbation-resilient representations—crucial for robust multimodal sentiment understanding.

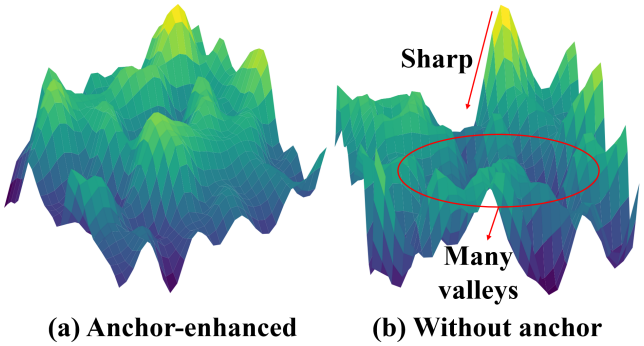


Figure 5: Loss surface visualization with (left) and without (right) semantic anchor. The anchor yields a smoother landscape, indicating improved optimization stability.

Experiments

Datasets and Metrics

We evaluate on two standard MSA benchmarks: CMU-MOSI (Zadeh et al. 2016) (2,199 utterances from 93 videos, scores in $[-3, 3]$) and CMU-MOSEI (Zadeh et al. 2018) (23,453 utterances from 1,000+ speakers). We follow the standard train/validation/test splits and report four metrics.

- **ACC2**: Binary accuracy (positive vs. negative sentiment).
- **F1-score**: Harmonic mean of precision and recall.
- **ACC7**: 7-class accuracy based on discretized sentiment scores.
- **MAE**: Mean Absolute Error for continuous sentiment regression.

Experiments run on 8×NVIDIA A100 (512 GB RAM) with PyTorch 1.8.2/CUDA 11.1 and spaCy 3.5.0. We fix seed

CMU-MOSEI				
Model	ACC2↑	F1↑	ACC7↑	MAE↓
LLaMA-3.1-8B	86.72%	86.62%	53.30%	0.52
Qwen-2.5-7B	86.55%	86.47%	53.12%	0.52
Mistral-7B	85.33%	85.20%	52.84%	0.56
Gemma-2B	84.15%	84.05%	51.35%	0.59
SSU (Ours)	**87.93%**	**87.72%**	**55.29%**	**0.51**
CMU-MOSI				
Model	ACC2↑	F1↑	ACC7↑	MAE↓
LLaMA-3.1-8B	87.65%	87.55%	44.30%	0.72
Qwen-2.5-7B	87.43%	87.30%	44.00%	0.73
Mistral-7B	86.21%	86.12%	43.55%	0.75
Gemma-2B	84.73%	84.65%	41.90%	0.80
SSU (Ours)	**89.32%**	**89.28%**	**51.89%**	**0.57**

Table 2: Comparison of SSU and LLM-based graph constructors on CMU-MOSEI and CMU-MOSI. Metrics include accuracy (ACC2, ACC7), F1, and MAE. SSU outperforms all LLMs while incurring minimal runtime cost.

CMU-MOSI				
Method	ACC2↑	F1↑	ACC7↑	MAE↓
w/o Semantic Anchor	86.89%	86.86%	48.54%	0.5990
Full model (SSU)	**89.32%**	**89.28%**	**51.89%**	**0.5666**
CMU-MOSEI				
Method	ACC2↑	F1↑	ACC7↑	MAE↓
w/o Semantic Anchor	87.37%	87.27%	53.94%	0.5175
Full model (SSU)	**87.93%**	**87.72%**	**55.29%**	**0.5090**

Table 3: Effect of semantic anchor. Performance drops without the anchor confirm its importance in semantic alignment and stable graph learning.

68 and train with batch size 128, sequence length 128, hidden size 128, and learning rate 1×10^{-5}.

Comparison with SOTA Methods

We compare SSU with recent state-of-the-art methods across both attention-based and graph-based multimodal sentiment models.

Attention-based models (e.g., CIA (Chauhan et al. 2019), GATE (Kumar and Vepa 2020), MPT (Cheng et al. 2021), UniMSE (Hu et al. 2022), SPECTRA (Yu et al. 2023)) focus on dynamic cross-modal alignment using attention or contrastive objectives. **Graph-based methods** (e.g., MM-Graph (Mai et al. 2020), CJTF-BERT (Lu et al. 2024), MoSARe (Moradinasab et al. 2025)) explicitly model structural dependencies, often relying on static graphs.

As shown in Table 1, SSU achieves new state-of-the-art results on both CMU-MOSI and CMU-MOSEI across all metrics. Compared to the graph-based baseline MoSARe, SSU improves ACC2 by $+1.6\%$ on MOSI and $+1.3\%$ on MOSEI, while reducing MAE and improving ACC7.

Loss Combination	ACC2↑	F1↑	ACC7↑	MAE↓
CMU-MOSI				
$\mathcal{L}_{\text{reg}}$ only	86.45%	86.38%	46.30%	0.7031
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{align}}$	87.32%	87.25%	47.40%	0.6837
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{self}}$	88.10%	88.00%	48.55%	0.6622
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$	88.54%	88.46%	49.30%	0.6455
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{self}}$ + $\mathcal{L}_{\text{align}}$	88.73%	88.65%	50.10%	0.6233
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$ + $\mathcal{L}_{\text{align}}$	88.91%	88.84%	50.66%	0.6074
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$ + $\mathcal{L}_{\text{self}}$	88.97%	88.90%	51.02%	0.5901
Full (Ours)	**89.32%**	**89.28%**	**51.89%**	**0.5666**
CMU-MOSEI				
$\mathcal{L}_{\text{reg}}$ only	86.15%	86.08%	52.13%	0.6034
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{align}}$	86.94%	86.90%	53.21%	0.5893
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{self}}$	87.22%	87.16%	54.00%	0.5756
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$	87.47%	87.41%	54.31%	0.5632
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{self}}$ + $\mathcal{L}_{\text{align}}$	87.61%	87.55%	54.65%	0.5528
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$ + $\mathcal{L}_{\text{align}}$	87.74%	87.68%	54.91%	0.5392
$\mathcal{L}_{\text{reg}}$ + $\mathcal{L}_{\text{sup}}$ + $\mathcal{L}_{\text{self}}$	87.83%	87.71%	55.13%	0.5264
Full (Ours)	**87.93%**	**87.72%**	**55.29%**	**0.5090**

Table 4: Ablation study on the multi-view contrastive objective. The full model combines all three contrastive losses and the main regression loss.

SSU versus LLM-based Graph Constructors

We compare SSU with representative large language models (LLMs) that construct modality graphs via offline prompting (Touvron et al. 2023; Bai et al. 2025; Hamzah and Sulaiman 2024; Team et al. 2024). While LLMs encode strong linguistic priors, they suffer from high graph construction latency and require multi-billion parameter models, limiting their practicality for real-time applications.

SSU constructs modality-specific graphs online and without supervision. For fair comparison, all LLMs are 8-bit quantized and receive concatenated text, audio, and visual embeddings. Hidden states from their penultimate layers are linearly projected to form adjacency matrices. SSU completes this process in 0.015s, significantly faster than LLaMA-3.1-8B (8.45s), Qwen-2.5-7B (8.14s), Mistral-7B (8.82s), and Gemma-2B (3.77s), as shown in Fig. 4. Despite having only 0.3B parameters, SSU consistently outperforms these LLMs on CMU-MOSI and CMU-MOSEI benchmarks (Table 2).

Effect of Semantic Anchor

We assess the impact of the semantic anchor via ablation. As shown in Table 3, removing the anchor consistently degrades performance on both CMU-MOSI and CMU-MOSEI—e.g., a 2.42% F1 drop on MOSI—accompanied by lower ACC2/ACC7 and higher MAE, indicating weaker semantic discrimination and alignment. The loss-surface visualization in Fig. 5 further shows a smoother landscape with the anchor, whereas its absence yields sharper, less stable minima, underscoring the anchor's role in stabilizing optimization and promoting cross-modal semantic coherence.

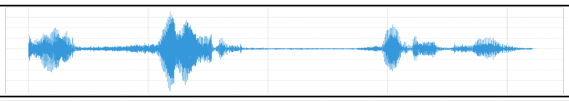
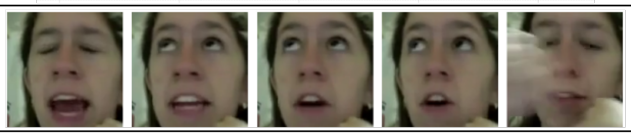
Modality	Behaviors	SSU Prediction	w/o Semantic Anchor	Ground Truth
Text	and i hate home on the range	**-2.21**	-0.28	**-2.60**
Audio				
Video				

Figure 6: Prediction comparison on a negative utterance. Anchors enhance modality alignment and semantic consistency, leading to better sentiment estimation. Predictions are raw regression outputs before activation (i.e., pre-logit values).

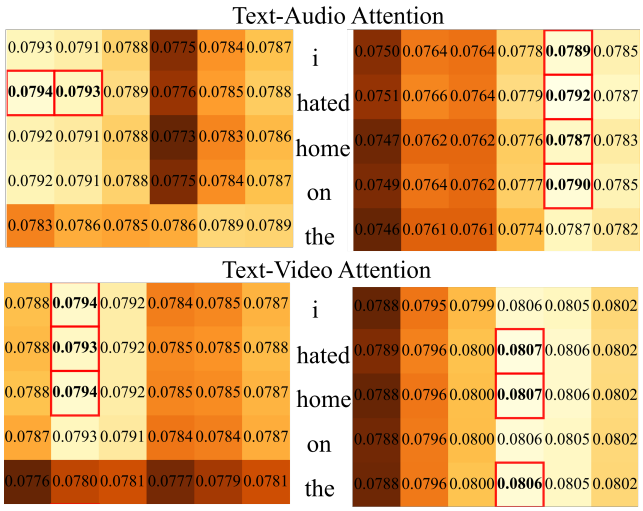


Figure 7: Cross-modal attention maps with (right) and without (left) semantic anchor. Anchors produce more concentrated and semantically meaningful attention.

Analysis of Multi-view Contrastive Objective

We evaluate the multi-view contrastive framework with an ablation that progressively adds contrastive losses to the regression objective under identical backbones and hyperparameters. As shown in Table 4, using only $\mathcal{L}_{\text{reg}}$ is competitive, and performance improves monotonically as $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{sup}}$, and $\mathcal{L}_{\text{self}}$ are introduced. The full combination attains the best results on both datasets, yielding consistent gains in ACC2, F1, and ACC7 alongside reduced MAE. These trends indicate complementary roles: $\mathcal{L}_{\text{align}}$ strengthens cross-modal agreement, $\mathcal{L}_{\text{sup}}$ sharpens class separability, and $\mathcal{L}_{\text{self}}$ enhances robustness to structural perturbations, leading to more stable and semantically aligned representations.

Case Study

To qualitatively assess our framework, we present two analyses: attention visualization and prediction comparison.

Prediction Visualization. Fig. 6 compares sentiment predictions with and without anchors. In the first case (*"and I hated home on the range"*), the full SSU model outputs a score of -2.21 (vs. ground truth -2.60), whereas the anchor-free variant yields a much weaker response (-0.28). With anchors, attention over text is sharply focused on sentiment-bearing words like *"hated"*, and aligns closely with expressive audio and visual cues (e.g., stressed vocal segments, negative facial expressions). In the second case, the anchor enables the model to correctly associate semantically rich tokens (e.g., *"Star Wars"*) with nonverbal sentiment indicators, reflecting improved generalization across modalities.

Cross-Modal Attention Analysis. Fig. 7 shows the text-audio and text-video attention maps for the first case. Without anchors, attention is scattered and lacks clear semantic focus. With anchors, attention becomes more concentrated and sentiment tokens align with salient nonverbal regions. Red boxes highlight the top-N attention weights, indicating the strongest cross-modal connections between text and audio/video segments. These results confirm that semantic anchor sharpen inter-modal alignment and enhance interpretability.

Conclusion

We propose SSU, a unified and lightweight framework for multimodal sentiment analysis that integrates intra-modal structural modeling with cross-modal semantic alignment. SSU builds modality-specific graphs using syntax and text-guided attention, and introduces semantic anchors to align heterogeneous modalities and stabilize training. A multi-view contrastive objective further enhances semantic coherence, structural integrity, and task discriminability. Extensive experiments on CMU-MOSI and CMU-MOSEI demonstrate that SSU achieves superior accuracy and efficiency, while case studies highlight its interpretability in capturing fine-grained sentiment signals.

Overall, SSU offers a scalable and interpretable approach that effectively bridges structure and semantics in multimodal sentiment understanding.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant 2024YFC3308500, National Natural Science Foundation of China under Grant 62406036, Beijing Municipal Natural Science Foundation under Grant L251042, the State Key Laboratory of Networking and Switching Technology under Grant NST20250110.

References

Anand, S.; Devulapally, N. K.; Bhattacharjee, S. D.; and Yuan, J. 2023. Multi-label emotion analysis in conversation via multimodal knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6090–6100.

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

Chauhan, D. S.; Akhtar, M. S.; Ekbal, A.; and Bhattacharyya, P. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 5647–5657.

Chen, Q.; Huang, G.; and Wang, Y. 2022. The weighted cross-modal attention mechanism with sentiment prediction auxiliary task for multimodal sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 2689–2695.

Cheng, J.; Fostiropoulos, I.; Boehm, B.; and Soleymani, M. 2021. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2447–2458.

Delbrouck, J.-B.; Tits, N.; Brousmiche, M.; and Dupont, S. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*.

Delbrouck, J.-B.; Tits, N.; and Dupont, S. 2020. Modulated fusion using transformer for linguistic-acoustic emotion recognition. *arXiv preprint arXiv:2010.02057*.

Gong, P.; Liu, J.; Zhang, X.; Li, X.; Wei, L.; and He, H. 2024. Adaptive Multimodal Graph Integration Network for Multimodal Sentiment Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hamzah, F.; and Sulaiman, N. 2024. Multimodal integration in large language models: A case study with mistral llm.

Han, W.; Chen, H.; and Poria, S. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv preprint arXiv:2107.06779*.

Huddar, M. G.; Sannakki, S. S.; and Rajpurohit, V. S. 2021. Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN.

Kumar, A.; and Vepa, J. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4477–4481. IEEE.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2023. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550: 126427.

Li, M.; Yang, D.; Lei, Y.; Wang, S.; Wang, S.; Su, L.; Yang, K.; Wang, Y.; Sun, M.; and Zhang, L. 2024. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 10074–10082.

Liang, B.; Gui, L.; He, Y.; Cambria, E.; and Xu, R. 2024. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*, 15(4): 1874–1888.

Lin, R.; and Hu, H. 2023. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11: 1686–1702.

Lu, Q.; Sun, X.; Gao, Z.; Long, Y.; Feng, J.; and Zhang, H. 2024. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. *Information Processing & Management*, 61(1): 103538.

Mai, S.; Xing, S.; He, J.; Zeng, Y.; and Hu, H. 2020. Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. *arXiv preprint arXiv:2011.13572*.

Mai, S.; Xing, S.; and Hu, H. 2021. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1424–1437.

Moradinasab, N.; Sengupta, S.; Liu, J.; Syed, S.; and Brown, D. E. 2025. Towards robust multimodal representation: A unified approach with adaptive experts and alignment. *arXiv preprint arXiv:2503.09498*.

Peng, W.; Hong, X.; and Zhao, G. 2021. Adaptive modality distillation for separable multimodal sentiment analysis. *IEEE Intelligent Systems*, 36(3): 82–89.

Reza, M. K.; Patil, A.; Solh, M.; and Asif, M. S. 2025. Robust Multimodal Learning via Cross-Modal Proxy Tokens. *arXiv preprint arXiv:2501.17823*.

Su, L.; Hu, C.; Li, G.; and Cao, D. 2020. Msaf: Multimodal split attention fusion. *arXiv preprint arXiv:2012.07175*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.

Wu, J.; Mai, S.; and Hu, H. 2021. Graph capsule aggregation for unaligned multimodal sequences. In *Proceedings of the 2021 international conference on multimodal interaction*, 521–529.

Wu, Z.; Gong, Z.; Koo, J.; and Hirschberg, J. 2023. Multimodal multi-loss fusion network for sentiment analysis. *arXiv preprint arXiv:2308.00264*.

Yang, X.; Feng, S.; Zhang, Y.; and Wang, D. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 328–339.

Yu, T.; Gao, H.; Lin, T.-E.; Yang, M.; Wu, Y.; Ma, W.; Wang, C.; Huang, F.; and Li, Y. 2023. Speech-text dialog pre-training for spoken dialog understanding with explicit cross-modal alignment. *arXiv preprint arXiv:2305.11579*.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12): 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.