SDG-MLLM: Injecting Structured Dialogue Graphs into MLLM for Multimodal Conversational Aspect-Based Sentiment Analysis

Xinjing Liu
Beijing University of Posts and Telecommunications
Beijing, China
liuxj_ai@bupt.edu.cn

Pengyue Lin
Beijing University of Posts and Telecommunications
Beijing, China
linpengyue@bupt.edu.cn

Xinyu Tu
Beijing University of Posts and Telecommunications
Beijing, China
tuxinyu@bupt.edu.cn

Wenqi Jia
Beijing University of Posts and Telecommunications
Beijing, China
jiawenqi@bupt.edu.cn

Chen Jiang
Beijing University of Posts and Telecommunications
Beijing, China
jiangchen@bupt.edu.cn

Ruifan Li*
Beijing University of Posts and Telecommunications
Beijing, China
rfli@bupt.edu.cn

Abstract

Multimodal Conversational Aspect-based Sentiment Analysis (MCABSA) is a challenging task for multimodal dialogue understanding. Existing works often treat the entire dialogue as a flat sequence and feed it into Large Language Models (LLMs) for pipeline-style generation. However, these methods sometimes accumulate errors and overlook critical discourse structure and fine-grained inter-word relations that are essential for accurate sentiment reasoning. To address these limitations, we propose *SDG-MLLM*, a unified generative framework that integrates Structured Dialogue Graphs into Multimodal LLM (MLLM) for an end-to-end MCABSA. Specifically, we construct heterogeneous dialogue graphs that capture diverse structural relations, including syntactic dependencies, coreference links, speaker turns, reply flow, semantic role labeling, and sentiment propagation paths. These graphs are encoded using a heterogeneous dialogue graph encoder, and the resulting structure-aware graph features are injected into the embedding layer of LLM. Furthermore, SDG-MLLM incorporates aligned multimodal features such as image, audio, and video cues at the utterance level to enable unified and context-aware multimodal reasoning. Experiments on the MCABSA dataset show that SDG-MLLM significantly outperforms strong baselines across multiple tasks. In addition, our method also achieved top performance in the ACM MM 2025 Grand Challenge of MCABSA. Our code is available at https://github.com/Liuxj-Anya/SDG-MLLM.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

*Corresponding Author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762071

Keywords

Heterogeneous Dialogue Graph, Multimodal Large Language Model, Multimodal Conversational Aspect-Based Sentiment Analysis.

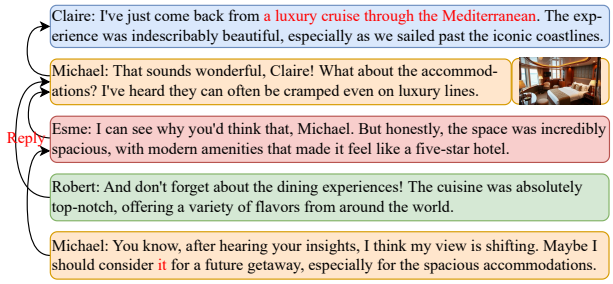
ACM Reference Format:

Xinjing Liu, Pengyue Lin, Xinyu Tu, Wenqi Jia, Chen Jiang, and Ruifan Li. 2025. SDG-MLLM: Injecting Structured Dialogue Graphs into MLLM for Multimodal Conversational Aspect-Based Sentiment Analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3762071

1 Introduction

Understanding fine-grained sentiment in multimodal social dialogues is important for analyzing public discourse in online interactions. Unlike traditional sentiment analysis tasks, Multimodal Conversational Aspect-based Sentiment Analysis (MCABSA) [20] involves multi-turn conversations, where each utterance may be accompanied by additional modalities such as images, audio, or video. Effective understanding of such dialogues requires not only analyzing the textual content, but also capturing the dialogue flow between speakers and the corresponding multimodal cues. MCABSA includes two subtasks. **1)** Panoptic Sentiment Sextuple Extraction extracts sentiment sextuple including the holder, target, aspect, opinion, sentiment polarity, and rationale. **2)** Sentiment Flipping Analysis detects dynamic changes in a speaker's sentiment toward a given aspect-target pair and identifies the underlying trigger of the flip. In Figure 1, speakers express opinions on different aspects of a shared target (a luxury cruise through the Mediterranean), with some sentiments shifting over the course of the conversation.

Existing methods [20] often treat the entire dialogue as a flat sequence and feed it into LLMs for pipeline-style generation. This method easily accumulates errors and overlooks the rich structural information embedded in dialogues. As a result, the model struggles to capture nuanced sentiment cues, speaker-intent shifts, and long-distance semantic dependencies essential for fine-grained sentiment reasoning. However, MCABSA remains a challenging task due to the complexity of dialogue text. **First**, multi-turn conversations are inherently structured, with non-sequential reply relations. For example, in Figure 1, Michael initially expresses a negative sentiment on the target: a luxury cruise through the Mediterranean, but later



Subtask-1: Panoptic Sentiment Sextuple Extraction

Holder	Target	Aspect	Opinion	Sentiment	Rationale
Claire	a luxury cruise through the Mediterranean	scenic beauty	indescribably beautiful	positive	we sailed past the iconic coastlines
Michael		accommodations	cramped	negative	even on luxury lines
Esme		accommodations	incredibly spacious	positive	made it feel like a five-star hotel
Robert		dining experiences	absolutely top-notch	positive	offering a variety of flavors
Michael		accommodations	worth considering	positive	the spacious accommodations

Subtask-2: Sentiment Flipping Analysis

Holder	Target	Aspect	Initial Sentiment	Flipped Sentiment	Trigger type
Michael	a luxury cruise through the Mediterranean	accommodations	negative	positive	Participant Feedback and Interaction

Figure 1: An illustration of MCABSA task. Each utterance may contain image, audio, or video modalities. Subtask-I extracts full sentiment tuples, i.e., (holder, target, aspect, opinion, sentiment, rationale). Subtask-II jointly extracts aspect-target pairs, sentiment shifts, and triggers.

revises his view after hearing from Esme. This shift, along with speaker reply and long-range dependencies, cannot be effectively modeled by treating dialogues as flat text. **Second**, key sentiment elements such as target, aspects, and rationales are often implicitly expressed or dispersed across multiple utterances, requiring structural reasoning beyond surface-level token sequences. **Third**, fine-grained sentiment understanding often depends on relational cues such as syntactic dependencies, co-reference links, and sentiment propagation paths. For example, in Figure 1, Michael's use of "it" likely refers to "a luxury cruise through the Mediterranean" mentioned earlier by Claire. Several sentiment expressions are also contextually linked across utterances, which are not explicitly encoded in standard language model inputs.

To address these challenges, we propose SDG-MLLM, a unified generative framework that injects structured dialogue graphs into Multimodal LLM (MLLM) for an end-to-end sentiment reasoning. Instead of flattening dialogues into plain sequences, we construct heterogeneous dialogue graphs that capture structural relations, such as syntactic dependencies, coreference links, speaker turns, reply flow, semantic role labeling, and sentiment propagation paths. These graphs are encoded using a heterogeneous dialogue graph encoder to obtain structure-aware graph features, which provide token-level features infused with structural knowledge. To seamlessly integrate structural information into LLM, we employ a gating mechanism to inject the structure-aware features into its embedding layer, obtaining graph-augmented features. This enables the model to reason jointly over textual and structural context. Finally, our model aligns multimodal features at the utterance level. These features are integrated with the graph-augmented features and jointly fed into LLM, enabling unified and end-to-end generation. By combining structural knowledge with MLLMs, SDG-MLLM gains a comprehensive understanding of dialogue context.

Our contributions are summarized as follows. **1)** We construct heterogeneous dialogue graphs that encode discourse-level structural relations. These graphs are encoded by a heterogeneous dialogue graph encoder, and the resulting structure-aware features are injected into the embedding layer of LLM. **2)** We propose SDG-MLLM, a unified generative end-to-end framework that integrates structured dialogue graphs into MLLM, effectively combining the strengths of graph-based reasoning and large-scale language modeling. **3)** Extensive experiments on the MCABSA dataset show that SDG-MLLM consistently outperforms strong baselines on two subtasks, demonstrating its effectiveness.

2 Related Work

2.1 Multimodal Conversational ABSA

Aspect-based Sentiment Analysis (ABSA) [4, 14, 26] aims to extract fine-grained aspect-sentiment pairs. Early ABSA methods focused on monologue textual data such as reviews [15, 28], using pipeline [3] or dependency-based approaches [23]. Recent advances leverage deep learning and Transformer architectures [37] for joint extraction of semantic relations. Multimodal ABSA extends this task by incorporating visual inputs alongside text, as in the Twitter-15/17 dataset [39]. Various multimodal fusion techniques [16, 17, 27, 42] demonstrate that visual context improves sentiment disambiguation and aids aspect detection. However, extending ABSA to multimodal, multi-turn dialogues remains underexplored. Prior works [12, 25, 35] mainly perform utterance-level coarse-grained sentiment classification, lacking the capacity to capture fine-grained aspect-level sentiments.

Therefore, Sentica [20] introduced the MCABSA benchmark, extending MABSA into the dialogue domain by formulating two fine-grained tasks: Panoptic Sentiment Sextuple Extraction and Sentiment Flipping Analysis. However, their method treats the entire dialogue as a flat sequence and feeds it into LLMs for pipeline-style generation, which easily accumulates errors and fails to explicitly model structural relations. To address these limitations, we propose a novel end-to-end framework that integrates structured dialogue graphs and multimodal cues into LLM for MCABSA.

2.2 Graph-Augmented Language Models

LLMs such as GPT [1] and LLaMA [31] have achieved strong performance on a wide range of tasks. However, they inherently process text as linear token sequences and often struggle with capturing discrete structural relationships, such as dependencies, speaker turns, and relational semantics that are better represented by graphs.

Recent studies inject graph-based knowledge into LLMs to address their structural understanding limitations. One line of work incorporates knowledge graphs by embedding triples or subgraphs into the input [21, 22, 29]. Another focuses on syntactic or semantic dependency graphs, using GCNs or GATs to enhance input representations before or during LLM encoding [24, 40]. Recent hybrid approaches inject graph representations via adapters [10, 11, 41] or graph-aware attention [6, 30, 33], improving performance on relation extraction, event reasoning, and commonsense inference.

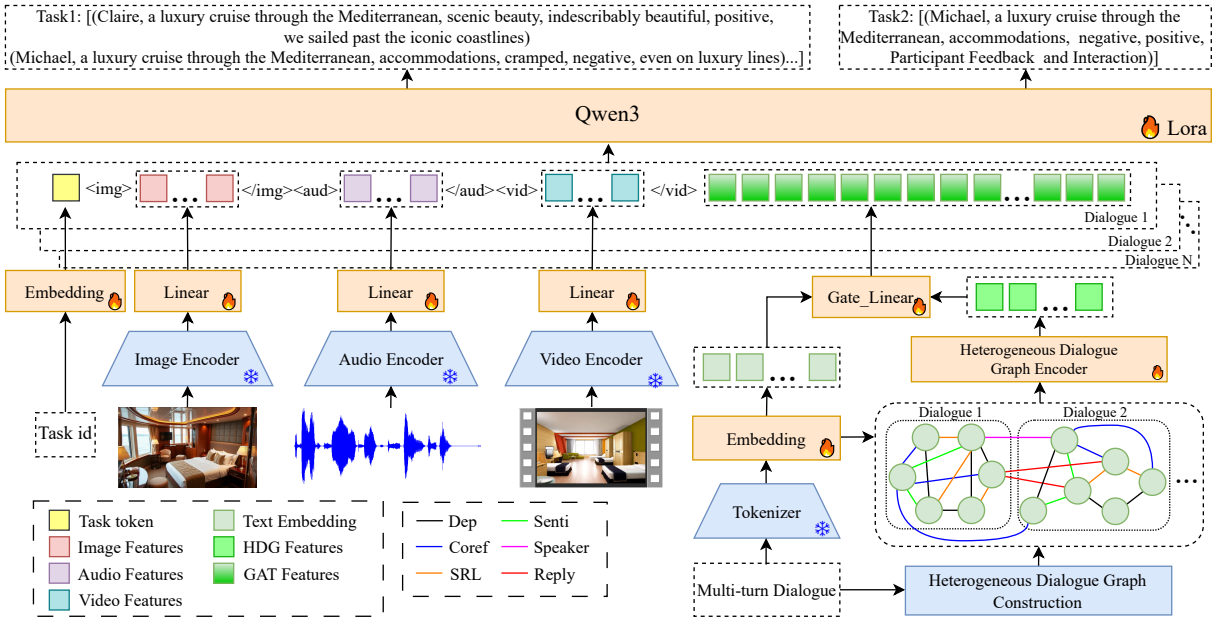


Figure 2: An overview of our SDG-MLM architecture. The model consists of four modules: 1) Multimodal Feature Extraction encodes utterance-level image, audio, and video features using pretrained encoders, 2) Heterogeneous Dialogue Graph Construction builds graphs capturing six types of structural relations, 3) Heterogeneous Dialogue Graph Encoder uses a relation-specific graph transformer to encode the graph and obtain structure-aware graph features, and 4) Graph-Augmented MLLM injects the multimodal features and Graph-Augmented Text features into a pretrained Qwen3.

We construct and encode heterogeneous dialogue graphs to inject structure-aware representations into LLMs, enabling explicit structural reasoning in multimodal dialogue understanding.

3 Methodology

In this section, we present our SDG-MLLM, a unified generative end-to-end framework designed to address the challenges of MCABSA. As illustrated in Figure 2, SDG-MLLM integrates heterogeneous structural knowledge and multimodal signals into a pretrained LLM, enabling sentiment reasoning over multi-turn dialogue contexts.

3.1 Problem Formulation

MCABSA is a sextuple prediction task over multi-turn multimodal social dialogues. Given a dialogue $D = \{u_1, u_2, ..., u_n\}$ consisting of n utterances, where each utterance u_i is composed of a sequence of tokens and may be accompanied by other modalities (image I_i, audio A_i, or video V_i). The goal is decomposed as two subtaks. **Subtask-I**, Panoptic Sentiment Sextuple Extraction, aims to extract sentiment sextuple of the form (h, t, a, o, s, r), representing the holder, target, aspect, opinion, sentiment polarity (from positive, negative, neutral), and rationale. These elements may appear explicitly in a single utterance or be implicitly conveyed across multiple turns and modalities, posing significant challenges for reasoning. **Subtask-II**, Sentiment Flipping Analysis, focuses on identifying sentiment transitions within a dialogue. The model is required to detect tuples $(h, t, a, \zeta, \phi, \tau)$, where the same speaker's sentiment toward a particular target-aspect pair flips from an initial sentiment ζ to a new sentiment ϕ. The trigger τ describes the underlying reason for the change. It is categorized into four predefined types: introduction of new information, logical argumentation, participant feedback and interaction, or personal experience and self-reflection.

3.2 Multimodal Feature Extraction

To enable fine-grained sentiment reasoning in multimodal dialogues, we extract modality-specific features at the utterance level, involving text, image, audio, and video. **1)** Text Features. We employ the pretrained Qwen3-8B [38] to encode the textual content of each utterance. Given an utterance $u_i = \{w_1, w_2, \ldots, w_m\}$, we obtain text embeddings $\mathbf{E}_{\text{text},i} = \{\mathbf{e}_{i1}, \ldots, \mathbf{e}_{ij}, \ldots, \mathbf{e}_{im}\} \in \mathbb{R}^{m \times d}$ from Qwen3's embedding layer. These embeddings are later fused with graph-informed and multimodal signals. **2)** Image Features. For utterances paired with images, we use the SigLIP2 vision encoder [32] to extract patch-level visual features, yielding $\mathbf{E}_{\text{img},i} \in \mathbb{R}^{576 \times 1536}$. These features encode both global and local visual semantics aligned with the utterance. **3)** Audio Features. For utterances paired with audio, we use WavLM's encoder [2] to extract audio features, yielding $\mathbf{E}_{\text{aud},i} \in \mathbb{R}^{273 \times 768}$. **4)** Video Features. For utterances paired with video clips, we uniformly sample frames and use SigLIP2 to extract frame-level visual embeddings. These are then averaged across frames to form video features $\mathbf{E}_{\text{vid},i} \in \mathbb{R}^{576 \times 1536}$.

3.3 Heterogeneous Dialogue Graph Construction

To capture the complex structural cues in multimodal conversations, we construct a heterogeneous dialogue graph for each dialogue instance. Each heterogeneous dialogue graph explicitly models multiple types of relational edges between tokens across the dialogue.

This method captures structural cues essential for fine-grained sentiment reasoning beyond surface-level token sequences. **Graph Definition.** Formally, for a dialogue $D = \{u_1, u_2, ..., u_n\}$, we define a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where each node $v \in \mathcal{V}$ corresponds to a token in the dialogue, and each edge $e = (v_p, v_q) \in \mathcal{E}$ denotes a relation $r \in \mathcal{R}$ from node v_p to node v_q. The graph is heterogeneous in the sense that it includes multiple edge types representing different discourse-level relations.

Edge Types. We include six types of edges in the dialogue graph, each capturing a distinct relation of conversational structure and meaning. **1)** Dependency Edges. We apply syntactic dependency parsing using spaCy [7] to extract grammatical relations (e.g., nsubj, dobj, and amod) within each utterance. These edges help capture local syntactic structure and modifier relationships between sentiment expressions and their targets. **2)** Coreference Edges. We utilize AllenNLP's [5] coreference resolution module to identify coreferent mentions within and across utterances (e.g., "it" ↔ "a luxury cruise through the Mediterranean"), linking tokens that refer to the same underlying entity. These edges are essential for resolving implicit references to sentiment targets or aspects mentioned earlier in the dialogue. **3)** Speaker Turn Edges. We connect the first tokens of adjacent utterances to reflect speaker transitions. **4)** Reply Flow Edges. This dataset provides explicit reply-to relations between utterances (e.g., $u_i \rightarrow u_j$), which we leverage to construct reply connections. Specifically, we add edges between the main verbs and nouns in the replying utterance u_i and those in the replied-to utterance u_j, allowing the model to capture predicate-argument interactions across turns. **5)** Sentiment Propagation Edges. Based on a sentiment lexicon, we connect sentiment-laden opinion words to candidate target and aspects within or across utterances, modeling the sentiment influence and contextual cues that guide sentiment inference. **6)** Semantic Role Labeling (SRL) Edges. Using AllenNLP's SRL parser, we extract predicate-argument structures and add edges between verbs and their arguments. These semantic frames enrich the graph with role-based relationships useful for understanding who expresses what sentiment and toward what.

Graph Construction Process. Each utterance is tokenized using the same tokenizer as Qwen3. Edges are created between subword tokens based on their corresponding word-level relations. For efficiency, we use adjacency lists with relation-type annotations and represent the final graph as $(\mathcal{E}, \mathcal{R})$, where $\mathcal{E} \in \mathbb{N}^{2|\mathcal{E}|}$ stores source and target indices for each edge, $\mathcal{R} \in \mathbb{N}^{|\mathcal{E}|}$ encodes relation type r for each edge. These dialogue graphs provide the foundation for subsequent graph encoding and fusion into the language model.

3.4 Heterogeneous Dialogue Graph Encoder

To capture the diverse structural relations in the heterogeneous dialogue graph, we employ the Heterogeneous Graph Transformer (HGT) [9] to encode structure-aware graph features. Unlike standard GAT-based models that treat all edges uniformly, HGT models each relation with distinct transformation and attention mechanisms, making it well-suited for our multi-relation dialogue graph.

Formally, given a heterogeneous dialogue graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where $\mathcal{V}$ is the set of token-level nodes corresponding to subword tokens from the original conversation. In addition, each node $p \in \mathcal{V}$ is initialized with its embedding $\mathbf{E}_{\text{text}}$, $\mathcal{E}$ is the set of edges, and $\mathcal{R}$ is the set of relation types. We compute relation-specific attention scores between a node p and its neighbor $q \in \mathcal{N}_p^{(r)}$ as follows,

$$\alpha_{pq}^{(r)} = \frac{(\mathbf{W}_q^{(\tau_p,r)}\mathbf{h}_p^{(l)})^\top (\mathbf{W}_k^{(\tau_q,r)}\mathbf{h}_q^{(l)})}{\sqrt{d}}. \tag{1}$$

Here, τ_p and τ_q denote the types of nodes p and q, and $\mathbf{W}_q$, $\mathbf{W}_k$ are type and relation-specific query and key projections. The message from node q is then transformed as follows,

$$\mathbf{m}_{pq}^{(r)} = \alpha_{pq}^{(r)} \mathbf{W}_v^{(\tau_q,r)} \mathbf{h}_q^{(l)}. \tag{2}$$

All incoming messages are aggregated across all relations and neighbors, and passed through a type-specific feedforward network with residual connection and layer normalization, i.e.,

$$\mathbf{h}_p^{(l+1)} = \text{LN}(\mathbf{h}_p^{(l)} + \text{FFN}_{\tau_p}(\sum_{r\in\mathcal{R}} \sum_{q\in\mathcal{N}_p^{(r)}} \mathbf{m}_{pq}^{(r)})). \tag{3}$$

Stacking multiple layers of HGT allows each token node to iteratively integrate multi-hop, multi-relational context from the dialogue graph. The output features after the final layer, denoted as $\mathbf{h}_{ip}^{(L)}$, form the Heterogeneous Dialogue Graph (HDG) features $\mathbf{E}_{\text{HDG},i} = \{\mathbf{h}_{i1}^{(L)}, \ldots, \mathbf{h}_{ip}^{(L)}, \ldots, \mathbf{h}_{im}^{(L)}\}$. These structure-aware features are injected into the embedding layer of the LLM to support structural sentiment reasoning.

3.5 Graph-Augmented MLLM

To enable fine-grained sentiment reasoning in complex conversational contexts, we propose a graph-augmented input strategy that injects structure-aware features and multimodal cues into a pretrained LLM. The key idea is to construct the model's input at the utterance level, where each utterance is enriched with both aligned multimodal embeddings and graph-augmented textual features.

For each utterance u_i, the multimodal features $\mathbf{E}_{\text{img},i}$, $\mathbf{E}_{\text{aud},i}$, and $\mathbf{E}_{\text{vid},i}$ are projected into the same representation space as the text. Formally, we define a linear layer for each modality as follows,

$$\tilde{\mathbf{E}}_{m,i} = \mathbf{W}_{m,i}\mathbf{E}_{m,i} + \mathbf{b}_{m,i}, \quad m \in \{\text{img}, \text{aud}, \text{vid}\}. \tag{4}$$

And they are inserted before the corresponding utterance text, wrapped with modality-specific boundary tokens (e.g., $\langle m \rangle$, $\langle /m \rangle$). Thus, these form a modality block $\mathcal{M}_{m,i} = [\langle m \rangle, \tilde{\mathbf{E}}_{m,i}, \langle /m \rangle]$. This sequential integration enables the model to jointly attend to multimodal and textual information within a unified space.

Next, we obtain the text-enhanced features of u_i. For each token with a graph node, we apply a gated fusion to combine the original text embedding $\mathbf{e}_{ip}$ and its HDG features $\mathbf{h}_{ip}^{(L)}$. Thus, we obtain Graph-Augmented Text (GAT) features $\mathbf{e}_{ip}^{\text{GAT}} \in \mathbf{E}_{\text{GAT,i}}$ that capture both contextual semantics and explicit dialogue structure,

$$\begin{cases} \lambda_{ip} = \sigma\left(\mathbf{W}_g[\mathbf{e}_{ip}; \mathbf{h}_{ip}^{(L)}]\right) \\ \mathbf{e}_{ip}^{\text{GAT}} = \lambda_i \mathbf{e}_{ip} + (1-\lambda_{ip})\mathbf{h}_{ip}^{(L)} \end{cases} \tag{5}$$

We prepend a special task token $\langle$task1$\rangle$ or $\langle$task2$\rangle$ at the beginning of the input sequence to explicitly indicate which subtask the model should perform. Let $\hat{u}_i$ denote the enriched representation of utterance u_i, defined as $\hat{u}_i = [\mathcal{M}_{img,i}; \mathcal{M}_{aud,i}; \mathcal{M}_{vid,i}; \mathbf{E}_{\text{GAT,i}}]$.

	Dialogue			Sextuple		Modality			
	Dia.	Utt.	Spk.	Sext.	Flip.	Txt.	Img.	Aud.	Vid.
Total	6783	32065	30417	32404	2354	32065	2157	854	518
train	4226	20321	18968	20074	1460	20321	1224	405	208
val	1057	5045	4661	4997	364	5045	305	90	50
test	1500	6699	6788	7333	530	6699	628	359	260

Table 1: Main statistics of PanoSent dataset.

The full input sequence is then constructed by concatenating all enriched utterances in natural dialogue order,

$$\hat{\mathcal{D}} = [\ \langle \text{task} \rangle; \hat{u}_1; \hat{u}_2; \ldots; \hat{u}_n\ ]. \tag{6}$$

This design preserves utterance structure and multimodal alignment while allowing global attention across the entire dialogue.

The final multimodality-enhanced input sequence $\hat{\mathcal{D}}$ is fed into a pretrained Qwen3-8B, which is responsible for generating structured sentiment outputs for both tasks, i.e.,

$$y_t = Qwen3(y_{<t}, \hat{\mathcal{D}}). \tag{7}$$

We adopt a causal language modeling objective, optimizing the model to autoregressively generate target sequences given the dialogue input. The training loss is defined as follows,

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, X), \tag{8}$$

where X denotes the input sequence, $y_{1:T}$ is the target output sequence (i.e., sentiment tuples), and $P(y_t \mid y_{<t}, X)$ is the probability of generating token y_t at time step t.

To efficiently adapt the LLMs, we apply LoRA-based [8] finetuning to the Qwen3. In contrast, all projection layers for image, audio, and video modalities, as well as the Heterogeneous Dialogue Graph Encoder and gate linear, are trained in a fully end-to-end manner to enable rich multimodal and structural integration.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on the PanoSent dataset [20], which is designed to evaluate multimodal conversational sentiment analysis. The statistics of this dataset are summarized in Table 1.

Evaluation Metrics. We follow the evaluation established in Sentica [20]. Subtask-I is evaluated at three levels of granularity: 1) element-level, 2) pairs-level, and 3) full sextuple-level. For explicit elements such as Holder, Target, and Sentiment, we apply exact-match F1. For implicit elements, we adopt a binary matching approach, using GPT-4 to assess semantic equivalence (1 if equivalent and 0 otherwise). Rationale evaluation uses proportional overlap-based F1 to allow partial but meaningful matches, as precise span matching is not always needed. In compound-level evaluations (pairs-level or sextuple-level), a prediction is correct only if all elements are accurate. For rationales in this setting, we consider them correct if the proportional score exceeds 0.5. Sextuple evaluation is further divided into micro-F1, which assesses exact matches of the complete sextuple, and identification F1, which ignores sentiment polarity. For Subtask-II, we evaluate the exact-match F1 of the jointly predicted Initial-Flipped Sentiment and Trigger.

Model	Element-level					Pairs-level				Sextuple	
	H	T	A	O	R	T-A	H-O	S-R	O-S	Micro	Iden.
DiaASQ	69.56	58.61	52.04	44.39	22.90	33.07	33.52	18.98	40.26	13.49	19.07
UGF	71.17	61.83	55.25	47.68	25.87	35.39	36.08	22.37	42.80	15.85	20.12
Unified-IO 2	75.82	65.81	59.50	51.57	29.03	39.41	40.36	26.16	47.03	18.95	22.03
NExT-GPT	76.07	66.25	59.97	52.12	29.95	40.23	41.24	27.07	47.89	20.01	24.98
Sentica	84.30	76.51	71.16	62.47	43.23	51.09	52.20	39.50	60.25	32.18	35.72
SDG-MLLM	**96.14**	**77.86**	**81.17**	**75.46**	**87.48**	**61.35**	**74.82**	**84.60**	**74.20**	**48.79**	**49.34**

Table 2: Main results of Subtask-I, Panoptic Sentiment Sextuple Extraction. H/T/A/O/R/S represents Holder, Target, Aspect, Opinion, Rationale, and Sentiment, respectively.

NExT-GPT	Sentica	**SDG-MLLM**
55.80	69.39	**76.70**

Table 3: Results of Subtask-II, Sentiment Flipping Analysis.

Subtask	Team A	Our Team	Team C	Team D	Team E
Subtask-I	49.65	49.07	46.26	47.38	34.66
Subtask-II	76.18	76.70	74.46	74.12	79.04

Table 4: Results of the MM2025 Grand Challenge.

Baselines. We compare our model against several strong baselines across both subtasks. UGF [36] unifies all ABSA subtasks into a single generative framework. DiaASQ [13] formulates a dialogue-level ABSA task and proposes an end-to-end model. Unified-IO2 [19] unifies vision, language, audio, and action tasks in an instruction-based Transformer. NExT-GPT [34] enables any multimodal generation by combining an LLM with modality adaptors and diffusion decoders. Sentica [20] benchmarks MCABSA and a MLLM framework. All baselines are evaluated under the same experimental setting on the MCABSA benchmark.

Implementation Details. We train model on four A6000 GPUs with AdamW [18] (batch size 8, learning rate 5×10^{-5}) for 3 epochs.

4.2 Main Results

The results for Subtask-I are reported in Table 2. SDG-MLLM consistently outperforms all baselines across element-level, pair-level, and sextuple-level evaluations. In particular, our model achieves substantial gains in overall sextuple F1. The improvements are especially notable on aspects, opinions, and rationales that require inference across utterances. These results demonstrate that explicitly modeling discourse structures and inter-utterance relations through structured dialogue graphs significantly enhances the LLM's ability to understand complex sentiment in conversation.

The results for Subtask-II are reported in Table 3. SDG-MLLM achieves the highest performance in the jointly predicted Initial-Flipped Sentiment and Trigger. Our model better captures sentiment shifts and their causes by integrating structural dialogue graphs. Unlike Unified-IO 2 and NExT-GPT, which lack structural modeling, our approach achieves more accurate sentiment interpretation.

In addition, the overall results of all participating teams in the MM2025 Grand Challenge are summarized in Table 4, where our SDG-MLLM ranks among the top-performing systems.

Model	Subtask-I						Subtask-II
	T-A	H-O	S-R	O-S	Micro	Iden.	Flip-Trig
w/o Dialogue Graph	57.07	73.14	81.34	72.38	45.18	45.75	74.49
w/o Graph & Modalities	55.90	72.19	80.95	71.90	44.01	44.52	74.08
w/o Relation. Types	59.20	72.53	81.78	72.91	45.64	46.19	75.43
w/o Gated Fusion	60.26	72.80	83.80	72.71	46.92	47.48	75.77
Full SDG-MLLM	**61.35**	**74.82**	**84.60**	**74.20**	**48.79**	**49.34**	**76.70**

Table 5: Performance of our full model and its variants.

4.3 Ablation Study

To verify the effectiveness of each component in SDG-MLLM, we conduct various ablation experiments. The results are reported in Table 5. We first remove the heterogeneous dialogue graph and feed only the original dialogue text into the Qwen3 model. This leads to a significant decline in performance on both subtasks, especially on targets and aspects extraction. This suggests structural dialogue information is essential for fine-grained sentiment reasoning.

We further remove both the graph and multimodal components, using only the plain dialogue text along with the task prompt. This results in a substantial performance drop across both subtasks, highlighting the crucial role of structural and multimodal signals.

To assess the importance of relation heterogeneity, we collapse all edge types into a single generic relation. This leads to a notable performance decline, particularly in tasks that require fine-grained structural understanding such as pair-level extraction and flipping trigger classification. These results underscore the importance of explicitly modeling diverse relation types to preserve the structural nuances essential for accurate sentiment reasoning.

Lastly, we evaluate directly concatenates the graph features with the text embeddings instead of fusing them through the gated mechanism. This naive concatenation results in performance drop, highlighting that simple feature stacking is less effective than our proposed fusion strategy. The result supports our choice of using a gated mechanism at the embedding layer, which better captures the complementary nature of structural and linguistic information.

4.4 Case Study

To illustrate the impact of injecting heterogeneous structural knowledge, we conduct qualitative comparisons between SDG-MLLM and other baselines on representative MCABSA test examples. As shown in Figure 3, our model demonstrates a superior ability in reasoning over complex dialogue structures involving coreference, syntactic dependency, and reply relations.

In the left case, a user refers to "it" when discussing an entity mentioned earlier by another speaker. Baseline model fails to resolve this cross-utterance coreference, leading to incorrect aspect identification and sentiment attribution. In contrast, SDG-MLLM leverages coreference edges in the dialogue graph to accurately resolve the pronoun. As a result, it successfully extracts the full sentiment sextuple, including holder, target, aspect, opinion, sentiment, and rationale. This highlights the strength of our method of injecting Structured Dialogue Graphs into MLLMs in handling implicit references and cross-turn entity tracking.

In the right case, the sentiment is implied through subtle multi-relational links across utterances. Baseline model fails to capture the shift due to lack of structural graph context. In contrast, SDG-MLLM

Figure 3: Comparison between SDG-MLLM and its variant without the dialogue graph. The left case is Subtask-I result, while the right case is Subtask-II. Green text highlights incorrect predictions by the model without dialogue graph.

leverages reply edges, speaker turns, and sentiment propagation edges in the dialogue graph to trace inter-utterance influence. It successfully detects the sentiment flipping and correctly identifies the trigger as "Participant Feedback and Interaction". These structural relations help maintain sentiment continuity and model sentiment shifts driven by others' input. This enables SDG-MLLM to capture subtle affective shifts missed by flat-text models.

These cases show that integrating structural cues via heterogeneous graphs enables SDG-MLLM to outperform structure-agnostic baselines in fine-grained sentiment reasoning.

5 Conclusion and Future Work

In this paper, we presented SDG-MLLM, a unified generative framework that integrates structured dialogue knowledge into MLLM for end-to-end MCABSA. Our approach constructs heterogeneous dialogue graphs to capture rich relational structures. These graphs are encoded by a heterogeneous graph encoder and injected into the LLM's embedding space to infuse structural awareness. In addition, SDG-MLLM incorporates aligned multimodal features at the utterance level, enabling unified and context-aware multimodal reasoning. Experiments on the MCABSA benchmark demonstrate that SDG-MLLM consistently outperforms strong baselines, demonstrating the effectiveness of structural graph integration.

In the future, we plan to explore dynamic graph construction to better model evolving sentiment states over time, and to extend SDG-MLLM to other dialogue scenarios. We further explore fine-grained multimodal alignment and prompt-based adaptation.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (NSFC) No. 62076032, by the China Computer Federation of Zhipu Foundation No. CCF-Zhipu202407, and by BUPT Kunpeng & Ascend Center of Cultivation. The authors thank the organizers of the MCABSA grand challenge and their comments.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (Oct. 2022), 1505–1518. doi:10.1109/jstsp.2022.3188113

[3] Zhuang Chen and Tieyun Qian. 2022. Retrieve-and-Edit Domain Adaptation for End2End Aspect Based Sentiment Analysis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 30 (Jan. 2022), 659–672. doi:10.1109/TASLP.2022.3146052

[4] Rui Fan, Shu Li, Tingting He, and Yu Liu. 2025. Aspect-Based Sentiment Analysis with Syntax-Opinion-Sentiment Reasoning Chain. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 3123–3137. https://aclanthology.org/2025.coling-main.210/

[5] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform. arXiv:arXiv:1803.07640

[6] Zhong Guan, Likang Wu, Hongke Zhao, Ming He, and Jianpin Fan. 2024. Enhancing collaborative semantics of language model-driven recommendations via graph-aware learning. *arXiv preprint arXiv:2406.13235* (2024).

[7] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685

[9] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*. 2704–2710.

[10] Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms?. In *Proceedings of the ACM Web Conference 2024*. 893–904.

[11] Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Gnns as adapters for llms on text-attributed graphs. In *The Web Conference 2024*.

[12] Ye Jing and Xinpei Zhao. 2024. DQ-Former: Querying Transformer with Dynamic Modality Priority for Cognitive-aligned Multimodal Emotion Recognition in Conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 4795–4804. doi:10.1145/3664647.3681599

[13] Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13449–13467. doi:10.18653/v1/2023.findings-acl.849

[14] Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6319–6329. doi:10.18653/v1/2021.acl-long.494

[15] Xin Li and Wai Lam. 2017. Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2886–2892. doi:10.18653/v1/D17-1310

[16] Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-language pre-training for multimodal aspect-based sentiment analysis. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers)* (2022), 2149–2159.

[17] Xinjing Liu, Ruifan Li, Shuqin Ye, Guangwei Zhang, and Xiaojie Wang. 2025. Multimodal Aspect-Based Sentiment Analysis under Conditional Relation. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 313–323. https://aclanthology.org/2025.coling-main.22/

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[19] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26439–26455.

[20] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. 2024. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7667–7676.

[21] Zihan Luo, Xiran Song, Hong Huang, Jianxun Lian, Chenhao Zhang, Jinqi Jiang, and Xing Xie. 2024. Graphinstruct: Empowering large language models with graph understanding and reasoning capability. *arXiv preprint arXiv:2403.04483* (2024).

[22] Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8449–8456.

[23] Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu, and Lijie Wen. 2023. AMR-based Network for Aspect-based Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 322–337. doi:10.18653/v1/2023.acl-long.19

[24] Diego Marcheggiani and Ivan Titov. 2017. Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 1506–1515. doi:10.18653/v1/D17-1159

[25] Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. 2024. Ada2I: Enhancing Modality Balance for Multimodal Conversational Emotion Recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 9330–9339. doi:10.1145/3664647.3681648

[26] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8600–8607.

[27] Tianshuo Peng, Zuchao Li, Ping Wang, Lefei Zhang, and Hai Zhao. 2024. A novel energy based model mechanism for multi-modal aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18869–18878.

[28] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Preslav Nakov and Torsten Zesch (Eds.). Association for Computational Linguistics, Dublin, Ireland, 27–35. doi:10.3115/v1/S14-2004

[29] Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting Structured Knowledge in Text via Graph-Guided Representation Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8980–8994. doi:10.18653/v1/2020.emnlp-main.722

[30] Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 491–500.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786 [cs.CV] https://arxiv.org/abs/2502.14786

[33] Xinfeng Wang, Jin Cui, Fumiyo Fukumoto, and Yoshimi Suzuki. 2024. Enhancing high-order interaction awareness in llm-based recommender model. *arXiv preprint arXiv:2409.19979* (2024).

[34] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

[35] Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. 2025. A Dual Contrastive Learning Framework for Enhanced Multimodal Conversational Emotion Recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven

Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 4055–4065. https://aclanthology.org/2025.coling-main.272/

[36] Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A Unified Generative Framework for Aspect-based Sentiment Analysis. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2416–2429. doi:10.18653/v1/2021.acl-long.188

[37] Zehong Yan, Wynne Hsu, Mong-Li Lee, and David Bartram-Shaw. 2024. Modeling Complex Interactions in Long Documents for Aspect-Based Sentiment Analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 23–34. doi:10.18653/v1/2024.wassa-1.3

[38] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[39] Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. IJCAI.

[40] Jiawei Zhang. 2023. Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt. *arXiv preprint arXiv:2304.11116* (2023).

[41] Jiasheng Zhang, Jialin Chen, Ali Maatouk, Ngoc Bui, Qianqian Xie, Leandros Tassiulas, Jie Shao, Hua Xu, and Rex Ying. 2024. LitFM: A Retrieval Augmented Structure-aware Foundation Model For Citation Graphs. *arXiv preprint arXiv:2409.12177* (2024).

[42] Ru Zhou, Wenya Guo, Xumeng Liu, Shenglong Yu, Ying Zhang, and Xiaojie Yuan. 2023. AoM: Detecting Aspect-oriented Information for Multimodal Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 8184–8196. doi:10.18653/v1/2023.findings-acl.519