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

Xinjing Liu, Pengyue Lin, Xinyu Tu,
Wenqi Jia, Chen Jiang, Ruifan Li*

Beijing University of Posts and Telecommunications

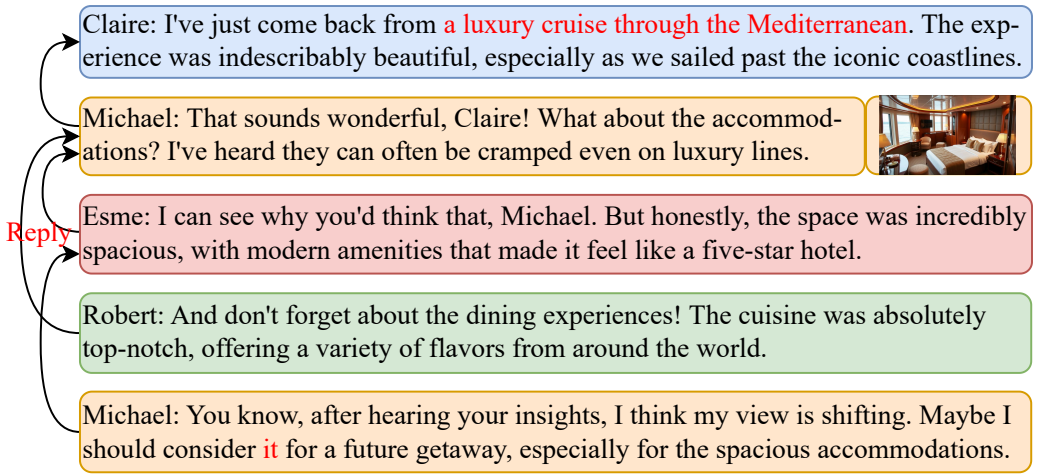
October 30, 2025

Outline

Outline

MCABSA Task

Multimodal Conversational Aspect-based Sentiment Analysis (MCABSA) includes two subtasks.



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.

Related Works

The issues with previous work regarding MCABSA are as follows:

1. Error Propagation Problem
   - Existing methods often treat the entire dialogue as a flat sequence and feed it into LLMs for pipeline-style generation. This method easily accumulates errors.
2. Ignoring structural information
   - Failing to fully leverage the rich structural information embedded in dialogues, making it difficult for the model to capture subtle sentiment cues, shifts in speaker intent, and long-distance semantic dependencies, thereby impairing the effectiveness of fine-grained sentiment reasoning.

Contributions

We propose SDG-MLLM, a unified generative end-to-end framework.

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.

Outline

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.
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.

Model Overview

The model consists of four modules: 1) Multimodal Feature Extraction, 2) Heterogeneous Dialogue Graph Construction, 3) Heterogeneous Dialogue Graph Encoder, and 4) Graph-Augmented MLLM.

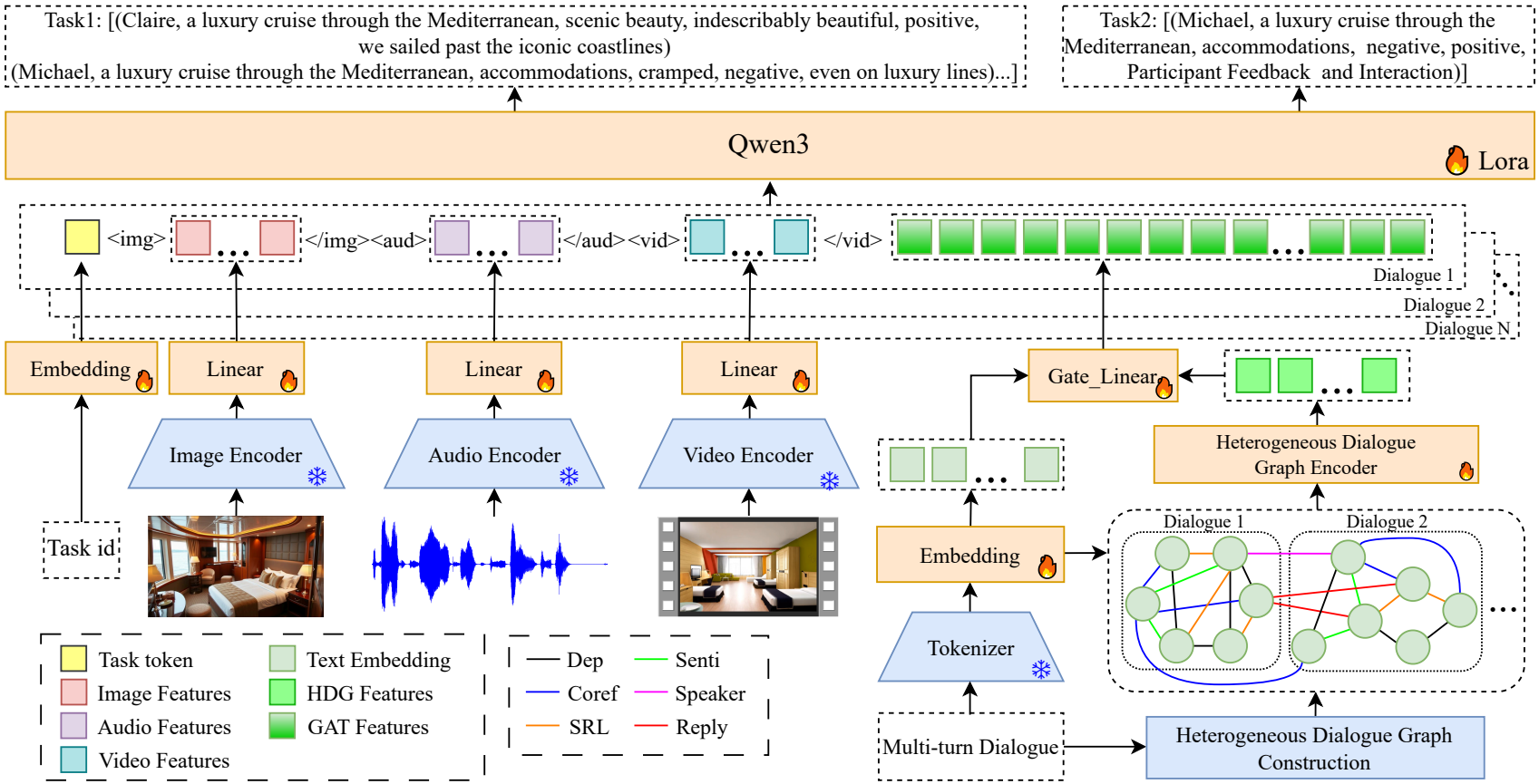


Figure 2: An overview of our SDG-MLM architecture.

Multimodal Feature Extraction

1) Text Features. We employ the pretrained Qwen3-8B 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.

2) Image Features. We use the SigLIP2 vision encoder to extract patch-level visual features, yielding $\mathbf{E}_{\text{img},i} \in \mathbb{R}^{576 \times 1536}$.

3) Audio Features. We use WavLM's encoder to extract audio features, yielding $\mathbf{E}_{\text{aud},i} \in \mathbb{R}^{273 \times 768}$.

4) Video Features. We uniformly sample frames and use SigLIP2 to extract frame-level visual embeddings. These are then averaged across frames to form video features $\mathbf{E}_{\text{vid},i} \in \mathbb{R}^{576 \times 1536}$.

Heterogeneous Dialogue Graph Construction

Each heterogeneous dialogue graph explicitly models multiple types of relational edges between tokens across the dialogue. **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.

Edge Types. We include six types of edges in the dialogue graph.

1) Dependency Edges. We apply spaCy to extract grammatical relations (e.g., nsubj, dobj, and amod) within each utterance.

2) Coreference Edges. We utilize AllenNLP's coreference resolution module to identify coreferent mentions within and across utterances, linking tokens that refer to the same underlying entity.

Heterogeneous Dialogue Graph Construction

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.

5) Sentiment Propagation Edges. Based on a sentiment lexicon, we connect sentiment-laden opinion words to candidate target and aspects within or across utterances.

6) Semantic Role Labeling (SRL) Edges. Using AllenNLP's SRL parser, we extract predicate-argument structures and add edges between verbs and their arguments.

Heterogeneous Dialogue Graph Encoder

We employ the Heterogeneous Graph Transformer (HGT) which models each relation with distinct transformation and attention mechanisms. We compute relation-specific attention scores between a node p and its neighbor $q \in \mathcal{N}_p^{(r)}$ as follows,

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

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,

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

The output features after the final layer $\mathbf{h}_{ip}^{(L)}$, form the Heterogeneous Dialogue Graph (HDG) features $\mathbf{E}_{\mathrm{HDG},i} = \{\mathbf{h}_{i1}^{(L)}, \ldots, \mathbf{h}_{ip}^{(L)}, \ldots, \mathbf{h}_{im}^{(L)}\}$.

Graph-Augmented MLLM

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.

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

And they are inserted before the corresponding utterance text forming a modality block $\mathcal{M}_{m,i} = [\langle m \rangle, \tilde{\mathbf{E}}_{m,i}, \langle /m \rangle]$.

For the text-enhanced features of u_i. 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,

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

Graph-Augmented MLLM

We prepend a special task token $\langle \text{task1} \rangle$ or $\langle \text{task2} \rangle$ at the beginning of the input sequence to explicitly indicate which subtask.
Let $\hat{u}_i$ denote the enriched representation of utterance u_i, defined as $\hat{u}_i = [\mathcal{M}_{img,i}; \mathcal{M}_{aud,i}; \mathcal{M}_{vid,i}; \mathbf{E}_{\mathrm{GAT,i}}]$.
The full input sequence is then constructed by concatenating all enriched utterances in natural dialogue order,

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

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.,

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

We apply LoRA-based fine-tuning to the Qwen3.

Outline

Experiments Settings

- Datasets.
  - We evaluate our method on the PanoSent dataset (see Table 1).
- Baselines.
  - We compare our model against several strong baselines across both subtasks. 1) UGF 2) DiaASQ 3) Unified-IO2 4) NExT-GPT 5) Sentica
- Implementation Details.
  - We train model on four A6000 GPUs with AdamW (batch size 8, learning rate 5×10^{-5}) for 3 epochs.

	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.

Main Results

The results are reported in Table 2, 3, 4.

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.

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.

Case Study

As shown in Figure 3, our model demonstrates a superior ability involving coreference, syntactic dependency, and reply relations.

Input	**Valerie**:The **Orient Express** truly redefines travel as luxury. The cabin interiors are stunning, when I stepped inside, the polished wood made it feel like a palace on rails.  **Raj**: Absolutely, but what amazed me most was the cuisine. The quality of each course felt exceptional because chefs source ingredient fresh at every stop.(reply Valerie) **Isabelle**:I was less impressed by the route though;**it** by passess a lot of the smaller towns I hoped to see, which made the itinerary slightly disappointing.(reply Valerie) **Theo**:But don't you think the onboard service is impeccable? The staff anticipate your needs, they made me feel noticeably pampered the entire journey. (reply Raj)	**Ella**:Have you all seen the latest blockbuster? The storytelling is phenomenal; it captivated me from start to finish.  **Sophie**:I felt the pacing was a bit rushed at times, particularly during key plot moments that could have used more depth.(replyMike) **Mike**:I agree the storytelling is engaging.However,I thought the characters were quite underdeveloped which is disappointing. (reply Ella) **Ella**: While I initially found the characters compelling,I think I understand you both mean. Their depth could have been improved which elevate the entire film experience.(reply Mike)
Ground Truth	(Valerie, Orient Express, cabin interiors, stunning, positive,...) (Raj, Orient Express, cuisine, exceptional, positive, ...) (Isabelle, Orient Express, route, slightly disappointing, negative,..) (Theo, Orient Express, onboard service, impeccable, positive...)	(Ella, latest blockbuster, characters, Initial Sentiment: positive, Flipped Sentiment: negative, Participant Feedback and Interaction)
SDG-MLLM	(Valerie, Orient Express, cabin interiors, stunning, positive, ...) (Raj, Orient Express, cuisine, exceptional, positive, ...) (Isabelle, Orient Express, route, slightly disappointing, negative,..) (Theo, Orient Express, onboard service, impeccable, positive...)	(Ella, latest blockbuster, characters, Initial Sentiment: positive, Flipped Sentiment: negative, Participant Feedback and Interaction)
SDG-MLLM (w/o Graph)	(Valerie, Orient Express, cabin interiors, stunning, positive, ...) (Raj, **the cuisine**, **quality**, exceptional, positive, ...) (Isabelle, **the route**, route, slightly disappointing, negative, ...) (Theo, **the onboard**, **service**, impeccable, positive ...)	(Ella, latest blockbuster, characters, Initial Sentiment: positive, **Flipped Sentiment: neutral,** **Introduction of new information**)

Figure 3: The left case is Subtask-I result, while the right case is Subtask-II.

Outline

Conclusion & Future Work

- Conclusion
  - 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.
  - Experiments on the MCABSA benchmark demonstrate that SDG-MLLM consistently outperforms strong baselines.
- Future Work
  - We plan to explore dynamic graph construction to better model evolving sentiment states over time, and to extend SDG-MLLM to other dialogue scenarios.

Q&A

Thank You for Your Attention.