Temporal Prompting with Vision-Language Models for Consistent Video Scene Graph Generation

Dohyeong Lee*†, Choulsoo Jang*, Sang-Kyu Lim*, Chang-Eun Lee*†
*Defense & Safety Convergence Research Division,
Electronics and Telecommunications Research Institute, Daejeon 34129, Korea
†Department of Artificial Intelligence,
University of Science and Technology, Daejeon 34113, Korea
e-mail: {dohlee, jangcs, sklim, celee}@etri.re.kr

***Abstract*—Video scene graphs capture objects, represented as nodes and their temporal relationships, providing a structured representation for video understanding. However, constructing large-scale annotated datasets is challenging in specialized domains where real-world data collection is infeasible, such as the military domain. In this paper, we propose a temporal prompting framework for vision-language models to generate consistent video scene graphs. Specifically, our framework ensures the temporal consistency of objects and their relationships, preserving the meaning of each frame and the overall temporal context. Our temporal prompting strategy leverages previous-frame tracking, object states, and relational context to enable consistent relational annotations. The framework comprises two steps: detection and alignment, and relation estimation and alignment. We evaluate the framework using two metrics, frame-level recall and continuity recall, to assess the consistency of relationships in video scene graphs. Experiments on a public video dataset demonstrate that our temporal prompting approach yields more continuous relationship predictions and achieves better overall performance compared to baseline prompting. These results highlight the framework's potential for producing training-ready video scene graph datasets, particularly in domains with scarce annotated data.**

***Index Terms*—Video Scene Graph, Scene Graph Generation, Vision-Language Model, Prompt Engineering**

I. Introduction

Video scene graphs extend traditional scene graphs by representing not only object relationships but also temporal and action-based interactions [1]. While scene graphs are limited to single-frame understanding, video scene graphs incorporate spatiotemporal information, enabling richer comprehension of the entire scene. This spatiotemporal information is maintained across frames by assigning unique identifiers to objects using tracking algorithms [2], ensuring the continuity of object information over time.

Training data for video scene graphs typically include both detected objects and their tracking information, along with relationships defined across frame segments [3]. An example of inter-frame relational information from the *VidOR* dataset [4] is shown in Fig. 1. Since object actions and movements are relatively time-invariant in consecutive frames, their relationships should also remain time-consistent. However, single-image-based relationship extraction using vision–language models introduces inconsistencies: semantically identical relationships can be expressed differently across consecutive frames [5].

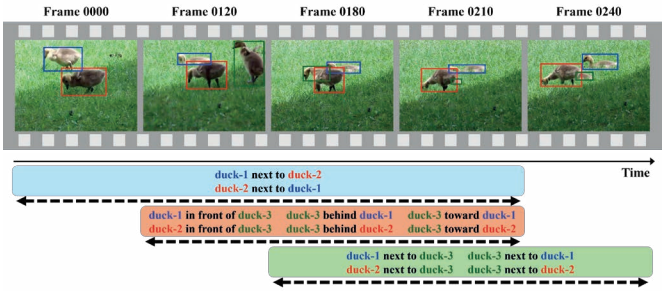

Fig. 1. Example of inter-frame relational information of objects in a video scene graph.

For example, the spatial relationship between two objects may be labeled as "near" in one frame and "next to" in the next, reducing label consistency and increasing noise.

To address this issue, we propose a novel temporal prompting framework that operates in two main steps. First, our framework performs detection and alignment to extract object and tracking information from video frames. This allows us to use an open-vocabulary detection algorithm while maintaining object continuity [6]. Second, our framework performs relation estimation and alignment. By leveraging both current and previous frame information, the framework explicitly incorporates temporal continuity during relation inference. As a result, our approach can effectively maintain the consistency of relationship representations across frames.

The constructed temporal prompts were compared against a baseline approach that does not utilize information from previous frames. The performance was evaluated on a public video dataset using both frame-level recall and continuity recall to assess the accuracy and temporal consistency of the predicted relationships. Our experimental results demonstrate that the temporal prompting approach significantly outperforms the baseline, highlighting its potential for generating high-quality training data for video scene graph tasks.

II. Related Work

A. Scene Graph Generation

Scene Graph Generation (SGG) aims to represent an image as a structured graph, where nodes correspond to objects and edges denote semantic relationships [7]. Early approaches

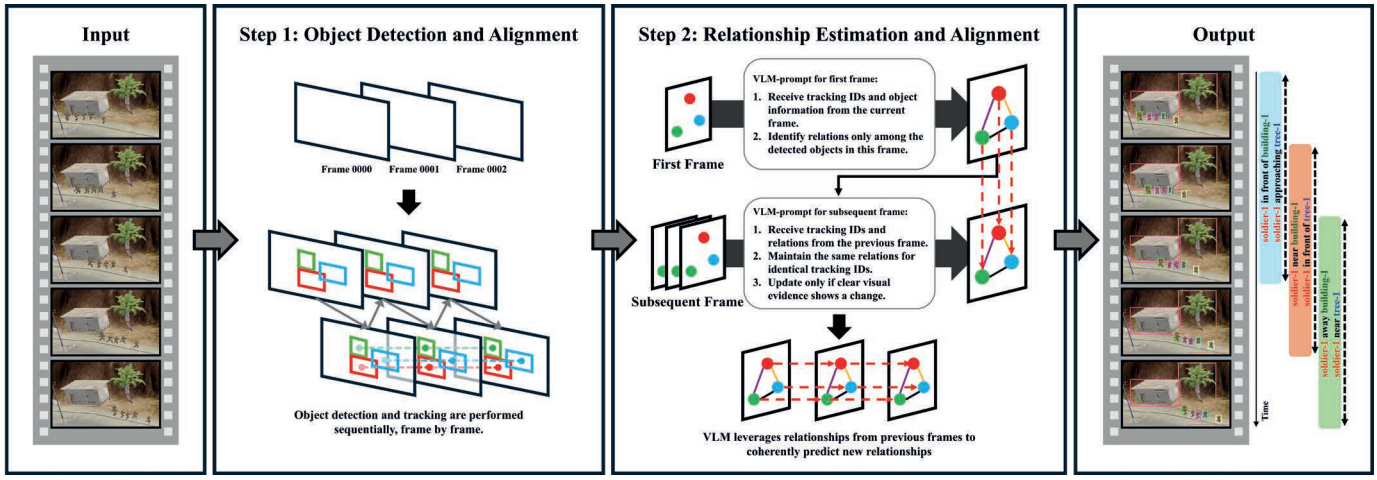

Fig. 2. Overview of our proposed framework.

relied primarily on supervised learning with large-scale annotated datasets such as Visual Genome [8]. While subsequent studies focused on improving relationship prediction accuracy and mitigating long-tail distributions, these methods are fundamentally limited to single-frame understanding and do not account for temporal dynamics essential for video analysis.

B. Video Scene Graph Generation

Extending SGG to videos introduces additional challenges, including temporal consistency and dynamic interactions. Video Scene Graph Generation (VidSGG) methods aim to capture evolving object states and relationships across frames. Recent works address temporal coherence through trajectory-level reasoning, spatio-temporal graph modeling, and leveraging motion cues to enhance relation prediction. However, many approaches require large-scale annotated video datasets or specialized architectures [9], whereas our framework leverages vision-language model (VLM) prompting to achieve a temporally consistent relation inference in a data-efficient manner.

C. Open-Vocabulary Object Detection

Traditional object detection models are constrained to fixed categories defined by training datasets. Open-Vocabulary Object Detection (OVOD) leverages vision-language pretraining (e.g., *CLIP*, *OWL-ViT*, *Grounding DINO*) to recognize unseen categories specified via text prompts [10]–[12]. This paradigm enables zero-shot or few-shot recognition, facilitating scalable scene graph construction without reliance on closed-set detectors. In our framework, OVOD allows flexible detection of objects without additional model retraining, thereby reducing computational overhead.

D. Object Tracking

Object tracking ensures consistent identification of objects across video frames, which is crucial for maintaining coherent video scene graphs. Classical methods include correlation filters and Siamese-based trackers, while modern approaches integrate deep learning with appearance and motion cues. Trackers such as *DeepSORT* remain widely used due to their simplicity, stability, and ability to assign consistent object IDs across frames [13], supporting reliable relation inference over time.

E. Vision-Language Models and Temporal Prompting for Relation Consistency

Vision-language models have demonstrated effectiveness in refining scene understanding by aligning visual content with textual descriptions. Prompt engineering techniques further improve zero-shot generalization and facilitate explainable reasoning [14]. However, existing prompting methods often treat frames independently and fail to explicitly enforce temporal consistency of relationships. Our work introduces a temporal prompting strategy that incorporates previous-frame information into VLM prompts, ensuring consistent relational predictions across video sequences and addressing this gap.

III. Proposed Method

The proposed temporal prompting framework consists of two main steps: *detection and alignment*, and *relation estimation and alignment*. An overview of the framework is illustrated in Fig. 2. The following sections describe each step in detail.

A. Step 1: Detection and Alignment

The first step aims to extract object information and maintain temporal alignment across consecutive frames. For object detection, we adopt an open-vocabulary approach, allowing recognition of a wide range of object classes without additional training. A predefined set of object classes is provided as input, enabling the extraction of class labels and bounding boxes for each frame. We employ *Grounding DINO* and *OWL-ViT* for detection and find that *Grounding DINO* exhibits better compatibility with the tracking model.

To ensure temporal continuity, each detected object is assigned a unique tracking ID using the *DeepSORT* tracker. Both the object detection outputs and tracking information,

including object trajectories and unique IDs, are temporarily stored. This temporary storage enables efficient retrieval of previous frame information for subsequent relation inference in Step 2. While *DeepSORT* is used in this work, more advanced tracking algorithms could be employed in future extensions to further enhance temporal alignment.

B. Step 2: Relation Estimation and Alignment

In the second step, the temporarily stored object and tracking information from Step 1 are utilized to construct prompts for the VLM, specifically *Gemma-3* [15]. By incorporating both current frame information and previously stored data, the framework explicitly accounts for temporal continuity during relation inference.

The temporal prompt construction is designed as follows:

- **First-frame prompt**: Since no prior frame exists, the VLM infers relationships solely based on the current frame's object information.
- **Subsequent-frame prompt**: For all subsequent frames, the VLM receives tracked object IDs, bounding boxes, and previously inferred relationships from the temporary storage. This enables the model to reason about relationships in a temporally coherent manner, maintaining consistency while adapting to dynamic changes in the scene.

Through this two-step process, the framework ensures that relationship representations remain coherent across frames, effectively preserving temporal consistency while capturing dynamic interactions in video sequences.

IV. Experiments and Results

To evaluate the performance of our proposed framework, we conducted experiments comparing two prompting strategies under a consistent configuration to ensure reproducibility.

A. Experimental Setup

We evaluated our framework using a subset of the public *VidOR* dataset, which provides ground truth (GT) annotations for relational descriptions essential for evaluation. Video frames were extracted according to the original FPS of each video. The following configuration was applied:

- **Object Detection and Tracking**: *Grounding DINO* was used for open-vocabulary object detection, and *DeepSORT* was adopted to assign consistent tracking IDs across frames.
- **Relation Inference**: *Gemma-3-4B VLM* was employed to extract relational descriptions.

B. Experimental Design

We compared two primary prompting strategies to investigate the effect of temporal consistency on relation prediction accuracy:

1) **Baseline Prompt (Frame-Only)**: Relations were inferred using only the current frame's object and tracking information, without leveraging any previous frame data. This approach is analogous to conventional single-image-based methods.
2) **Temporal Prompt (with History)**: Relations were inferred by integrating previously predicted relation information along with the current frame's object and tracking IDs. This strategy allows the model to capture changes in relationships over time while maintaining temporal continuity.

Assuming reasonably accurate object detection and tracking, our evaluation focuses on the quality of the relations predicted by the VLM. To measure the effect of temporal consistency explicitly, we employed *Continuity Recall*, in addition to the standard frame-level *Recall*. *Continuity Recall* quantitatively assesses how consistently the same object relationships are predicted across consecutive frames, providing a direct measure of the benefit of temporal prompting.

C. Evaluation Metrics

To evaluate the performance of our video-based scene graph generation framework, we adopt recall-based metrics that quantify the accuracy of predicted relations against the ground truth (GT). Recall is calculated as:

$$\text{Recall} = \frac{\text{Number of True Positives (TP)}}{\text{Number of True Positives (TP)} + \text{Number of False Negatives (FN)}}. \quad (1)$$

Note that conventional video scene graph models often use *Recall@k*, which relies on model confidence scores to rank predictions. However, our VLM-based approach does not output confidence values for predicted relations, making *Recall@k* inapplicable. Instead, we report frame-level predicted recall, which provides a similar measure of prediction accuracy per frame without requiring confidence scores.

1) Frame-level Recall: Frame-level recall measures prediction accuracy on each video frame.

a) Micro Frame-level Recall: aggregates all frames:

$$\text{Frame-level Recall}_{\text{micro}} = \frac{\sum_{t \in \mathbf{T}} |GT_t \cap Pred_t|}{\sum_{t \in \mathbf{T}} |GT_t|}. \quad (2)$$

b) Macro Frame-level Recall: averages per-frame recall:

$$\text{Frame-level Recall}_{\text{macro}} = \frac{1}{|\mathbf{T}|} \sum_{t \in \mathbf{T}} \frac{|GT_t \cap Pred_t|}{|GT_t|}. \quad (3)$$

Here, $\mathbf{T}$ denotes the set of all frames in the video. Micro recall reflects the overall proportion of correctly predicted relations, while macro recall captures the average performance per frame. In these definitions, the intersection $GT_t \cap Pred_t$ represents the set of relations in frame t that are correctly predicted by the model, i.e., relations that appear both in the ground truth annotations and in the model's predictions.

2) Continuity Recall: Continuity Recall evaluates how consistently the predicted relations are maintained across consecutive frames. This metric allows comparison of the model's performance when previous frame information is incorporated versus when it is not, highlighting the effect of temporal consistency rules.

a) Micro Continuity Recall: considers all relations across frames:

$$\text{Continuity Recall}_{\text{micro}} = \frac{\sum_{r \in \mathbf{R}} \text{Number of consecutive frames correctly predicted}}{\sum_{r \in \mathbf{R}} \text{Number of GT frames}}. \tag{4}$$

b) Macro Continuity Recall: averages the consistency per relation instance:

$$\text{Continuity Recall}_{\text{macro}} = \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \frac{\text{Number of consecutive frames correctly predicted}}{\text{Number of GT frames for } r}. \tag{5}$$

Here, $\mathbf{R}$ denotes the set of all relation instances in the video. Micro continuity recall reflects the total proportion of correctly maintained relations, while macro continuity recall captures the average consistency per relation instance.

D. Results and Discussion

Table I presents the experimental results. Frame-level recall indicates that the temporal prompting strategy slightly outperforms the baseline (frame-only) prompting approach. More notably, continuity recall shows a substantial improvement, demonstrating that incorporating information from previous frames significantly enhances temporal consistency.

TABLE I
COMPARISON OF BASELINE (FRAME-ONLY) AND TEMPORAL PROMPTING STRATEGIES IN TERMS OF FRAME-LEVEL AND CONTINUITY RECALL.

Metric	Baseline	Temporal
Frame-level Recall (Micro)	0.1690	0.1960
Frame-level Recall (Macro)	0.1825	0.2088
Continuity Recall (Micro)	0.0704	0.1595
Continuity Recall (Macro)	0.0216	0.1418

Although improvements in frame-level recall are modest, continuity recall demonstrates the advantage of temporal prompting in maintaining consistent relational predictions. As shown in Figure 3, the VLM often infers relations that, while ignoring exact object positions, provide richer and more semantically meaningful descriptions of the scene, indicating that the proposed framework captures relational information beyond simple spatial arrangements.

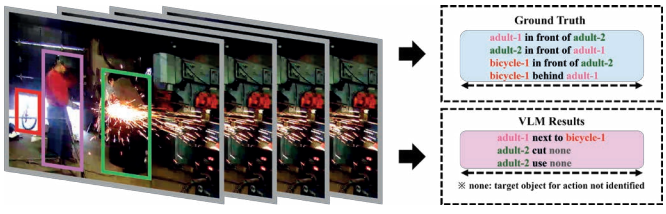

Fig. 3. Comparison of scene graphs generated by the VLM-based approach and the ground truth over a specific segment of the evaluation dataset.

V. CONCLUSION

In this study, we propose a framework for video scene graph generation using temporal prompting with vision-language models. While improvements in frame-level recall are modest, our results show a clear advantage in maintaining consistent relational predictions, as evidenced by significant gains in continuity recall. Moreover, the VLM demonstrates the ability to infer relations that are semantically richer and more interpretable, highlighting its potential to provide valuable relational information for video scene graph generation. Although our experiments were conducted on public video datasets, we also confirmed the framework's applicability on simulation-generated videos, indicating a promising direction for future work. Leveraging larger-scale simulation datasets could not only expand training resources but also support the stable training of larger models, paving the way for broader real-world applications.

ACKNOWLEDGMENT

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) under the funding from the Korean government (Defense Acquisition Program Administration) in 2023 (KRIT-CT-23-021).

REFERENCES

[1] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation " in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1300–1308.

[2] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial-temporal transformer for dynamic scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16372–16382.

[3] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10236–10247.

[4] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua, "Annotating objects and relations in user-generated videos," in *Proc. 2019 Int. Conf. Multimedia Retr.*, 2019, pp. 279–287.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[6] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14393–14402.

[7] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, 2021.

[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[9] Y. Teng, L. Wang, Z. Li, and G. Wu, "Target adaptive context aggregation for video scene graph generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13688–13697.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[11] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, *et al.*, "Simple open-vocabulary object detection," in *Euro. Conf. Comput. Vis.*, 2022, pp. 728–755.

[12] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Euro. Conf. Comput. Vis.*, 2024, pp. 38–55.

[13] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 3645–3649.

[14] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[15] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.