A Closer Look at Dynamic Scene Graph Generation In the Era of Multimodal Large Language Models

Xuanming Cui[1] **Jaiminkumar Ashokbhai Bhoi**[1] **Chionh Wei Peng**[2] **Adriel Kuek**[2] **Ser Nam Lim**[1]

[1]University of Central Florida [2]DSO National Laboratories
{xuanming.cui, ja882177, sernam}@ucf.edu
{weipeng, kyongjie}@dso.org.sg

Abstract

Dynamic Scene Graph Generation (DSGG) aims to capture objects and their evolving relations in videos. Despite recent progress, the practicality and quality of generated scene graphs remain limited compared to the rapid advances in Multimodal Large Language Models (MLLMs). In this work, we revisit DSGG from two fundamental perspectives: task setup and model design. From the task setup perspective, we identify two key limitations of the current recall-oriented evaluation protocol: (i) a severe precision–recall trade-off, and (ii) uninformative and redundant relation generation. To better assess practical usefulness, we introduce five additional metrics that provide a more comprehensive evaluation of the quality of generated dynamic scene graphs. From the model design perspective, we explore directly using MLLMs for scene graph generation and establish a strong MLLM-based DSGG baseline through three design changes. First, we replace the conventional bottom-up pipeline with a top-down *reason-then-locate* strategy. Second, we reformulate frame-wise dynamic graphs as Temporal Relation Set (TRS) prediction, improving both efficiency and performance. Third, we introduce Importance-Aware Finetuning (IAF) to encourage more relevant and diverse relation generation. Extensive experiments on Action Genome, VidVRD, and PVSG show that our approach consistently achieves state-of-the-art performance.

1 Introduction

Understanding complex visual scenes requires not only recognizing objects but also reasoning about the relationships between them. Scene Graph Generation (SGG) addresses this problem by representing visual content as a structured graph where nodes correspond to objects and edges describe their relations. Such structured representations provide an interpretable intermediate representation and have been shown to benefit a range of downstream tasks, including image retrieval [4, 53, 46, 29], visual question answering (VQA) [4, 53, 46, 29], and controllable image generation [11]. On top of static image-based SGG, Dynamic Scene Graph Generation (DSGG) [5, 49, 63] aims to capture both objects and their evolving relationships over time, enabling structured understanding of dynamic scenes through temporal dimension.

Despite steady progress in DSGG research, the practical quality of generated scene graphs remains limited, compared to the wide adaptation of Multimodal Large Language Models (MLLMs). Recent advances in MLLMs have dramatically improved visual understanding capabilities, demonstrating strong performance on a wide range of practical visual tasks such as video captioning, reasoning [1, 15], and retrieval [8]. This contrast raises a natural question:

Why does DSGG lag behind the broader progress in multimodal video understanding?

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

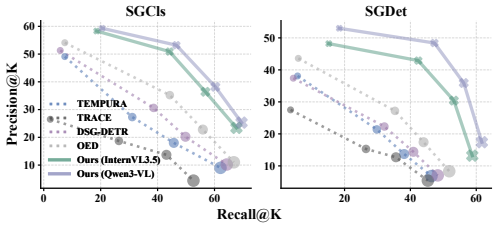


Figure 1: Precision-Recall [1]curve for previous DSGG [5, 49, 63] and our MLLM-based methods for two DSGG tasks (SGCls and SGDet) on Action Genome [23]. We plot top-k with $k \in \{1, 10, 20, 50\}$, indicated by the size of the markers. We observe severe precision-recall trade-off on existing methods such as OED [63], as opposed to our MLLM-based methods, whose precision are notably better.

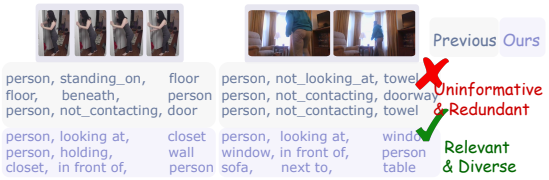


Figure 2: Visualization of top-3 prediction comparison with previous DSGG method [63] and our MLLM-based method. Two sample video clips are from Action Genome [23]. While previous DSGG method predicts correct but uninformative and redundant triplets (e.g. $\langle \texttt{person}, \texttt{standing_on}, \texttt{floor} \rangle$ and $\langle \texttt{floor}, \texttt{beneath}, \texttt{person} \rangle$), ours predicts $\langle \texttt{person}, \texttt{looking_at}, \texttt{closet} \rangle$, which is the primary activity of the scene, while covering more semantics of the scene.

In this work, we re-examine the task of DSGG to better understand this discrepancy from two fundamental perspectives: *1) task setup*, and *2) model design*.

Rethinking DSGG Task Setup. Existing DSGG methods are evaluated exclusively using recall-based metrics. We argue that recall alone provides an incomplete view of scene graph quality.

Precision-Recall Trade-off. Recall-only evaluation ignores the precision–recall trade-off. Existing DSGG models can artificially inflate recall by predicting a large number of relations at the expense of prediction precision. This behavior frequently results in scene graphs that contain numerous incorrect and noisy relations. While such trade-off is prevalent in regular retrieval systems, it brings substantially more impact to DSGG, as the goal of DSGG is not retrieval, but to provide accurate spatial-temporal information for downstream applications such as video QA and robotic planning [22]. In such scenarios, severely damaging the performance of downstream tasks (Figure 1 and 9).

uninformative and redundant relation generation. Existing literature typically views DSGG as an unordered retrieval problem, where the relative importance of predicted relations and the diversity of the generated scene graphs are ignored. However, not all relations contribute equally to understanding a scene. Semantically meaningful interactions (e.g., $\langle \texttt{person}, \texttt{riding}, \texttt{horse} \rangle$) are often far more informative than trivial spatial relations (e.g., $\langle \texttt{floor}, \texttt{beneath}, \texttt{person} \rangle$). Moreover, as uninformative relations (e.g. `not looking at`) are more prevalent in training data [23], traditional DSGG methods tend to produce higher probability for such relations, resulting in uninformative relations dominating top-ranked predictions (Figure 2). While this may seem relevant to the well-studied long-tail problem in DSGG, we note that whether a relation is uninformative or redundant depends on the video and generated relations. A long-tail relation (*e.g.* ‘‘`have it on the back`’’) can be uninformative, whereas a common relation (*e.g.* ‘‘`in`’’) may still be informative.

To better assess the practical utility of generated scene graphs, we introduce five complementary metrics. In addition to precision, we introduce **Importance Aware Ranking**, which treat DSGG as a ranking task to consider triplet importance and diversity. We further incorporate two downstream tasks as evaluation metrics: **Scene Graph Retrieval** (SGRet), which measures the effectiveness of scene graphs via video retrieval, and **Scene Graph Question Answering** (SGQA), a curated Multi Choice Question (MCQ) dataset designed to evaluate reasoning over generated scene graphs. Lastly, we propose **SGCls**$^{+}$, a modified DSGG task setting, which provides more realistic assessment of relation predictions. Together, these metrics provide a more comprehensive assessment of DSGG quality beyond the widely-used recall-based metrics [2].

Rethinking DSGG Model Design. Most prior DSGG approaches follow a bottom-up pipeline inherited from image-based SGG: objects are first detected, the regional features are subsequently passed through specialized spatial–temporal aggregation modules to infer object and relation types. While effective to some extent, this design introduces several challenges.

[1]While MLLM-based DSGG's precision also seems to degrade as recall increases, this is due to how evaluation is performed: we fill MLLM's predictions with random triplets when the number of output is less than k. In reality, due to the autoregressive nature, MLLMs *"knows"* to stop generation when it does not find further plausible predictions.

[2]Note that while the above discussion could also apply to image-based SGG, we focus on the video modality as it is substantially more complex than image-based SGG, and has wider downstream applications.

(1) Bottom-up pipelines are inherently bottlenecked by the object detector, as relation prediction relies entirely on detected regions. Errors or missing detections can propagate through the pipeline and degrade final scene graph quality. **(2)** Existing DSGG models [5, 61, 12, 63] typically rely on hand-crafted spatial–temporal modules, leading to increasingly complex architectures that are difficult to generalize across datasets or tasks. **(3)** Detector-based pipelines typically operate on cropped object regions, which may discard important contextual information needed to understand interactions in a scene. **(4)** Existing DSGG models are mostly trained from scratch for each individual task, limiting their ability to leverage broader knowledge from large-scale multimodal data.

In contrast, modern MLLMs adopt a fundamentally different paradigm for visual reasoning. Rather than building visual understanding from localized detections, MLLMs perform holistic reasoning over the entire visual input, leveraging large-scale pretraining and multimodal knowledge. This top-down process better resembles how humans interpret visual scenes, where interactions and activities are often understood before individual components are explicitly localized. Such capabilities suggest that MLLMs may provide a promising alternative foundation for scene graph generation.

Motivated by the above observation, we provide the first comprehensive study of adopting MLLMs for direct video scene graph generation. We first reveal that while SOTA MLLMs excel at single prediction verification (Table 1), they do not naturally work for dense visual generation (Figure 4). Subsequently, we introduce three novel design changes that adapt MLLMs to the DSGG task. **(1)** We depart from the traditional bottom-up pipeline and adopt a top-down *reason-then-locate* strategy, where interactions are first inferred semantically before being grounded in objects. **(2)** We reformulate frame-wise DSGG as Temporal Relation Set (TRS) prediction, which improves both efficiency and temporal coherence for MLLM-based DSGG. **(3)** We propose Importance-Aware Finetuning (IAF) to guide the model toward generating ranked relations that are both relevant and diverse.

Our contributions are summarized as follows:

1) Rethinking DSGG Task Setup. We identify key limitations of the current recall-based evaluation protocol and introduce five complementary metrics that better reflect the practical quality of generated scene graphs. In addition, we curate a high-quality MCQ dataset to effectively evaluate SGQA.

2) Rethinking DSGG model design. We are the first to formalize a *Reason-then-Locate* paradigm for DSGG. To effectively adapt MLLM as scene graph generators, we propose TRS and IAF to improve the performance, efficiency, and quality of the predicted scene graphs.

3) A strong baseline for MLLM-based DSGG. We provide the first systematic study of adopting MLLMs for direct DSGG. Through extensive evaluation across multiple datasets, tasks, and metrics, our MLLM-based approach demonstrates significant improvements over existing SOTAs.

2 Related Work

Scene Graph Generation. Scene Graph Generation (SGG) provides a graph-based representation of the visual content. Traditional approaches typically exploit the graph property of the scene graph via Message Passing [67, 37, 70, 73] and Graph Convolutional Network [41]. Another line of works focuses on handling the long-tailed problem due to coarse labeling [60, 58, 76]. Given to the advances in Vision Language Models [54, 32] and Multimodal Large Language Models [43, 1], several recent works utilize pre-trained multimodal foundational models for open-vocabulary SGG [74, 35]. For instance, VS^3 utilizes the pre-trained GLIP [33] for region-word alignment, and [35] treats SGG as an image-to-sequence task by finetuning BLIP [32] for entity/relation prediction and localization.

Dynamic Scene Graph Generation. Dynamic Scene Graph Generation (DSGG) extends SGG with the additional temporal dimension. Existing methods in DSGG typically resemble similar bottom-up architectural designs in traditional SGG models. For instance, STTran [5] adds an attention module as the temporal decoder on top of the spatial encoder. OED [63] follows the similar architectural design but replace Faster-RCNN with DETR [3] for one-stage DSGG. A few other works [49, 68, 27] focus on the long-tail problem in DSGG. On the other hand, several recent works [21, 48, 51, 22] explore utilizing MLLMs to enhance DSGG performance. These works typically exploit MLLM as synthetic data generator for weakly-supervised DSGG [21, 22], as a post-processing tool [48], or as a downstream application module [51]. For instance, HyperGLM [51] generates video scene graph with traditional bottom-up approach, and pass the generated scene graph to text-only LLM for

reasoning; LASER [21] and ESCA [22] utilizes MLLM for synthetic data generation. However, none of these previous works explore using MLLM directly as video scene graph generator.

Multimodal Large Language Models. Following LLaVA [43]'s success, a large amount of visual instruction-tuned Large Multimodal Models (MLLMs) have been developed [9, 77, 42, 31]. These models are typically trained for captioning, standard VQA, object grounding, and complex visual reasoning [1, 66]. Despite the wide range of evaluation suites [19, 13, 62], there is a lack of assessment of MLLMs' capacity on fine-grained dense generation for video tasks such as DSGG.

3 Method

3.1 Problem Formulation

Dynamic Scene Graph Generation (DSGG) aims to represent objects in a video and their interactions over time as a sequence of structured graphs. Given a video $V = \{I_t\}_{t=1}^{T}$ consisting of T frames, the goal of DSGG is to predict a *dynamic scene graph* $G = \{G_t\}_{t=1}^{T}$, where each frame-level graph contains a set of scene graph triplet $\langle \texttt{sub}, \texttt{rel}, \texttt{obj} \rangle$, where `sub`, `rel` and `obj` refer to subject, relation and object respectively. Compared to image scene graph generation, DSGG additionally requires modeling the temporal dynamics of object interactions.

3.2 Evaluating the Practicability of DSGG

Existing DSGG works primarily evaluate models using recall or mean recall. While useful for measuring coverage in retrieval tasks, we argue that recall alone does not provide a reliable assessment of the practical quality of generated scene graphs. We empirically observe two critical limitations:

(1) Precision–recall trade-off. Existing DSGG methods often exhibit severe precision degradation as the number of sampled scene graph triplets increases (Figure 1). As a result, the generated graphs frequently contain a large number of inaccurate or noisy relations. Such noise significantly degrades the usefulness of scene graphs in downstream applications such as Video QA or retrieval (Figure 9).

(2) Uninformative and redundant relations. Due to the prevalence of trivial relations in existing DSGG datasets [23], existing DSGG models tend to assign high confidence to trivial or repetitive relations (as each triplet prediction is made separately) (Figure 2), leading to scene graphs that are unnecessarily large yet semantically weak. Such redundancy becomes problematic when scene graphs are used as compact video representations (e.g., for metadata storage) or when downstream algorithms are sensitive to graph size (e.g., graph neural networks or neuro-symbolic methods [39]).

To systematically investigate these issues, we introduce five additional metrics that evaluate DSGG from the perspective of practical utility.

(1) Precision with MLLM-as-a-Judge. Ground-truth annotations often do not exhaustively cover all valid scene graph relations, making direct precision estimation difficult. To address this issue, we adopt an MLLM-as-a-Judge strategy to assess the validity of predicted relations. Specifically, we first remove predicted triplets that already appear in the ground-truth annotations. For the remaining predictions, we prompt a more powerful MLLM to determine whether each triplet represents a valid subject-object interaction within the corresponding temporal interval.

To validate the reliability of *MLLM-as-a-Judge* over MLLM's generation, we benchmark MLLM's performance on verifying whether a given triplet is valid according to the video over 1K randomly sampled triplets from both Action Genome [23] and VidVRD [56]. We find that although MLLMs' zero-shot performance on dense DSGG is substantially lower than finetuned ones (Figure 4), their performance on individual triplet verification is exceptionally strong (Table 1). This confirms the reliability of utilizing *MLLM-as-a-Judge* for DSGG. We further show that SGCls^{+} $P@10$ on AG is consistent across judges from different model families (Section 4.4).

(2) Scene Graph Retrieval (SGRet). We introduce *Scene Graph Retrieval (SGRet)* to measure how effectively predicted graphs support video retrieval. Specifically, we convert predicted scene graphs into natural language descriptions and use an off-the-shelf multimodal embedding model (e.g., Qwen3VL-Embedding [34]) to perform text-to-video retrieval. Retrieval accuracy then serves as a practical proxy for the semantic quality of generated scene graphs.

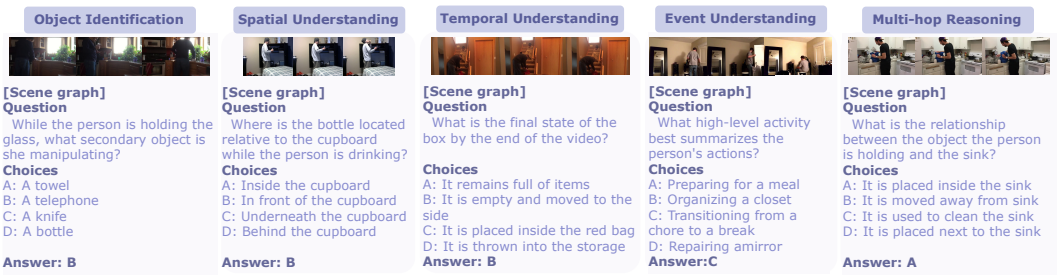


Figure 3: Samples of our curated SGQA dataset based on Action Genome.

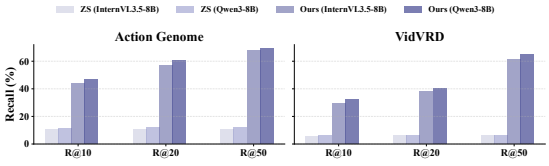


Figure 4: Recall comparison on SGCls$^+$. **ZS**: Zero-Shot prediction; **Ours**: MLLMs finetuned with our proposed procedure.

Table 1: Individual scene graph triplet zero-shot verification performance. "Truthful" indicates accuracy on ground-truth triplets ("`yes`" responses), and "Adv." denotes rejection rate on adversarial triplets ("`no`" responses). Higher is better (max = 100).

	ActionGenome		VidVRD	
Method	Truthful (%)	Adv. (%)	Truthful (%)	Adv. (%)
Zero-Shot (InternVL3.5-8B)	96.4	98.2	98.6	98.1
Zero-Shot (Qwen3-8B)	97.0	98.5	98.3	99.5

(3) Scene Graph QA (SGQA). To assess the utility of generated scene graphs for downstream tasks, we introduce SGQA, a diagnostic benchmark comprising 1,000 multiple-choice questions derived from 200 Action Genome videos [23]. Unlike existing benchmarks like AGQA [18] that focus on compositional logic over fixed symbolic structures, SGQA employs a frontier MLLM (*e.g.*, Gemini 2.5 Pro [15]) to curate "hard" questions conditioned on both raw video and ground-truth scene graphs.

We focus on five reasoning dimensions: object/action identification, spatial, temporal, event-level, and multi-hop reasoning. We intentionally targets implicit inferences that go beyond simple graph lookups, requiring a scene graph to preserve rich semantic nuances to support open-ended reasoning. During evaluation, an LLM answers the question using only the predicted scene graph as context; the resulting accuracy serves as a proxy for scene graph quality. By shifting from fixed query execution to holistic LLM reasoning, SGQA provides a more stringent measure of the scene graphs' utility.

(4) DSGG with Importance-aware Ranking. In addition to accuracy, the quality of scene graphs also depends on the *importance* of the relations it contains. We therefore evaluate DSGG as a ranking problem based on relation importance, where we consider two criteria: *informativeness* and *diversity*:

Informativeness score. Prior work [59, 40, 7] shows that LLMs can serve as effective zero-shot rerankers. We adopt a similar strategy to estimate how informative a triplet is with respect to a video clip. Given a triplet and the corresponding video segment, we prompt an MLLM with the query: `"[video][triplet] Does the triplet represent an important action/event/visual component in the video? Yes or No."` We then use the normalized probability of the `"Yes"` token as the informativeness score.

Diversity score. To encourage diversity among predicted relations, we measure how distinct each triplet is from previous ones, by encoding each triplet using a text encoder and compute the average cosine similarity between the new triplet and existing ones. Lower similarity indicates higher diversity.

The final triplet importance score is therefore a combination of informativeness and diversity scores. Given these scores, we propose to evaluate DSGG as a ranking task using Normalized Discounted Cumulative Gain (nDCG), a standard metric in information retrieval [47]:

$$\mathrm{DCG}_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad \mathrm{nDCG}_p = \frac{\mathrm{DCG}_p}{\mathrm{IDCG}_p}$$

where p denotes the rank position, rel_i is the importance score of the i-th triplet, and IDCG_p is the ideal DCG obtained from the ground-truth ranking.

(5) From SGCls to SGCls$^+$. We propose a practical variant of the SGCls task. In the standard SGCls setting, models predict scene graphs given *ground-truth* bounding boxes. However, such annotations are unavailable in real-world scenarios and therefore overlook the impact of detection errors on relation prediction. To better consider practical applications (VQA, retrieval, metadata

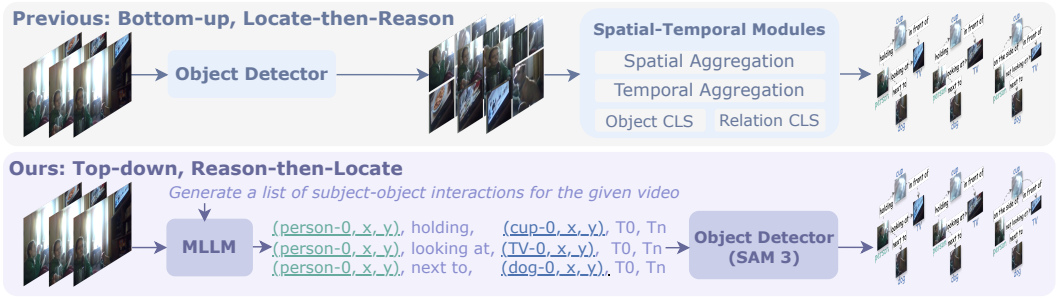


Figure 5: Conventional DSGG pipelines (top) and ours (bottom). Existing methods typically follow a bottom-up approach: the detector extracts regional features, which are then processed by specialized modules. In contrast, our method adopts a top-down *reason-then-locate* paradigm, where a MLLM first perform holistic reasoning, then grounded using a open-vocabulary detector.

storage, etc.) where grounding (*i.e.* SGDet) is not required, we replace ground-truth bounding boxes with the model's predicted boxes during evaluation. We denote this setting as SGCls^{+}.

3.3 Adapting MLLM for DSGG

Reason-then-Locate. Traditional DSGG pipelines follow a bottom-up paradigm, where objects are first detected and relations are subsequently inferred. In contrast, we adopt a top-down *reason-then-locate* strategy enabled by MLLMs. Specifically, the MLLM first performs holistic reasoning over the video to generate scene graph, and the grounding stage is performed afterward. Figure 5 illustrates the conceptual difference between conventional DSGG pipelines and our approach.

DSGG as Temporal Relation Set Prediction. Existing DSGG methods typically model dynamic scene graphs as frame-wise graphs, where each frame contains a separate snapshot of object relations. Directly applying this formulation to MLLM-based generation leads to substantial redundancy, as the model must repeatedly generate identical triplets for every frame where the relation persists. To address this issue, we reformulate DSGG as a *Temporal Relation Set (TRS)* prediction problem. Instead of frame-wise relations, each interaction is represented as a 5-ary, temporally grounded relation

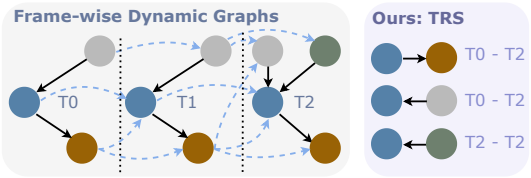


Figure 6: Illustration of traditional frame-wise dynamic graphs (left) and the proposed Temporal Relation Set (TRS, right).

$$\langle \texttt{sub}, \texttt{rel}, \texttt{obj}, t_{\texttt{start}}, t_{\texttt{end}} \rangle.$$

As illustrated in Figure 6, this formulation eliminates duplicated predictions across frames and substantially improves both training and inference efficiency (Figure 8). It also improves model accuracy, as excessive duplication during language-model finetuning can degrade generation quality.

Grounding via Open-Vocabulary Detector. Although recent MLLMs [1, 66] possess basic grounding capabilities, they are not optimized for multi-object, temporally consistent video tracking. We therefore employ an off-the-shelf segmentation model. For each predicted temporal relation, the MLLM outputs the starting location (x, y) of the subject and object. We then utilize the referring expression capacity of frontier segmentation model (SAM 3 [2]) to provide accurate grounding.

Training MLLM-based DSGG with Importance-Aware Finetuning. The autoregressive nature of MLLMs provides a simple mechanism to incorporate relation importance during generation. Specifically, during training we order the scene graph triplets according to their importance scores so that more important relations appear earlier in the sequence. The model is then finetuned using the standard negative log-likelihood objective:

$$\mathcal{L} = -\sum_{t=1}^{T} \log \mathrm{P}(x_t | V, x_{<t}; \theta_{\texttt{MLLM}})$$

where V denotes the input video, x_t denotes the t-th token.

Table 2: SGCls$^+$ performance under With/No Constraints settings on AG. Best/2nd best are highlighted in **bold**/underlined. Deltas are computed based on the SoTA baseline.[3]

Method	With Constraint			No Constraint		
	R@10	R@20	R@50	R@10	R@20	R@50
STTran	30.2	37.6	44.2	31.1	45.7	62.1
TEMPURA	33.6	36.4	42.8	37.8	48.5	59.7
TRACE	20.3	21.7	21.8	26.4	43.1	52.6
OED	41.2	52.3	63.6	44.3	55.9	66.8
Ours (InternVL-3.5)	43.5 (+2.3)	56.6 (+4.3)	63.8 (+0.2)	44.2 (-0.1)	56.8 (+0.9)	67.7 (+0.9)
Ours (Qwen3VL)	**47.3** (+6.1)	**61.1** (+8.8)	**71.7** (+8.1)	**46.6** (+2.3)	**60.3** (+4.4)	**69.5** (+2.7)

Table 3: SGCls$^+$ and SGDet results on VidVRD dataset. Deltas are computed against the strongest baseline per task.

Method	Recall			Precision		
	@10	@20	@50	@1	@10	@20
			SGCls$^+$			
VrdONE	19.1	27.8	43.0	**80.5**	46.2	33.6
OED[4]	24.4	31.4	55.2	61.4	49.7	38.2
Ours (InternVL-3.5)	29.2 (+4.8)	38.4 (+7.0)	61.5 (+6.3)	69.1 (-11.4)	61.3 (+11.6)	53.4 (+15.2)
Ours (Qwen3VL)	**32.1** (+7.7)	**40.3** (+8.9)	**65.2** (+10.0)	72.3 (-8.2)	**66.2** (+16.5)	**55.3** (+17.1)
			SGDet			
VrdONE	10.1	13.6	18.2	53.0	24.5	16.4
OED	19.4	24.3	22.8	37.5	31.3	24.9
Ours (InternVL-3.5)	26.6 (+7.2)	33.7 (+9.4)	41.3 (+18.5)	54.8 (+1.8)	45.6 (+14.2)	41.2 (+16.3)
Ours (Qwen3VL)	**31.5** (+12.1)	**38.2** (+13.9)	**48.6** (+25.8)	**56.1** (+3.1)	**51.4** (+20.2)	**45.2** (+20.3)

4 Experiments

4.1 Experimental Setup

Datasets. Following prior DSGG works [5, 49, 63], we use the Action Genome (AG) dataset [23] as the primary benchmark, which contains ˜10K videos annotated with 35 object and 25 relation categories. In addition, we evaluate our method on ImageNet-VidVRD [56] (VidVRD) and Panoptic Video Scene Graph (PVSG) [71], which extends DSGG to panoptic segmentation masks.

Evaluation Metrics. For AG and VidVRD, we evaluate our method under the standard scene graph detection (SGDet) setting as well as our proposed SGCls$^+$ protocol. Following previous works [5, 12, 63], we consider both the *With Constraint* and *Without Constraint* settings and report recall@k with $k \in \{10, 20, 50\}$. To assess the correctness of predicted relations, we additionally report precision@k using the proposed *MLLM-as-a-Judge* protocol with $k \in \{1, 10, 20\}$. For PVSG, the evaluation protocol is similar to the SGDet setting but replaces bounding boxes with segmentation masks. Following [28, 71], we report recall and mean recall@k with $k \in \{20, 50\}$.

Implementation Details. We conduct experiments using two recent multimodal large language models: Qwen3VL [1] and InternVL3.5 [66]. For visual grounding and tracking, we employ SAM 3 [2]. The MLLMs are fine-tuned using the standard LoRA [20] procedure with rank and scaling factor set to 32. For all datasets, models are trained for one epoch with a learning rate of 1×10^{-5}. Additional implementation details are provided in Appendix B.

4.2 Main Results

Results on Action Genome (AG). As shown in Table 4 and 2, our MLLM-based approach consistently outperforms existing DSGG models across both SGDet and SGCls$^+$. Under the SGDet *With Constraint* setting, Ours (Qwen3VL) achieves $49.6/57.9/62.2$ at $R@10/20/50$, surpassing the SOTA (UNO) by $+10.3$, $+12.7$, and $+8.4$, respectively. Similar gains occur under *No Constraint*, with $R@50$ reaching 61.7 ($+4.6$ over UNO). Furthermore, Table 8 shows significant precision gains; e.g., under SGDet, $P@10$ improves from 27.2 (OED) to 48.4. These results indicate that our method generates fewer redundant relations, yielding a superior precision-recall trade-off (Figure 1).

Results on VidVRD. Table 3 shows that our approach maintains its performance edge on VidVRD. Under SGCls$^+$, Ours (Qwen3VL) achieves $32.1/40.3/65.2$ at $R@10/20/50$, outperforming the strongest baseline (OED) by $+7.7$, $+8.9$, and $+10.0$, respectively. In SGDet, our method achieves an $R@50$ of 48.6, consistently leading in both recall and precision across all evaluated metrics.

Table 4: Results under With/No Constraints settings for SGDet on AG. Deltas are computed based on the SoTA baseline (UNO).

Method	With Constraint			No Constraints		
	R@10	R@20	R@50	R@10	R@20	R@50
VRD [45]	19.2	24.5	26.0	19.1	28.8	40.5
GPS-Net [41]	24.7	33.1	35.1	24.4	35.7	47.3
STTran [5]	25.2	34.1	37.0	24.6	36.2	48.8
TRACE [61]	13.9	14.5	14.5	26.5	35.6	45.3
TEMPURA [49]	28.1	33.4	34.9	29.8	38.1	46.4
TR^2 [64]	26.8	35.5	38.3	27.8	39.2	50.0
APT [38]	26.3	36.1	38.3	25.7	37.9	50.1
RelTR [6]	19.7	23.4	25.9	20.9	24.6	28.2
DSG-DETR [12]	30.3	34.8	36.1	32.1	40.9	48.3
TPT [75]	-	-	-	32.0	39.6	51.5
OED [63]	33.5	40.9	48.9	35.3	44.0	51.8
UNO [28]	39.3	45.2	53.8	40.8	49.7	57.1
ARN [65]	35.1	41.8	47.2	37.6	46.8	54.1
RS-Net [24]	30.5	35.0	36.3	-	-	-
Ours (InternVL-3.5)	43.1 (+3.8)	55.8 (+10.6)	59.3 (+5.5)	42.4 (+1.6)	53.2 (+3.5)	58.8 (+1.7)
Ours (Qwen3VL)	**49.6** (+10.3)	**57.9** (+12.7)	**62.2** (+8.4)	**47.3** (+6.5)	**56.2** (+6.4)	**61.7** (+4.6)

Results on PVSG. Table 7 reports the results on the PVSG benchmark under different vIoU thresholds. Our method consistently outperforms prior approaches across both recall and mean recall metrics. In par-

[4]Note that since we need to re-evaluate baselines for SGCls$^+$, we only report models which provide codebase and for which we can reproduce the reported results.

Table 5: Effect of Importance-Aware Finetuning (IAF) on ranking performance. Deltas are computed *w.r.t.* w/o IAF.

Model	w/ IAF	nDCG@5	nDCG@10
TEMPURA	-	27.1	31.6
DSG-DETR	-	29.4	32.8
OED	-	32.3	35.2
Ours (InternVL-3.5)	✗	35.7	38.4
	✓	43.0 (+7.3)	45.2 (+6.8)
Ours (Qwen3VL)	✗	37.9	41.6
	✓	44.3 (+6.4)	47.6 (+5.9)

Table 6: Performance on SGRet and SGQA. Best results are **bold**, second best are underlined.

	SGRet		SGQA
Method	R@5	R@10	Acc (%)
GT	68.4	80.2	63.8
STTran	8.41	14.3	25.5
TEMPURA	14.3	20.6	31.6
OED	16.3	22.3	36.2
Ours (InternVL3.5)	35.9 (+19.5)	44.5 (+22.1)	47.8 (+11.75)
Ours (Qwen3VL)	**36.8** (+20.44)	**46.4** (+24.1)	**48.9** (+12.8)

ticular, Ours (Qwen3VL) achieves $17.8/17.1$ R/mR@20 and $19.2/16.3$ R/mR@50 under vIoU $= 0.5$, improving over the SOTA (UNO) by large margins. These results demonstrate that the proposed MLLM-based framework generalizes well to panoptic scene graph generation.

4.3 Ablations

Effect of Importance-Aware Finetuning (IAF). Table 5 evaluates the impact of our proposed IAF on relation ranking. Without IAF, both prior DSGG methods and our MLLM-based models show limited awareness of triplet importance, yielding lower ranking quality; specifically, our baseline results are comparable to bottom-up methods like OED. This confirms that standard MLLM finetuning does not inherently prioritize informative triplets. Incorporating IAF yields notable gains: the InternVL3.5-based model improves by $+7.3$ (43.0) in nDCG@5 and $+6.8$ (45.2) in nDCG@10, while the Qwen3VL-based model reaches 44.3 and 47.6 in nDCG@5/10, respectively. These improvements demonstrate that explicitly modeling importance during finetuning enables more informative and diverse relation generation, as qualitatively validated in Figure 2.

Downstream Task Evaluation (SGRet and SGQA). Table 6 evaluates the quality of generated scene graphs through two downstream tasks: SGRet and SGQA. Our MLLM-based approach substantially outperforms prior DSGG methods on both tasks. Notable improvements are observed on SGQA, where our method reaches 48.9% accuracy compared to 36.2% from OED. Notably, the improvements are much larger than those observed on traditional recall metrics, suggesting that existing DSGG models often generate relations that match annotations but fail to capture the key semantics required for downstream reasoning. In contrast, our approach produces more informative and coherent scene graphs, which directly translate into stronger retrieval and question answering performance.

Table 7: Performance comparison of PVSG under different vIoUs.

	vIoU = 0.5		vIoU = 0.1	
Method	R/mR@20	R/mR@50	R/mR@20	R/mR@50
VPS [71]	0.42 / 0.61	0.73 / 0.76	6.50 / 5.75	9.64 / 8.25
IPS+T [71]	3.88 / 2.81	5.66 / 4.12	9.01 / 6.69	14.9 / 11.3
VISA [36]	- / 8.82	- / 8.94	-	-
MCL [50]	0.84 / 0.98	1.26 / 1.22	8.18 / 8.00	12.9 / 11.5
UNO [28]	9.44 / 8.25	10.8 / 9.72	17.5 / 14.8	23.8 / 18.3
(Ours) InternVL3.5	13.4 / 12.2	15.8 / 14.6	21.1 / 20.6	26.5 / 25.4
(Ours) Qwen3-VL	**17.8 / 17.1**	**19.2 / 16.3**	**26.5 / 23.2**	**31.3 / 30.8**

Figure 9 further analyzes SGRet performance as a function of the number of top-k scene graph triplets used during retrieval. Overall, MLLM-based approaches consistently outperform traditional DSGG models across different values of k. Particularly, existing methods exhibit an obvious concave trend: retrieval performance is poor when only the top-1 triplet is used and also deteriorates when a large number of triplets ($k = 50$) are included. The weak performance at $k = 1$ suggests that the highest-confidence triplets produced by these models are often not the most informative ones. Meanwhile, the drop at larger k reflects the accumulation of noisy and redundant relations, which lowers retrieval precision.

Table 8: Precision on SGDet and SGCls$^+$ under AG. Deltas are computed based on the SoTA baseline (OED).

	SGDet			SGCls$^+$		
Method	P@1	P@10	P@20	P@1	P@10	P@20
STTran	26.3	17.6	13.1	32.7	22.6	16.9
TRACE	27.5	15.3	12.7	26.5	18.7	13.7
TEMPURA	38.1	21.4	13.7	49.1	27.3	18.0
DSG-DETR	37.4	22.3	14.4	51.3	30.6	20.2
OED	43.6	27.2	17.4	54.1	35.2	22.8
Ours (InternVL-3.5)	48.2 (+4.6)	42.9 (+15.7)	30.4 (+13.0)	58.3 (+4.2)	50.9 (+15.7)	36.4 (+13.6)
Ours (Qwen3VL)	**53.0** (+9.4)	**48.4** (+21.2)	**35.9** (+18.5)	**59.1** (+5.0)	**53.1** (+18.0)	**38.2** (+15.4)

In contrast, our approach maintains strong performance across a wide range of k. When IAF is disabled, our model exhibits a similar degradation at $k = 1$, confirming that standard training objectives lack awareness of triplet importance. After incorporating IAF, however, the model learns to

[4]While the original OED [63] does not test on VidVRD, we run training using OED's codebase with the default configurations from AG for VidVRD.

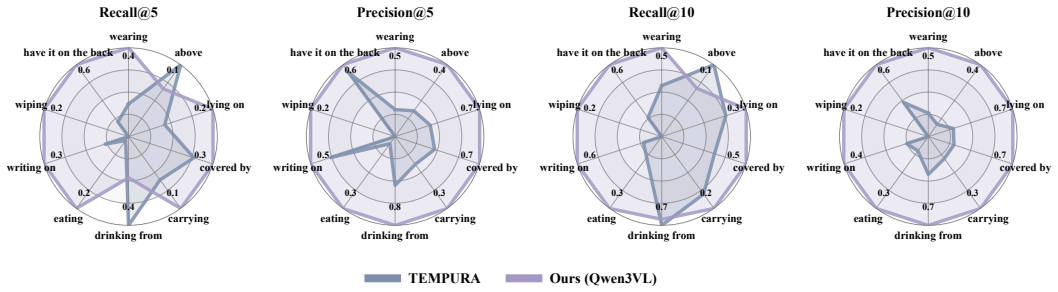


Figure 7: Per relation category Recall/Precision performance between our MLLM-based method (purple) and TEMPURA (blue) on Action Genome dataset, under the SGCls$^+$ setting. We show the 15 ***least***-frequent relation categories.

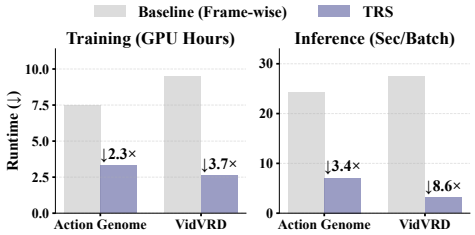


Figure 8: Training/Inference speed w/ or w/o TRS.

Table 9: Performance comparison on Action Genome with or without adopting the proposed Temporal Relation Set (TRS) representation of dynamic scene graphs.

Backbone	w/ TRS	Action Genome		VidVRD	
		R@10	P@1	R@10	P@1
InternVL3.5	✗	42.9	57.6	28.4	70.3
	✓	44.2 (+1.3)	58.3 (+0.7)	29.2 (+0.8)	69.1 (–1.2)
Qwen3-VL	✗	45.3	58.3	30.4	73.1
	✓	**46.6** (+1.3)	**59.1** (+0.8)	**32.1** (+1.7)	**73.9** (+0.8)

prioritize informative relations, leading to strong retrieval performance even when only a few triplets are used. Moreover, our method remains robust when k increases, as MLLMs naturally "know" when to stop generation. This behavior contrasts with conventional DSGG models, which tend to over-generate relations and therefore suffer significant precision loss when larger triplet sets are used.

Impact of Temporal Relation Set (TRS). Figure 8 and Table 9 evaluate the efficiency and accuracy gains of our TRS representation over the redundant, frame-wise DSGG baseline. By representing relations as temporally grounded tuples, TRS eliminates duplicated predictions across frames, drastically reducing sequence length. Consequently, training speed increases by $2.3\times$ on AG, while inference latency drops by up to $8.6\times$. Beyond efficiency, TRS consistently improves performance; for instance, $R@10$ increases by up to $+1.3$ and $+1.7$ on Action Genome and VidVRD, respectively. These results suggest that removing frame-wise redundancy provides a cleaner supervision signal, allowing MLLMs to achieve higher accuracy with significantly lower computational overhead.

Long-tailed prediction with MLLM-based DSGG. Figure 7 compares per-category accuracy between our MLLM-based approach and TEMPURA, a method specifically designed to mitigate long-tail issues. Nevertheless, TEMPURA still exhibits highly uneven behavior, achieving moderate performance on a few relations but failing on many others. In contrast, our MLLM-based model maintains consistent performance across most relation categories, including rare predicates such as *wiping*, *writing on*, and *have it on the back*. This behavior suggests that MLLMs benefit from their large-scale pretraining and broader visual-semantic knowledge, allowing them to generalize beyond the frequency biases present in DSGG datasets. As a result, our approach acts as a more balanced scene graph generator, improving both coverage and reliability for long-tail relations.

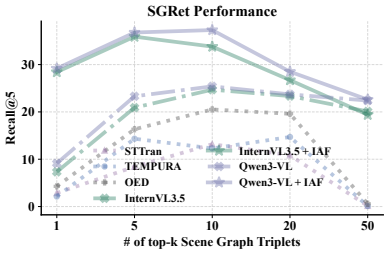


Figure 9: SGRet's recall@5 performance on Action Genome dataset as a function of the number of top-k triplets used during retrieval.

4.4 Further Analysis

Robustness of MLLM-as-a-Judge. As our default judge (Qwen3-VL-30B-A3B) belongs to the same model family as one of our generators, we examine whether precision depends on the judge. Table 10 re-evaluates SGCls$^+$ $P@10$ on AG with two judges from independent model families, Gemini 3.5 Flash-Lite [17] and Gemma 4 26B-A4B [16]. The ranking of methods is identical under all three judges, and each method's score varies by at most 1.8. Notably, the same-family judge scores Ours (Qwen3VL) at 53.1, between the two independent judges (54.3 and 52.5), showing no sign of same-family score inflation. Moreover, generation (greedy decoding with fixed frames and prompts) and judging (a single forward pass per triplet) are both deterministic, ensuring reproducibility.

Table 10: SGCls$^+$ P@10 on AG under three MLLM judges: Qwen3-VL-30B-A3B (default), Gemini 3.5 Flash-Lite, and Gemma 4 26B-A4B. Std: standard deviation across judges.

Method	Qwen3-VL	Gemini	Gemma	Std
DSG-DETR	30.6	29.9	31.2	0.65
OED	35.2	33.6	33.4	0.99
Ours (InternVL-3.5)	50.9	50.1	51.7	0.80
Ours (Qwen3VL)	53.1	54.3	52.5	0.92

Table 11: Error analysis. **FP**: false positives due to wrong temporal intervals (Temp.), wrong relations (Rel.), or other errors. **FN**: missed relations with below-mean informativeness (Low-Info.) or duration (Short), which can overlap.

	FP (%)			**FN (%)**	
Method	Temp.	Rel.	Other	Low-Info.	Short
OED	44.1	38.7	17.2	48.2	56.8
Ours (InternVL-3.5)	74.6	23.1	2.3	83.6	67.3
Ours (Qwen3VL)	69.5	22.3	8.2	80.4	60.1

Table 12: Comparison with zero-shot (ZS) proprietary MLLMs on AG (SGCls$^+$, No Constraint).

	Precision			**Recall**			**nDCG**	
Method	@1	@10	@20	@10	@20	@50	@5	@10
OED	54.1	35.2	22.8	44.3	55.9	66.8	32.3	35.2
GPT-5.6 Luna (ZS)	63.9	55.4	38.2	24.5	45.2	59.1	16.6	18.4
Gemini 3.5 Flash-Lite (ZS)	**68.4**	**57.6**	**40.7**	27.3	43.7	56.4	17.9	18.2
Ours (Qwen3VL)	59.1	53.1	38.2	**46.6**	**60.3**	**69.5**	**44.3**	**47.6**

Table 13: Open-vocabulary results on VidVRD (Novel split) under the OpenVidVRD protocol [44], from which baseline results are taken. 50%/100%: fraction of training data.

	SGDet			**SGCls**		
Method	mAP	R@50	R@100	mAP	R@50	R@100
ALPro [30]	0.98	2.79	4.33	3.69	7.27	8.92
VidVRD-II [55]	3.11	7.93	11.38	5.70	13.22	18.34
RePro [14]	5.87	12.75	16.23	10.32	19.17	25.28
MMP [72]	12.15	13.75	15.21	17.57	21.98	28.43
OpenVidVRD [44]	17.95	17.69	19.01	18.25	28.98	31.85
Ours (Qwen3VL, 50%)	**32.13**	**33.85**	**33.87**	**37.63**	**40.29**	**40.32**
Ours (Qwen3VL, 100%)	29.54	32.66	32.69	32.41	36.83	36.87

Error analysis. Table 11 categorizes false positives (triplets rejected by the judge) and false negatives (missed ground-truth relations). The false positives of OED spread across temporal, relation, and other (*e.g.*, entity) errors, whereas 69.5–74.6% of ours are correct triplets with imprecise temporal intervals. This suggests that MLLM-based DSGG is substantially more semantically consistent, and its main remaining weakness is temporal boundary estimation, which is bounded by the dense temporal localization ability of MLLMs. Meanwhile, our models disproportionately miss short and less informative relations: unlike traditional models that over-generate, MLLMs produce fewer spurious edges at the cost of coverage on less important triplets. Consistently, our SGQA gains are largest on event-level and multi-hop questions, which require multi-edge coherence (Table 14).

Comparison with zero-shot frontier MLLMs. Open-source MLLMs perform poorly on zero-shot DSGG (Figure 4). To test whether proprietary MLLMs remove the need for finetuning, we evaluate GPT-5.6 Luna [52] and Gemini 3.5 Flash-Lite zero-shot (Table 12). Both achieve high precision, as *MLLM-as-a-Judge* credits valid triplets beyond the annotations, yet their margin over ours is modest; most of their errors are correct triplets with imprecise temporal intervals. Their recall and nDCG are substantially lower (*e.g.*, 18.4/18.2 *vs.* 47.6 nDCG@10), as their predictions are misaligned with the target relation distribution; frontier MLLMs thus do not remove the need for finetuning.

Open-vocabulary DSGG. This work follows the closed-vocabulary setting of mainstream DSGG benchmarks. As a preliminary step beyond it, we adopt the OpenVidVRD [44] open-vocabulary protocol on VidVRD (Table 13). Ours (Qwen3VL, 50%) improves every metric over the strongest baseline by a wide margin, *e.g.*, SGDet mAP from 17.95 to 32.13 and SGCls mAP from 18.25 to 37.63. Training on 50% of the data outperforms using the full set: with a small, fixed vocabulary, longer training overfits the MLLM to it and hurts generalization. Faithfully evaluating open-vocabulary capacity thus calls for open-world, expressive DSGG datasets, which we leave for future work. Additional analysis and discussion are provided in Appendix E.

5 Conclusion

In this work, we re-examine Dynamic Scene Graph Generation (DSGG) from the perspective of practical utility in the era of Multimodal Large Language Models (MLLMs). We identify key limitations of existing DSGG approaches, and introduce a set of practicability-oriented evaluation metrics that better reflect the usefulness of generated scene graphs in downstream tasks. Building on these insights, we explore the use of MLLMs for DSGG and propose several key designs. Extensive experiments across multiple datasets demonstrate that our approach not only achieves SOTA recall performance but also produces more informative, balanced, and practically useful scene graphs.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025.

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer International Publishing, 2020.

[4] Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. (2.5+1)D spatio-temporal scene graphs for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 444–453, 2022.

[5] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16372–16382, 2021.

[6] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023.

[7] Xuanming Cui, Hong-You Chen, Hao Yu, Hao Yuan, Zihao Wang, Shlok Kumar Mishra, Hanchao Yu, Yonghuan Yang, Jun Xiao, Ser-Nam Lim, Jianpeng Cheng, Qi Guo, and Xiangjun Fan. Reason to contrast: A cascaded multimodal retrieval framework, 2025.

[8] Xuanming Cui, Jianpeng Cheng, Hong you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding, 2026.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc., 2023.

[10] Xiaoyu Deng, Zhengjian Kang, Xintao Li, Yongzhe Zhang, and Tianmin Guo. Covis: A collaborative framework for fine-grained graphic visual understanding. In *2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 298–303. IEEE, 2025.

[11] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Björn Ommer, and Nassir Navab. SceneGenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 88–98, 2023.

[12] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5130–5139, January 2023.

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025.

[14] Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[16] Gemma Team. Gemma 4 technical report, 2026.

[17] Google DeepMind. Gemini 3.5 Flash-Lite model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/`, 2026.

[18] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11287–11297, 2021.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[21] Jiani Huang, Ziyang Li, Mayur Naik, and Ser-Nam Lim. LASER: A neuro-symbolic framework for learning spatio-temporal scene graphs with weak supervision. In *International Conference on Learning Representations*, pages 41977–41999, 2025.

[22] Jiani Huang, Amish Sethi, Matthew Kuo, Mayank Keoliya, Neelay Velingker, JungHo Jung, Ser-Nam Lim, Ziyang Li, and Mayur Naik. ESCA: Contextualizing embodied agents via scene-graph generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[23] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action Genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10236–10247, 2020.

[24] Hae-Won Jo and Yeong-Jun Cho. RS-Net: Context-aware relation scoring for dynamic scene graph generation. *Pattern Recognition*, 177:113352, 2026.

[25] Zhengjian Kang, Qi Chen, Rui Liu, Kangtong Mo, Xingyu Zhang, Xiaoyu Deng, and Ye Zhang. Causality-aware temporal projection for video understanding in video-llms. *arXiv e-prints*, pages arXiv–2601, 2026.

[26] Zhengjian Kang, Ye Zhang, Xiaoyu Deng, Xintao Li, and Yongzhe Zhang. Lp-detr: Layer-wise progressive relation for object detection. In *International Conference on Intelligent Computing*, pages 144–156. Springer Nature Singapore, Singapore, 2025.

[27] Anant Khandelwal. FloCoDe: Unbiased dynamic scene graph generation with temporal consistency and correlation debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2516–2526, 2024.

[28] Huy Le, Nhat Chung, Tung Kieu, Jingkang Yang, and Ngan Le. UNO: Unifying one-stage video scene graph generation via object-centric visual representation learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2797–2807, 2026.

[29] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. Symbolic replay: Scene graph as prompt for continual learning on VQA task. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1250–1259, 2023.

[30] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022.

[31] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3D in large multimodal models. In *International Conference on Learning Representations (ICLR)*, pages 81182–81199, 2025.

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022.

[33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022.

[34] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026.

[35] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28076–28086, June 2024.

[36] Yanjun Li, Zhaoyang Li, Honghui Chen, and Lizhi Xu. Unbiased video scene graph generation via visual and semantic dual debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19047–19056, 2025.

[37] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1347–1356, 2017.

[38] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13864–13873, 2022.

[39] Ziyang Li, Jiani Huang, and Mayur Naik. Scallop: A language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1463–1487, 2023.

[40] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *International Conference on Learning Representations (ICLR)*, pages 44215–44234, 2025.

[41] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. GPS-Net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3743–3752, 2020.

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, 2023.

[44] Qi Liu, Weiying Xue, Yuxiao Wang, and Zhenao Wei. OpenVidVRD: Open-vocabulary video visual relation detection via prompt-driven semantic space alignment, 2025.

[45] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9905 of *Lecture Notes in Computer Science*, pages 852–869. Springer, 2016.

[46] Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Zhifan Feng, Yajuan Lyu, Hong Liu, and Yong Zhu. Dynamic multistep reasoning based on video scene graph for video question answering. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3894–3904, Seattle, United States, July 2022. Association for Computational Linguistics.

[47] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.

[48] Sayak Nag, Udita Ghosh, Calvin-Khang Ta, Sarosij Bose, Jiachen Li, and Amit K. Roy-Chowdhury. Conformal prediction and MLLM aided uncertainty quantification in scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11686, 2025.

[49] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K Roy-Chowdhury. Unbiased scene graph generation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22803–22813, 2023.

[50] Thong Thanh Nguyen, Xiaobao Wu, Yi Bin, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. Motion-aware contrastive learning for temporal panoptic scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6218–6226, 2025.

[51] Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. HyperGLM: HyperGraph for video scene graph generation and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29150–29160, 2025.

[52] OpenAI. GPT-5.6 system card. `https://deploymentsafety.openai.com/gpt-5-6`, 2026.

[53] Tianwen Qian, Jingjing Chen, Shaoxiang Chen, Bo Wu, and Yu-Gang Jiang. Scene graph refinement network for visual question answering. *IEEE Transactions on Multimedia*, 25:3950–3961, 2023.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[55] Xindi Shang, Yicong Li, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Video visual relation detection via iterative inference. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3654–3663, 2021.

[56] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, pages 1300–1308, New York, NY, USA, 2017. Association for Computing Machinery.

[57] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024.

[58] Shuzhou Sun, Shuaifeng Zhi, Qing Liao, Janne Heikkilä, and Li Liu. Unbiased scene graph generation via two-stage causal modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12562–12580, 2023.

[59] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore, 2023. Association for Computational Linguistics.

[60] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3722, 2020.

[61] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13688–13697, 2021.

[62] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. SOK-Bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13384–13394, 2024.

[63] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. OED: Towards one-stage end-to-end dynamic scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27938–27947, 2024.

[64] Jingyi Wang, Jinfa Huang, Can Zhang, and Zhidong Deng. Cross-modality time-variant relation learning for generating dynamic scene graphs. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 8231–8238, 2023.

[65] Liwei Wang, Yanduo Zhang, Tao Lu, Fang Liu, Huiqin Zhang, Jiayi Ma, and Huabing Zhou. End-to-end entity-predicate association reasoning for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17729–17738, October 2025.

[66] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025.

[67] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, July 2017.

[68] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In *Computer Vision – ECCV 2022*, pages 374–390. Springer Nature Switzerland, 2022.

[69] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin

Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[70] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018.

[71] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18675–18685, 2023.

[72] Shuo Yang, Yongqi Wang, Xiaofeng Ji, and Xinxiao Wu. Multi-modal prompting for open-vocabulary video visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6513–6521, 2024.

[73] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

[74] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2915–2924, 2023.

[75] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. End-to-end video scene graph generation with temporal propagation transformer. *IEEE Transactions on Multimedia*, 26:1613–1625, 2024.

[76] Zijian Zhou, Miaojing Shi, and Holger Caesar. HiLo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21637–21648, October 2023.

[77] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

```
### Task
Generate 5 diagnostic multiple-choice questions (one per reasoning dimension) based on the provided Video and its corresponding
Ground-Truth Scene Graph.

### Constraints
1. Each question must be non-trivial: it cannot be solved by a simple keyword lookup from the provided scene graph. However, it should
also be able to reasonably inferred from the scene graph.
2. Distractors must be semantically plausible and present in the scene graph but contextually incorrect.
3. Use the following reasoning dimensions:
  - Object/Action Identification: Identifying subtle attributes or non-standard interactions.
  - Spatial Reasoning: Understanding 3D/relative arrangements (e.g., behind, between, furthest).
  - Temporal Reasoning: Identifying event sequences or anticipate state changes over time.
  - Event-Level Reasoning: Inferring high-level intent or the nature of a multi-step activity.
  - Multi-Hop Reasoning: Requiring a traversal of at least 3 nodes/edges (e.g., A -> B -> C).

### Output Format
Return ONLY a JSON object with the following structure:
{
 "video_id": "string",
 "questions": [
  {
   "dimension": "string",
   "question": "string",
   "options": {"A": "string", "B": "string", "C": "string", "D": "string"},
   "answer": "string (A/B/C/D)",
   "cot": "Explain which triplets/video segments are needed to solve this."
  }
 ]
}

### Input Data
[VIDEO FILE]
[GROUND-TRUTH SCENE GRAPH]
```

Figure 10: Prompt for generating the proposed SGQA dataset.

A SGQA Dataset Construction

To evaluate the downstream utility and semantic fidelity of generated video scene graphs (SGs), we introduce **SGQA**, a diagnostic benchmark comprising multiple-choice questions (MCQs) grounded in scene graph structures. We utilize Gemini 2.5 Pro [15] to curate these questions, conditioned on both the video and the ground-truth dynamic scene graphs represented in our proposed Temporal Relation Set (TRS) format: $\langle sub, rel, obj, t_{\text{start}}, t_{\text{end}} \rangle$. Specifically, we define five core reasoning dimensions to assess the structural and temporal completeness of the SGs:

- **Object/Action Identification:** Measures the ability to verify the presence of fine-grained entities or subtle interactions that exceed simple label matching. *E.g., "Which specific object is being manipulated during the primary interaction?"*
- **Spatial Reasoning:** Evaluates the understanding of 3D relative positioning and orientation between entities within the scene. *E.g., "What is the relative position of the person with respect to the shelf at the start of the clip?"*
- **Temporal Reasoning:** Tests the capacity to track state transitions, sequential dependencies, and the chronological order of relations. *E.g., "What action did the person perform immediately before interacting with the cupboard?"*
- **Event-Level Reasoning:** Requires high-level abstraction of discrete triplets into a coherent understanding of intent or activity context. *E.g., "Which overarching activity best summarizes the sequence of interactions observed?"*
- **Multi-hop Reasoning:** Serves as a complexity stress test requiring the model to traverse multiple relational dependencies and nodes simultaneously to reach a conclusion.

For each dimension, we randomly sample 200 video clips from Action Genome [23], with each clip represented by 8 keyframes, totaling 1,000 diagnostic samples. To ensure the benchmark is sufficiently challenging, we prompt the generator to craft "hard" questions with semantically plausible distractors that cannot be resolved by simply traversing the scene graph. The prompt used for generating dataset is shown in Figure 10.

Evaluation. To evaluate the performance of DSGG with our constructed MCQ, we first take the top-20 generated scene graph triplets and convert into TRS format. Then we use Qwen3-8B [69] as the fixed evaluator to answer MCQs based on generated SGs.

```
### Task
Analyze the provided video and generate a Dynamic Scene Graph (DSGG) representing the interactions.

### Output Format
Return ONLY a Python list of tuples. Each entry must strictly follow this schema:
(subject-id, relation, object-id, start_second, end_second)

### Definitions
- subject_id / object_id: The class label of the entity (e.g., 'person', 'cup'). Choose from the following available entity names: {entity-vocabulary}
- relation: The action or spatial predicate (e.g., 'holding', 'sitting_on', 'behind'). Choose from the following available relation types: {relation types}

### Requirements

- Be as comprehensive as possible. Cover all plausible interactions.
- Strictly follow the vocabulary for entity and relation.
```

Figure 11: Prompt for zero-shot DSGG experiments.

B Implementation Details

MLLM Backbone. We adopt two state-of-the-art MLLMs: InternVL3.5 [66] and Qwen3VL [1] with 8B model size as the scene graph generator. We use Qwen3VL-2B for calculating informativeness score. We use Jina V3 text embedding model [57] for calculating diversity score. For *MLLM-as-a-Judge*, we choose the `Qwen/Qwen3-VL-30B-A3B-Instruct-FP8` model, balancing performance and efficiency. For the cross-judge analysis (Table 10), we additionally use Gemini 3.5 Flash-Lite [17] and Gemma 4 26B-A4B [16] as judges.

Frame Sampling. For frame sapling, we use consecutive frames whenever there is scene graph annotation, and each sample consists of 8 frames.

SGRet setup. We adopt Qwen3VL-Embedding-2B [34] as the multimodal embedding model for embedding both textual scene graph and the video clip. We randomly sample 1000 video clips as the retrieval pool.

Grounding with Off-the-Shelf Detector. To perform accurate grounding with an off-the-shelf open-vocabulary detector (*e.g.* SAM 3 [2]), we use a combination of textual labels and points generated by the MLLM scene graph generator. Specifically, we first pass each triplet as referring expression (*e.g.* `person holding cup`) to the detector. In case when there are multiple predictions, we select the mask whose centroid is closest to the generated point.

Prompts. The prompt for zero-shot DSGG experiment is shown in Figure 11. For *MLLM-as-a-Judge*, we use the prompt:

"`Given the video and the following subject-relation-object triplet, determine if the triplet is valid based on the video. Answer Yes or No.`".

For the MLLM scene graph generator, we adopt the prompt:

"`Given the video, generate a python list of tuples in the form of ((subject-id, start-x, start-y), relation, (object-id, start-x, start-y), start_time, end_time).`"

C Training details

We apply LoRA [20] finetune for all experimented backbone MLLMs with two Nvidia H100 GPUs. The LoRA rank is set to 32 with an alpha value of 64 across all experiments. We train all models for a single epoch. The learning rates are set to be 4×10^{-5} for Qwen3-VL and InternVL3.5. Batch sizes are set to 1 with gradient accumulation steps equal to 8.

D Additional results

Top-k generation's informativeness score. Figure 12 compares the informativeness scores of the top-k generated triplets across different methods on Action Genome and VidVRD. Overall, MLLM-based approaches consistently produce more informative relations than existing DSGG models. Traditional methods such as STTran and TEMPURA tend to generate triplets with significantly lower informativeness scores, suggesting that their high-confidence predictions often correspond to trivial or redundant relations. In contrast, our MLLM-based models generate relations that better

Table 14: Performance across different MCQ question types. Task abbreviation is as follow: **Spatial**: spatial understanding; **Temporal**: temporal understanding; **Event**: event understanding; **Multi-hop**: multi-hop reasoning.

Model	Obj	Spatial	Temporal	Event	Multi-hop	Overall
STTran	26.5	25.8	25.1	25.5	24.8	25.5
TEMPURA	36.9	33.6	27.9	33.6	25.7	31.5
OED	44.1	40.2	28.5	40.3	27.2	36.1
Ours (InternVL3.5)	52.4 (+8.3)	51.3 (+11.1)	38.4 (+9.9)	52.6 (+12.3)	44.4 (+17.2)	47.8 (+11.7)
Ours (Qwen3-VL)	**53.2** (+9.1)	**51.8** (+11.6)	**39.6** (+11.1)	**53.4** (+13.1)	**46.6** (+19.4)	**48.9** (+12.8)

capture the key semantics of the video. Furthermore, applying Importance-Aware Finetuning (IAF) further improves the ranking quality of generated triplets, especially for the top-ranked predictions. This effect is most noticeable at small k, where IAF encourages the model to prioritize highly informative relations earlier in the generation process. As k increases, the informativeness scores gradually decrease for all methods, reflecting the diminishing semantic value of lower-ranked triplets. Nevertheless, our approach consistently maintains higher informativeness scores across both datasets, demonstrating the effectiveness of the proposed ranking-aware training strategy.

SGQA Performance Across Question Types. Table 14 reports the SGQA accuracy across different question categories. Our MLLM-based approaches significantly outperform prior DSGG methods across all question types. In particular, **Ours (Qwen3-VL)** achieves the best overall accuracy of 48.9%, improving over the strongest baseline (OED) by $+12.8$. The improvements are consistent across all reasoning categories, including object identification, spatial reasoning, temporal understanding, event comprehension, and multi-hop reasoning. Notably, the largest gains are observed on the more complex reasoning tasks such as *Multi-hop* and *Event* questions, where our method improves by up to $+19.4$ and $+13.1$, respectively. These results indicate that the scene graphs generated by our MLLM-based framework capture richer semantic and temporal information, enabling more effective reasoning over video content.

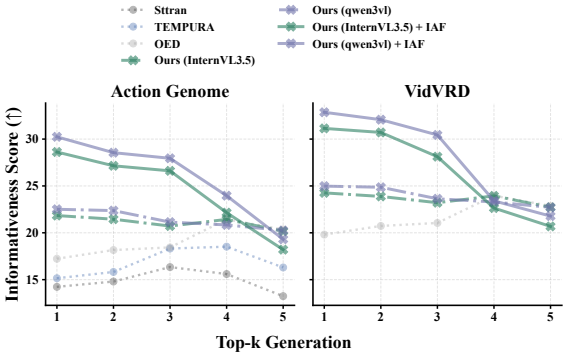


Figure 12: Top-5 triplets' informativeness score generated by STTran [5], TEMPURA [49], ours, and ours with Importance-Aware Finetuning (IAF) [23] on Action Genome and VidVRD dataset.

Qualitative comparison between LMM and Previous methods. Figure 13 and Figure 14 compare scene graph generated from STTran [5], TEMPURA [49], Ours, and Ours with Importance-Aware Finetuning (IAF). We can observe STTran and TEMPURA are able to generate more triplets but also exhibit a higher rate of false positives, including the generation of context-based triplets—instances where triplets are produced even when the corresponding objects are absent from the scene. In contrast, Ours-based results are noticeably more accurate. We can observe that as the number of predictions increases, STTran and TEMPURA tend to generate more false positives, leading to a decline in precision at higher values of K. In contrast, Ours effectively aligns triplet generation with the scene context, maintaining higher precision. Meanwhile, we can also observe from both Figure 13 and Figure 14 that previous methods like STTran and TEMPURA tends predict prevalent objects and relations such as "floor" and "not looking at" as the highest confidence entities and relations. However, in the given examples, these entities and relations are either not visible in the scene, or are not important to the primary activity. On the other hand, Ours with IAF is able to generate the most relevant triplets first. For instance, for the right most sample of Figure 14, where the person is sitting on a chair while reading a book, while STTran predicts $\langle$*floor*, *beneath*, *person*$\rangle$, Ours is able to predict $\langle$*person*, *looking at*, *cabinet*$\rangle$, which is more relevant to the person's action. Meanwhile, by training with IAF, the results also involve less redundant triplets at a rank position. Take the same example for instance, while STTran predicts redundant triplets: $\langle$*chair*, *beneath*, *person*$\rangle$, $\langle$*person*, *sitting on*, *chair*$\rangle$, $\langle$*chair*, *behind*, *person*$\rangle$, Ours + IAF is able to predict distinct distinct triplets up to a rank position.



Figure 13: Qualitative comparison between Tempura, STTran, and Ours. Predictions/objects highlighted in red represent false predictions. Triplets are displayed (top-down) in the order of their model predictions (sorted by probability for STTran and TEMPURA, or by their order of generation.

E Additional Analysis and Discussion

Evaluation protocol and reproducibility. Both scene graph generation and *MLLM-as-a-Judge* are deterministic in our evaluation. For generation, we use greedy decoding with fixed frames and prompts. For judging, we run a single forward pass per triplet with the fixed prompt in Appendix B and extract the logit of the `"Yes"` token. Replacing the default Qwen3-VL-30B-A3B judge with Gemini 3.5 Flash-Lite or Gemma 4 26B-A4B preserves the ranking of all evaluated methods, with a maximum within-method difference of 1.8 in $P@10$ (Table 10). We will release the code, datasets, and evaluation pipeline so that future methods can be evaluated under the same protocol.

Restricting predictions to the target taxonomy. Although MLLMs and SAM 3 are open-vocabulary models, we find that finetuned MLLMs are faithful to the target vocabulary. Soft-matching the generated labels to the dataset vocabulary with a text-embedding model brings only a marginal ($< 1\%$) gain in recall, as the finetuned MLLMs rarely generate novel labels that are synonyms of existing categories. When they do generate truly novel labels, these predictions are assessed by the *MLLM-as-a-Judge* precision. SAM 3 is applied only after entity and relation generation and does not alter the predicted labels: for each predicted triplet, we pass the triplet as a referring expression (*e.g.*, `person holding cup`) to SAM 3 and, when multiple masks are returned, select the mask whose centroid is closest to the subject/object point predicted by the MLLM (Appendix B).

Identity association and occlusion. The MLLM assigns an identifier to each entity (*e.g.*, `person-0` in Figure 5), which associates the same entity across different triplets. We do not explicitly model occlusions or entities that temporarily leave the frame. Instead, such entities are inferred implicitly through the holistic video understanding of the MLLM, while SAM 3 tracks them through partial or short-term occlusions with its memory. This is another advantage of the reason-then-locate paradigm: locate-then-reason methods operate on detected regional features and are therefore inherently limited when entities are invisible; accordingly, the standard DSGG evaluation protocol excludes occluded objects (*e.g.*, by filtering invisible objects out of the ground truth).

Comparison with traditional DSGG models. Our MLLM-based models benefit from the large-scale pretraining of MLLMs and SAM 3, whereas most traditional DSGG models are trained from scratch, which leads to an asymmetry in model scale and pretraining data. However, leveraging



Figure 14: Qualitative examples of generations from traditional DSGG method (STTran [5]), versus ours, and ours with Importance-Aware Finetuning (IAF). Predictions/objects highlighted in red represent false predictions. Triplets are displayed (top-down) in the order of their model predictions (sorted by probability for STTran and TEMPURA, or by their order of generation.

pretrained foundation models is an established practice in scene graph generation: VS^3 [74] exploits pretrained vision-language models for image SGG, and LASER [21] and ESCA [22] distill the knowledge of commercial MLLMs to train DSGG models. Moreover, MLLMs do not naturally excel at DSGG: open-source MLLMs perform poorly on zero-shot DSGG (Figure 4), and even frontier proprietary MLLMs used zero-shot attain substantially lower recall and nDCG than finetuned models (Table 12). Rather than proposing another specialized module within the traditional DSGG framework, our goal is to analyze the failure modes of existing DSGG models, to establish practical evaluation metrics for DSGG, and to provide a systematic study of MLLM-based DSGG that can serve as a solid foundation for future work.

Open-vocabulary evaluation protocol. For the results in Table 13, we follow the open-vocabulary protocol of OpenVidVRD [44] on ImageNet-VidVRD [56], which splits object and predicate categories by frequency into base (common) and novel (rare) categories as in RePro [14]. Models are trained on base categories, and results are reported on the test set under the Novel split, which includes all object categories and the novel predicate categories. We report mAP and $R@50/100$ for SGDet and SGCls, where SGCls uses ground-truth trajectories, and a predicted triplet is considered correct if it matches a ground-truth triplet with a trajectory IoU above 0.5. The 50% and 100% settings denote the fraction of training data used for finetuning, and baseline results are taken from [44].

Details of the error analysis. For false positives (Table 11), we categorize the predicted triplets rejected by *MLLM-as-a-Judge* into correct triplets with incorrect temporal intervals (Temp.), correct

subjects and objects with incorrect relations (Rel.), and other errors such as incorrect entities (Other). For false negatives, we report the fraction of missed ground-truth relations whose informativeness score (Section 3) or duration is below the mean; the two attributes are independent and can overlap. Beyond these statistics, a qualitative audit of SGRet failures shows that traditional models repeatedly emit the same frequent, low-information relations across distinct videos, making their graph embeddings less discriminative, whereas the residual failures of our models more often involve attributes and fine-grained relations outside the dataset taxonomy. Qualitative comparisons of false predictions are shown in Figures 13 and 14.

F Limitation

While this work studies adopting MLLMs for direct DSGG, the current scope mainly focuses on the closed-vocabulary setting. Although our preliminary open-vocabulary results are promising (Table 13), faithfully evaluating open-vocabulary, in-the-wild DSGG requires open-world and expressive DSGG datasets, which we leave for future work. In addition, temporal boundary estimation remains the main error source of MLLM-based DSGG (Table 11) and is hard to address with lightweight finetuning on small-scale DSGG datasets, which motivates improving the dense temporal localization ability of MLLMs. Finally, since the current pipeline mainly considers short video clips (e.g. 8 frames), it will be helpful to extend MLLM-based DSGG to long-video understanding, considering downstream applications of DSGG (e.g. long-video summarization and reasoning).