Synthetic Visual Genome 2: Extracting Large-scale Spatio-Temporal Scene Graphs from Videos

Ziqi Gao[1,2], Jieyu Zhang[1,2], Wisdom Oluchi Ikechukwu[1,2], Jae Sung Park[1], Tario G You[2], Daniel Ogbu[4], Chenhao Zheng[1,2], Weikai Huang[1,2], Yinuo Yang[2], Winson Han[1], Quan Kong[3], Rajat Saini[3], and Ranjay Krishna[1,2]

[1] Allen Institute for AI, Seattle, WA, USA
[2] University of Washington, Seattle, WA, USA
[3] Woven by Toyota, Inc., Tokyo, Japan
[4] Microsoft, Redmond, WA, USA

Abstract. We introduce **Synthetic Visual Genome 2** (**SVG2**), a large-scale panoptic video scene graph dataset. SVG2 contains over 636K videos with 6.6M objects, 52.0M attributes, and 6.7M relations, providing an order-of-magnitude increase in scale and diversity over prior spatio-temporal scene graph datasets. To create SVG2, we design a fully automated pipeline that combines multi-scale panoptic segmentation, online–offline trajectory tracking with automatic new-object discovery, per-trajectory semantic parsing, and GPT-5-based spatio-temporal relation inference. Human verification of SVG2 annotation accuracy confirms its reliability (objects: 93.8%, attributes: 88.3%, relations: 85.4%). Building on this resource, we train TraSeR, a trajectory-grounded video scene graph generation model. TraSeR augments VLMs with a trajectory-aligned token arrangement mechanism and new modules: an object-trajectory resampler and a temporal-window resampler to convert raw videos and panoptic trajectories into compact spatio-temporal scene graphs in a single forward pass. The temporal-window resampler binds visual tokens to short trajectory segments to preserve local motion and temporal semantics, while the object-trajectory resampler aggregates entire trajectories to maintain global context for objects. On PVSG, VIPSeg, VidOR, and $\text{SVG2}_{\text{test}}$, TraSeR outperforms the strongest open-source baselines by 15∼20 points in relation detection, 20∼40 points in object prediction, and 13 points in attribute prediction. It also surpasses GPT-5 by 13 points in object prediction and 3 points in attribute prediction. When TraSeR's generated scene graphs are sent to a VLM for *video* question answering, it delivers a +1.5∼4.6 absolute accuracy gain over using video alone or video augmented with Qwen2.5-VL's generated scene graphs, demonstrating the utility of explicit spatio-temporal scene graphs as an intermediate representation[5].

Keywords: Video scene graph · VLM

[5] The SVG2 data, model checkpoints, and code are available at https://uwgzq.github.io/papers/SVG2/

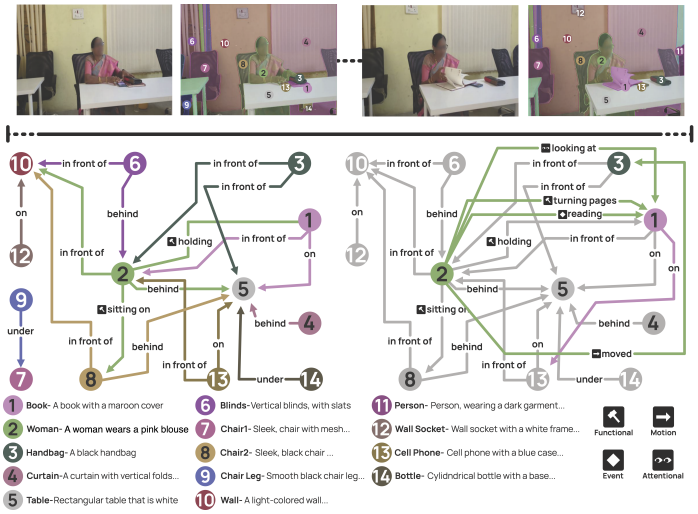

Fig. 1: Synthetic Visual Genome 2 (SVG2), a large-scale synthetic panoptic video scene graph dataset. SVG2 provides dense panoptic trajectories, fine-grained object categories and attributes, and temporally grounded spatiotemporal relations across more than 636K videos, which is an order-of-magnitude increase in scale and diversity over prior datasets.

1 Introduction

Decades of work in human cognition show that people do more than recognize objects in an image; they rapidly extract structured relational information between those objects [5, 6, 19, 24]. Image scene graphs make this structure explicit and have emerged as a core abstraction in computer vision [27, 29]. A scene graph is a structured, graphical representation of a visual scene, where nodes correspond to objects (e.g., person, bicycle) and their attributes (*e.g.* blue, rectangular, old), and directed edges represent their pairwise relationships (*e.g.* riding). This representation, which bridges the gap between vision and language, has proven to be a powerful intermediate representation [9, 14, 16, 20, 23, 64], improving captioning [3], image retrieval [27], image generation [26], vision-language evaluation [22, 36, 63], AI auditing [30], and robotic navigation [28]. Spatio-temporal scene graphs [25] expand this abstraction to video reasoning [10, 15], capturing scene dynamics and parsing events into changes in relationships between objects (*e.g.* riding a bicycle to walking).

Despite their utility, there are very limited high-quality spatio-temporal scene graphs available. Due to the high cost and difficulty of dense frame-level manual annotation, current datasets remain limited in scale and difficult to expand [33]. Attempts to annotate only a handful of frames per video have led to incomplete annotations that do not capture new objects that appear or disappear [20, 25]. As a result, models trained on such datasets tend to overfit to dataset-specific distributions and predicate classifiers. This is exacerbated by the severe long-

tail bias, sparse and inconsistent annotations, noise, and insufficient temporal labeling [39, 48, 53]. Consequently, scene graph prediction models struggle to generalize to new scenes with an open vocabulary of new objects, attributes, and relationships [40, 61].

We address these limitations with **Synthetic Visual Genome 2 (SVG2), a large-scale panoptic synthetic video scene graph dataset**. SVG2 contains over *636K* videos, *6.6M* objects, *52M* attributes, and *6.7M* relations. This dataset is an order-of-magnitude increase in scale and diversity over prior datasets (Table 1). Human verification of SVG2 confirms the reliability and accuracy of its annotations, yielding 93.8% accuracy for objects, 88.3% for attributes, and 85.4% for relations. To create SVG2, we design a holistic, scalable, fully automated pipeline for synthesizing high-quality video scene graphs. Our pipeline unifies multi-scale panoptic segmentation [49], trajectory tracking with automatic new-object discovery, per-trajectory semantic parsing, and GPT-5-based [43] spatio-temporal relation inference into a robust video scene graph generation framework. Our heuristic online-offline two-stage tracking mechanism enables (i) proactively detecting newly emerging instances without human prompts or external detectors and (ii) preserving identity consistency across the entire video.

With SVG2, **we design and train TraSeR, a trajectory-grounded video scene graph parsing model** that converts raw videos and panoptic trajectories into spatio-temporal scene graphs in a single forward pass. Video scene graph parsing is technically challenging for two key reasons. First, object categories are globally stable, but attributes and relations are temporally local and can change rapidly over time. So the model must capture both long-range semantics and fine-grained temporal variation. Second, long videos with many objects produce extremely long token sequences and a combinatorial number of relation candidates, making naive dense encoding and decoding computationally infeasible.

TraSeR addresses these challenges by augmenting VLMs with two modules: a *temporal-window resampler* and an *object-trajectory resampler*. The temporal-window resampler binds visual tokens to individual object trajectory segments, compacting dense patch-level tokens into short, trajectory-aligned, time-aware streams that preserve relationship and attribute changes. The object-trajectory resampler summarizes across the tokens of a specific object's trajectory to encode global object context.

Coupled with strong pretrained VLM priors and training on SVG2, TraSeR generates structured, temporally consistent scene graphs. Our resampler designs translate into clear gains: the object-trajectory resampler yields substantial improvements in object recognition, while the temporal-window resampler drives large gains in relation prediction. Empirically, TraSeR improves relation prediction by 15∼20 percentage points over the strongest open-source models [17, 21, 58], boosts object prediction by 20 to 40 points over open-source baselines [4, 21, 57, 62] and by 13 points over GPT-5 [43], and raises attribute prediction by 13 points relative to open-source state of the art [21, 21, 58].

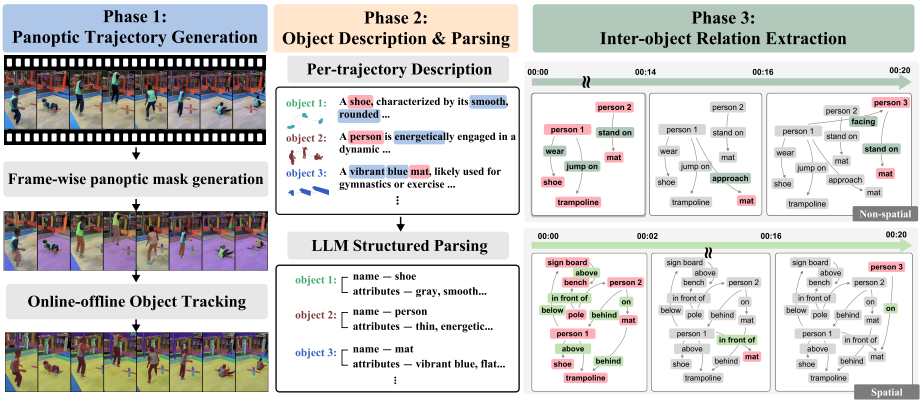

Fig. 2: Overview of SVG2 synthesis pipeline. **Phase 1**: panoptic trajectory generation with online–offline object tracking mechanism that discovers new objects and preserves identity consistency. **Phase 2**: per-trajectory description and semantic parsing. **Phase 3**: GPT-5–based spatiotemporal relation inference to produce the final video scene graph.

Finally, we evaluate the utility of TraSeR's generated scene graphs using video question answering. For this, we use the AGQA dataset [20] and the Perception-Test [45]. When we feed TraSeR's predicted video scene graphs into GPT-4.1, we obtain absolute accuracy gains of +1.5∼4.6 over baselines using video only or video augmented with Qwen2.5-VL's scene graphs. This result suggests that TraSeR is able to surface and summarize useful information that even Qwen2.5-VL misses.

2 Related work

Scene graph datasets. Early scene graph research was enabled by the introduction of Visual Genome [29], which provided large-scale object, attribute, and relational annotations. Follow-up datasets focused on improving annotation quality, coverage, or grounding: VRD [35] offered cleaner predicate labels for relationship detection, Open Images V6 [31] expanded to mask-grounded relations, and PSG [59] introduced panoptic segmentation as the grounding for scene graph nodes. Most recently, Synthetic Visual Genome (SVG) [44] scales image scene graphs to the million-image level and provides dense relational annotations, approximately 4 times more relations per object than Visual Genome. In the video domain, several Video Scene Graph datasets have been introduced [13, 18, 25, 41, 50–52]. However, large-scale datasets with dense and temporally structured annotations remain a challenge [32, 61]. Our work aims to close these gaps by proposing a robust pipeline to create large-scale video scene graph datasets.

Video scene graph generation. Prior efforts in video scene graph generation have attempted to develop unified frameworks capable of constructing temporally

grounded object–relation graphs directly from raw video data [12, 39–41, 47, 52, 54, 61]. Despite these advances, existing approaches exhibit notable limitations. Early methods relying on bounding-box-based scene graphs [39–41, 47, 52] struggle to capture fine-grained spatial details, particularly for non-rigid objects and amorphous background regions, leading to incomplete or imprecise relational reasoning. More recent work [60, 61] addresses this problem by introducing segmentation-and-tracking pipelines approach but fail to achieve dense object relationships and attributes.

3 Synthetic Visual Genome 2

We introduce SVG2, a large-scale synthetic video dataset built upon our newly designed automated VSG generation pipeline. Unlike static image scene graphs, video scene graphs must encode temporal dynamics, capturing both the continuous motion of visual entities and the evolving interactions among them. Building such representations requires solving three core challenges: (1) Consistent and accurate instance-level tracking across the entire video, including object emergence and disappearance. (2) Exhaustive identification of all objects and their attributes. (3) Precise extraction and temporal localization of inter-object relationships, covering both spatial and temporal interactions.

3.1 Automatic pipeline

To address these challenges, we propose a fully automated synthesis pipeline (Fig. 2) that integrates SAM2 [49], Describe Anything (DAM) [34], and GPT-5 [43], producing dense, temporally grounded, and semantically rich video scene graphs. More details are provided in Appx. **??**.

Phase 1: Panoptic trajectory generation. To obtain dense and temporally consistent object trajectories, we leverage SAM2 to produce high-quality panoptic masks on each frame using multi-scale grid prompts. These per-frame masks serve as candidates for object tracking. Because SAM2 cannot proactively detect newly appearing objects and is sensitive to prompt initialization, we introduce a two-stage online–offline tracking framework that achieves both dynamic object discovery and global temporal consistency. In the *online stage*, we propagate initial masks forward while continuously monitoring uncovered regions. When candidate masks cover new or previously untracked areas beyond a threshold, the pipeline identifies object emergence, assigns new object IDs, and resumes propagation. A conservative asymmetric-overlap matching strategy prevents identity switches under occlusion or reappearance. To restore full temporal history, the *offline stage* replays the video once more, reinitializing each object at its first appearance and tracking all instances jointly in a single forward pass. This produces globally stable, complete trajectories. A lightweight post-filtering module removes redundant overlapping tracks and corrects minor mask artifacts (Appx. **??**). At IoU = 0.5, our tracking achieves AR of 0.754 (VIPSeg [37]), 0.686

(PVSG [61]), and 0.623 (VidOR [51]), demonstrating robust performance across diverse benchmarks.

Phase 2: Object description and structured parsing. After obtaining panoptic segmentation trajectories, we generate detailed textual descriptions for each object track using DAM-3B-Video [34]. These descriptions are then processed by GPT-4.1-nano [1] to extract the object name and associated adjectival attributes (e.g., color, shape, etc.). To mitigate synthetic-label noise, we introduce a SAM3-based verification step after Phase 2. We prompt SAM3 [8] with object labels extracted from the structured annotations to obtain SAM3 instance trajectories, then match each original trajectory to SAM3 trajectories using video-level spatiotemporal IoU with a Hungarian assignment. SAM3 is used purely as a verifier: we keep the original masks and discard objects whose labels are not supported by the matched SAM3 trajectories.

Phase 3: Inter-object spatiotemporal relation extraction. We then pass sampled frames along with object IDs, names, and bounding boxes to GPT-5 for inter-object relation inference. Building on prior image scene-graph taxonomies [20, 25], we define a video-adapted relation space covering seven types: spatial (geometric/positional context), functional (contact and manipulation), stateful (persistent attachment or carrying), motion (relative movement over time), social (animate-to-animate interaction), attentional (gaze or camera focus), and event-level (temporally extended, goal-directed dynamics). Considering that spatial relations naturally dominate over non-spatial relations in diverse video environments, we split each video into two batches and query GPT-5 separately for spatial and non-spatial relation extraction. For spatial relations, we suppress trivial 2D cues (e.g., left of, right of) derivable directly from mask layout, prompting GPT-5 to focus on non-trivial relations that require visual evidence.

Human verification. To assess dataset quality, we randomly sample 1.2K object trajectories, 2K attributes and 1.1K relations for manual verification. Human annotators report 93.8% accuracy for objects, 88.3% for attributes and 85.4% for relations. Among relation errors, about 86% stem from predicate errors under occlusion or crowded scenes, and 14% from temporal range errors. We further verify 350 sampled complete triplets and obtain 84.3% accuracy. We find that predicate-correct but subject/object-wrong cases are rare; endpoint and predicate errors are often correlated.

3.2 Extracted dataset

Leveraging our pipeline, we create SVG2, which contains 636K automatically annotated videos, and **SVG2$_{\text{test}}$**, a high-quality expert-annotated diagnostic benchmark of 100 videos (More dataset statistics are provided in Appx. **??**).

SVG2. We sample **43K** videos from SA-V [49] and **593K** videos from PVD [7]. SA-V provides strong manually annotated trajectories, but they are temporally sparse and limited in panoptic completeness. To address this, we introduce a hybrid refinement step during tracking in phase 1: we first propagate the human-annotated masks across all frames, then align and compare the resulting trajectories with those produced by our pipeline. We preserve synthetic masks in

Table 1: Comparison with related benchmarks. Subscripts $_{S}/_{D}$ denote sparse (sampled) vs. dense (per-frame) annotations. 'Cls' indicates the number of categories. (*) AG reports total frame-level relation instances, whereas others report unique trajectory-level instances.

Data	#Vid	Ann.	Type	Frame/ Vid	#Obj/ Traj	Obj Cls	#Rel	Rel Cls	#Attr
SA-V [49]	50.9K	SAM-2 + Human	Seg_S	330	0.6M	-	-	-	-
VIPSeg [37]	2.8K	Human	Seg_S	23	38.2K	124	-	-	-
VidVRD [52]	0.8K	Human	Box_S	304	2.4K	35	25.9K	132	-
AG [25]	7.8K	Human	Box_S	808	26.2K	35	1.4M*	26	-
VidOR [51]	7.0K	Human	Box_D	1K	34.6K	80	0.3M	50	-
PVSG [61]	338	Human	Seg_S	375	6.3K	125	3.6K	62	-
$\text{VIPSeg}_{\text{val}}$ [37]	343	Human	Seg_S	23	4.7K	124	-	-	-
$\text{VidVRD}_{\text{test}}$ [52]	0.2K	Human	Box_D	258	0.6K	35	4.8K	132	-
AG_{test} [25]	1.8K	Human	Box_S	679	7.9K	35	0.5M*	26	-
$\text{VidOR}_{\text{val}}$ [51]	835	Human	Box_D	1K	4.0K	80	30.1K	50	-
PVSG_{val} [61]	62	Human	Seg_S	365	1.3K	108	1.0K	58	-
SVG2 (Ours)	**636K**	Our pipeline	Seg_D	479	**6.6M**	**54.2K**	**6.7M**	**35.3K**	**52M**
$\text{SVG2}_{\text{test}}$(Ours)	100	Human	Seg_D	160	3.2K	749	3.3K	249	9.7K

regions that lack human annotations. This fusion procedure allows us to retain the precision of manual labels while significantly expanding coverage to achieve dense panoptic trajectories. We then apply our attribute and relation extraction stages, resulting in **0.66M** object instances, **5.01M** attributes and **0.72M** spatiotemporal relations. For the **593K** sampled PVD subset, we operate the pipeline fully automatically. We then generate **5.89M** object instances, **46.95M** attributes and **5.99M** spatiotemporal relations. This large-scale synthetic stage substantially broadens the diversity of SVG2 (Tab. 1).

$\text{SVG2}_{\text{test}}$. Existing benchmarks support only closed-set object and relation evaluation [51,61], with no benchmark jointly assessing objects, attributes and relations over an open vocabulary. We therefore construct $\text{SVG2}_{\text{test}}$, a high-quality expert-annotated diagnostic benchmark. We select 100 diverse videos from SA-V [49] and VSPW [38]. Reflecting how humans perceive scenes hierarchically, annotators follow a multi-granularity panoptic protocol: coarse instance masks are produced first, then decomposed into salient parts (e.g., a bicycle into frame, wheels, and handlebars), with open-vocabulary categories, multi-class attributes, and time-localized relations assigned and consistency-verified throughout. As shown in Tab. 1, this multi-stage, hierarchical labeling process yields substantially denser and more diverse per-video annotations than existing splits (with vocabularies 4× to 6× larger and explicit multi-class attributes), enabling evaluation of fine-grained compositional VSG prediction beyond closed-set graphs.

4 TraSeR

The integration of multiple VLMs [34,42,43] in our pipeline ensures high-quality trajectory-level annotations but incurs considerable computational overhead and complexity. Yet, existing VLMs lack native support for panoptic mask trajectory conditioning (typically trained on bbox/point data). To this end, we propose **TraSeR**, a trajectory-grounded VLM that produces structured video scene

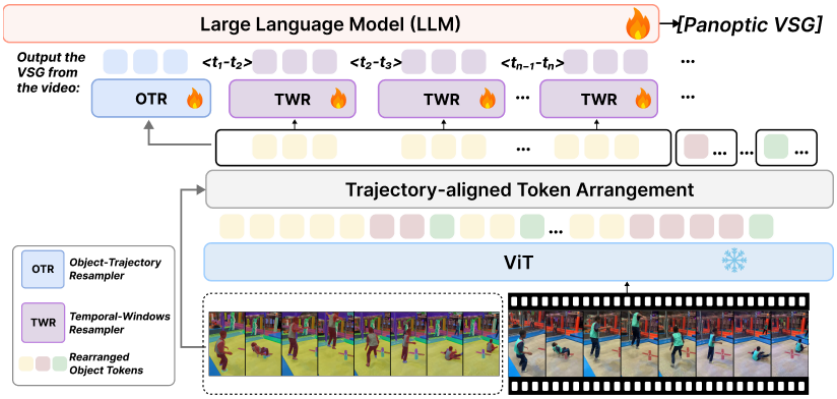

Fig. 3: TraSeR architecture. TraSeR first performs **trajectory-aligned token arrangement**, grounding ViT tokens to instance trajectories to form identity-preserving token streams. It then applies an **object-trajectory resampler** to aggregate global semantics over each full trajectory, and a **temporal-window resampler** to preserve fine-grained motion and temporal cues. The resulting tokens are decoded by the language model into a structured video scene graph. TraSeR predicts object labels, attributes, relations, and triplets given object trajectories; it does not jointly discover and track objects from raw video.

graphs in one pass from raw videos and panoptic object trajectories (Figure 3). Unlike free-form video-language tasks (*e.g.* VideoQA), VSG generation requires accurate alignment between visual semantics and spatio-temporal scene structure, structured prediction of hierarchical graph elements (objects, attributes, relations) rather than natural language, and fine-grained localization with temporal reasoning over video patches to capture interactions, motion, and state changes. To address these challenges, we first bind video patches directly to object trajectories via trajectory-aligned token arrangement, restoring and preserving consistent object identities over time. We then compress aligned tokens through a dual resampler module: an object-trajectory resampler aggregates global semantic context, and a temporal-window resampler retains fine-grained local motion and temporal dynamics.

4.1 Trajectory-Aligned Token Arrangement

We first arrange visual tokens based on object trajectory. Given the sampled frame at time t, we define $M_{o,t} \in \{0,1\}^{H\times W}$ as the binary segmentation mask for object o. In ViT, each output token corresponds to a spatiotemporal region obtained by merging g consecutive frames and an $(m \times m)$ grid of base ViT patches (each of size $P \times P$ pixels). Thus, a token indexed by (t_g, h_m, w_m) covers a pixel volume of size $(g \times mP \times mP)$ in the original video. We compute a coverage score $\mathcal{C}$ for object o at token (t_g, h_m, w_m) by averaging the mask values over the exact pixel support of each token, followed by a temporal max over the

frames:

$$\mathcal{C}_{o,t_g,h_m,w_m} = \max_{k\in[0,g-1]} \left(\text{avgpool}_{mP}(M_{o,t_g\cdot g+k})[h_m, w_m]\right), \tag{1}$$

where avgpool$(\cdot, mP)$ denotes 2D average pooling on the pixel-level mask. The kernel size and stride are set to $(m \times P)$, matching the token's exact spatial pixel footprint. Tokens are assigned to object o if their coverage score exceeds a threshold $\tau_{\text{eff}} \in [0, 1]$:

$$I_o = \left\{(t_g, h_m, w_m) \;\middle|\; \mathcal{C}_{o,t_g,h_m,w_m} \geq \tau_{\text{eff}}\right\}. \tag{2}$$

This yields a deterministic set of object-associated visual tokens grounded directly by segmentation evidence. After obtaining the set of object-associated tokens, we temporally sort them according to their t_g indices to form an ordered token trajectory for each object. We concatenate these trajectories across all objects, and introduce a special [TRJ] token to separate objects and explicitly encode trajectory boundaries. This arrangement produces a structured token stream where each object's trajectory is explicitly segmented and identity-conditioned, providing the LLM with clear object grounding and temporal continuity priors before decoding.

4.2 Dual Resampler Module

After trajectory-aligned token arrangement, each object o yields a variable-length token set $X_o \in \mathbb{R}^{|I_o|\times D_{\text{in}}}$. Directly feeding all tokens into a language model is computationally prohibitive and can bias attention toward densely sampled frames and repeatedly selected tokens, which affects stable training. To obtain compact yet semantically faithful representations, we introduce two resampler modules that progressively aggregate tokens from pixel-level to object-level and temporal-window-level summaries.

Object-trajectory Resampler. Given X_o, we introduce M learnable latent queries $L \in \mathbb{R}^{M\times D_{\text{latent}}}$ and resample all arranged tokens of object o using a Perceiver-Resampler [2] f_θ:

$$Z_o = f_\theta(X_o, L), \qquad Z_o \in \mathbb{R}^{M\times D_{\text{out}}}. \tag{3}$$

In this stage, Z_o captures the global semantics of object o over its entire temporal span.

Temporal-window Resampler. While object-Level resampler reduces token count and abstracts global semantics for each video, it may lose time sensitivity, which is crucial for capturing per-object dynamics and inter-object temporal relations. We therefore partition the video into disjoint windows $\{w\}$ of length Δt. Let $X_o^{(w)}$ be the subset of tokens of object o that fall into window w. Using another group of M' latent queries L', we resample each window of each object independently:

$$Z_o^{(w)} = f_{\theta'}(X_o^{(w)}, L'), \qquad Z_o^{(w)} \in \mathbb{R}^{M'\times D_{\text{out}}}. \tag{4}$$

If object o is absent in w, no summary is produced for that window, yielding a presence-adaptive, variable-length sequence across time. To expose temporal positions explicitly to the language model, we also add timestamp embeddings before each temporal resampled token. Finally, the output for object o concatenates the global summary and all window summaries in temporal order:

$$\mathcal{Z}_o = [Z_o; Z_o^{(1)}; Z_o^{(2)}; \dots]. \tag{5}$$

The two resamplers are complementary: object-trajectory resampler provides global object-centric semantics, while temporal-window resampler restores temporal granularity. The resulting representation $\mathcal{Z}_o$ for each object is compact, object-aware, and time-aware, being well-suited for structured decoding in the language model.

4.3 Training

Training datasets. We train TraSeR by combining our synthetic SVG2, which provides complete scene-graph supervision, with established human-annotated benchmarks. Specifically, we integrate video instance segmentation datasets (LV-VIS [55], OVIS [46], VIPSeg [37]) for robust temporal grounding under occlusion, and video relation datasets (VidOR [51], VidVRD [52]) to enrich relational reasoning. To ensure architectural compatibility, we use SAM2 to convert bounding box annotations from VidOR and VidVRD into segmentation masks. Finally, we unify these heterogeneous datasets using task-specific templates that serialize annotations into a standardized autoregressive format, effectively blending dense synthetic structures with complex real-world dynamics.
Implementation Details. We build on the Qwen2.5-VL-3B and set τ_{eff}=0.5 during trajectory-aligned token arrangement. In the resampling stage, we use two three-layer Perceiver-Resamplers with 32 learnable latent queries each. We finetune the model for one epoch (37K steps), freezing the Qwen2.5-ViT backbone to preserve pretrained visual priors while jointly training the vision–language projector, object-trajectory and temporal-window resamplers, and the language model with learning rates of 5×10^{-5}, 1×10^{-4}, and 2×10^{-5}, respectively. This schedule enables the new resampling modules to adapt more aggressively without destabilizing pretrained components. Models are trained on a single node with 8×H100 GPUs.

5 Experiments

We compare our model against various proprietary and open-source VLM baselines using a standardized setup in which all models receive identical object trajectories and structured prompting templates. We evaluate open-vocabulary video scene graph generation, assessing object, attribute, relation, and triplet prediction across multiple academic benchmarks and our $SVG2_{test}$. We introduce an LLM-based judge that enables open-vocabulary evaluation through semantic, hierarchy-aware

Table 2: We evaluate VLMs on four VidSGG benchmarks using panoptic object trajectories. We report performance across four metrics: triplet recall, relation recall, object accuracy, and attribute recall. The results in this table are reported under the **lenient** semantic criterion, with a temporal IoU threshold of 0.5 for relations and triplets. TraSeR outperforms all open-source baselines and surpasses GPT-5 in object and attribute prediction.

	Triplet			Relation			Object				Attribute	
Model	PVSG	VidOR	$SVG2_{test}$	PVSG	VidOR	$SVG2_{test}$	VIPSeg	PVSG	VidOR	$SVG2_{test}$	$SVG2_{test}$	**AvgRank**
API call only												
GPT-4.1 [1]	6.0	10.8	6.4	7.3	11.9	7.5	59.9	51.4	86.6	58.5	15.8	3
Gemini-2.5-PRO [11]	7.4	9.8	8.7	8.8	11.0	9.9	56.7	31.7	82.9	49.8	13.6	4
GPT-5 [43]	**16.6**	19.7	**17.9**	**18.3**	21.7	**19.4**	68.1	54.2	88.5	65.5	24.1	2
Open weights only												
Qwen2.5-VL-3B [4]	0.1	0.2	0.2	0.1	0.4	0.3	22.1	10.4	45.0	24.2	1.4	12
MiniCPM-V 4.5 [62]	0.1	3.0	1.1	0.2	4.0	2.4	40.0	14.3	59.1	38.5	8.4	8
InternVL3.5-4B [57]	0.2	0.4	0.1	0.3	0.5	0.2	33.7	20.4	66.4	35.0	7.4	11
GLM-4.1-9B-Thinking [17]	0.3	3.9	1.8	0.5	5.0	2.9	46.5	17.8	61.1	28.5	9.1	6
Qwen3-VL-4B [58]	0.1	0.7	1.4	0.1	0.8	1.6	34.1	21.8	65.8	35.6	8.3	10
Qwen3-VL-4B-Thinking [58]	0.1	2.3	3.3	0.4	3.4	3.6	35.8	18.3	67.6	37.1	8.8	7
Fine-tuned baselines												
FT-Qwen2.5-VL-3B (First Bbox)	0.1	1.6	0.1	0.9	4.5	0.5	25.5	25.9	51.7	36.7	10.4	9
FT-Qwen2.5-VL-3B (Bbox Traj.)	0.5	1.8	1.4	1.6	4.2	3.0	35.1	33.6	56.9	46.1	13.4	5
TraSeR	16.1	**22.9**	16.7	16.9	**25.0**	18.7	**86.5**	**72.7**	**91.4**	**79.0**	**27.1**	1

matching, overcoming the closed-set constraints of traditional metrics. Finally, we perform extensive ablations on training data composition and architectural components to isolate their individual contributions.

Baselines. We compare TraSeR with leading proprietary and open-source VLMs, including Gemini-2.5-Pro [11], GPT-4.1 [1], GPT-5 [43], Qwen-3-VL-4B [56], InternVL-3.5-4B [57], GLM-4.1-9B-Thinking [17], and MiniCPM-4.5 [62]. Since existing VLMs cannot effectively process panoptic mask trajectories, we supply the baselines with dense bounding box trajectories to maximize their supported spatial-temporal context. We construct instruction-aligned prompting templates for baselines and require all models to output structured scene-graph predictions following a JSON schema. To disentangle the benefits of our architecture from the data effects, we fine-tune Qwen2.5-VL-3B on either first-appearance or full bounding-box trajectories, determining whether standard fine-tuning alone is sufficient to generate high-quality VSGs.

Evaluation Setup. We evaluate video scene graph generation under the open-vocabulary (OV-VSGGen) setting, where the model generates a complete scene graph from ground-truth object trajectories. This entails four prediction sub-tasks: (1) object (labeling each trajectory), (2) attribute (assigning associated characteristics to objects), (3) relation (detecting interactions between trajectories), and (4) triplet (jointly identifying subject-relation-object pairs).

Table 3: Validation of the LLM semantic aligner. We report Cohen's κ scores between GPT-4o-mini, Gemini 3 Flash and human annotators. The high agreement confirms the reliability of the automated aligner. "Id+Syn" denotes *Identical* or *Synonym* matches; "Lenient" includes all valid semantic categories.

	Object		Relation	
	Id+Syn	**Lenient**	**Id+Syn**	**Lenient**
Human vs. **GPT-4o-mini**	0.901	0.877	0.842	0.791
Gemini 3 Flash vs. **GPT-4o-mini**	0.846	0.724	0.750	0.765

Table 4: Model architecture ablation study. Combining the object-trajectory and temporal-window resamplers yields the best VSG generation performance. Segmentation-aligned token arrangement further provides superior grounding compared to bbox-based signal. **(*)**Variants without resamplers cannot support long-video inference; therefore, PVSG results are reported on the PVSG VidOR subset only.

Variants			**Triplet**		**Relation**		**Object**	
Temporal-window resampler	Object-trajectory resampler	Supervision	PVSG*	VidOR	PVSG*	VidOR	PVSG*	VidOR
✗	✗	Seg.	7.03	13.56	7.15	14.58	73.34	90.79
✗	✓	Seg.	2.44	8.23	2.68	9.04	77.10	89.62
✓	✗	Seg.	8.52	15.00	8.63	16.78	77.84	89.76
✓	✓	Bbox	8.69	15.01	9.39	15.19	73.05	89.76
✓	✓	Seg.	10.14	15.02	11.38	16.30	78.21	90.15

We evaluate on PVSG [61], VidOR [51], VIPSeg [37], and our $\text{SVG2}_{\text{test}}$. While VIPSeg is limited to object prediction and PVSG/VidOR focus on objects and relations, Our $\text{SVG2}_{\text{test}}$ provides complete object–attribute–relation annotations, enabling holistic VidSGG evaluation.

Since VLMs perform open-vocabulary autoregressive generation, traditional closed-set metrics such as Recall@k become inadequate. Standard metrics may unfairly penalize valid predictions due to lexical diversity (e.g., human vs. person). Because curating a manual synonym dictionary is impossible for unbounded vocabularies, we introduce a two-tiered evaluation framework. First, we compute a rule-based **strict score** requiring exact string matches between ground-truth and predicted elements. Second, to accurately reflect the model's true semantic understanding without unfairly penalizing language diversity, we compute a **lenient score** using an LLM-assisted semantic aligner. To prevent the self-referential bias of free-form grading, we strictly constrain GPT-4o-mini [42] to act purely as a pairwise lexical matcher. Guided by a rigidly structured prompt, it classifies ground-truth and predicted elements into five categories: *Identical*, *Synonym*, *Hypernym/Hyponym*, *Semantic Overlap*, or *Mismatch*. Any category other than *Mismatch* is considered a correct prediction in lenient settings. This dual-metric framework applies to objects, attributes, and relations. Relations additionally require strict temporal grounding (interval IoU $>$ threshold). Triplets are only considered correct if the subject, object, and relation all meet the semantic match criteria and the temporal intersection rules. This evaluation approach ensures a rigorous, bias-mitigated assessment of open-vocabulary scene graphs.

To validate our automated aligner (GPT-4o-mini), we assess its alignment with human evaluators and an independent model (Gemini 3 Flash). Evaluating a stratified subset of 250 objects and 150 relations, we observe substantial inter-annotator agreement (Cohen's κ) between the judge and humans (Tab. 3). This confirms that our constrained matching approach effectively mitigates hallucination, establishing the automated metric as a highly reliable proxy for human judgment.

Table 5: Training-data ablation study. Incorporating SVG2 improves triplet and relation prediction, and boosts object accuracy and attribute recall over academic-only training. Training with the full configuration achieves the best overall performance.

Training	Triplet			Relation			Object				Attr.	Rank
	PVSG	VidOR	$SVG2_t$	PVSG	VidOR	$SVG2_t$	VIPSeg	PVSG	VidOR	$SVG2_t$	$SVG2_t$	
Acad	1.35	13.61	2.68	12.55	28.66	18.74	32.71	21.50	61.09	25.68	–	4
SA-V	8.29	15.84	13.58	9.86	17.12	16.39	81.61	72.92	90.52	77.64	25.96	3
Acad+SA-V	10.70	19.23	13.00	11.12	21.06	14.69	83.10	73.46	90.96	78.14	22.72	2
Acad+SA-V+PVD	16.91	22.87	16.63	16.13	25.01	18.74	86.53	72.74	91.42	79.01	27.10	1

5.1 Scene graph generation results.

Consistent with our evaluation framework, Table 2 reports model performance under the **lenient** semantic criterion with a temporal IoU threshold of 0.5. Results under the **strict** exact-match score and a relaxed IoU threshold are detailed in Appx. **??**. Our light-weight model achieves strong improvements across all benchmarks (Tab. 2), outperforming open-source baselines by substantial margins, approximately +**15** points in triplet detection, +**15** points in relation detection, +**25** points in object prediction, and +**13** points in attribute recognition. TraSeR also surpasses proprietary models such as GPT-4.1 and Gemini-2.5-Pro, and even outperforms GPT-5 on object and attribute prediction, highlighting the effectiveness of our model design and training data. We further observe a pronounced performance gap between open-source and proprietary VLMs, especially on longer and more complex videos such as those in PVSG that require precise temporal localization of dynamic relations. This underscores the importance of our large-scale, high-quality open video scene graph dataset (SVG2) that provides the rich temporal and dense object–attribute–relation annotations. Both fine-tuned baselines still underperform TraSeR, confirming that our performance gains stem from our object-centric temporal design rather than mere task-specific fine-tuning. Standard fine-tuning alone is insufficient to capture temporal scene dynamics. As detailed in Appx. **??**, TraSeR maintains its substantial lead over all baselines even when evaluated under the rule-based strict score and a lower IoU threshold.

5.2 Video question answering with scene graphs.

We further assess the role of structured video scene graphs in video QA. We sample 1K open-ended questions from the AGQA 2.0 benchmark [20] and 500 multiple-choice questions from the Perception-Test benchmark [45]. For each associated video, we first use our pipeline (Phase 1) to generate panoptic object trajectories, and then predict the corresponding video scene graphs using both our trained TraSeR and the baseline Qwen2.5-VL-3B model.

During testing, we sample video frames at 1 fps and feed them into GPT-4.1 under three conditions: (1) video-only input, (2) video plus scene graphs generated by Qwen2.5-VL, and (3) video plus scene graphs generated by our TraSeR. The results, summarized in Tab. 6, show that incorporating high-quality video scene graphs (*e.g.* those generated by TraSeR) reliably improves VideoQA accuracy (+0.4 on AGQA 2.0 and +4.6 on Perception Test).

We also conduct a blind-video evaluation on AGQA. We instruct a blind LLM (GPT-4.1-mini) to answer questions using *only* the text-based scene graph, without visual input. TRASER's graphs achieve 13.22% accuracy, outperforming Qwen2.5-VL's graphs by +6.5%. We attribute this gain to the structured and fine-grained spatiotemporal information captured by the video scene graphs, which complements the raw visual input and allows the VLM to reason more effectively.

Table 6: GPT-4.1 VQA performance on AGQA 2.0 and Perception-Test. Using TRASER's video scene graphs consistently improves performance over baselines.

Input	AGQA 2.0	Perception-Test
video-only	25.9	66.8
Video + Qwen2.5-VL-3B's VSG	24.8	68.6
Video + TRASER's VSG	**26.3**	**71.4**

5.3 Ablation Study

Training-data ablations. We compare four training configurations: academic datasets only (Acad) (Sec. 4.3), the SVG2 SA-V subset (SA-V), their combination (Acad + SA-V), and the full setup adding the PVD subset (Acad + SA-V + PVD). Since academic datasets lack attribute annotations, we evaluate models trained on them only for object and relation outputs. As shown in Tab. 5, the full configuration achieves the best overall performance. Training solely on Acad overfits to specific benchmarks like VidOR. This results in poor generalization to PVSG and $\text{SVG2}_{\text{test}}$, alongside low object accuracy. Incorporating SA-V and PVD yields substantial gains, demonstrating that our dataset's dense, temporally aligned annotations, diverse categories, and visual variability are crucial for learning structured scene graphs.

Model ablations. Table 4 reports ablations on our architectural choices, trained for three epochs on the SVG2 SA-V subset. Using only the object-trajectory resampler eases model training but collapses temporal grounding, while the temporal-window resampler is essential for accurate dynamic relation localization. Combining both yields the best balance of object-level and temporal alignment. We also compare supervision signals for token arrangement: replacing segmentation trajectories with bounding boxes significantly degrades performance on PVSG, as longer, more dynamic videos make coarse box signals too noisy for stable grounding.

6 Limitations and Discussion

Dual-use considerations. Large-scale spatio-temporal scene graph extraction can benefit video retrieval, accessibility, robotics, and structured video understanding, but may also be repurposed for surveillance, behavioral analysis, or monitoring interactions at scale. Such risks are especially relevant for videos containing identifiable individuals or sensitive activities, even without explicit

face recognition or biometric inference. The broader impact of this technology therefore depends on the provenance of the input videos, the deployment context, and the governance of the resulting structured representations.

Limitations and future work. While SVG2 inherits the limitations of the automated models used in its construction, this approach unlocks a scale, density, and consistency that is practically unattainable through manual annotation. Importantly, it initiates a virtuous cycle: as foundation models continue to improve, they can be continuously leveraged to refine and elevate future iterations of the dataset. Architecturally, TraSeR is currently designed as a trajectory-grounded model. A compelling direction for future work is to extend this approach into a unified, fully end-to-end framework capable of natively proposing and tracking objects alongside VSG generation. Finally, spatio-temporal scene graphs remain a foundational representation for complex video reasoning. With the introduction of SVG2, we see a promising opportunity to integrate large-scale, structured VSG data directly into the training pipelines of general VLMs, potentially unlocking much deeper compositional and temporal understanding.

7 Conclusion

We present **Synthetic Visual Genome 2** (**SVG2**), a large-scale synthetic panoptic video scene graph dataset of over 636K videos with dense object, attribute, and relation annotations. Built entirely through an automated pipeline, SVG2 expands the scale of prior resources by an order of magnitude. Leveraging this dataset, we introduce TraSeR, which augments a VLM with object-trajectory and temporal-window resamplers to produce complete video scene graphs in a single forward pass.

Acknowledgements

This work is funded by Toyota Motor, Inc.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 6, 11
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022), https://arxiv.org/abs/2204.14198 9
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European conference on computer vision. pp. 382–398. Springer (2016) 2

4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923 3, 11
5. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychological review **94**(2), 115 (1987) 2
6. Biederman, I.: On the semantics of a glance at a scene. In: Perceptual organization, pp. 213–253. Routledge (2017) 2
7. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, D., Dollár, P., Feichtenhofer, C.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv:2504.13181 (2025) 6
8. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), https://arxiv.org/abs/2511.16719 6
9. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: Scene graphs: A survey of generations and applications. arXiv preprint arXiv:2104.01111 **2** (2021) 2
10. Cherian, A., Hori, C., Marks, T.K., Le Roux, J.: (2.5+ 1) d spatio-temporal scene graphs for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 444–453 (2022) 2
11. Comanici, G., Bieber, E., Schaekermann, M., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261 11
12. Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., Yang, M.Y.: Spatial-temporal transformer for dynamic scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16372–16382 (October 2021) 5
13. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations (2022), https://arxiv.org/abs/2209.13064 4
14. Farshad, A., Yeganeh, Y., Chi, Y., Shen, C., Ommer, B., Navab, N.: Scenegenie: Scene graph guided diffusion models for image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 88–98 (2023) 2
15. Gao, K., Chen, L., Niu, Y., Shao, J., Xiao, J.: Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19497–19506 (2022) 2
16. Gao, Z., Huang, W., Zhang, J., Kembhavi, A., Krishna, R.: Generate any scene: Scene graph driven data synthesis for visual generation training (2025), https://arxiv.org/abs/2412.08221 2
17. GLM, T., Zeng, A., Xu, B., Wang, B., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools (2024), https://arxiv.org/abs/2406.12793 3, 11
18. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Others: Ego4d: Around the world in 3,000 hours of egocentric video (2022), https://arxiv.org/abs/2110.07058 4

19. Greene, M.R.: Scene perception and understanding (03 2023). https://doi.org/10.1093/acrefore/9780190236557.013.883, https://oxfordre.com/psychology/view/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-883 2
20. Grunde-McLaughlin, M., Krishna, R., Agrawala, M.: Agqa: A benchmark for compositional spatio-temporal reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11287–11297 (2021) 2, 4, 6, 13
21. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025) 3
22. Hsieh, C.Y., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. Advances in neural information processing systems **36**, 31096–31116 (2023) 2
23. Huang, W., Zhang, J., Jia, T., Zheng, C., Gao, Z., Park, J.S., Krishna, R.: Sos: Synthetic object segments improve detection, segmentation, and grounding (2025), https://arxiv.org/abs/2510.09110 2
24. Hummel, J.E., Biederman, I.: Dynamic binding in a neural network for shape recognition. Psychological review **99**(3), 480 (1992) 2
25. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10236–10247 (2020) 2, 4, 6, 7
26. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018) 2
27. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015) 2
28. Kim, J., Oh, M., Myung, H.: Tacs-graphs: Traversability-aware consistent scene graphs for ground robot indoor localization and mapping. arXiv preprint arXiv:2506.14178 (2025) 2
29. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016), https://arxiv.org/abs/1602.07332 2, 4
30. Künzel, S., Munz-Körner, T., Tilli, P., Schäfer, N., Vidyapu, S., Thang Vu, N., Weiskopf, D.: Visual explainable artificial intelligence for graph-based visual question answering and scene graph curation. Visual Computing for Industry, Biomedicine, and Art **8**(1), 9 (2025) 2
31. Kuznetsova, A., et al.: The open images dataset v6. International Journal of Computer Vision (2020) 4
32. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Video k-net: A simple, strong, and unified baseline for video segmentation (2022), https://arxiv.org/abs/2204.04656 4
33. Li, Y., Li, Z., Chen, H., Xu, L.: Unbiased video scene graph generation via visual and semantic dual debiasing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19047–19056 (2025) 2
34. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video captioning. arXiv preprint arXiv:2504.16072 (2025) 5, 6, 7
35. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (ECCV) (2016) 4

36. Ma, Z., Hong, J., Gul, M.O., Gandhi, M., Gao, I., Krishna, R.: Crepe: Can vision-language foundation models reason compositionally? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10910–10921 (2023) 2
37. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21033–21043 (June 2022) 5, 7, 10, 12
38. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4133–4143 (June 2021) 7
39. Nag, S., Min, K., Tripathi, S., Chowdhury, A.K.R.: Unbiased scene graph generation in videos (2023), https://arxiv.org/abs/2304.00733 3, 5
40. Nguyen, T.T., Nguyen, P., Cothren, J., Yilmaz, A., Luu, K.: Hyperglm: Hypergraph for video scene graph generation and anticipation (2025), https://arxiv.org/abs/2411.18042 3, 5
41. Nguyen, T.T., Nguyen, P., Luu, K.: Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding (2024), https://arxiv.org/abs/2312.03050 4, 5
42. OpenAI: Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (2024), accessed: 2025-02-19 7, 12
43. OpenAI: Gpt-5 system card. Tech. rep., OpenAI (Aug 2025), https://cdn.openai.com/gpt-5-system-card.pdf 3, 5, 7, 11
44. Park, J.S., Ma, Z., Li, L., Zheng, C., Hsieh, C.Y., Lu, X., Chandu, K., Kong, Q., Kobori, N., Farhadi, A., Choi, Y., Krishna, R.: Synthetic visual genome (2025), https://arxiv.org/abs/2506.07643 4
45. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems **36**, 42748–42761 (2023) 4, 13
46. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H.S., Bai, S.: Occluded video instance segmentation: A benchmark (2021), https://arxiv.org/abs/2102.01558 10
47. Qian, X., Zhuang, Y., Li, Y., Xiao, S., Pu, S., Xiao, J.: Video relation detection with spatio-temporal graph. In: Proceedings of the 27th ACM international conference on multimedia. pp. 84–93 (2019) 5
48. Qiu, Y., Nagasaki, Y., Hara, K., Kataoka, H., Suzuki, R., Iwata, K., Satoh, Y.: Virtualhome action genome: A simulated spatio-temporal scene graph dataset with consistent relationship labels. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3351–3360 (2023) 3
49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 3, 5, 6, 7
50. Rodin, I., Furnari, A., Min, K., Tripathi, S., Farinella, G.M.: Action scene graphs for long-form understanding of egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18622–18632 (2024) 4

51. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. p. 279–287 (2019) 4, 6, 7, 10, 12
52. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: Proceedings of the 25th ACM International Conference on Multimedia. pp. 1300–1308 (2017) 4, 5, 7, 10
53. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020) 3
54. Teng, Y., Wang, L., Li, Z., Wu, G.: Target adaptive context aggregation for video scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13688–13697 (October 2021) 5
55. Wang, H., Yan, C., Wang, S., Jiang, X., Tang, X., Hu, Y., Xie, W., Gavves, E.: Towards open-vocabulary video instance segmentation (2023), https://arxiv.org/abs/2304.01715 10
56. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution (2024), https://arxiv.org/abs/2409.12191 11
57. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), https://arxiv.org/abs/2508.18265 3, 11
58. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 3, 11
59. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation (2022), https://arxiv.org/abs/2207.11247 4
60. Yang, J., Cen, J., Peng, W., Liu, S., Hong, F., Li, X., Zhou, K., Chen, Q., Liu, Z.: 4d panoptic scene graph generation. Advances in Neural Information Processing Systems **36**, 69692–69705 (2023) 5
61. Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C.C., Liu, Z.: Panoptic video scene graph generation (2023), https://arxiv.org/abs/2311.17058 3, 4, 5, 6, 7, 12
62. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Minicpm-v: A gpt-4v level mllm on your phone (2024), https://arxiv.org/abs/2408.01800 3, 11
63. Zhang, J., Huang, W., Ma, Z., Michel, O., He, D., Gupta, T., Ma, W.C., Farhadi, A., Kembhavi, A., Krishna, R.: Task me anything. ArXiv **abs/2406.11775** (2024), https://api.semanticscholar.org/CorpusID:270560772 2
64. Zhang, J., Xue, L., Song, L., Wang, J., Huang, W., Shu, M., Yan, A., Ma, Z., Niebles, J.C., Savarese, S., Xiong, C., Chen, Z., Krishna, R., Xu, R.: Provision: Programmatically scaling vision-centric instruction data for multimodal language models (2024), https://arxiv.org/abs/2412.07012 2