MULTI-MODAL OPEN-VOCABULARY VIDEO VISUAL RELATIONSHIP DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary Video Visual Relationship Detection (OV-VidVRD) aims to detect and localize subject-predicate-object relationship triplets in videos, even for relationships unseen during training. While pivotal for comprehensive video understanding, existing methods are largely confined to the visual modality. This uni-modal focus neglects rich contextual cues from other modalities, which are often vital for disambiguating complex relationships in real-world videos. To bridge this gap, we introduce the new task of Multi-modal Open-vocabulary Video Visual Relationship Detection (MM-OV-VidVRD) and present the first-of-its-kind benchmark VaM-VidVRD to catalyze research in this area. Our benchmark features numerous videos where each instance pairs the visual stream with a heterogeneous set of auxiliary modalities, such as audio, textual descriptions, and 3D data. Importantly, the benchmark encompasses many predicate categories that are visually indistinguishable yet semantically distinct—for instance, *sing_to* and *speak_to* share nearly identical visual appearances but are readily separable given audio cues—making auxiliary modalities essential rather than merely supplementary. This design mirrors the diverse and source-dependent nature of real-world multi-modal data. To tackle this new challenge, we propose a novel and flexible framework that effectively processes arbitrary combinations of modalities. The core of this framework is a Multi-modal Synergistic Prompting (MMSP) mechanism, which learns to dynamically align and fuse features from diverse sources into a unified and relationship-aware representation. Extensive experiments on our benchmark show that our method significantly outperforms vision-only approaches.

1 INTRODUCTION

Open-vocabulary Video Visual Relationship Detection Gao et al. (2023); Yang et al. (2024); Wang et al. (2025b) (OV-VidVRD) represents a key frontier in achieving a fine-grained and structured comprehension of dynamic scenes. It aims to detect and localize subject-predicate-object triplets by moving beyond predefined categories to include relationships unseen during training. This open-vocabulary capability promises a scalable and practical approach to interpreting the immense complexity of real-world visual data. Such a deep understanding of dynamic interactions is a core requirement for a spectrum of real-world applications, including autonomous driving Weng et al. (2024); Zhang et al. (2024); Yang et al. (2025), robotics Ni et al. (2024); Tang et al. (2025), and video moderation Tang et al. (2021); Lee et al. (2024).

However, the current paradigm for OV-VidVRD is fundamentally limited by its uni-modal nature, relying exclusively on the visual stream. This forces models to infer complex interactions from visual cues alone, which are often ambiguous. Many predicate categories are visually indistinguishable yet semantically distinct—for instance, *person-sing/speak_to-person* share nearly identical appearances but are readily separable given audio cues, while spatial predicates such as *above* and *in_front_of* become unambiguous only with 3D geometric information. By ignoring these complementary signals, current methods are inherently limited to a superficial understanding of the events unfolding in a video.

To address this limitation, we introduce the task of Multi-modal Open-Vocabulary Video Visual Relationship Detection (MM-OV-VidVRD), which requires detecting and localizing open-vocabulary

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

subject-predicate-object triplets by fusing the visual stream with available auxiliary modalities—audio, textual descriptions, or 3D data—to resolve ambiguities. Throughout this paper, we use *auxiliary modalities* to refer to the non-visual input signals, and reserve *multi-modal* for the combined setting that integrates vision with one or more auxiliary modalities.

To catalyze research on this task, we introduce the Variable Multi-modal VidVRD (VaM-VidVRD) benchmark. Its design is driven by two observations: (1) many predicates are visually ambiguous and require specific auxiliary modalities for reliable recognition, and (2) available modalities are inherently source-dependent—e.g., autonomous driving provides 3D LiDAR, while video moderation offers textual descriptions. Accordingly, VaM-VidVRD encompasses a rich set of visually ambiguous predicate pairs that demand auxiliary modal cues for disambiguation, and pairs each video with a variable set of auxiliary modalities based on its source, moving beyond the common assumption of fixed inputs and challenging models to operate flexibly across different modal compositions.

To tackle this new challenge, we propose a novel and flexible framework centered on a mechanism we term Multi-modal Synergistic Prompting (MMSP). The core philosophy of MMSP is to leverage a pre-trained vision-language model (VLM) such as CLIP Radford et al. (2021), treating the visual stream as the primary modality and utilizing other available modalities as dynamic "prompts" to enrich it.

The MMSP mechanism operates in three steps. First, visual features are extracted using the VLM's visual encoder, while all available auxiliary modalities are processed by their respective pre-trained encoders. Second, to handle the variable and incomplete nature of the inputs, a masking strategy is employed as all available features are dynamically fused into a synergistic prompt vector. Finally, this fused prompt is injected into the VLM in a dual manner: it is integrated with the visual tokens and simultaneously embedded within the textual input tokens. This dual-prompting strategy ensures that both the visual and textual branches are conditioned on the same fused multi-modal context, generating a unified, relationship-aware representation that is far more robust than vision-alone cues.

In summary, our contributions are threefold: (1) We introduce the task of Multi-modal Open-vocabulary Video Visual Relationship Detection (MM-OV-VidVRD) and its corresponding VaM-VidVRD benchmark. It features a heterogeneous, source-dependent set of auxiliary modalities and a rich collection of visually ambiguous predicates that demand multi-modal reasoning. (2) We propose a flexible framework centered on the Multi-modal Synergistic Prompting (MMSP) mechanism, which effectively fuses arbitrary and incomplete modal combinations into a unified and relationship-aware representation. (3) We provide extensive experiments demonstrating significant improvements over vision-only baselines, along with ablation studies that validate the MMSP design and quantify the contribution of each auxiliary modality.

2 Related Work

Video Visual Relationship Detection Benchmarks. Foundational VidVRD benchmarks, such as ImageNet-VidVRD Shang et al. (2017) and VidOR Shang et al. (2019), suffer from two key limitations: they rely on noisy, auto-generated object tracks and are fundamentally uni-modal, ignoring critical complementary cues. While recent efforts like HyperGLM Nguyen et al. (2025) expand evaluation along the *task* dimension by unifying diverse video understanding tasks, they remain limited to visual inputs. In contrast, our VaM-VidVRD extends the field along the *modality* dimension. Through a rigorous annotation pipeline, it establishes high-precision, manually-validated trajectories and is the first to introduce synchronized auxiliary modalities (audio, text, and 3D data) from diverse origins, facilitating robust multi-modal relationship detection.

Open-Vocabulary Video Visual Relationship Detection. To move beyond fixed object and predicate categories, OV-VidVRD methods Gao et al. (2023); Yang et al. (2024); Wu et al. (2024b); Wang et al. (2025a); Liu et al. (2025); Wang et al. (2025b) leverage vision-language models (VLMs) like CLIP Radford et al. (2021) and ALPro Li et al. (2022) to detect unseen relationships. However, constrained by the aforementioned uni-modal datasets, these approaches only leverage visual features from the video instance. While they effectively use text to define their open-vocabulary categories, they neglect complementary signals accompanying the video, such as audio, textual descriptions, or 3D data. These signals are often crucial for disambiguating visually similar or subtle interactions,

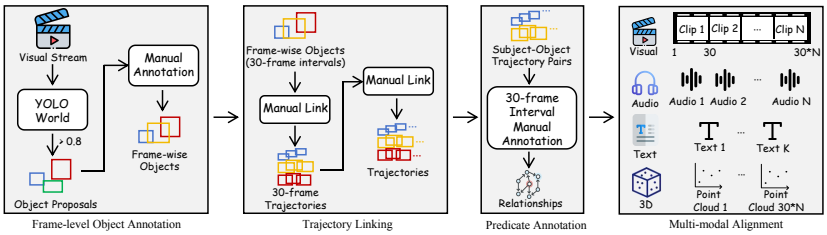

Figure 1: The four-stage high-fidelity annotation pipeline for VaM-VidVRD.

and their omission motivates our definition of the new MM-OV-VidVRD task, which explicitly incorporates these multi-modal contexts.

Multi-modal Fusion. Multi-modal fusion has been extensively studied in audio-visual Arandjelovic & Zisserman (2017), vision-language Lu et al. (2019); Tan & Bansal (2019), and vision-3D perception Liu et al. (2022). However, existing approaches are not directly applicable to OV-VidVRD, as they are designed for fixed modality settings or specific downstream tasks rather than open-vocabulary structured relationship prediction. Although recent studies explore robustness to missing modalities Tsai et al. (2019); Wu et al. (2024a); Wang et al. (2023), they still assume predefined modality combinations and cannot handle arbitrary source-dependent inputs. USG Wu et al. (2025) recently introduces a universal scene graph representation by aligning objects from different modalities into a unified graph. However, it focuses on merging independently generated graphs into a closed-vocabulary representation, whereas our goal is to leverage auxiliary modalities to enhance video relationship detection with unseen object and predicate categories. Therefore, flexible multi-modal fusion for OV-VidVRD remains unexplored. To address this challenge, we propose the Multi-modal Synergistic Prompting (MMSP) mechanism, which enables flexible fusion of arbitrary modality combinations.

3 The VaM-VidVRD Benchmark

3.1 Data Curation and Composition

A core contribution of VaM-VidVRD is its heterogeneous and source-dependent modality composition. Rather than enforcing a single, fixed set of modalities, our benchmark is constructed by aggregating several data sources, including VidVRD Shang et al. (2017), VidOR Shang et al. (2019), MammAlps Gabeff et al. (2025), VggSound Chen et al. (2020), KITTI Geiger et al. (2013), nuScenes Caesar et al. (2020), and ActivityNet Caba Heilbron et al. (2015), each reflecting a distinct real-world scenario. We curate videos from these sources based on several key criteria, including the presence of clear dynamic interactions, stable footage without severe camera shake, and the high-quality and synchronized availability of their corresponding auxiliary modalities. Since no existing VidVRD benchmark provides vision, audio, text, and 3D data for every video, enforcing complete modality coverage would require synthesizing unavailable signals, increasing annotation costs and potentially introducing artifacts. We instead preserve the modalities provided by each source and apply a unified, manually validated annotation and alignment pipeline. As a result, VaM-VidVRD covers all four modalities at the benchmark level, while each video retains a source-dependent subset. The resulting four modality compositions are summarized in Table 1.

3.2 Annotation Pipeline

Establishing a high-fidelity benchmark for MM-OV-VidVRD demands an annotation pipeline of the highest rigor. Our VaM-VidVRD benchmark is therefore built upon a multi-stage manually-validated process to ensure data quality across all components. This high-fidelity pipeline consists of four key stages: frame-level object annotation, trajectory linking, predicate annotation, and multi-modal alignment, as illustrated in Figure 1,

Frame-level Object Annotation. This stage combines automated proposals with meticulous manual refinement. First, object proposals are generated by YOLO-World Cheng et al. (2024), retaining

Table 1: Composition of the VaM-VidVRD benchmark.

Modalities	Source	Scenario	Videos
V-Only	VidVRD	General	600
V + A	VidOR	General	1,000
	MammAlps	Wildlife	150
	VggSound	General	750
V + 3D	KITTI	Autonomous Driving	50
	nuScenes	Autonomous Driving	250
V + A + T	ActivityNet	Human Activity	800
Total			**3,600**

Table 2: Comparison of VidVRD, VidOR, and VaM-VidVRD.

Metric	**VidVRD**	**VidOR**	**VaM-VidVRD**
Videos	1,000	10,000	3,600
Predicate Categories	132	50	56
Object Categories	35	80	54
Object Trajectories	3,017	52,344	12,775
Avg. Trajectory Duration	102	662	115
Predicate Instances	30,752	377,931	54,339
Avg. Relationship Duration	43	243	102
Avg. Valid Frames Per Video	120	898	140
Modalities	V	V+A	V+A+T+3D

those above a $\tau = 0.8$ confidence threshold. These proposals are then meticulously corrected by human annotators, including refining boundaries, deleting false positives, and adding missing ground-truth objects. To ensure high precision and recall, each video is fully annotated and subsequently cross-validated by a separate annotator.

Trajectory Linking. The annotated frame-wise objects are then manually linked into temporally-consistent trajectories. To manage this complexity, annotators first link objects within 30-frame intervals; these segments are then linked at their boundaries to form complete trajectories. All trajectories are cross-validated to ensure the absence of any identity switches or temporal inconsistencies.

Predicate Annotation. Using the high-fidelity object trajectories, we identify all pairs co-occurring for more than 30 frames. Their co-occurrence duration is then partitioned into 30-frame intervals for annotation. To ensure high accuracy, we employ a focused strategy where each annotator is assigned a specific subset of only 5 predicate categories to identify within each interval. All predicate annotations are then cross-validated for consistency.

Multi-modal Alignment. Finally, all auxiliary modalities are rigorously synchronized with the visual data at different temporal granularities. Audio is segmented into 30-frame intervals, strictly aligned with the corresponding visual clip. Textual Descriptions, which often cover longer spans, are aligned at the clip-level, associating a single description with all 30-frame clips within its temporal window. 3D Data is aligned at the frame-level, with each point cloud matched to its corresponding video frame.

3.3 Data Statistics and Vocabulary Split

Dataset statistics are summarized in Table 2 The VaM-VidVRD benchmark comprises 3,600 videos (2,800 for training, 800 for testing), averaging over 140 valid annotated frames per video. In total, it contains 12,775 high-precision object trajectories and 54,339 predicate instances across 54 object and 56 predicate categories. The instance distributions exhibit a natural long-tail characteristic that mirrors real-world data, which motivates our open-vocabulary setting. Consequently, we partition the vocabularies based on frequency into base categories (the top 37 objects and 36 predicates) and novel categories (the remaining 17 objects and 20 predicates). Further details are provided in Appendix B.

4 Our Method

4.1 Problem Formulation

Multi-modal Open-vocabulary Video Visual Relationship Detection (MM-OV-VidVRD) is defined over a multi-modal video instance, which consists of a visual stream (V) and a heterogeneous set of auxiliary modalities ($\mathcal{A}$). This auxiliary set $\mathcal{A}$ is instance-specific (e.g., audio, textual descriptions, 3D data).

The objective is to detect a set of relationship instances, where each instance $\mathcal{R}$ is defined as a tuple:

$$\mathcal{R} = (c^s, c^r, c^o, T^s, T^o), \tag{1}$$

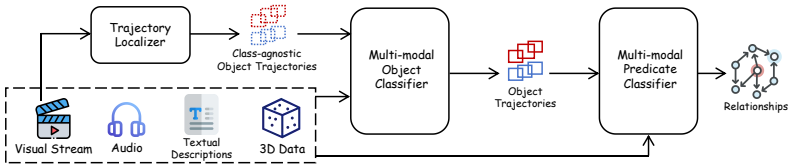

Figure 2: An overview of the proposed framework. For brevity, the details of the two classifiers are not shown here.

where $c^s, c^o \in \mathcal{C}^O$ are the subject/object categories from the object vocabulary $\mathcal{C}^O$, and $c^r \in \mathcal{C}^R$ is the predicate category from the predicate vocabulary $\mathcal{C}^R$. The terms T^s and T^o represent the subject and object trajectories, defined as a sequence of bounding boxes over a temporal segment.

The open-vocabulary constraint partitions the vocabularies into disjoint base and novel sets. Models are exposed only to base categories during the training phase, while at inference, they must generalize to detect all valid relationships, including those involving novel categories.

4.2 Framework

As illustrated in Figure 2, our framework processes multi-modal video data through a sequential, three-component structure. The process begins with a Trajectory Localizer Wang et al. (2025b), which parses the input video stream V to generate a set of class-agnostic object trajectories $\{T^k\}$.

These class-agnostic trajectories, along with the auxiliary modalities $\mathcal{A}$ and the visual stream V, are then processed by two distinct classification components: a Multi-modal Object Classifier and a Multi-modal Predicate Classifier. The object classifier assigns a semantic object category to each individual trajectory. Subsequently, the categorized trajectories are paired based on temporal co-occurrence, and the predicate classifier identifies the interaction predicate between each pair.

Both classifiers are powered by the same MMSP-driven classification module. This module's operation begins by extracting initial visual feature f_v^{in} from the video stream V_{in} using the CLIP visual encoder. This feature, along with the auxiliary set $\mathcal{A}$, is fed into our MMSP mechanism (Section 4.3) to generate a dedicated visual prompt p_v and textual prompt p_t. These prompts are then used to generate the multi-modal-aware representations in two parallel streams. On the visual side, the original visual data is processed by the CLIP visual encoder, which is dynamically guided by the visual prompt p_v injected into each of its layers. Specifically, the visual prompt $p_v \in \mathbb{R}^d$ is prepended to the patch token sequence at the first Transformer layer and propagated through all subsequent L layers via the standard self-attention mechanism. This produces the multi-modal-aware visual feature f_v. On the textual side, the textual prompt p_t is integrated with the learnable textual tokens T_c, where c represents the corresponding category. We adopt a CoOp-style Zhou et al. (2022) prompting strategy extended with our multi-modal textual prompt. For each candidate category c, the textual input to the CLIP text encoder is constructed as:

$$T_c = [p_t;\ t_1;\ t_2;\ \ldots;\ t_M;\ e_c], \tag{2}$$

where $p_t \in \mathbb{R}^d$ is the textual prompt projected from the MMSP mechanism, $\{t_1, \ldots, t_M\}$ are M learnable context tokens, and e_c denotes the CLIP word embedding of the category name c. The learnable context tokens are shared across all categories and optimized end-to-end during training. The combined tokens are fed into the CLIP text encoder to produce a set of textual features $\{f_t^c\}$.

The Multi-modal Object Classifier applies this module once per trajectory, using the trajectory's visual data as its V_{in}. The final object category is then determined by an argmax over the resulting:

$$\text{score}(c) = \text{sim}(f_v, f_t^c). \tag{3}$$

The multi-modal predicate classifier leverages this core module in a three-stream approach. For each trajectory pair, it applies the MMSP-driven module three times independently, using the subject visual data (V_s), the object visual data (V_o), and the global video context (V_g) as the V_{in}, respectively. This process yields three distinct visual features (f_v^s, f_v^o, f_v^g) and three corresponding sets of textual features ($\{f_t^{s,c}\}, \{f_t^{o,c}\}, \{f_t^{g,c}\}$). These features are then aggregated via concatenation to form a

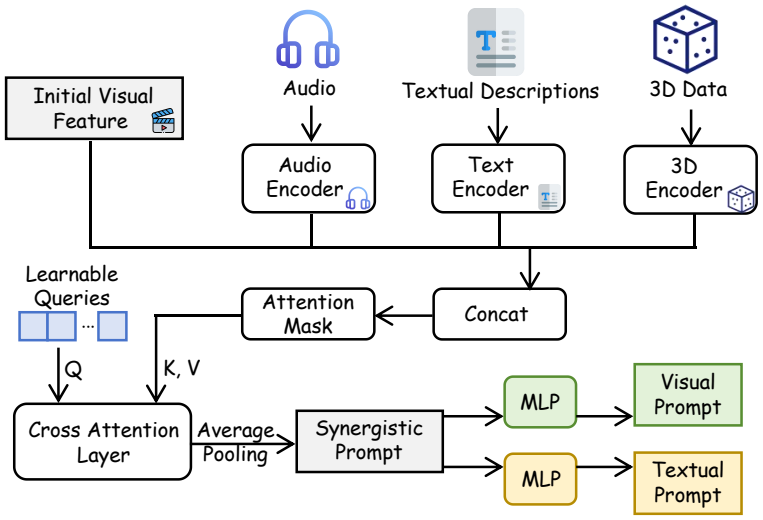

Figure 3: The architecture of our Multi-modal Synergistic Prompting (MMSP) mechanism.

final visual vector $f_v^{rel} = [f_v^s; f_v^o; f_v^g]$ and a final textual vector $f_t^{rel,c} = [f_t^{s,c}; f_t^{o,c}; f_t^{g,c}]$ for each predicate category. These aggregated vectors are then projected by two separate projection functions ($\mathcal{H}_v$ and $\mathcal{H}_t$, implemented as MLPs) before the final similarity calculation, which determines the score for each predicate category c:

$$\text{score}(c) = \text{sim}(\mathcal{H}_v(f_v^{rel}), \mathcal{H}_t(f_t^{rel,c})). \tag{4}$$

4.3 Multi-modal Synergistic Prompting

Synergistic Fusion. All available auxiliary modalities are processed by their encoders to produce auxiliary features $\{f_A, f_T, f_{3D}\}$. These features are concatenated with the initial visual feature f_v^{in} to form the input context:

$$C_{in} = [f_v^{in};\ \hat{f}_A;\ \hat{f}_T;\ \hat{f}_{3D}], \tag{5}$$

where $\hat{f}_m = f_m$ if modality m is available for the current instance, and $\hat{f}_m = \mathbf{0} \in \mathbb{R}^d$ otherwise. This fusion module employs a small set of learnable fusion queries, Q_{fuse}. A cross-attention layer is then applied, where Q_{fuse} serves as the query, and the entire input context C_{in} serves as the source for the key and value. Concretely, the attention mask is constructed based on the ground-truth modality availability of each instance. For positions in C_{in} corresponding to absent modalities, the attention logits are set to $-\infty$ prior to the softmax operation in the cross-attention layer, ensuring that missing modalities contribute zero weight. During training, this masking is naturally determined by the heterogeneous composition of VaM-VidVRD. This exposes the model to diverse masking patterns across the training set without requiring any data augmentation or dropout-based simulation. At inference, the identical masking logic is applied based on whichever auxiliary modalities accompany the test instance, requiring no separate missing-modality protocol. The output of this layer is then pooled into the single, synergistic prompt vector p_{syn}.

Dual Prompt Projection. The synergistic vector p_{syn} is projected through two separate prompt generation functions ($\mathcal{G}_v$ and $\mathcal{G}_t$, implemented as MLPs) to generate distinct prompts for the visual (p_v) and textual (p_t) branches:

$$p_v = \mathcal{G}_v(p_{syn}), p_t = \mathcal{G}_t(p_{syn}). \tag{6}$$

4.4 Training Objective

We adopt a progressive three-stage training paradigm, keeping the CLIP model and auxiliary encoders frozen throughout the process. First, following EOV-MMP Wang et al. (2025b), we pre-train the visual trajectory localizer utilizing standard localization and objectness losses. Next, we freeze

Table 3: Comparison results on VaM-VidVRD benchmark.

Split	Method	SGDet				SGCls				PredCls		
		mAP	R@50	R@100	mAP_o	mAP	R@50	R@100	mAP_o	mAP	R@50	R@100
Novel	Qwen3-VL-235B-A22B	2.42	1.91	1.91	4.32	3.38	3.18	3.18	19.97	5.76	5.04	5.04
	Gemini2.5 Pro	4.14	6.86	6.86	3.82	6.95	12.15	12.15	21.64	8.71	13.36	13.36
	CLIP	-	-	-	-	0.31	2.77	3.53	13.38	3.35	12.22	17.53
	RePro	1.02	4.99	10.87	3.18	2.66	17.03	28.43	15.52	4.39	35.84	46.65
	OV-MMP	3.34	22.61	30.35	5.18	4.37	35.36	46.71	16.26	5.81	47.63	57.27
	EOV-MMP	4.85	28.89	36.11	6.86	6.69	37.78	47.78	17.40	7.70	48.33	58.44
	Ours	**10.33**	**32.22**	**40.22**	**11.92**	**12.19**	**44.44**	**51.56**	**23.83**	**14.16**	**52.78**	**62.11**
All	Qwen3-VL-235B-A22B	1.62	4.46	4.46	8.25	2.88	6.34	6.34	42.18	3.53	7.66	7.66
	Gemini2.5 Pro	3.68	6.99	6.99	7.51	5.69	13.44	13.44	46.20	7.15	19.15	19.15
	CLIP	-	-	-	-	0.22	0.32	0.41	13.91	2.13	4.55	8.04
	RePro	4.89	7.12	10.01	10.49	22.26	28.39	39.16	38.85	31.72	41.33	49.50
	OV-MMP	20.22	22.98	25.79	24.21	34.84	42.83	52.11	47.78	47.23	55.62	68.22
	EOV-MMP	24.98	28.11	31.56	26.73	38.88	44.51	53.38	49.70	49.47	57.27	69.94
	Ours	**28.16**	**32.64**	**36.18**	**32.29**	**44.58**	**49.93**	**59.98**	**55.17**	**54.82**	**62.05**	**74.34**

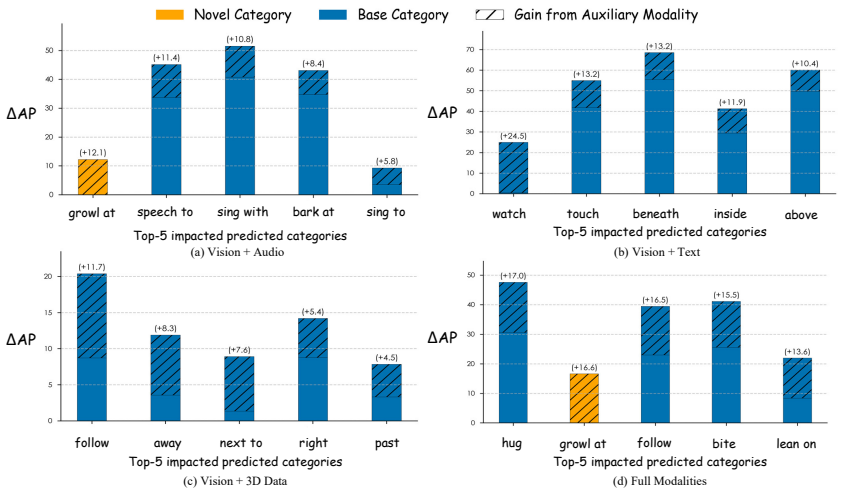

Figure 4: Absolute AP change ($|\Delta AP|$) for the top-5 most impacted predicate categories.

this localizer to train the Multi-modal Object Classifier via a Cross-Entropy (CE) loss. Finally, with both preceding components frozen, we optimize the Multi-modal Predicate Classifier using a Binary Cross-Entropy (BCE) loss. Note that during all training stages, the parameters of the pre-trained CLIP model and all auxiliary modality encoders are kept frozen.

5 Experiments

5.1 Experimental Setup

Evaluation Protocol. We conduct a comprehensive set of experiments on our newly introduced VaM-VidVRD benchmark. We evaluate performance on three tasks: (1) Predicate Classification **(PredCls)** (2) Scene Graph Classification **(SGCls)** (3) Scene Graph Detection **(SGDet)** We use mean Average Precision (mAP) and Recall@K (R@K) with K = 50, 100 as evaluation metrics for relationship classification. Following EOV-MMP Wang et al. (2025b), we also use mean Average Precision of object trajectory (mAP_o) to evaluate the quality of object detection.

Implementation Details. We adopt the pre-trained CLIP ViT-L/14@336px Radford et al. (2021) as our visual encoder. For auxiliary modalities, we utilize domain-specific pre-trained models for feature extraction: wav2vec2-base-960h Baevski et al. (2020) for audio, unsup-simcse-bert-base-uncased Gao et al. (2021) for text, and PointPillars Lang et al. (2019) for 3D point clouds. Further implementation details and hyperparameter settings are provided in Appendix A.

Table 4: Ablation study on the impact of auxiliary modalities on the SGDet task. We compare the V-Only baseline against configurations augmented with a single auxiliary modality: Audio (A), Textual Descriptions (T), or 3D Data (3D). "Full" denotes the configuration using all available modalities.

Modality	Novel-split		All-split	
	mAP	mAP$_o$	mAP	mAP$_o$
V-Only	5.79	6.05	25.40	26.79
V + A	8.61	8.33	26.68	28.75
V + T	7.98	9.82	26.35	28.76
V + 3D	6.37	6.72	26.47	27.91
Full Modalities	**10.33**	**11.92**	**28.16**	**32.29**

Table 5: Ablation on which component to apply multi-modal (MM) fusion on the SGDet task. We analyze fusing MM data into the Object Classifier (MM-Obj) and/or the Predicate Classifier (MM-Pred). (✓= Fused; blank = V-Only component)

Fusion Component		Novel-split		All-split	
MM-Obj	MM-Pred	mAP	mAP$_o$	mAP	mAP$_o$
		5.79	6.05	25.40	26.79
✓		7.96	**11.92**	26.89	**32.29**
	✓	8.89	6.05	27.22	26.79
✓	✓	**10.33**	**11.92**	**28.16**	**32.29**

Table 6: Ablation study on fusion mechanisms on SGDet.

Fusion Method	Novel-split		All-split	
	mAP	mAP$_o$	mAP	mAP$_o$
V-Only	5.79	6.05	25.40	26.79
Simple Fusion	5.57	3.43	25.39	25.47
Cross-Attention	6.59	4.99	25.68	26.74
MMSP	**10.33**	**11.92**	**28.16**	**32.29**

5.2 Comparison with Existing Methods

To validate the effectiveness of our multi-modal framework, we conduct a comparison against vision-only (V-Only) methods and multi-modal large language models (MLLMs). The V-Only baselines include CLIP Radford et al. (2021); RePro Gao et al. (2023); OV-MMP Yang et al. (2024); and EOV-MMP Wang et al. (2025b). Additionally, we compare against MLLMs, specifically Qwen3-VL-235B-A22B Bai et al. (2025) and Gemini 2.5 Pro Comanici et al. (2025). Detailed experimental settings are provided in the Appendix F.

As shown in Table 3, our method consistently surpasses all V-Only methods and MLLMs. While MLLMs exhibit strong object classification with ground-truth trajectories (reflected by high SGCls mAP$_o$), they severely struggle with precise spatio-temporal localization and fine-grained reasoning. This reveals a fundamental gap between general semantic recognition and the structured understanding required for MM-OV-VidVRD. Moreover, against the state-of-the-art EOV-MMP Wang et al. (2025b), our method yields absolute gains of 5.48% in mAP and 5.06% in mAP$_o$ on the challenging Novel-split SGDet task, demonstrating that multi-modal fusion simultaneously improves localization and open-vocabulary generalization.

5.3 Ablation Studies

Impact of Auxiliary Modalities. To validate our multi-modal premise, we establish a vision-only baseline by masking all auxiliary modalities during MMSP fusion. Table 4 demonstrates that augmenting the baseline with any single modality consistently boosts performance, while the "Full" configuration achieves the best results. Further diagnostics in Appendix D.2 confirm that these improvements arise from genuine multi-modal feature learning rather than source-modality shortcuts.

To analyze the source of these gains, Figure 4 visualizes the top-5 predicate categories with the largest absolute AP improvements ($|\Delta\text{AP}|$) per modality, revealing distinct patterns. Specifically, Audio (a) significantly boosts sound-related actions; Textual Descriptions (b) excel on complex, action-oriented predicates; and 3D Data (c) best resolves spatial relations. Furthermore, our "Full" model (d) achieves robust gains across novel categories, confirming its ability to synergistically leverage all inputs for effective generalization.

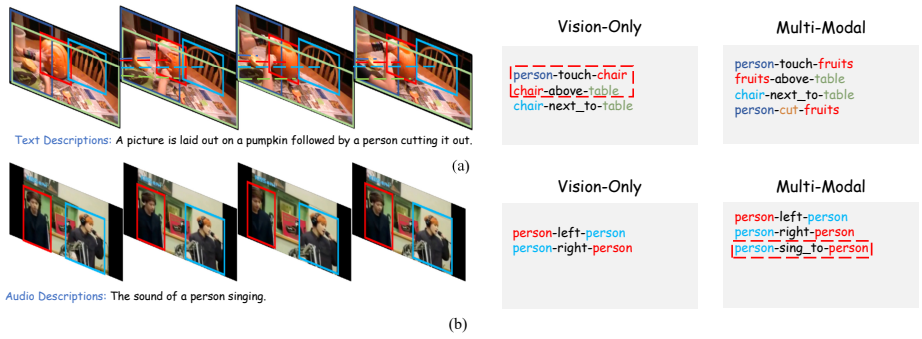

Figure 5: Qualitative comparison between the V-Only baseline (left) and our Multi-Modal method (right). Incorrect relationship triplets are enclosed in red boxes, and orange predicates denote novel categories.

Ablation on Fusion Components. We analyze the impact of multi-modal (MM) fusion on two key components: the Object Classifier (MM-Obj) and the Predicate Classifier (MM-Pred). The results in Table 5 show that while applying MM fusion to either component individually yields a clear performance improvement over the V-Only baseline, the dual-fusion configuration achieves the best overall performance. This demonstrates a clear synergistic benefit, validating our choice to leverage multi-modal information throughout the entire classification pipeline.

Efficacy of MMSP. We compare MMSP against two conventional baselines equipped with standard learnable prompts Zhou et al. (2022): Simple Fusion (concatenation prior to an MLP) and Cross-Attention (visual features querying modality keys/values). As shown in Table 6, naive fusion strategies yield marginal or even negative gains. Specifically, Simple Fusion drastically degrades Novel mAP_o from 6.05% to 3.43%. Although Cross-Attention slightly raises Novel mAP (6.59%), it compromises mAP_o and stagnates on the All-split. In contrast, MMSP consistently elevates all metrics, achieving 10.33% Novel mAP and 11.92% Novel mAP_o, thereby proving its superior efficacy in exploiting auxiliary modalities over generic fusion mechanisms.

5.4 Qualitative Analysis

Figure 5 qualitatively compares our multi-modal method against the V-Only baseline on VaM-VidVRD. In (a), the V-Only baseline hallucinates objects (e.g., a chair) and yields incorrect relationships. Conversely, guided by text descriptions (pumpkin", cutting"), our method correctly identifies the fruits and detects the novel predicate *cut*. Furthermore, (b) illustrates a partial failure case: while two people are singing, only one microphone is active, leading our model to predict *sing_to* for just one pair. Nonetheless, it still surpasses the V-Only baseline, which completely misses the singing predicates due to the action's non-visual nature. Additional qualitative examples are provided in Appendix B, and a discussion of potential label leakage is provided in Appendix D.1.

6 Conclusion

In this paper, we introduce Multi-modal Open-vocabulary Video Visual Relationship Detection (MM-OV-VidVRD) to overcome the limitations of uni-modal approaches. To this end, we present the VaM-VidVRD benchmark, a high-fidelity dataset featuring heterogeneous, source-dependent auxiliary modalities. We propose Multi-modal Synergistic Prompting (MMSP), a novel framework that effectively fuses arbitrary and incomplete modalities to guide VLM-based object and predicate classification. Extensive evaluations demonstrate that MMSP substantially outperforms state-of-the-art vision-only baselines, particularly on novel splits. Supported by detailed ablations, our findings confirm the critical role of auxiliary modalities and the efficacy of our design, laying a robust foundation for holistic video understanding.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with manuscript preparation, including drafting, rewriting, grammar correction, and language polishing, and for assisting with the development, debugging, and documentation of research code. We also used large multimodal language models, specifically Qwen3-VL-235B-A22B and Gemini 2.5 Pro, as zero-shot baselines for video visual relationship detection experiments. In addition, an LLM-based semantic matching procedure was used in the label-leakage diagnostic reported in Appendix D.1 to identify and mask paraphrases and implicit semantic expressions related to object and predicate categories.

We have not used generative AI tools to replace human annotation of the benchmark, to make final annotation decisions, or to independently substantiate the scientific claims of this work.

All AI-assisted text and code were reviewed by the authors. AI-assisted code was inspected, adapted to the intended implementation, and tested through the reported experiments and ablation studies. The outputs produced by Qwen3-VL and Gemini 2.5 Pro were treated as experimental predictions and were evaluated using the same benchmark protocol as the other baselines. The LLM-based semantic masking procedure was checked against the intended category and predicate semantics, and the resulting analyses were reviewed by the authors. We also manually checked the AI-assisted manuscript content, citations, mathematical descriptions, experimental settings, and reported results. We take responsibility for the final content of this work, including all text, claims, code, experimental results, and other artifacts produced with the aid of generative AI.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725, 2020.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbul, Alexander Mathis, and Devis Tuia. Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13854–13864, 2025.

Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*, 2023.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697–12705, 2019.

Dokyun Lee, Sangeun Seo, Chanwoo Park, Sunjun Kim, Buru Chang, and Jean Y Song. Exploring intervention techniques to alleviate negative emotions during video content moderation tasks as a worker-centered task design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 1701–1721, 2024.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH 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*, pp. 4953–4963, 2022.

Qi Liu, Weiying Xue, Yuxiao Wang, and Zhenao Wei. Openvidvrd: Open-vocabulary video visual relation detection via prompt-driven semantic space alignment. *arXiv preprint arXiv:2503.09416*, 2025.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

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 Computer Vision and Pattern Recognition Conference*, pp. 29150–29160, 2025.

Zhe Ni, Xiaoxin Deng, Cong Tai, Xinyue Zhu, Qinghongbing Xie, Weihang Huang, Xiang Wu, and Long Zeng. Grid: Scene-graph-based instruction-driven robotic task planning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 13765–13772. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

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*, pp. 1300–1308, 2017.

Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 279–287, 2019.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28694–28698, 2025.

Tan Tang, Yanhong Wu, Yingcai Wu, Lingyun Yu, and Yuhong Li. Videomoderator: A risk-aware framework for multimodal video moderation in e-commerce. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):846–856, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, volume 2019, pp. 6558, 2019.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Yongqi Wang, Xinxiao Wu, and Shuo Yang. Metor: A unified framework for mutual enhancement of objects and relationships in open-vocabulary video visual relationship detection. *arXiv preprint arXiv:2505.06663*, 2025a.

Yongqi Wang, Xinxiao Wu, Shuo Yang, and Jiebo Luo. End-to-end open-vocabulary video visual relationship detection using multi-modal prompting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024a.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14158–14168, 2025.

Ziyue Wu, Junyu Gao, and Changsheng Xu. Open-vocabulary video scene graph generation via union-aware semantic alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8566–8575, 2024b.

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*, volume 38, pp. 6513–6521, 2024.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, et al. Drivearena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025.

Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv preprint arXiv:2403.19098*, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

A Implementation Details

We provide further details on our experimental setup, covering the pre-trained encoders, the multi-stage training process and the specific architectures of our proposed modules.

A.1 Pre-trained Encoders

As mentioned in the main paper, our framework is built upon several pre-trained encoders, the parameters of which are kept frozen during all training stages. A key design choice is to project all feature representations into a unified feature space of 768 dimensions.

- **Visual Encoder:** We use the CLIP ViT-L/14@336px model Radford et al. (2021) to extract initial visual feature (f_v^{in}) and to produce the final multi-modal-aware visual features (f_v). The native feature dimension extracted by this model is 768.
- **Text Encoder:** We use the CLIP ViT-L/14@336px text encoder to produce the final textual features (f_t^c). The output dimension is 768.
- **Audio Encoder:** Audio features (f_A) are extracted using the pre-trained wav2vec2-base-960h model Baevski et al. (2020). The output feature dimension of this model is 768.
- **Text Description Encoder:** Textual description features (f_T) are encoded via the unsup-simcse-bert-base-uncased model Gao et al. (2021).
- **3D Data Encoder:** The 3D data (f_{3D}) is initially encoded using the PointPillars model Lang et al. (2019), which transforms the sparse point cloud into a dense, gridded 2D feature map. This map is structured with a shape of $[C, W, H]$, where the channel dimension C is 364, and W and H are the spatial grid dimensions defined by the model's pre-processing. To effectively aggregate the spatial information captured in this representation, we apply an average pooling operation across the spatial dimensions, yielding an intermediate 364-dimensional feature vector $f_{3D}^{inter} \in \mathbb{R}^{364}$. Finally, to project and align the 3D features with the target 768-dimensional space shared by the other encoders, we introduce a lightweight projection layer, which maps the 364-dimensional vector to the final 768-dimensional 3D feature vector f_{3D}.

A.2 Training Strategy Details

As mentioned in Section 4.4, our framework is optimized using a three-stage training strategy. All our models are trained on a single NVIDIA 4090 GPU. The specific hyper-parameters for each stage are as follows:

- **Stage 1: Trajectory Localizer.** We pre-train the trajectory localizer for 20 epochs. We use an AdamW optimizer with an initial learning rate of 1e-4, a weight decay of 1e-4, and a global batch size of 16.
- **Stage 2: Multi-modal Object Classifier.** With the localizer frozen, we train the object classifier for 10 epochs. We use an AdamW optimizer with a learning rate of 1e-2 and a batch size of 4.
- **Stage 3: Multi-modal Predicate Classifier.** Next, with all other components frozen, we train the predicate classifier for 40 epochs. We use an AdamW optimizer with a learning rate of 1e-4 and a batch size of 32.

Note that as stated in the main paper, the parameters of the pre-trained CLIP model and all auxiliary modality encoders are kept frozen throughout all training stages.

A.3 Model Architecture Details

Here we provide the specific architectural details for our proposed modules.

MMSP Mechanism. The core MMSP mechanism (Section 4.3) is implemented as follows:

- **Synergistic Fusion:** The cross-attention-based fusion module consists of 1 Transformer decoder layers. The set of learnable fusion queries Q_{fuse} has a dimension of 768 and

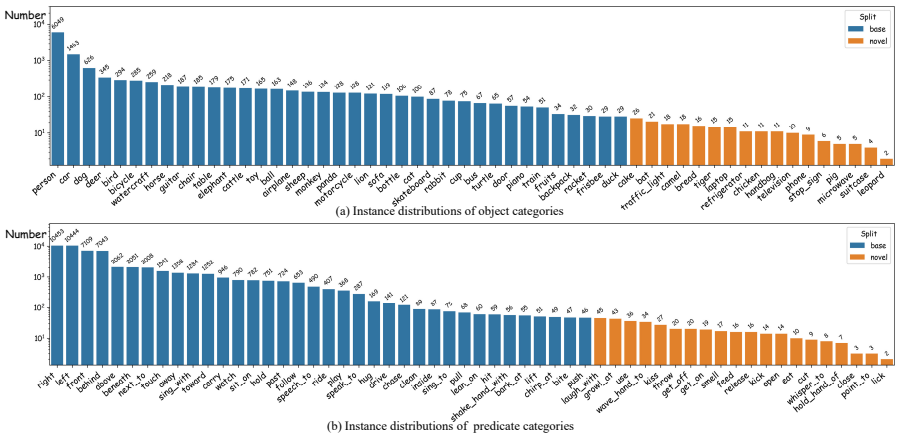

Figure 6: Instance distributions of (a) object categories and (b) predicate categories in the VaM-VidVRD benchmark, plotted on a logarithmic scale. The open-vocabulary split into base (blue) and novel (orange) categories is based on instance frequency.

Table 7: The complete object vocabulary (54 categories), split into Base (37) and Novel (17) sets.

Base Objects (37)							
airplane	backpack	ball	bicycle	bird	bottle	bus	cake
car	cat	cattle	chair	cup	deer	dog	door
duck	elephant	frisbee	fruits	guitar	horse	lion	monkey
motorcycle	panda	person	rabbit	racket	sheep	skateboard	sofa
table	toy	train	turtle	watercraft			
Novel Objects (17)							
bat	bread	camel	chicken	handbag	laptop	leopard	microwave
phone	piano	pig	refrigerator	stop_sign	suitcase	television	tiger
traffic_light							

contains 4 tokens. The resulting synergistic prompt vector p_{syn} is the mean-pooled output of these tokens, with a dimension of 768.

- **Dual Prompt Projection:** The two prompt generation functions, $\mathcal{G}_v$ and $\mathcal{G}_t$ (Eq. 6), are both implemented as 2-layer MLPs with a hidden dimension of 1024. They project p_{syn} to the final prompt dimensions p_v (768) and p_t (768).

Predicate Classifier Projections. In the multi-modal predicate classifier (Section 4.2), the final aggregated vectors (f_v^{rel} and $f_t^{rel,c}$) are projected by two separate MLPs ($\mathcal{H}_v$ and $\mathcal{H}_t$, Eq. 4). These are implemented as 2-layer MLPs that map the concatenated feature dimension of 2304 (768×3) down to the CLIP embedding dimension of 768 before the final similarity calculation.

B Details on the VaM-VidVRD Benchmark

This section provides more details for the VaM-VidVRD benchmark, complementing the main paper. We provide the complete category splits, further statistical analyses, and qualitative examples of our high-fidelity annotation pipeline.

B.1 Category Split

As described in Section 3.3 of the main paper, we partition our vocabulary into base and novel categories based on instance frequency. This split is central to the open-vocabulary task. Table 7 and

Table 8: The complete predicate vocabulary (56 categories), split into Base (36) and Novel (20) sets.

Base Predicates (36)							
above	away	bark_at	behind	beneath	bite	carry	chase
chirp_at	clean	drive	follow	front	hit	hold	hug
inside	lean_on	left	lift	next_to	past	play	pull
push	ride	right	shake_hand_with	sing_to	sing_with	sit_on	speak_to
speech_to	touch	toward	watch				
Novel Predicates (20)							
close	cut	eat	feed	get_off	get_on	growl_at	hold_hand_of
kick	kiss	laugh_with	lick	open	point_to	release	smell
throw	use	wave_hand_to	whisper_to				

Table 9: Comparison of Video Visual Relationship Detection Datasets: VidVRD, VidOR, and VaM-VidVRD

Metric	**VidVRD**	**VidOR**	**VaM-VidVRD**
Videos	1000	10000	3600
Predicate Categories	132	50	56
Object Categories	35	80	54
Object Trajectories	3017	52344	12775
Average Trajectory Duration	102	662	115
Predicate Instance	30752	377931	54339
Average Relation Triplet Duration	43	243	102
Average Valid Frames per Video	120	898	140
Modalities	Vision	Vision+Audio	Vision+Audio+Text+3D

Table 8 list the complete vocabularies. Figure 6 visualizes the instance count and split assignment of each object and predicate category.

B.2 Further Statistics

Table 9 provides a comprehensive statistical comparison between our proposed VaM-VidVRD benchmark and existing datasets, namely VidVRD Shang et al. (2017) and VidOR Shang et al. (2019).

- **Annotation Quality vs. Scale.** While VidOR operates on a massive scale with 10,000 videos and over 52,000 object trajectories, it heavily relies on semi-automatic generation pipelines to achieve this volume. Such automated approaches, while efficient, often introduce noise in object localization and tracking, which can propagate errors into the visual relationship labels. In contrast, VaM-VidVRD prioritizes annotation fidelity. Although our dataset contains 3,600 videos and 12,775 trajectories, these annotations have undergone rigorous human verification. This strict quality control ensures that the relationship triplets are grounded on precise object trajectories, providing a more reliable benchmark for evaluating complex reasoning capabilities, rather than testing a model's resilience to noisy ground truth.
- **Category Diversity and Balance.** VaM-VidVRD offers a balanced taxonomy with 54 object and 56 predicate categories. In comparison, while VidVRD lists 132 predicate categories, this number is significantly inflated by the mechanical combination of intransitive verbs (e.g., run, stand) with various spatial prepositions (e.g., left, behind) to form compound relations. This approach fragments a single semantic action into multiple distinct classes (e.g., treating *run-left* and *run-right* as separate predicates). Our benchmark avoids such artificial redundancy by focusing on distinct interaction types, providing a more concise and semantically structured distribution for learning core visual relationships.



Figure 7: Qualitative comparison of trajectory annotations. (a) Ground truth from original benchmarks (e.g., VidOR) often suffers from coarse localization and missing annotations. (b) VaM-VidVRD provides tight, high-fidelity, and temporally consistent trajectories. **Note:** An asterisk (*) marks objects where the original bounding boxes were severely inaccurate or drifted significantly. A red dashed box highlights objects (e.g., static furniture) that were entirely missing in the original annotations but are recovered in our benchmark.

- **Temporal Density.** Our dataset features an average trajectory duration of 115 frames and an average relation triplet duration of 102 frames. While VidOR has longer average durations (662 frames), the validity of these long-range dependencies is often compromised by the tracking drift inherent in automatic labeling. Our average valid frame count of 140 represents a dense and meaningful temporal window, significantly richer than VidVRD (120 frames), ensuring that relationships are sustained and observable.

- **Multi-Modal Integration.** VaM-VidVRD is the first benchmark to explicitly enable and evaluate research across 4 distinct modalities (Visual, Text, Audio, 3D). Regarding VidOR, although its source videos contain audio tracks (as they are collected from the web), the benchmark treats relationship detection as a strictly uni-modal (visual) task, offering neither audio-specific features nor baselines that leverage this information. In contrast, our framework actively integrates 3D point cloud features, audio embeddings, and rich textual descriptions, establishing a new standard for holistic, multi-modal video understanding.

B.3 Trajectory Annotation Quality Comparison

As outlined in the main paper (Section 3.1), our diverse video corpus is curated from multiple existing sources. However, while we adopt the raw video data, we do not rely on their original annotations. Existing large-scale benchmarks, such as VidOR, often utilize automated or semi-automated labeling pipelines. As a result, their provided ground truth frequently suffers from significant noise, including bounding box drift, loose localization, and the omission of critical objects.

To guarantee high-fidelity benchmarks, we performed a comprehensive re-annotation of all object trajectories (Section 3.2). Figure 7 visualizes the qualitative leap in annotation standards between the original benchmarks (left) and our VaM-VidVRD (right), highlighting two major improvements:

- **Precision and Localization.** Original annotations often exhibit "jittery" or loose bounding boxes that fail to tightly encompass the objects. As marked by the asterisks (*) in the top two rows of Figure 7, the original boxes for the performers and rider are loosely fitted or drifted. In contrast, VaM-VidVRD provides tight, human-verified bounding boxes that accurately track the object boundaries throughout the sequence.

- **Completeness of Context.** A critical limitation of previous benchmarks is the frequent omission of relevant objects. As shown in the bottom row of Figure 7, while the original dataset annotates the people, it misses the furniture they interact with. In VaM-VidVRD, we explicitly recover these missing objects (highlighted in the red dashed box, e.g., table, chair), as they are indispensable for detecting contact-based relationships.

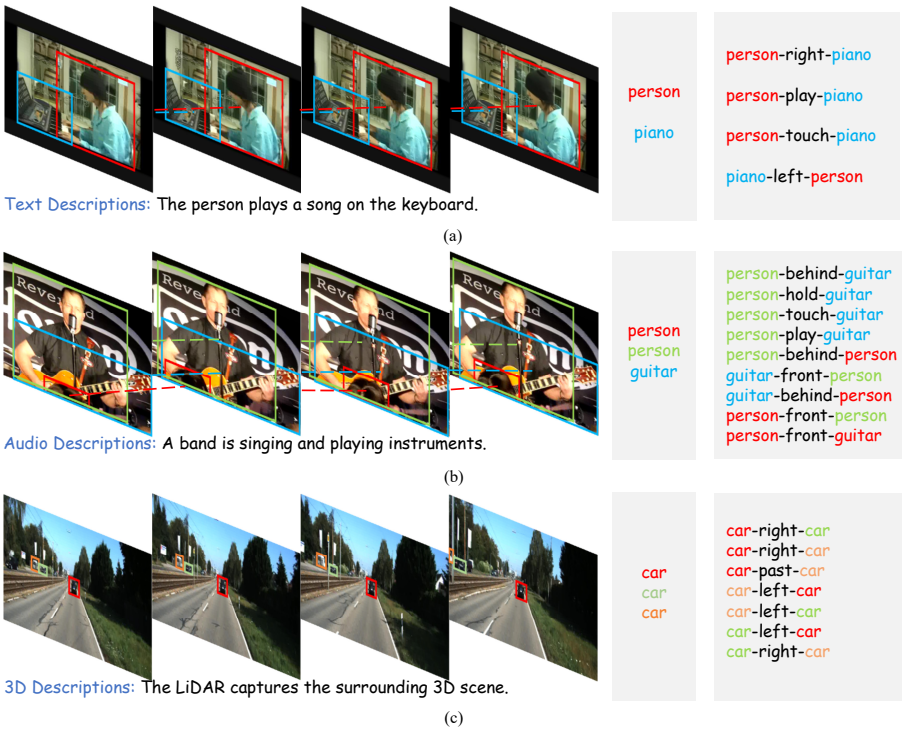

Figure 8: Qualitative examples of multi-modal relationship annotations in VaM-VidVRD. The figure showcases three distinct scenarios enriched with different modality descriptions.

B.4 Relationship Annotation Instances

We provide qualitative examples of our final relationship annotations in Figure 8. A distinguishing feature of VaM-VidVRD is that, beyond the standard visual ground truth, it uniquely associates each video sequence with multi-modal information (Text, Audio, and 3D). As illustrated in the figure, the benchmark covers a wide spectrum of scenarios:

- **Scenario (a) & (b)** show human-object interactions in indoor and performance settings. These clips are paired with textual and audio information respectively, providing high-level semantic context that complements the fine-grained visual predicate annotations.
- **Scenario (c)** presents an outdoor autonomous driving scene accompanied by 3D information. This highlights the dataset's inclusion of 3D perception data, which is critical for spatial reasoning tasks in traffic scenarios.

Each instance consists of high-fidelity object trajectories and precise predicate labels, offering a dense and detailed representation of the visual scene.

C Qualitative Analysis

To provide a deeper understanding of our model's behavior and the impact of multi-modal integration, we present a additional qualitative comparison in Figure 9. We analyze predictions on the VaM-VidVRD test set, contrasting the results from a baseline Vision-Only model against our proposed Multi-Modal framework.

The comparison highlights how supplementary modalities (Audio, Text, and 3D) serve as critical correctives when visual cues are ambiguous, occluded, or generic. We detail four representative scenarios shown in Figure 9:

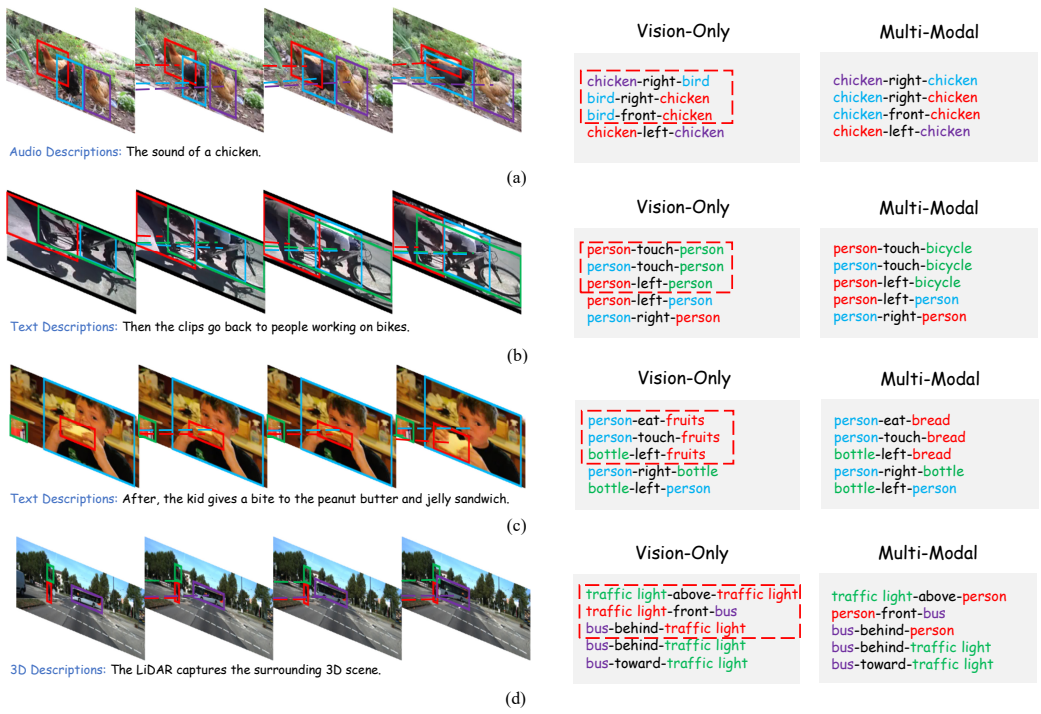

Figure 9: Qualitative case studies comparing the baseline Vision-Only predictions (left column) with our Multi-Modal framework (right column). The red dashed boxes highlight erroneous detections made by the vision-only baseline due to visual ambiguity, occlusion, or lack of depth information.

- **Audio-Driven Disambiguation (a):** In this scenario, the visual features of the animal are somewhat generic, leading the Vision-Only model to predict the hypernym bird. However, the accompanying audio (The sound of a chicken) provides a decisive acoustic cue. Our Multi-Modal model successfully leverages this to refine the prediction from bird to the specific category chicken.

- **Text-Guided Object Recovery (b):** Here, the visual inputs are challenging, likely due to motion blur or occlusion, causing the Vision-Only model to misclassify the object as person. The textual narration ("...working on bikes"), however, explicitly mentions the object of interest. By integrating this semantic context, our model corrects the hallucinated person label to the accurate bicycle category.

- **Fine-Grained Semantic Refinement (c):** The visual appearance of the object in the child's hand is ambiguous, which the Vision-Only model incorrectly classifies as fruits. The text description ("...peanut butter and jelly sandwich") offers precise semantic details that visual features alone struggle to resolve. Consequently, the Multi-Modal model rectifies the object classification to bread.

- **3D-Aided Spatial Reasoning (d):** In this complex traffic scene, the Vision-Only model suffers from severe hallucinations, predicting incorrect traffic light relations, likely due to depth confusion or background clutter. The integration of 3D information, allows our model to eliminate these visual artifacts and correctly identify the spatial relationships between the person and the bus.

These cases demonstrate that while visual features are foundational, the fusion of text, audio, and 3D data significantly enhances the model's robustness against visual ambiguity and improves the precision of relationship detection.

Table 10: Label leakage diagnostic on the SGDet task. "Original" denotes the original textual descriptions. "Masked" removes explicit object and predicate category names as well as their lexical variants, while "Semantic Masked" further removes semantically related expressions identified by an LLM-based matching strategy.

Model	Text	**Novel-split**		**All-split**	
		mAP	mAP$_o$	mAP	mAP$_o$
V-Only	–	5.79	6.05	25.40	26.79
V+T	Original	7.98	9.82	26.35	28.76
V+T	Masked	7.44	9.71	26.19	28.68
V+T	Semantic Masked	6.80	8.85	26.04	28.34
Full MMSP	Original	10.33	11.92	28.16	32.29
Full MMSP	Masked	10.24	11.66	28.05	32.11
Full MMSP	Semantic Masked	9.69	11.22	27.56	31.84

D Reliability Analysis

We conduct two analyses on the SGDet task to assess whether MMSP enhances video relationship detection by leveraging task-relevant knowledge from aligned multi-modal evidence, rather than relying on explicit label leakage or source-modality priors.

D.1 Label Leakage Diagnostic

Relationship annotations are manually curated from visual trajectory pairs following the VidVRD protocol, rather than derived from text or audio. The audio encoder receives raw acoustic signals; textual renderings of audio content shown in figures are provided only for visualization and are not used as model input. To further examine whether MMSP exploits label-related semantic cues from textual inputs, we conduct both lexical-level and semantic-level masking experiments. Specifically, we remove explicit category names and use an LLM-based semantic matching strategy to identify and mask paraphrases or implicit expressions associated with object and predicate categories.

Table 10 evaluates whether MMSP relies on label-related information in textual inputs. Removing explicit category names causes only minor performance degradation for both V+T and Full MMSP, indicating that the model does not directly exploit lexical overlap between text descriptions and target categories. Further applying semantic masking introduces a stricter perturbation by removing implicit expressions related to object and predicate categories. Although performance decreases due to the loss of useful semantic context, Full MMSP still consistently outperforms the V-Only baseline, demonstrating that its improvements mainly arise from effective multi-modal evidence integration rather than textual label leakage.

D.2 Source-Modality Shortcut Analysis

Because modality availability is source-dependent in VaM-VidVRD, we test whether the model exploits source identity rather than instance-level auxiliary evidence. Specifically, we shuffle auxiliary features across videos from the same source, preserving the source, domain, and modality composition while breaking video-level cross-modal alignment. We compare this control with aligned Full MMSP, V-Only, and a prior-only baseline using only source/modality identity. To further analyze whether this behavior is consistent across different domains, we also report source-wise SGDet mAP on the All split.

As shown in Table 11, same-source shuffling reduces Novel-split mAP/mAP$_o$ by 4.56/5.96 points and All-split mAP/mAP$_o$ by 2.53/5.74 points relative to aligned Full MMSP. The shuffled and prior-only models both remain close to V-Only. Table 12 further shows that aligned auxiliary features consistently improve All-split mAP across all six sources, whereas same-source shuffling reduces performance to near or below the corresponding V-Only result. These results indicate that source or modality priors alone cannot explain the gains; correctly aligned instance-level auxiliary evidence is necessary.

Table 11: Same-source auxiliary-shuffling diagnostic on SGDet. "Aligned" uses the auxiliary features paired with each video, whereas "Same-source shuffle" replaces them with features from another video from the same source. "Prior only" uses source/modality identity without instance-level auxiliary content.

Model	Auxiliary Input	Novel-split		All-split	
		mAP	mAP$_o$	mAP	mAP$_o$
V-Only	–	5.79	6.05	25.40	26.79
Full MMSP	Aligned	10.33	11.92	28.16	32.29
Full MMSP	Same-source shuffle	5.77	5.96	25.63	26.55
Prior only	Source/modality	5.81	6.09	25.82	27.24

Table 12: Source-level performance breakdown on the SGDet All split. "V-Only" uses only the visual stream, "Aligned" uses the instance-aligned auxiliary modalities, and "Same-source Shuffle" replaces the auxiliary features with features from another video from the same source while preserving the source and modality composition.

Source	**Modality**	**V-Only**	**Aligned**	**Same-source Shuffle**
VidOR	V+A	30.34	33.92	29.15
MammAlps	V+A	31.55	33.81	31.69
VGGSound	V+A	24.07	28.51	22.60
KITTI	V+3D	14.95	17.08	13.83
nuScenes	V+3D	15.23	18.85	15.11
ActivityNet	V+A+T	21.19	25.96	20.07

E Discussion on Noise Robustness

A natural concern for multi-modal methods is their robustness to noisy or partially incorrect auxiliary inputs. We argue that our framework possesses inherent noise robustness owing to both the nature of the training data and the design of the MMSP mechanism.

Inherent Noise in Training Data. The auxiliary modalities in VaM-VidVRD exhibit varying levels of alignment quality. Audio and 3D data are well-aligned with the visual stream, synchronized at the frame level or 30-frame interval level. Textual descriptions, however, are aligned at a coarser clip-level granularity: a single description may span multiple 30-frame intervals and thus does not precisely correspond to each annotated relationship segment. This temporal misalignment naturally introduces noise, as the description may refer to events before or after the target interval. Furthermore, audio streams from real-world sources (e.g., web videos in VidOR and activity videos in ActivityNet) inevitably contain background noise, overlapping speakers, and irrelevant sounds. Consequently, our model is trained on a realistic mixture of clean and noisy auxiliary inputs, which implicitly encourages robustness to imperfect signals.

Robustness via MMSP Design. The cross-attention-based fusion in MMSP further contributes to noise tolerance. Rather than treating all auxiliary features equally (as in simple concatenation), the learnable fusion queries in MMSP selectively attend to the most informative features in the input context C_{in}. This attention-based selection naturally down-weights noisy or irrelevant signals. The consistent improvements over the V-Only baseline across all modality configurations (see Table 4 in the main text), despite the presence of the aforementioned noise, empirically support this claim.

F MLLMs Evaluation Details

In addition to the learned models evaluated in the main paper, we also assess the performance of frontier multimodal large language models (MLLMs). For all MLLM-based evaluations, we directly query the official APIs of Qwen3-VL-235B-A22B Bai et al. (2025) and Gemini2.5 Pro Comanici et al. (2025). The goal of this evaluation is to examine how far strong general-purpose MLLMs can go on MM-OV-VidVRD when prompted in a zero-shot manner.

F.1 TASK-SPECIFIC PROMPTING PROTOCOL

The three evaluation settings require different levels of information to be predicted. Accordingly, we use different prompting pipelines for PredCls, SGCls, and SGDet.

PredCls. In the PredCls setting, the ground-truth subject trajectory, object trajectory, and their object categories are provided. Therefore, the MLLM only needs to predict the predicate labels for a given subject-object pair over time. We use a predicate classification prompt for this setting:

Task
Video Visual Relation Detection with Vision, Audio and Text Descriptions.

Information

- **Vision Content**: [Vision Input]
- **Audio Content**: [Audio Input]
- **Text Descriptions**: [Text Input]

The video contains the following tracked objects:

- TID 0: [Subject Category] (Box Color: [xx, xx, xx])
- TID 1: [Object Category] (Box Color: [xx, xx, xx])

Instructions

1. **Visual Observation**: Observe movements and spatial interactions of tracked objects.
2. **Audio Analysis**: Use audio cues to confirm actions (e.g., sound of barking, crashing).

- **CRITICAL**: Only detect predicates from this specific subset: [Predicate Category List]
- If no predicate from this subset is found, return an empty list `[]`.

Timing Constraints (30 FPS Base)

- Video Frames: Sampled at 2 FPS.
- Original Video: 30 FPS.
- Requirement: Output `start_time` and `end_time` as frame indices (30 FPS scale).
- Calculation: 1 sampled frame = 15 units.

Output Requirement
Output strictly as a valid Python/JSON list of lists.
Example format: `[[0, "above", 1, 20.0, 50.0]]`

SGCls. In the SGCls setting, the ground-truth class-agnostic object trajectories are given, but the object categories and predicate categories must both be predicted. Instead of separating the two steps, we directly use a joint prompt that asks the MLLM to infer the object trajectory categories together with the predicates between trajectory pairs. We refer to this as the object trajectory classification + predicate classification prompt:

Task
Video Visual Relation Detection with Vision, Audio, and Text Descriptions.

Information

- **Vision Content**: [Vision Input]
- **Audio Content**: [Audio Input]
- **Text Descriptions**: [Text Input]

The video contains the following tracked objects identified by TID and bounding-box colors. You must first identify the category of each object based on its visual appearance.

- TID 0: (Box Color: [xx, xx, xx])
- TID 1: (Box Color: [xx, xx, xx])

Instructions

1. **Visual Observation**: Observe the movements and spatial interactions of the tracked objects.
2. **Audio Analysis**: Use audio cues to confirm actions (e.g., sounds of barking or crashing).

- **CRITICAL**: Only detect predicates from this specific subset: [Predicate Category List]
- If no predicate from this subset is found, return an empty list `[]`.

Timing Constraints (30 FPS Base)

- Video Frames: Sampled at 2 FPS.
- Original Video: 30 FPS.
- Requirement: Output `start_time` and `end_time` as frame indices on the 30 FPS scale.
- Calculation: 1 sampled frame = 15 units.

Output Requirement
Output strictly as a valid Python/JSON list containing two elements:

1. A dictionary mapping each TID (as an integer) to its identified category string.
2. A list of relation triplets with temporal intervals.

DO NOT include any conversational text, explanations, or Markdown code blocks.
Example format: `[0: "dog", 1: "person", [[0, "chase", 1, 20.0, 50.0], [1, "run_away", 0, 25.0, 55.0]]]`

SGDet. In the SGDet setting, the model must start from the raw video and complete the entire pipeline. We therefore decompose the prompting process into two consecutive stages. First, the MLLM is prompted to perform trajectory localization and object category prediction. Second, based on the detected trajectories and predicted categories, the MLLM is prompted again for predicate classification (See **PredCls**). The prompt for trajectory localization and object category prediction is as follows:

Task
Video Object Trajectory Annotation. Identify the main objects and track them.

Information

- **Vision Content**: [Vision Input]
- **Audio Content**: [Audio Input]
- **Text Descriptions**: [Text Input]

Constraints

- Candidate Categories: [Object Categories]
- Bounding Box: Use `[xmin, ymin, xmax, ymax]` format, normalized to 0–1000.

Output Requirement
Output strictly in the following JSON format:

```
{"categories": {"TID": "category_name"},"frames": {"TID":
↪  [[frame_index, [xmin, ymin, xmax, ymax]],[frame_index, [xmin,
↪  ymin, xmax, ymax]]]}}
```

F.2 Multi-label Predicate Prediction by MLLMs

A practical challenge when using MLLMs for MM-OV-VidVRD is that predicate prediction is inherently a multi-label classification problem. For the same subject-object pair within the same temporal window, multiple predicates may co-exist. However, when all predicate categories are presented at once, MLLMs tend to predict only a few highly salient and visually obvious predicates, while miss-

Table 13: Comparison of different predicate grouping strategies for zero-shot MLLM-based predicate classification on a randomly sampled subset of 100 test videos. "All-in-One" means that all predicate categories are included in a single prompt. "Group-of-5" means that 5 predicate categories are included in each prompt. "Single-Predicate" means that each prompt contains only 1 predicate category. "Cost" denotes the relative token/API consumption compared with the All-in-One setting.

Model	Group	mAP	R@50	R@100	Cost
Qwen3-VL-235B-A22B	All	1.73	0.74	0.74	1×
	Group-5	3.43	5.27	5.27	∼12×
	Single	4.06	6.12	6.12	∼56×
Gemini 2.5 Pro	All	4.72	2.73	2.73	1×
	Group-5	6.94	11.28	11.28	∼12×
	Single	7.55	12.90	12.90	∼56×
Ours	–	**52.20**	**58.88**	**66.67**	–

ing more fine-grained, subtle, or less visually dominant predicates. This issue becomes more severe for predicates that require careful temporal reasoning or weak evidence accumulated over multiple frames.

A simple yet effective mitigation strategy is to split the predicate vocabulary into several smaller groups and query the MLLM group by group. For example, if there are 56 predicate categories, they can be divided into 12 groups with roughly 5 categories per group, or even into finer groups. For each group, the MLLM is only asked to determine whether each predicate exists and, if so, during which time interval it occurs. This group-wise querying strategy usually improves recall, especially for fine-grained predicates, because it substantially reduces the competition among labels within a single prompt.

Nevertheless, this improvement comes at a significant computational cost. If the full predicate set is divided into groups of 5 categories, then 12 separate MLLM calls are needed to cover all 56 predicates, leading to approximately 12× higher token consumption than using a single prompt containing the full label set. Thus, there is a clear trade-off between prediction quality and inference cost.

F.3 Grouping Strategies for Predicate Classification

As discussed above, directly asking an MLLM to predict predicates from the full predicate vocabulary often leads to a strong bias toward only a few salient and obvious relations, while many subtle or fine-grained predicates are easily overlooked. To better analyze the trade-off between prediction quality and inference cost, we compare several predicate grouping strategies, including: (1) putting **all predicate categories** into a single prompt, (2) splitting the predicate categories into groups of **5 predicates per prompt**, and (3) using **1 predicate per prompt**.

Since calling frontier MLLMs through official APIs is expensive, we conduct this analysis on a randomly sampled subset of 100 test videos from the test set. Moreover, because our goal here is to study the effect of grouping strategies on predicate classification itself, we only evaluate on the PredCls setting, where the ground-truth object trajectories and object categories are given, and the MLLM only needs to predict the predicates. In addition, since all MLLM evaluations in this section are conducted in a zero-shot manner, we do not distinguish between the Novel and All splits here; instead, we report the overall performance on the sampled subset.

The quantitative results are summarized in Table 13, which also includes the performance of our proposed model on this subset for comparison. For Gemini 2.5 Pro, transitioning from an All-in-One to a Group-of-5 configuration significantly boosts performance, improving mAP from 4.72 to 6.94 and R@50 from 2.73 to 11.28. This strategy reduces candidate competition within the prompt, facilitating the recall of subtle relations. Further reducing the granularity to a single predicate per prompt yields only marginal gains, improving mAP from 6.94 to 7.55 and R@50 from 11.28 to 12.90 at a substantially higher computational cost. Balancing effectiveness and efficiency, we adopt the Group-of-5 configuration for the MLLM in our primary evaluations and report these results across the entire test set in the main paper. However, it is noteworthy that even with this optimized

prompting strategy, there remains a substantial performance gap between frontier MLLMs and our proposed model. This disparity primarily stems from the inherent difficulties current MLLMs face when reasoning about fine-grained and subtle spatio-temporal relations in videos.