Neural Networks 203 (2026) 109183

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

SAGE: Semantic-guided framework with decoupled optimization for open-vocabulary video visual relationship detection

Shiqi Wang, Weiying Xue, Shuyi Hu, Haowen Li, Qi Liu *

School of Future Technology, South China University of Technology, Guangzhou, 511442, China

ARTICLE INFO

Keywords:
Open-vocabulary video visual relationship detection
Semantic-guided framework
Multimodal large language model
Cross-attention mechanism

ABSTRACT

Open-vocabulary video visual relationship detection (VidVRD) aims to expand video visual relationship detection beyond annotated categories by detecting unseen relationships between both seen and unseen objects in videos. Existing approaches primarily focus on adapting static image-text models (e.g., CLIP) via visual prompting but pay limited attention to the intrinsic visual-semantic gap and optimization instability in dynamic video contexts. To overcome this, we propose a Semantic-Guided Framework with Decoupled Optimization (SAGE) to decouple explicit semantic reasoning from robust classifier adaptation. Due to the static nature of pre-trained image encoders, low-level visual features often fail to capture subtle spatio-temporal action cues, leading to semantic ambiguity in distinguishing visually similar but semantically different interactions. We introduce a Multimodal LLM-based Semantic Teacher as a semantic information source to establish explicit semantic reasoning, extracting structured descriptions that are integrated with visual representations via cross-attention, thereby reducing the spatio-temporal gap. Furthermore, instance-level visual representations in videos are highly susceptible to visual noise (e.g., motion blur, occlusion). In existing instance-conditioned methods, this noise propagates into learnable prompts, causing semantic drift, classification inconsistency, and poor generalization to novel categories. To mitigate this noise sensitivity, we propose a Decoupled Class-Aware Prompting strategy. Unlike instance-conditioned methods, this module utilizes a Textual Knowledge Embedding network to transform stable class-level text embeddings into adaptive prompts, effectively mitigating semantic drift caused by visual noise. Extensive experiments on VidVRD and VidOR datasets validate that the proposed method achieves state-of-the-art performance, with significant gains on the challenging novel relationship categories.

1. Introduction

Video Visual Relationship Detection (VidVRD) represents a fundamental challenge in computer vision, aiming to parse complex scenes by identifying relational interactions between objects across temporal sequences (Shang et al., 2017). While traditional methods have shown success in closed-set scenarios (Gao et al., 2022a; Li et al., 2021; Liu et al., 2020), their inability to generalize beyond pre-defined categories has motivated the crucial shift towards Open-Vocabulary VidVRD (Open-VidVRD) (Gao et al., 2023).This generalization capability is particularly important given the growing demand for relationship understanding in applications such as hierarchical image retrieval (Ji et al., 2024), visual context learning for retrieval tasks (Qin et al., 2022), and visual question answering systems (Kim et al., 2021). The advent of large-scale vision-language models (VLMs) (Jia et al., 2021; Li et al., 2023, 2022b; Pham et al., 2023; Radford et al., 2021) like CLIP (Radford et al., 2021) has been a catalyst for this field, offering powerful, generalizable semantic representations learned from vast image-text corpora. However, adapting the static, image-centric knowledge of VLMs to the dynamic, compositional nature of video relationships presents a distinct set of formidable challenges (Gao et al., 2022b; Gu et al., 2021; Kuo et al., 2022; Ni et al., 2022; Weng et al., 2023; Xu et al., 2023).

Existing Open-VidVRD frameworks (Gao et al., 2022a; Ni et al., 2022), despite their progress, typically adopt a Visual Prompting paradigm to adapt CLIP. While effective in closed scenarios, these methods encounter two critical limitations when scaling to open-vocabulary settings, primarily due to inherent constraints in alignment and optimization (Herzig et al., 2023; Yuksekgonul et al., 2022; Zang et al., 2022; Zhou et al., 2022a), as summarized in Fig. 1. These challenges can be summarized as: *a) Semantic Ambiguity in Implicit Alignment*. A fundamental bottleneck arises because existing methods rely on CLIP's image encoder, which is pre-trained on static images and lacks the inherent ability to capture dynamic video relationships. Consequently, directly applying such context-blind encoders to videos often leads to semantic

* Corresponding author.
E-mail addresses: shiqi_wang1717@163.com (S. Wang), drliuqi@scut.edu.cn (Q. Liu).

https://doi.org/10.1016/j.neunet.2026.109183
Received 1 August 2025; Received in revised form 13 February 2026; Accepted 24 May 2026
Available online 26 May 2026

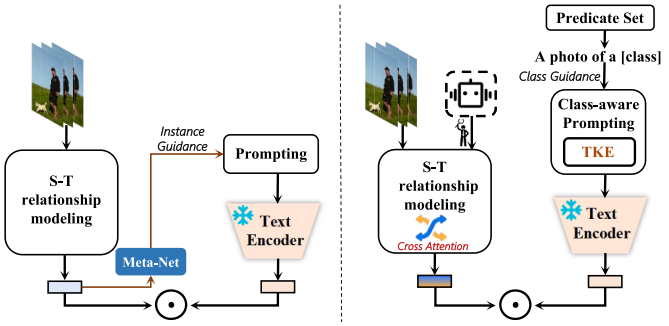


Fig. 1. Comparison of our semantic-guided framework with existing approaches. Previous methods (left) suffer from limited semantic context in visual features and unstable instance-based prompting. Our approach (right) introduces: (1) MLLM-based semantic teacher that extracts multi-perspective relational descriptions to enhance visual features through cross-attention, and (2) stable class-guided prompting mechanism. This design achieves improved robustness and generalization performance.

ambiguity, where the model struggles to distinguish visually similar but semantically distinct interactions (e.g., *stand_next_to* vs. *stop_next_to* distinguished by temporal motion history, or *play* vs. *fight* distinguished by interaction intent) due to the absence of temporal reasoning; *b) Semantic Drift due to Coupled Optimization*. Furthermore, existing methods (e.g., OV-MMP Yang et al., 2024) face optimization instability caused by Image-conditional Prompt Tuning. Relying on direct guidance from instance-level visual features (Zang et al., 2022; Zhou et al., 2022a) makes learnable prompts highly susceptible to visual noise (e.g., motion blur, occlusion). This coupling causes visual noise to propagate directly into the prompt embedding, triggering Semantic Drift,where the prompt absorbs irrelevant visual noise patterns instead of maintaining the stable class semantics. This leads to classification inconsistency and significantly weakens the generalization capability toward novel relationship categories.

To overcome these bottlenecks, we propose SAGE, a Semantic-Guided Framework with Decoupled Optimization. Inspired by the cognitive process of separating explicit reasoning from classification, our approach systematically decouples semantic understanding from classifier adaptation. This design philosophy raises two fundamental questions: *1) Why do we need an explicit Semantic Teacher instead of implicit feature refinement? 2) Why is Class-Aware Prompting superior to Image-conditional Prompt Tuning for video generalization?*

We answer the first question. In complex video scenes, relationship is often an abstract concept defined by intent and causality. Implicit alignment (Herzig et al., 2023; Yuksekgonul et al., 2022) fails to capture this. We employ a Multimodal LLM-based Semantic Teacher (e.g., Kan et al., 2023; Wang et al., 2024b) to serve as an explicit semantic bridge, providing structured relational descriptions as auxiliary semantic context rather than supervisory signals.Instead of relying solely on feature-level adaptation, this module extracts structured, multi-perspective relational descriptions. This high-level semantic information is incorporated into the visual pipeline via cross-attention, providing explicit semantic context that complements visual features and facilitates the modeling of subtle spatio-temporal cues that are otherwise easily overlooked. (***challenge a***).

We answer the second question. The limitation of Image-conditional Prompt Tuning lies in its sensitivity to visual variance (Zang et al., 2022; Zhou et al., 2022a). To mitigate this, we introduce a Decoupled Class-Aware Prompting strategy. Unlike methods that generate prompts from volatile visual instances, we utilize a Textual Knowledge Embedding (TKE) network (Yao et al., 2024a; Zhou et al., 2022b) to transform stable class-level Predicate Embeddings into adaptive prompts. This mechanism establishes independent and stable Semantic Anchors that remain invariant to visual noise, effectively mitigating the semantic drift inherent in coupled optimization and ensuring robust generalization (***challenge b***).

In summary, our contributions are threefold:

- We identify and analyze two critical bottlenecks in current Open-VidVRD methods: semantic ambiguity in implicit alignment and semantic drift in coupled optimization. We propose a novel framework that addresses these issues through a Decoupled Optimization strategy.
- To bridge the visual-semantic gap, we propose an MLLM-based Semantic Teacher. By leveraging explicit semantic reasoning to compensate for the lack of temporal dynamics, it significantly enhances the representation of fine-grained spatio-temporal interactions.
- To mitigate semantic drift induced by visual noise, we introduce a Decoupled Class-Aware Prompting strategy. By utilizing TKE to anchor prompts on stable class semantics, it effectively reduces noise sensitivity and achieves state-of-the-art performance, particularly on long-tail novel categories.

2. Related work

2.1. Video visual relationship detection

Video visual relationship detection aims to identify temporal object interactions within video sequences, requiring sophisticated understanding of both spatial configurations and temporal evolution patterns. The pioneering work (Shang et al., 2017) introduces the ImageNet-VidVRD dataset and establishes a three-stage detection framework that integrates trajectory features, spatial positions, and semantic classifications. Subsequently, the research landscape is dominated by spatio-temporal modeling approaches that capture evolving object interactions. Graph-based methodologies (Liu et al., 2020; Qian et al., 2019) represent videos as interconnected structures, employing graph convolutional networks for relationship reasoning, while recent advances (Cong et al., 2021) utilize spatial Transformer encoders coupled with temporal decoders for comprehensive spatiotemporal context modeling.Additionally, methods focusing on global context and pairwise-level fusion have shown promise in capturing complex interaction patterns (Wang et al., 2024a). As computational efficiency becomes paramount, architecture innovations shift toward unified frameworks. Query-based methodologies (Zheng et al., 2022) introduce autoregressive stage integration and single-stage approaches (Jiang et al., 2024) consolidate classification and segmentation to achieve end-to-end optimization.

In parallel, relationship refinement developments focus on improving detection quality through various mechanisms. Iterative approaches (Shang et al., 2021) enable triplet components to inform each other, while decomposition techniques (Chen et al., 2021) break complex multi-frame relationships into manageable single-frame interactions. To address dataset imbalances, researchers propose meta-learning frameworks (Xu et al., 2022) and adaptive weighting schemes (Lin et al., 2024) that mitigate bias issues. Contemporary research emphasizes dependency modeling sophistication, with methods (Cao & Huang, 2023; Zhang et al., 2024) capturing intricate correlations among predicate components and integrating contextual knowledge embedding for enhanced relationship understanding. Nevertheless, a critical limitation emerges across these approaches: they operate within predetermined categorical boundaries, thereby constraining their deployment in dynamic real-world environments where novel relationships frequently emerge.

2.2. Open-vocabulary visual relationship detection

The transition from closed-set to open-vocabulary recognition has fundamentally expanded the scope of visual understanding. Recent comprehensive surveys (Wu et al., 2024a) and generative frameworks like DetCLIPv3 (Yao et al., 2024b) have established solid foundations

for detecting novel concepts in static images. Building on this, Open-Vocabulary Video Visual Relationship Detection (Open-VidVRD) extends these capabilities to dynamic scenes, addressing the distributional imbalances and temporal complexities inherent in video domains.

Initial progress has emerged from image-based foundations that establish the conceptual framework for open-vocabulary relationship detection. Early works (He et al., 2022) have pioneered prompt-based fine-tuning approaches, while subsequent methods (Li et al., 2022b) leverage BLIP for relationship triplet generation through sophisticated prompting strategies. Further developments (Li et al., 2024; Yu et al., 2023) enhance visual relationship detection through composite visual cues and visually-prompted language models for fine-grained scene graph generation. Recent efforts (Zhao et al., 2023) have unified visual relationship detection through integrated vision-language model architectures.

Knowledge-centric solutions have attempted to bridge semantic gaps between training and deployment scenarios. Approaches (Cao & Huang, 2023) employ contextual knowledge embedding, while methods (Yuan et al., 2023) enhance relational language-image pre-training through improved scaling strategies.Recent advances in human-object interaction detection have also explored vision-language integration for zero-shot recognition (Xue et al., 2025), demonstrating the potential of cross-modal knowledge transfer for relationship understanding. However, fundamental challenges persist in knowledge base constraints that limit comprehensive semantic understanding capabilities.

Video-specific open-vocabulary solutions have remained nascent compared to their image counterparts. Foundational work [5] has established compositional prompt tuning with motion cues, while recent methods like OV-MMP (Yang et al., 2024) explore multi-modal prompting strategies. More recently, UASAN (Wu et al., 2024b) attempts to bridge the modality gap by explicitly aligning visual union regions with predicate concepts via a bridge encoder. However, a critical limitation persists in these approaches. While methods like UASAN incorporate union regions to enhance spatial context, they rely on implicitly learning semantic correspondence through feature concatenation. This approach remains susceptible to the semantic ambiguity inherent in visual features, lacking the explicit reasoning capabilities required to distinguish complex interaction intents (e.g., distinguishing aggressive hit from friendly touch). Unlike them, our SAGE framework introduces a semantic injection mechanism that incorporates external semantic context generated by an MLLM.The MLLM is used to produce structured relational descriptions that are integrated with visual representations via cross-attention, providing explicit semantic cues to alleviate semantic ambiguity.

2.3. Prompt learning for vision-language models

The democratization of large-scale vision-language models has revolutionized adaptation strategies, making sophisticated visual understanding accessible without extensive retraining requirements. Foundation models (Li et al., 2022b; Radford et al., 2021) and other architectures (Alayrac et al., 2022; Luo et al., 2020) have established powerful semantic representations through large-scale multi-modal learning, providing the foundation for various downstream adaptation techniques.

Textual prompt optimization has emerged as the primary adaptation strategy. Early works (Zhou et al., 2022b) have pioneered learnable prompt vectors that replace hand-crafted templates, while subsequent developments (Zhou et al., 2022a) address generalization concerns through conditional prompt generation with Meta-Net architectures for instance-specific adaptation. Knowledge-guided approaches (Yao et al., 2023) incorporate contextual optimization to preserve pre-trained semantic structures, and regularization mechanisms (Zhu et al., 2023a,b) have been developed for robust adaptation that prevents catastrophic forgetting.

Beyond textual approaches, multi-modal optimization has sought to leverage the full potential of vision-language models through joint adaptation strategies. Recent methods (Khattak et al., 2023) exemplify joint optimization across visual and textual encoders, improving cross-modal knowledge transfer for multi-task scenarios. Visual prompting methods (Jia et al., 2022) investigate learnable token integration into visual encoders as alternative adaptation strategies. Advanced prompting methodologies have emerged to address specific limitations in conventional approaches, with knowledge-enhanced methods (Kan et al., 2023) incorporating external knowledge for generalizable vision-language models and optimal transport approaches (Chen et al., 2023) applying theoretical frameworks for prompt learning optimization.

Concurrently, emerging research has focused on lightweight MLLMs (e.g., MobileVLM V2 Chu et al., 2024 and TinyLLaVA Jia et al., 2024) to enable direct video-to-text inference with reduced latency. However, we identify two critical limitations in these prevailing paradigms. First, regarding efficiency, Open-VidVRD is a dense prediction task. Even lightweight MLLMs incur prohibitive computational costs when executed repeatedly for numerous object pairs compared to visual encoders. Second, instance-conditional approaches (Yang et al., 2024) employ a coupled optimization strategy, where prompts are directly conditioned on visual inputs. As we argue in this work, this coupling makes the model highly vulnerable to visual noise(e.g., motion blur), leading to semantic drift. Distinct from methods that rely on run-time MLLM inference or coupled optimization, SAGE proposes a Training-time Semantic Injection and a Decoupled Class-Aware Prompting strategy. We strictly separate prompt generation from visual instances to mitigate drift and utilize the MLLM serves as an offline semantic generator, ensuring high inference efficiency.

3. Methodology

3.1. Problem definition

Open-vocabulary video visual relationship detection aims to identify relational triplets in video sequences beyond predefined categorical constraints. Given a video sequence $V = \{I_t\}_{t=1}^{T}$ with T frames, the objective is to detect relationship instances represented as quintuples $R = (s, p, o, T_s, T_o)$, where s, p, and o denote subject, predicate, and object categories respectively, while T_s and T_o represent their corresponding spatio-temporal trajectories defined as sequences of bounding boxes $\{B_t^s\}$ and $\{B_t^o\}$ spanning from t_{start} to t_{end}.

Following established open-vocabulary protocols, the categorical space is partitioned into base and novel subsets: base object categories C_o^b, novel object categories C_o^n, base relationship categories C_p^b, and novel relationship categories C_p^n. The training procedure exclusively utilizes base categories $\{C_o^b, C_p^b\}$, while evaluation encompasses both base and novel categories to assess generalization capabilities toward unprecedented relationship understanding.

3.2. Overview

We propose the Semantic-Guided Framework with Decoupled Optimization, a novel architecture engineered to systematically address the challenges of open-vocabulary video relationship detection. As illustrated in Fig. 2, our framework is composed of three specialized modules that progressively construct and refine relational understanding, thereby enhancing both model stability and generalization capabilities.

Semantic Teacher-Guided Enhancement serves as our foundational component, where the Φ_{sem} module leverages a Multimodal Large Language Model (MLLM) to generate comprehensive textual descriptions. This produces a “semantic teacher” F_{text} that provides structured guidance:

$$F_{text} = \Phi_{sem}(V) \tag{1}$$

Building upon this semantic foundation, our Advanced Spatio-Temporal Relationship Modeling component employs the Ψ_{st} module

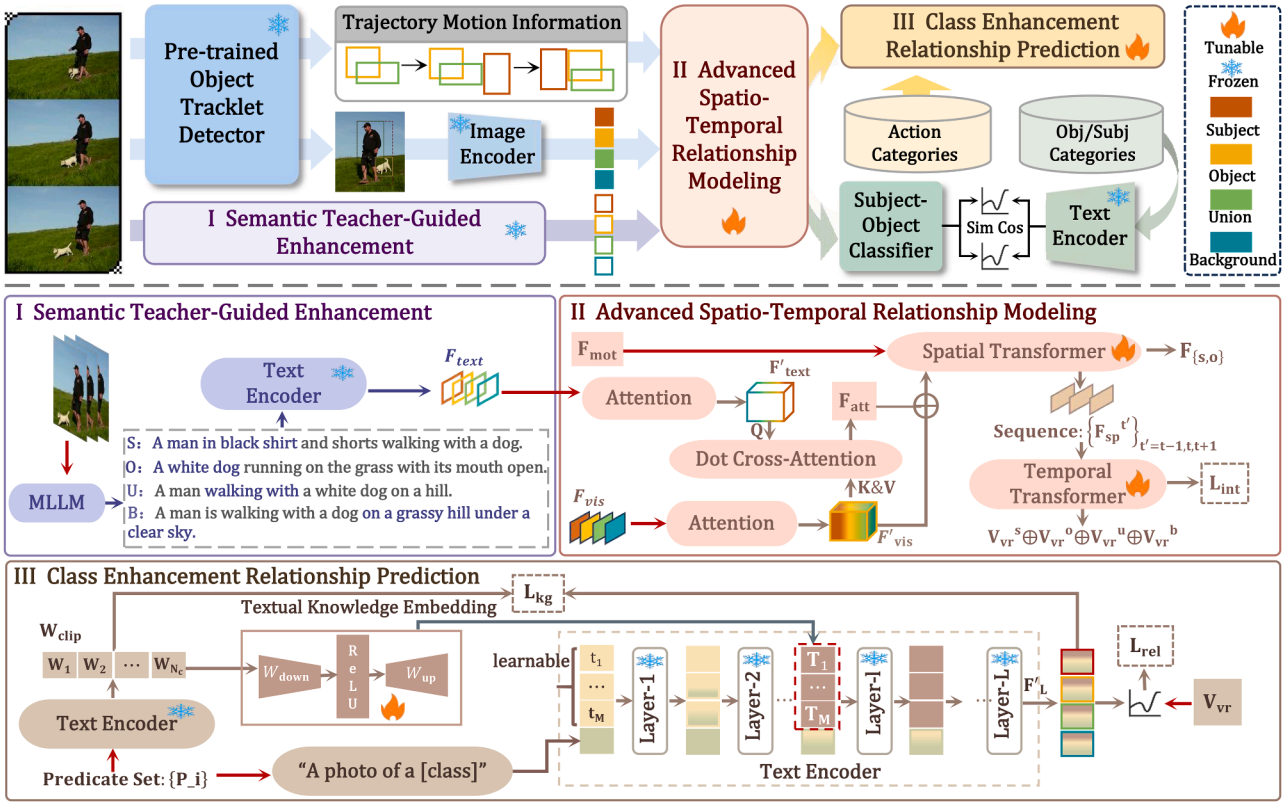


Fig. 2. The architecture of our semantic-guided framework with decoupled optimization (SAGE).The framework systematically separates semantic enhancement from classifier optimization. The (I) semantic teacher-guided enhancement module generates textual descriptions to use as guidance, which in turn conditions the (II) advanced spatio-temporal relationship modeling module's creation of a robust visual representation. The process culminates in the (III) class enhancement relationship prediction module, which employs a novel decoupled prompt tuning strategy to create a dynamic and accurate textual classifier.

to construct semantically-aware visual representations. The key innovation lies in using cross-attention mechanisms where F_{text} guides the enhancement of visual features F_{vis}, which are then processed by spatio-temporal transformers integrating motion cues F_{mot}:

$$V_{vr} = \Psi_{st}(\text{CrossAttn}(Q = F_{text}, K, V = F_{vis}), F_{mot}) \tag{2}$$

Finally, Class Enhancement Relationship Prediction represents our most novel contribution, where the Ω_{prompt} module generates dynamic textual classifiers through a sophisticated decoupled optimization strategy. This approach employs a dual-path mechanism: a TKE network processes class embeddings W_{clip} to distill textual knowledge embeddings, while learnable prompt tokens are contextualized through knowledge injection at intermediate Text Encoder layers:

$$F'_L = \Omega_{prompt}(P_{learn}, W_{clip}) \tag{3}$$

3.3. Semantic teacher-guided enhancement

Textual descriptions provide essential semantic information for understanding complex visual relationships in video sequences. In open-vocabulary relationship detection, high-quality semantic guidance is crucial for distinguishing subtle inter-object interactions and spatial configurations. Previous work on text-guided generation tasks (Shafir et al., 2023; Tevet et al., 2022; Zhang et al., 2022) has validated the effectiveness of textual guidance in enhancing feature representations and improving generation quality.

We employ ASMv2 (Wang et al., 2024b) as our semantic teacher to generate comprehensive relational descriptions. We extract multi-perspective regional information from input video sequences to capture comprehensive relational dynamics. Specifically, ASMv2 processes four critical perspectives: subject regions focusing on primary entities, object regions emphasizing secondary entities, union interaction areas capturing relational spaces, and global context providing environmental information.Targeted prompts guide ASMv2 to generate region-specific descriptions that emphasize both entity attributes and relational interactions. The multi-perspective descriptions are processed through a text encoder to produce unified semantic features F_{text}, which encode structured relational semantics from multiple visual perspectives.This semantic teacher provides structured guidance that maintains consistent relational understanding across diverse visual contexts, and is leveraged to condition subsequent spatio-temporal feature enhancement via cross-attention mechanisms.

3.4. Advanced spatio-temporal relationship modeling

This component processes multi-modal visual features to capture comprehensive spatio-temporal relationship dynamics through semantically-aware enhancement mechanisms.The input features consist of three distinct parts: Visual Features $F_{vis} = \{f_s, f_o, f_u, f_b\}$, corresponding respectively to the subject, object, union region enclosing both entities to capture interaction context, and background covering the full frame to provide global scene context, all extracted via RoI Align from the CLIP visual encoder;Motion Features F_{mot}, obtained by encoding the normalized bounding box trajectories and spatial configurations via a projection MLP;Semantic Guidance F_{text} from the teacher module. The architecture employs a sequential refinement approach to construct robust visual relationship representations.

Before cross-modal interaction occurs, both semantic and visual features undergo independent self-refinement steps. As shown in the diagram, the semantic teacher F_{text} and visual features F_{vis} are separately processed through attention blocks to produce refined representations F'_{text} and F'_{vis} respectively. This dual self-attention mechanism captures

internal contextual relationships within each modality, allowing the model to understand both semantic and visual information cohesively before their integration.

Building upon these refined features, the core of semantic enhancement lies in the dot cross-attention mechanism, where the refined semantic features F'_{text} serve as queries to interact with the refined visual features F'_{vis} and generate attention-weighted visual representations. The resulting attention feature F_{att} is then integrated back into the visual feature stream via a residual connection:

$$F_{att} = \text{CrossAttn}(Q = F'_{text}, K = F'_{vis}, V = F'_{vis}) \tag{4}$$

$$F_{enhanced} = F_{vis} + F_{att} \tag{5}$$

This design ensures that semantic guidance acts as an additive refinement to existing visual features, enriching them without overwriting crucial original visual information.

The fully enhanced, semantically-aware features are then processed to model relationships across space and time through a dual-transformer architecture, following the established design of existing methods (Gao et al., 2023; Yang et al., 2024). The spatial transformer serves a dual purpose: specialized portions of its output, corresponding to subject and object streams, are extracted to form feature $F_{\{s,o\}}$ which is used directly for subject and object classification tasks as indicated in the framework diagram. Simultaneously, the complete spatially-aware feature maps output by this transformer form a temporal sequence denoted as Sequence: $\{F^{t'}_{sp}\}_{t'=t-1,t,t+1}$, which serves as input for subsequent temporal modeling.

The temporal transformer receives this sequence of frame-level spatial features and analyzes their evolution over time, producing two critical outputs. First, an interaction confidence score L_{int} is computed to quantify the likelihood of meaningful interactions.It is obtained via a binary classification MLP $\psi(\cdot)$ applied to the aggregated representations: $\hat{y}^{int} = \sigma(\psi(V_{vr}))$.Second, the final visual relationship representation V_{vr} is generated as a compositional vector expressed as:

$$V_{vr} = V^s_{vr} \oplus V^o_{vr} \oplus V^u_{vr} \oplus V^b_{vr} \tag{6}$$

where $\oplus$ denotes the concatenation operation, and $V^s_{vr}, V^o_{vr}, V^u_{vr}$, and V^b_{vr} represent embeddings corresponding to the subject, object, union, and background streams defined in the input stage.

3.5. Class enhancement relationship prediction

This component mitigates the critical limitations of traditional prompt tuning methods in Open-VidVRD. Existing approaches typically employ an Instance-Conditional paradigm, where prompts are generated based on real-time visual features. While effective for static images, this creates a coupled optimization path in videos: visual noise (e.g., motion blur, occlusion) physically propagates into the prompt embedding, causing semantic drift where prompts overfit to noisy appearances rather than maintaining stable class semantics. To resolve this, we propose a Decoupled Class-Aware Prompting strategy.

The core innovation lies in structurally decoupling prompt generation from volatile visual instances. Instead of conditioning on unstable image features, we introduce a Textual Knowledge Embedding (TKE) network. This module performs dynamic prompt generation conditioned solely on the stable semantic definition of relationship categories. Given the pre-trained CLIP textual embeddings W_{clip} for relationship categories, our TKE transforms this general knowledge into discriminative, class-specific prompt tokens. This design enables the model to maintain robust performance across both seen and novel categories by leveraging inherent semantic structures.

Our TKE network employs a bottleneck architecture to efficiently transfer semantic knowledge. It consists of two sequential transformations: a downward projection that compresses the high-dimensional CLIP embeddings into a compact semantic representation, followed by an upward projection that expands this representation into prompt tokens suitable for text encoder integration:

$$\tau_1, \tau_2, \ldots, \tau_M = \text{TKE}(W_{clip}) \tag{7}$$

where τ_i represents the generated prompt tokens and M denotes the prompt length. Crucially, this design achieves structural decoupling. Unlike previous methods which rely on instance-specific visual inputs (modeling $P(\text{Prompt}|\text{Image})$), our TKE models the probability as $P(\text{Prompt}|\text{Class})$. Since the input W_{clip} is invariant to visual noise, the prompt generation process is mathematically independent of video perturbations. This ensures that the generated prompt acts as a stable Semantic Anchor, maintaining consistency across different video frames regardless of visual fluctuations.

The generated class-aware prompts are strategically inserted into intermediate layers of the text encoder, specifically at layer ℓ, where semantic representations have sufficiently evolved to benefit from class-specific enhancement. This insertion process can be formalized as:

$$F'_{\ell} = \text{TextEncoder}_{1:\ell-1}([P_{learn}, W_{class}]) \oplus [\tau_1, \tau_2, \ldots, \tau_M] \tag{8}$$

$$F'_L = \text{TextEncoder}_{(\ell+1):L}(F'_{\ell}) \tag{9}$$

where P_{learn} represents learnable prompt tokens, W_{class} denotes class tokens, $\oplus$ indicates the insertion operation, and F'_L represents the final enhanced textual embeddings.

During training, the core output of this component is the enhanced relationship textual embeddings F'_L, which serve as dynamic relationship classifiers. These enhanced embeddings enable relationship prediction through similarity computation with the input visual relationship representation V_{vr}. Additionally, to maintain consistency between the generated class-aware prompts and original CLIP knowledge, the system computes a knowledge-guided consistency loss, which measures the semantic alignment between the enhanced embeddings and the original CLIP textual embeddings. This design ensures that the model can generate relationship classifiers with stronger discriminative capability while maintaining coherence with pre-trained knowledge.

This decoupled optimization strategy maintains the stability of pre-trained components while allowing targeted enhancement of relationship-specific knowledge, resulting in improved generalization to novel relationship categories without compromising performance on familiar relationships.

3.6. Training objectives

The training of our framework consists of four parts: a relationship classification loss L_{rel}, an object classification loss L_{obj}, an interaction detection loss L_{int}, and a knowledge-guided consistency loss L_{kg}. The overall training loss is given by:

$$L = L_{rel} + \alpha L_{obj} + \beta L_{int} + \gamma L_{kg} \tag{10}$$

where α, β, and γ represent balance factors.

Relationship Classification Loss. Given the visual relationship representation V_{vr} and the enhanced textual embeddings F'_L, the prediction score of the relationship category r is calculated by:

$$\hat{y}^{rel}_r = \sigma\left(\frac{V_{vr} \cdot F'_{L,r}}{\tau}\right) \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function, τ is the temperature parameter, and both features are L2-normalized. The relationship classification loss is formulated by using the focal loss to address class imbalance:

$$L_{rel} = \frac{1}{N}\sum_{i=1}^{N} \text{FocalLoss}(\hat{y}^{rel}_i, y^{rel}_i) \tag{12}$$

where N is the number of training samples, $\hat{y}^{rel}_i$ is the predicted relationship score, and y^{rel}_i is the ground-truth relationship label.

Object Classification Loss. To avoid the visual feature drift caused by spatio-temporal modeling, we introduce an object classification loss to enforce the visual features to maintain object distinguishing capability. The similarity between the visual features and the text features of object category c is calculated by $\hat{y}_c^{obj} = \frac{V_{obj} \cdot W_{obj,c}}{\tau}$. The object classification loss is computed using the cross-entropy loss:

$$L_{obj} = \mathrm{CE}(\hat{y}^{sbj}, y^{sbj}) + \mathrm{CE}(\hat{y}^{obj}, y^{obj}) \tag{13}$$

where $\hat{y}^{sbj}$ and $\hat{y}^{obj}$ are the predicted subject and object similarities, and y^{sbj} and y^{obj} denote the ground-truth category labels.

Interaction Detection Loss. There may be no annotated relationships between some subjects and objects, that is, there is no interaction. For each pair of subject and object, if there are any relationship categories between them, we set the ground-truth interaction by $y^{int} = 1$, otherwise $y^{int} = 0$. To learn this interaction existence, we predict the interaction probability by $\hat{y}^{int} = \sigma(\psi(V_{ur}))$, where $\psi(\cdot)$ denotes the interaction classifier. The interaction loss is computed using the focal loss:

$$L_{int} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{FocalLoss}(\hat{y}_t^{int}, y_t^{int}) \tag{14}$$

where T represents the number of temporal clips.

Knowledge-Guided Consistency Loss. To maintain consistency between the generated class-aware prompts and original CLIP knowledge, the system computes a knowledge-guided consistency loss, which measures the semantic alignment between the enhanced embeddings and the original CLIP textual embeddings. The consistency loss is formulated as:

$$L_{kg} = 1 - \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{F'_{L,i} \cdot W_{clip,i}}{||F'_{L,i}|| \cdot ||W_{clip,i}||} \tag{15}$$

where N_c denotes the number of relationship categories, ensuring that the learning process does not deviate from the semantic space of the pre-trained model.

Two-Stage Training Strategy. Our training follows a progressive optimization schedule. In the first stage, we train the spatio-temporal relationship modeling components while keeping the text encoder frozen, enabling the visual encoder to learn effective relationship representations. In the second stage, we freeze the visual feature extraction modules and optimize the class enhancement components, allowing for targeted adaptation of textual representations without compromising learned visual features.

4. Experiment

4.1. Dataset and evaluation

Datasets. We conduct experiments on two benchmark datasets: VidVRD (Shang et al., 2017) and VidOR (Shang et al., 2019). The VidVRD dataset comprises 1000 video clips, with 800 videos allocated for training and 200 for testing, encompassing 35 object categories and 132 relationship categories. The VidOR dataset consists of 10,000 videos, distributed as 7000 for training, 835 for validation, and 2165 for testing, spanning 80 object categories and 50 relationship categories.

Evaluation Settings. Following the protocol established by OVMMP (Yang et al., 2024), we partition categories into base and novel sets based on their occurrence frequency. Frequent object and relationship categories are designated as base categories, while infrequent ones constitute novel categories. Model training is conducted exclusively on base categories. We employ two evaluation protocols: (1) Novel-split evaluation, which examines performance on all object categories paired with novel relationship categories; (2) All-split evaluation, which encompasses the complete set of object and relationship categories, serving as the comprehensive evaluation benchmark. Experiments are performed on the VidVRD test set and VidOR validation set, as ground-truth annotations for the VidOR test set remain unavailable. Additionally, to eliminate the influence of trajectory detection errors, we conduct supplementary evaluations using ground-truth object trajectories, thereby isolating the relationship detection performance with perfectly localized objects.

Evaluation Tasks. Adhering to the standard protocols of the VidVRD benchmark, we evaluate our framework on three tasks distinguished by the availability of ground-truth (GT) annotations. These tasks are designed to decouple the performance of trajectory localization, object recognition, and predicate prediction. **Scene Graph Detection (SGDet):** As the primary end-to-end task, it requires the model to jointly predict object trajectories, categories, and relationship predicates from raw videos without any ground-truth priors. We use SGDet to assess the framework's holistic capability in real-world scenarios. **Scene Graph Classification (SGCls):** A diagnostic setting where GT trajectories are provided. The model predicts both object categories and relationship predicates for the given tracks. This effectively isolates recognition performance from localization errors, evaluating the model's classification ability assuming perfect tracking. **Predicate Classification (PredCls):** A strictly diagnostic task where both GT trajectories and GT object labels are provided. The model solely predicts relationship predicates. This setting evaluates the pure semantic reasoning capability of the relationship classification module, eliminating interference from both localization and object classification errors.

Evaluation Metrics. We employ mean Average Precision (mAP) and Recall@K (R@K) with $K \in \{50, 100\}$ as our evaluation metrics for relationship detection performance. A detected relationship triplet is deemed correct when it matches a ground-truth triplet and achieves an Intersection over Union (IoU) score exceeding the threshold of 0.5.

4.2. Implementation details

We extract keyframes at 30-frame intervals across all video sequences to balance computational efficiency with temporal coverage. For dynamic object detection, we employ the MEGA framework (Chen et al., 2020) initialized with ResNet-50 (He et al., 2016) backbone parameters to perform per-frame object detection, subsequently applying the DeepSORT tracking algorithm (Wojke et al., 2017) to establish continuous spatiotemporal trajectories and generate coherent object tracklets for feature extraction. Our architecture adopts the frozen ViT-B/16 variant of CLIP as the visual encoder to preserve pre-trained visual-semantic representations.The Advanced Spatio-Temporal Relationship Modeling component integrates 1 Transformer block for VidVRD experiments and 2 blocks for VidOR dataset processing, each equipped with 8-head multi-attention mechanisms and 0.1 dropout probability for regularization. For language prompting, we allocate 8-token representations for both learnable continuous prompts and conditional prompts, positioning the [CLS] token at 75% of the total sequence length to optimize contextual information aggregation. Training adopts the AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of 1e-3, implementing a multi-step decay schedule that applies 0.1 reduction factors at training epochs 15, 20, and 25 respectively. All experimental procedures are conducted with a batch size of 32 on a single NVIDIA GeForce RTX 4090 GPU.

4.3. Comparison results

We conduct comprehensive experimental evaluation by comparing our proposed method against several state-of-the-art approaches in open-vocabulary video scene graph generation, including ALPro (Li et al., 2022a), VidVRD-II (Shang et al., 2021), RePro (Gao et al., 2023), UASAN (Wu et al., 2024b), and OV-MMP (Yang et al., 2024). The evaluation encompasses three primary tasks: Scene Graph Detection (SGDet), Scene Graph Classification (SGCls), and Predicate Classification (PredCls), assessed using standard metrics including mean Average Precision (mAP), Recall at 50 (R@50), and Recall at 100 (R@100) under both novel-split and all-split configurations. Experiments are conducted on

Table 1
Results of different methods on the VidVRD dataset.

Split	Method	SGDet			SGCls			PredCls		
		mAP	R@50	R@100	mAP	R@50	R@100	mAP	R@50	R@100
Novel	ALPro	0.98	2.79	4.33	3.69	7.27	8.92	4.09	9.42	10.41
	VidVRD-II	3.11	7.93	11.38	5.70	13.22	18.34	7.35	18.84	26.44
	RePro	5.87	12.75	16.23	10.32	19.17	25.28	12.74	25.12	33.88
	UASAN	11.05	13.88	18.35	14.50	23.14	30.58	17.62	28.93	36.53
	OV-MMP	12.15	13.72	15.21	17.57	21.98	28.43	21.14	30.41	37.85
	Ours	**17.82**	**17.67**	**19.17**	**18.33**	**28.49**	**31.72**	**22.31**	**31.57**	**39.51**
All	ALPro	3.03	2.57	3.11	3.92	3.88	4.75	4.97	4.50	5.79
	VidVRD-II	12.66	9.72	12.50	17.26	14.93	19.68	19.73	18.17	24.90
	RePro	21.12	12.63	15.42	30.15	19.75	25.00	34.90	25.50	32.49
	UASAN	23.57	15.90	19.23	32.24	25.03	31.07	38.43	30.01	37.13
	OV-MMP	22.10	13.26	16.08	29.38	23.56	28.89	38.08	30.47	37.46
	Ours	**28.61**	**17.21**	**19.94**	**30.45**	**23.69**	**30.92**	**39.05**	**31.46**	**38.69**

Table 2
Results of different methods on the VidOR dataset. For ALPro, VidVRD-II, and RePro, only the results of R@50 and R@100 on the **SGCls** and **PredCls** tasks are available from their original papers.

Split	Method	SGDet			SGCls			PredCls		
		mAP	R@50	R@100	mAP	R@50	R@100	mAP	R@50	R@100
Novel	ALPro	–	–	–	–	3.17	3.74	–	8.35	9.79
	VidVRD-II	–	–	–	–	1.44	2.01	–	4.32	4.89
	RePro	–	–	–	–	2.01	2.30	–	7.20	8.35
	UASAN	–	–	–	–	2.31	3.46	–	6.05	8.65
	OV-MMP	0.84	1.44	1.44	2.40	5.48	6.92	3.58	9.22	11.53
	Ours	**1.55**	**5.76**	**4.33**	**2.48**	**5.54**	**7.02**	**3.98**	**10.77**	**12.41**
All	ALPro	–	–	–	–	0.95	1.32	–	2.61	3.66
	VidVRD-II	–	–	–	–	9.40	12.78	–	24.81	34.11
	RePro	–	–	–	–	10.03	12.91	–	27.11	35.76
	UASAN	–	–	–	–	10.15	13.32	–	27.36	37.06
	OV-MMP	7.15	6.54	8.29	24.00	23.04	30.14	38.52	33.44	43.80
	Ours	**10.24**	**7.26**	**9.82**	**24.12**	**23.23**	**30.17**	**38.75**	**34.68**	**43.95**

two benchmark datasets: VidVRD and VidOR. Note that due to the unavailability of publicly accessible models or implementations for RePro (Gao et al., 2023) and UASAN (Wu et al., 2024b) on the SGDet task of the VidOR dataset, we do not report their results on this particular dataset.

Tables 1 and 2 present the quantitative comparison results on the VidVRD and VidOR datasets, respectively. Based on the experimental findings, we derive several important insights: (1) Our method achieves competitive or superior performance compared to baseline approaches on the majority of tasks and datasets, demonstrating the effectiveness of our semantic teacher-guided enhancement framework combined with advanced spatio-temporal relationship modeling. The integration of multimodal large language models (MLLM) for semantic guidance and learnable prompt-based text encoding enables more robust feature representations for both seen and unseen categories. (2) The most remarkable achievement of our approach lies in its strong generalization ability to unseen categories in open-vocabulary scenarios. On the VidVRD dataset's novel-split evaluation, our method attains 17.82% mAP on SGDet compared to UASAN's (Wu et al., 2024b) 11.05% (61% relative improvement), 18.33% mAP on SGCls versus UASAN's 14.50% (26% enhancement), and 22.31% mAP on PredCls against UASAN's 17.62% (27% gain). This superior zero-shot performance highlights the effectiveness of our class enhancement relationship prediction module, which leverages meta-networks and adaptive text encoders to bridge the semantic gap between visual features and textual concepts through learnable prompt tuning, enabling robust generalization to previously unseen relationship categories. (3) On the all-split evaluation, our method maintains strong performance with notable improvements on most metrics, achieving 28.61% mAP on SGDet (vs. UASAN's Wu et al., 2024b 23.57%) and 39.05% mAP on PredCls (vs. UASAN's 38.43%). However, on SGCls, UASAN (Wu et al., 2024b) slightly outperforms our approach (32.24% vs. 30.45%), indicating that our method may face challenges when dealing with complex scenes containing multiple overlapping relationships in seen categories. On the VidOR dataset, our approach demonstrates consistent improvements, achieving 1.55% mAP on novel-split SGDet (vs. OV-MMP's Yang et al., 2024 0.84%) and 10.24% mAP on all-split SGDet (vs. OV-MMP's 7.15%).The absolute numerical gains on VidOR appear suppressed due to two primary factors. First, VidOR focuses on compositional novelty, where the marginal returns of semantic reasoning are naturally lower compared to the intent-heavy VidVRD. Second, the dataset suffers from inherent annotation noise, characterized by coarse temporal boundaries and missing labels, which often penalize precise predictions. Despite these challenges, our method achieves a remarkable ~85% relative improvement over OV-MMP on SGDet and a 140% gain over UASAN on SGCls. This confirms that SAGE maintains superior performance and robustness even in noisy environments where universal metric suppression exists.

4.4. Ablation studies

Effectiveness of Progressive Component Integration. To validate the necessity of each core component in our framework, we conduct comprehensive ablation studies by systematically removing key modules from our complete model. As demonstrated in Table 3, the progressive integration of components reveals clear performance improvements, particularly evident in novel category scenarios. Spatial-temporal modeling forms the foundation of our approach. The removal of spatial modeling (w/o Spa) yields the most significant performance drop, demonstrating that understanding spatial relationships between objects is fundamental for accurate relationship detection. Without temporal modeling (w/o Tem), the system shows improved performance

Table 3
Ablation analysis of progressive component integration on the VidVRD dataset.Note that we add linear layers to keep similar amount parameters when spatial or temporal modules are absent. “Spa”, “Tem”, “SE”, and “CE” denote spatial transformer, temporal transformer, semantic enhancement, and class enhancement, respectively.

Component	Novel		All	
	SGDet	PredCls	SGDet	PredCls
w/o Spa	12.10	15.41	24.27	36.39
w/o Tem	13.92	17.54	26.79	36.73
w/o SE	15.04	19.83	26.35	37.10
w/o CE	16.70	19.88	27.39	35.83
Ours	**17.82**	**22.31**	**28.61**	**39.05**

compared to spatial-only modeling, indicating that capturing temporal dynamics provides additional valuable information for relationship understanding. The combination of both spatial and temporal modeling establishes a strong baseline for our framework. Semantic enhancement bridges visual and textual understanding. The integration of MLLM-guided semantic enhancement (comparing w/o SE to w/ SE) shows meaningful improvements, particularly in novel-split scenarios. This validates our hypothesis that explicit semantic guidance helps the model better understand relationship concepts beyond pure visual appearance, enabling more effective cross-modal knowledge transfer. Class enhancement enables open-vocabulary generalization. The final component, class enhancement through learnable prompt tuning and meta-network guidance, provides the crucial capability for handling unseen categories. The performance gain from w/o CE to the complete model is particularly pronounced in novel-split evaluation, confirming that adaptive text encoding and knowledge preservation mechanisms are essential for zero-shot relationship detection. The consistent performance improvements across all evaluation metrics validate our incremental design philosophy, where each module builds upon previous capabilities to create a comprehensive framework for open-vocabulary video scene graph generation.

Effect of Different MLLM Backbones. To investigate the impact of different multimodal architectures on relational understanding performance, we conduct comprehensive ablation studies comparing ASMv2 with BLIP-2 and baseline configurations without MLLM components. As presented in Table 4, experimental results consistently demonstrate that the specialized architecture of ASMv2 achieves superior performance compared to BLIP-2 across all evaluation metrics. This performance gain can be attributed to the unified relation modeling framework of ASMv2, which effectively integrates visual grounding capabilities with relational reasoning processes within a single architecture. In contrast to the conventional approach of BLIP-2 that predominantly emphasizes holistic image understanding, ASMv2 explicitly captures spatial relationships and object interactions through its structured relation representation mechanism. The ability of the model to concurrently process object entities alongside their relational contexts enables more robust comprehension of complex spatial configurations and dynamic interactions. Furthermore, we adopt ASMv2 over larger language models (e.g., Llama-2) due to computational efficiency considerations and output quality requirements, as larger models typically generate verbose and redundant descriptions that can introduce unwanted noise in downstream feature extraction applications.

Effectiveness of Multi-Modal Fusion Strategies. To determine the optimal approach for integrating semantic textual features with visual trajectory representations in spatio-temporal relationship modeling, we evaluate different fusion strategies within our framework. Our proposed method employs dot cross-attention mechanisms to effectively fuse semantic description features generated by ASMv2 with visual motion features from object trajectories. As presented in Fig. 3, we compare

Table 4
Comparison of different MLLM backbones for semantic enhancement on the VidVRD dataset.

MLLM Backbone	Novel		All	
	SGDet	PredCls	SGDet	PredCls
No MLLM	15.04	19.83	26.35	37.10
BLIP-2	17.61	21.12	27.79	37.94
ASMv2	**17.82**	**22.31**	**28.61**	**39.05**

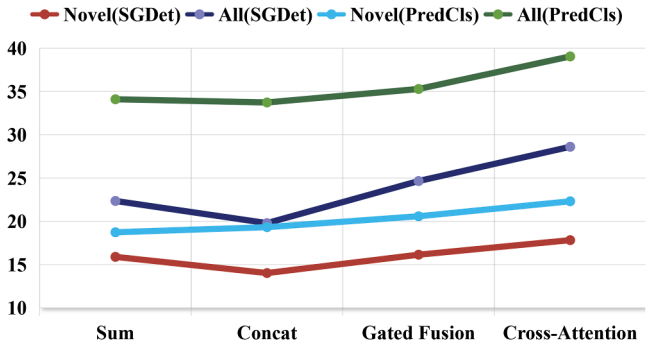


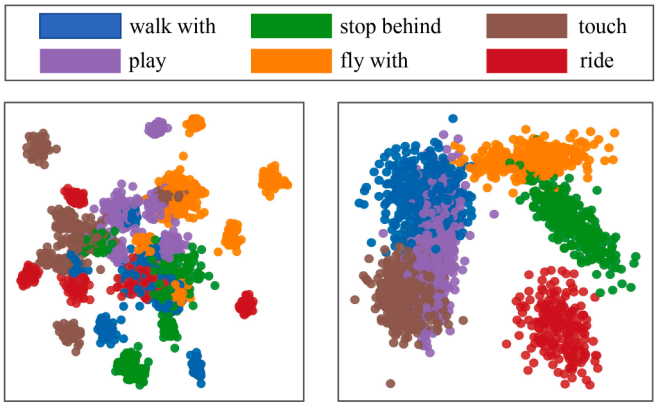
Fig. 3. Ablation study on different fusion strategies.

various fusion strategies on the VidVRD dataset. Baseline approaches include simple summation (Sum) and direct concatenation (Concat) of multi-modal features before feeding into the spatio-temporal transformer. The gated fusion method adaptively controls the contribution weights of different modalities through learnable gating mechanisms. Experimental results demonstrate that our cross-attention fusion strategy consistently outperforms all alternative approaches across evaluation scenarios. This superior performance stems from the cross-attention mechanism’s ability to dynamically select and weight relevant semantic textual features based on visual trajectory information, thereby enabling more precise relationship modeling and prediction in subsequent spatio-temporal transformers.

Visualizing Semantic Stability. To empirically substantiate the limitations of image-conditional prompts, we provide a t-SNE visualization on the test set (Fig. 4). We randomly selected six predicate categories and randomly sampled video frames within them, ensuring the inclusion of diverse environmental variations such as diverse lighting conditions, motion blur, and viewpoints. As illustrated, the baseline (a) exhibits scattered distributions and significant inter-class overlap, confirming that instance-specific visual features introduce substantial noise and semantic drift. Conversely, our class-aware approach (b) yields compact and well-separated clusters. This demonstrates that by prioritizing stable category-level semantics over volatile instance contexts, our method effectively maintains feature discriminability and enhances open-vocabulary generalization even under complex visual perturbations.

Effectiveness of Class-Aware Prompt Generation. We evaluate our decoupling strategy by comparing existing methods with the proposed class-aware strategy (Table 5). While existing methods rely on volatile instance-level visual features, our strategy anchors prompt generation to stable category-specific text embeddings. As shown, our method yields significant improvements, particularly in novel scenarios (+1.12 in SGDet and +2.43 in PredCls). This confirms that decoupling prompts from visual appearance effectively mitigates instance-level noise and semantic drift, thereby providing superior generalization for open-vocabulary relationship detection.

Effectiveness of Adapter Integration Mechanisms. To investigate the optimal configuration for incorporating adapter modules within our framework, we conduct ablation studies examining different combinations of visual and textual adapters. As presented in Table 6, we



(a)Image-conditional Prompt Tuning (b)Class-aware Prompt Tuning

Fig. 4. Comparison of t-SNE visualization between image-conditional prompt tuning and class-aware prompt tuning.

Table 5
Ablation study on prompt generation strategy on the VidVRD dataset.

Prompt Strategy	Novel		All	
	SGDet	PredCls	SGDet	PredCls
Image-Cond	16.70	19.88	27.39	35.83
Class-Aware	**17.82**	**22.31**	**28.61**	**39.05**
Improvement	*+1.12*	*+2.43*	*+1.22*	*+3.22*

Table 6
Ablation study for adapter integration mechanisms on the VidVRD dataset. "Vis" and "Txt" denote visual adapter and text adapter, respectively.

Vis	Txt	Novel		All	
		SGDet	PredCls	SGDet	PredCls
		17.29	20.51	26.58	34.54
✓		17.30	20.72	26.67	34.80
	✓	17.67	21.89	26.76	37.62
✓	✓	**17.82**	**22.31**	**28.61**	**39.05**

evaluate three distinct configurations: utilizing only the visual adapter, employing solely the textual adapter, and integrating both adapters simultaneously. The experimental results reveal that the textual adapter demonstrates superior performance compared to the visual adapter when employed independently. Specifically, the textual adapter achieves higher scores across all evaluation metrics, suggesting its enhanced capability in capturing semantic relationships within the spatio-temporal modeling framework. Furthermore, the simultaneous deployment of both visual and textual adapters yields the most substantial performance improvements across all evaluation scenarios. This synergistic effect indicates that the visual and textual adapters capture complementary aspects of multimodal information, with the visual adapter focusing on spatial-temporal dynamics while the textual adapter emphasizes semantic relationship understanding. The combined approach effectively leverages the strengths of both modalities, resulting in more comprehensive feature representations that enhance overall relationship prediction accuracy.

Effect of Meta-Network Bottleneck Dimension. We investigate the impact of bottleneck dimension in our meta-network architecture, with results presented in Table 7. Setting the dimension to 64 yields the best performance across evaluation metrics. Lower dimensions result in degraded performance as the compressed representation lacks sufficient capacity to preserve essential semantic information during knowledge transfer. Conversely, higher dimensions show diminishing returns due to increased parameter redundancy that may introduce noise and overfitting, particularly given the limited training data in few-shot scenarios. The 64-dimensional configuration strikes an optimal balance between representational adequacy and parameter efficiency.

Table 7
Ablation study for meta-network bottleneck dimension on the VidVRD dataset. The dimension represents the hidden layer size in the meta-network architecture.

Bottleneck Dim	Novel		All	
	SGDet	PredCls	SGDet	PredCls
32	16.47	21.84	28.12	38.82
64	**17.82**	**22.31**	**28.61**	**39.05**
128	17.34	21.72	27.66	38.44
256	16.56	21.64	27.74	38.90

Table 8
Ablation study for meta-network integration position on the VidVRD dataset. The layer indicates where class-aware prompts are inserted in the CLIP text encoder.

Integration Layer	Novel		All	
	SGDet	PredCls	SGDet	PredCls
2	16.98	21.28	27.42	38.32
4	17.03	21.41	27.59	38.35
6	**17.82**	**22.31**	**28.61**	**39.05**
8	17.11	21.81	27.66	38.79
10	17.12	21.59	27.32	38.27

Table 9
Ablation study on different prompt template initialization strategies.

Init Type	Novel		All	
	SGDet	PredCls	SGDet	PredCls
Fixed	16.29	21.86	27.97	38.96
Learnable	16.92	21.75	27.29	37.63
Domain-specific	**17.82**	**22.31**	**28.61**	**39.05**
Random Init	17.37	21.47	27.69	38.54

Effect of Meta-Network Integration Position. Our meta-network approach requires determining where to insert class-aware prompts within the text encoder layers. We examine prompt insertion across different transformer layers, with results presented in Table 8. Layer 6 delivers optimal performance for our framework. Inserting prompts at earlier layers fails to effectively leverage the evolved semantic representations, while insertion at later layers disrupts the final embedding stabilization process. Layer 6 represents the sweet spot where semantic features are sufficiently developed for meaningful class-aware enhancement.

Effectiveness of Prompt Template Initialization. Our analysis of different prompt initialization strategies is presented in Table 9. The results clearly indicate that a **domain-specific initialization**, using the template "Two entities in video sequence," consistently achieves the best performance across all metrics. Generic, hand-crafted prompts are less effective, regardless of being frozen or learnable, because their semantic starting point is poorly aligned with the video relationship task. While random initialization serves as a competitive baseline, it lacks the beneficial inductive bias of a domain-relevant template, resulting in suboptimal performance. These findings confirm that a carefully chosen, domain-specific prompt is crucial for providing an effective starting point that guides the optimization process toward a more robust solution.

Efficiency Comparison. To assess the computational efficiency of SAGE, we compare its parameter count and inference speed with the OV-MMP on the VidVRD dataset. As shown in Table 10, despite a moderate increase in model size, SAGE achieves substantially higher inference speed under both SGDet and PredCls settings. This improvement

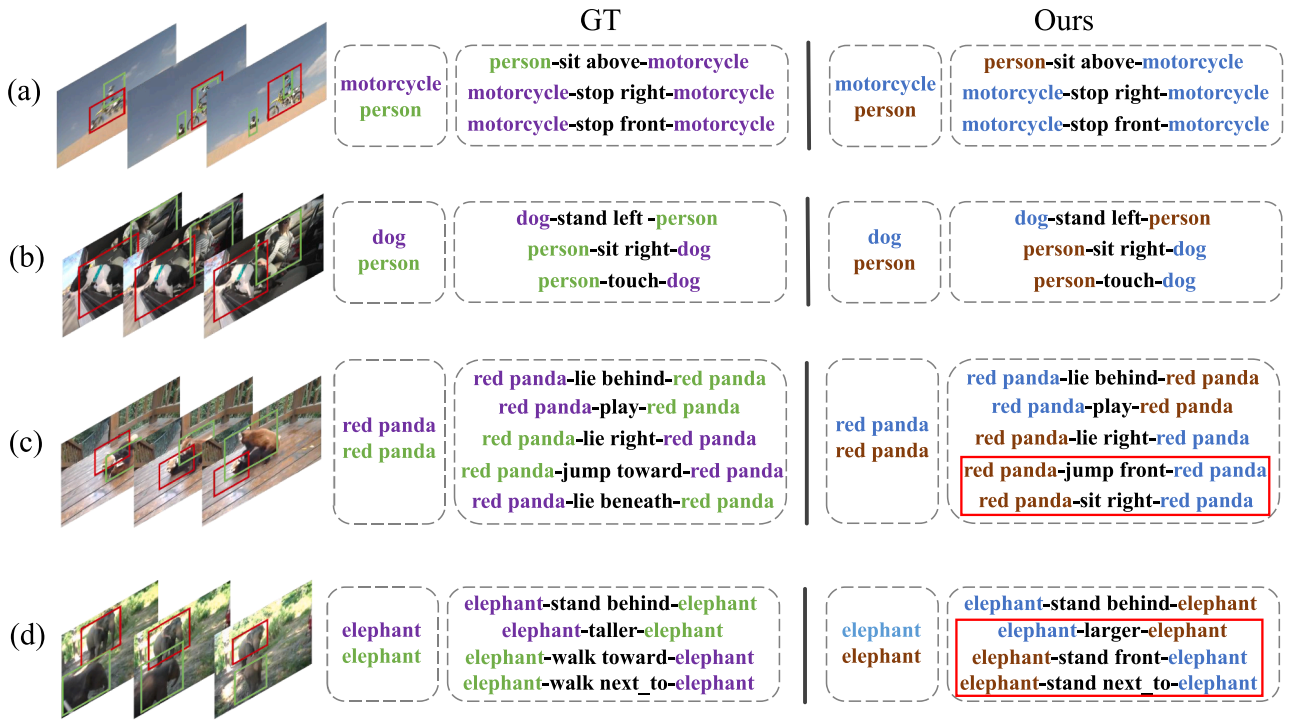


Fig. 5. Qualitative examples of our model's prediction results. For each sequence, we compare our model's predictions (Ours) against the ground truth (GT). Relationship triplets that are incorrectly detected or classified by our model are enclosed in red boxes.

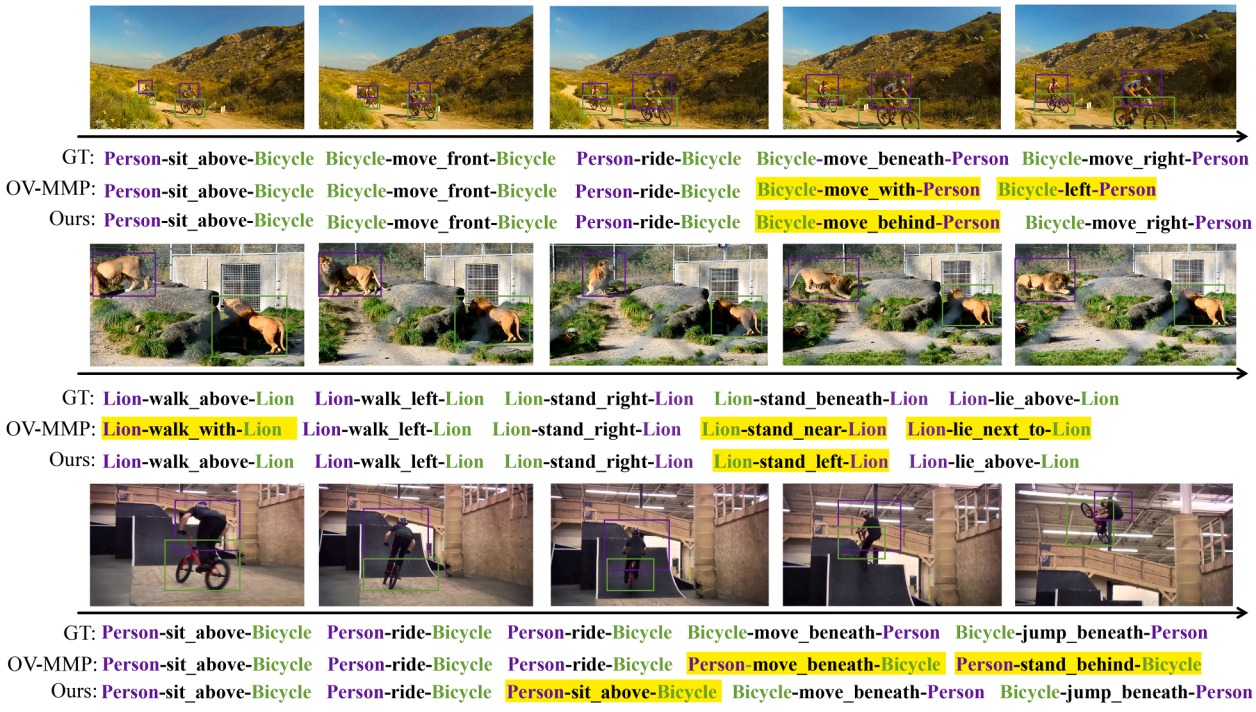


Fig. 6. Side-by-side comparison of failure cases.

primarily stems from the category-level semantic conditioning in SAGE, which avoids redundant per-pair computation required by the baseline's image-conditioned prompting. Notably, although SAGE leverages an MLLM (ASMv2) during training to generate semantic features, these features are pre-computed offline and do not incur additional runtime cost during inference.

4.5. Cross-dataset evaluation

To rigorously evaluate the model's generalization capability on truly unseen and rare relational concepts, we conduct a cross-dataset evaluation. We train the model on the base categories of the VidOR dataset and evaluate it directly on the VidVRD dataset, excluding overlapping

Table 10
Comparison of model complexity and inference speed between the OV-MMP and the proposed SAGE on the VidVRD dataset. Despite a moderately larger model size, SAGE achieves substantially higher inference speed under both SGDet and PredCls settings.

Model	Params (M)	Inference Speed (FPS)	
		SGDet	PredCls
OV-MMP	47.03	6.28	5.79
Ours	55.64	**121.62**	**139.24**

Table 11
Comparison of cross-dataset transferred models on the SGDet task of the VidVRD dataset. Models are trained on VidOR and tested on VidVRD.

Setting	Method	mAP
Cross dataset	ALPro	0.29
	VidVRD-II	0.88
	RePro	1.11
	OV-MMP	1.14
	Ours	**3.42**

training categories. This setting introduces substantial challenges: only 18 VidVRD object categories and 14 relationship categories appear in VidOR's novel split, while 17 objects and 118 relationships remain non-overlapping. Furthermore, the temporal dynamics differ vastly, with VidOR videos averaging 34.6 s compared to just 9.7 s in VidVRD. These significant discrepancies serve as a stringent test for generalizing to unfamiliar video scenes.

The results are presented in Table 11. SAGE achieves 3.42% mAP on this challenging split, significantly outperforming the strong baseline OV-MMP (1.14%) by approximately 200%. The substantial relative margin demonstrates that our Class-Aware Prompting mechanism learns robust, semantic-rich representations, effectively bridging the gap to unfamiliar and rare concepts.

4.6. Qualitative analysis

To provide an intuitive understanding of our model's capabilities, we present several qualitative examples in Fig. 5, which illustrate our model's performance in diverse scenarios by comparing its predictions (Ours) against the ground truth (GT). Our method demonstrates strong performance in recognizing multiple, simultaneous relationships. For instance, in Figure 5(a) and (b), the model successfully detects combinations of static spatial relations and fine-grained interactions, showcasing its robust comprehension abilities in common scenarios. However, the limitations of our model become apparent in scenes with significant visual ambiguity, such as those with heavy occlusion or highly similar objects (Figure 5(c) and dummyTXdummy-(5(d)). In these challenging situations, the model's ability to distinguish between highly similar predicates is compromised, occasionally confusing complex dynamic actions with simpler static states or other similar motions.To investigate these limitations systematically, we conducted a post-hoc manual inspection on a subset of 100 sampled failure cases. This analysis reveals that while the baseline model frequently suffers from semantic misalignment, our method effectively suppresses such errors, with the majority of remaining failures stemming from inevitable occlusion or spatial ambiguity. Complementing the qualitative results in Fig. 5, we present representative side-by-side comparisons selected from this analysis in Fig. 6. These examples specifically categorize the errors into occlusion-induced versus semantic misalignment types, providing a more comprehensive understanding of the model's failure modes.

5. Conclusion

We presents SAGE, a Semantic-Guided Framework with Decoupled Optimization, which addresses two fundamental challenges in open-vocabulary video visual relationship detection: semantic ambiguity caused by implicit visual-semantic alignment, and semantic drift arising from coupled, instance-conditioned prompt optimization.our approach decouples explicit semantic reasoning from classifier adaptation by introducing a training-time multimodal semantic teacher and a class-aware prompt optimization strategy. This design effectively improves semantic discriminability while maintaining high inference efficiency.Extensive experiments on the VidVRD and VidOR benchmarks demonstrate that SAGE achieves state-of-the-art performance, with particularly strong gains on novel relationship categories, validating its effectiveness in challenging open-vocabulary settings.Despite these improvements, our analysis reveals that while SAGE substantially reduces errors caused by high-level semantic misalignment, a large portion of remaining failures stem from low-level perceptual challenges, and spatial misalignment between interacting objects. These limitations suggest that further advances in spatio-temporal perception and geometry-aware reasoning are necessary.Future work will explore more robust modeling of occlusion and spatial dynamics, as well as extending the proposed decoupled semantic-guided paradigm to broader video-and-language tasks, such as dense video captioning and complex event understanding.

CRediT authorship contribution statement

Shiqi Wang: Writing – original draft; **Weiying Xue:** Methodology, Writing – original draft; **Shuyi Hu:** Investigation; **Haowen Li:** Formal analysis; **Qi Liu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. et al. (2022). Flamingo: A visual language model for few-shot learning. In *Proceedings of the advances in neural information processing systems (neurIPS)* (*vol. 35*). (pp. 23716–23736).

Cao, Q., & Huang, H. (2023). Video visual relation detection with contextual knowledge embedding. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), 13083–13095.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., & Zhang, K. (2023). Plot: Prompt learning with optimal transport for vision-language models. In *Proceedings of the international conference on learning representations (ICLR)*.

Chen, S., Shi, Z., Mettes, P., & Snoek, C. G. M. (2021). Social fabric: Tubelet compositions for video relation detection. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 13485–13494).

Chen, Y., Cao, Y., Hu, H., & Wang, L. (2020). Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10337–10346).

Chu, X., Qiao, L., Zhang, X., Xu, S., Wei, F., Yang, Y., Sun, Y., Hu, Y., Wang, X., & Zhang, B. (2024). MobileVLM v2: Faster and stronger baseline for vision language models. arXiv preprint arXiv:2402.03766.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 16372–16382).

Gao, K., Chen, L., Niu, Y., Shao, J., & Xiao, J. (2022a). Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 19497–19506).

Gao, K., Chen, L., Zhang, H., Xiao, J., & Sun, Q. (2023). Compositional prompt tuning with motion cues for open-vocabulary video relation detection. arXiv preprint arXiv:2302.00268.

Gao, M., Xing, C., Niebles, J. C., Li, J., Xu, R., Liu, W., & Xiong, C. (2022b). Open vocabulary object detection with pseudo bounding-box labels. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 266–282). Springer.

Gu, X., Lin, T.-Y., Kuo, W., & Cui, Y. (2021). Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

He, T., Gao, L., Song, J., & Li, Y.-F. (2022). Towards open-vocabulary scene graph generation with prompt-based finetuning. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 56–73). Springer.

Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., & Globerson, A. (2023). Incorporating structured representations into pretrained vision & language models using scene graphs. arXiv preprint arXiv:2305.06343.

Ji, Z., Li, Z., Zhang, Y., Wang, H., Pang, Y., & Li, X. (2024). Hierarchical matching and reasoning for multi-query image retrieval. *Neural Networks*, *173*, 106200.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the international conference on machine learning (ICML)* (pp. 4904–4916). PMLR.

Jia, J., Hu, Y., Weng, X. et al. (2024). TinyLLaVA factory: A modularized codebase for small-scale large multimodal models. arXiv preprint arXiv:2405.11788.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S.-N. (2022). Visual prompt tuning. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 709–727). Springer.

Jiang, X., Zheng, C., Xu, X. et al. (2024). VrdONE: One-stage video visual relation detection. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 1437–1446).

Kan, B., Wang, T., Lu, W. et al. (2023). Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 15670–15680).

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 19113–19122).

Kim, J.-J., Lee, D.-G., Wu, J., Jung, H.-G., & Lee, S. W. (2021). Visual question answering based on local-scene-aware referring expression generation. *Neural Networks*, *139*, 158–167.

Kuo, W., Cui, Y., Gu, X., Piergiovanni, A. J., & Angelova, A. (2022). F-VLM: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639.

Li, D., Li, J., Li, H., Niebles, J. C., & Hoi, S. C. H. (2022a). Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4953–4963).

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the international conference on machine learning (ICML)* (pp. 19730–19742). PMLR.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022b). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the international conference on machine learning (ICML)* (pp. 12888–12900). PMLR.

Li, L., Xiao, J., Chen, G., Shao, J., Zhuang, Y., & Chen, L. (2024). Zero-shot visual relation detection via composite visual cues from large language models. *Proceedings of the advances in neural information processing systems (NeurIPS)*, *vol. 36*.

Li, Y., Yang, X., Shang, X., & Chua, T.-S. (2021). Interventional video relation detection. In *Proceedings of the ACM international conference on multimedia (ACM MM)* (pp. 4091–4099).

Lin, X., Shi, C., Zhan, Y., Yang, Z., Wu, Y., & Tao, D. (2024). TD^2-Net: Toward denoising and debiasing for video scene graph generation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 3495–3503). (*vol. 38*).

Liu, C., Jin, Y., Xu, K., Gong, G., & Mu, Y. (2020). Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10840–10849).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., & Zhou, M. (2020). UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353.

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022). Expanding language-image pretrained models for general video recognition. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 1–18). Springer.

Pham, H., Dai, Z., Ghiasi, G., Kawaguchi, K., Liu, H., Yu, A. W., Yu, J., Chen, Y.-T., Luong, M.-T., Wu, Y. et al. (2023). Combined scaling for zero-shot transfer learning. *Neurocomputing*, *555*, 126658.

Qian, X., Zhuang, Y., Li, Y., Xiao, S., Pu, S., & Xiao, J. (2019). Video relation detection with spatio-temporal graph. In *Proceedings of the ACM international conference on multimedia (ACM MM)* (pp. 84–93).

Qin, Y., Gu, X., & Tan, Z. (2022). Visual context learning based on textual knowledge for image-text retrieval. *Neural Networks*, *152*, 434–449.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the international conference on machine learning (ICML)* (pp. 8748–8763). PMLR.

Shafir, Y., Tevet, G., Kapon, R., & Bermano, A. H. (2023). Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418.

Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., & Chua, T.-S. (2019). Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on international conference on multimedia retrieval* (pp. 279–287).

Shang, X., Li, Y., Xiao, J., Ji, W., & Chua, T.-S. (2021). Video visual relation detection via iterative inference. In *Proceedings of the ACM international conference on multimedia (ACM MM)* (pp. 3654–3663).

Shang, X., Ren, T., Guo, J., Zhang, H., & Chua, T.-S. (2017). Video visual relation detection. In *Proceedings of the ACM international conference on multimedia (ACM MM)* (pp. 1300–1308).

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., & Bermano, A. H. (2022). Human motion diffusion model. In *The eleventh international conference on learning representations*.

Wang, H., Yu, H., & Zhang, Q. (2024a). Human-object interaction detection via global context and pairwise-level fusion features integration. *Neural Networks*, *170*, 242–253.

Wang, W., Ren, Y., Luo, H., Li, T., Yan, C., Chen, Z., Wang, W., Li, Q., Lu, L., Zhu, X., Qiao, Y., & Dai, J. (2024b). The all-seeing project V2: Towards general relation comprehension of the open world. In *European conference on computer vision (ECCV)* (pp. 471–490).

Weng, Z., Yang, M., Li, A., Wu, Z., & Jiang, Y.-G. (2023). Open-VCLIP: Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. In *Proceedings of the international conference on machine learning (ICML)* (pp. 36978–36989). PMLR.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE international conference on image processing (ICIP)* (pp. 3645–3649).

Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X. et al. (2024a). Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(7), 5092–5113.

Wu, Z., Gao, J., & Xu, C. (2024b). Open-vocabulary video scene graph generation via union-aware semantic alignment. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 8566–8575).

Xu, L., Qu, H., Kuen, J., Gu, J., & Liu, J. (2022). Meta spatio-temporal debiasing for video scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 374–390). Springer.

Xu, M., Zhang, Z., Wei, F., Hu, H., & Bai, X. (2023). Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2945–2954).

Xue, W., Liu, Q., Wang, Y., Wei, Z., Xing, X., & Xu, X. (2025). Towards zero-shot human-object interaction detection via vision-language integration. *Neural Networks*, *187*, 107348.

Yang, S., Wang, Y., Ji, X., & Wu, X. (2024). Multi-modal prompting for open-vocabulary video visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (*vol. 38*). (pp. 6513–6521).

Yao, H., Zhang, R., & Xu, C. (2023). Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6757–6767).

Yao, H., Zhang, R., & Xu, C. (2024a). TCP: Textual-based class-aware prompt tuning for visual-language model. arXiv preprint arXiv:2311.18231.

Yao, L., Pi, R., Han, J., Liang, X., Xu, H., Zhang, W., Li, Z., & Xu, D. (2024b). DetCLIPv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27391–27401).

Yu, Q., Li, J., Wu, Y., Tang, S., Ji, W., & Zhuang, Y. (2023). Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 21560–21571).

Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., & Zhao, D. (2023). RLIPv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 21649–21661).

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., & Zou, J. (2022). When and why vision-language models behave like bags-of-words, and what to do about it? In *The eleventh international conference on learning representations (ICLR)*.

Zang, Y., Li, W., Zhou, K., Huang, C., & Loy, C. C. (2022). Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225.

Zhang, G., Tang, Y., Zhang, C. et al. (2024). Entity dependency learning network with relation prediction for video visual relation detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(12), 12425–12436.

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., & Liu, Z. (2022). Motion-diffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001.

Zhao, L., Yuan, L., Gong, B., Cui, Y., Schroff, F., Yang, M.-H., Adam, H., & Liu, T. (2023). Unified visual relationship detection with vision and language models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6962–6973).

Zheng, S., Chen, S., & Jin, Q. (2022). VRDFormer: End-to-end video visual relation detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 18836–18846).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 16816–16825).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, *130*(9), 2337–2348.

Zhu, B., Niu, Y., Han, Y., Wu, Y., & Zhang, H. (2023a). Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 15659–15669).

Zhu, B., Niu, Y., Lee, S., Hur, M., & Zhang, H. (2023b). Debiased fine-tuning for vision-language models by prompt regularization. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (pp. 3834–3842).