METOR: A Unified Framework for Mutual Enhancement of Objects and Relationships in Open-vocabulary Video Visual Relationship Detection

Yongqi Wang[1,2], **Xinxiao Wu**[1,2], **Shuo Yang**[2*]
[1]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
{3120230916, wuxinxiao}@bit.edu.cn, yangshuo@smbu.edu.cn

Abstract

Open-vocabulary video visual relationship detection aims to detect objects and their relationships in videos without being restricted by predefined object or relationship categories. Existing methods leverage the rich semantic knowledge of pre-trained vision-language models such as CLIP to identify novel categories. They typically adopt a cascaded pipeline to first detect objects and then classify relationships based on the detected objects, which may lead to error propagation and thus suboptimal performance. In this paper, we propose **M**utual **E**nhancemen**T** of **O**bjects and **R**elationships (METOR), a query-based unified framework to jointly model and mutually enhance object detection and relationship classification in open-vocabulary scenarios. Under this framework, we first design a CLIP-based contextual refinement encoding module that extracts visual contexts of objects and relationships to refine the encoding of text features and object queries, thus improving the generalization of encoding to novel categories. Then we propose an iterative enhancement module to alternatively enhance the representations of objects and relationships by fully exploiting their interdependence to improve recognition performance. Extensive experiments on two public datasets, VidVRD and VidOR, demonstrate that our framework achieves state-of-the-art performance. Codes are at https://github.com/wangyongqi558/METOR.

1 Introduction

Open-vocabulary video visual relationship detection (Open-VidVRD) [Gao *et al.*, 2023] focuses on detecting relationships between objects in videos, typically represented as triplets of the form $\langle subject, relationship, object \rangle$, following an open-vocabulary setting. In this setting, both object and relationship categories are divided into a base set and a novel set. The model is trained on the base set and is expected to generalize to the novel set during testing, enabling relationship detection in a wider range of open-vocabulary scenarios.

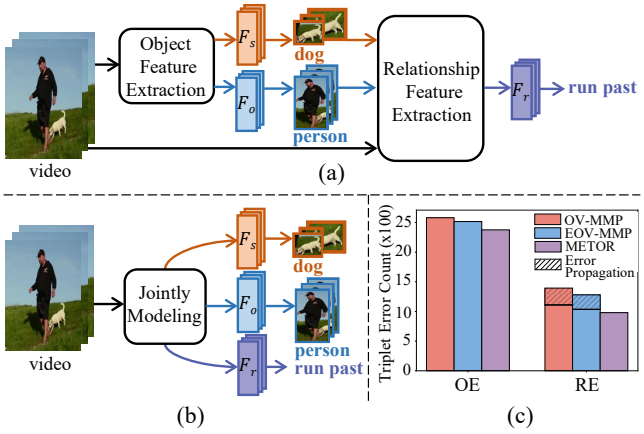

Figure 1: (a) Existing methods adopt a cascaded pipeline. (b) Our method jointly models objects and relationships for their mutually enhancement. (c) Statistics of triplet errors on the VidVRD test set, categorized into object errors (OE) and relationship errors (RE). We also highlight the relationship errors caused by error propagation in the cascaded methods (OV-MMP and EOV-MMP), where the relationship is correctly classified using ground-truth object trajectories but misclassified using detected object trajectories.

Recent advances in pre-trained vision-language models [Radford *et al.*, 2021; Li *et al.*, 2022; Li *et al.*, 2023] have demonstrated significant potential in enhancing open-vocabulary tasks [Liu *et al.*, 2024; Wu *et al.*, 2024b; Fan *et al.*, 2024]. By leveraging vast amounts of vision-language pairs during training, these models effectively encode rich semantic knowledge encompassing entities, actions, scenes, and relationships [Fang *et al.*, 2024; Wang *et al.*, 2021; Liang *et al.*, 2023; Gao *et al.*, 2023]. In the Open-VidVRD task, these pre-trained models have been used to recognize novel object and relationship categories, thereby improving the generalization capability beyond predefined categories. Existing Open-VidVRD methods [Gao *et al.*, 2023; Yang *et al.*, 2024; Wu *et al.*, 2024a; Wang *et al.*, 2025] typically adopt a cascaded pipeline, first detecting objects and then classifying relationships based on the detected objects, as illustrated in Fig. 1 (a). This design often leads to error

*Corresponding author.

propagation, as the inaccuracy of object detection adversely affects the relationship classification, resulting in suboptimal performance. Fig.1 (c) shows the number and causes of triplet errors for different methods on the VidVRD dataset. It is evident that the cascaded methods exhibit a significant increase in relationship errors due to error propagation, underscoring the inherent weaknesses of these methods.

In this paper, we propose **M**utual **E**nhancemen**T** of **O**bjects and **R**elationships (METOR), a query-based unified framework for Open-VidVRD, which jointly models and mutually enhances object detection and relationship classification, as shown in Fig. 1 (b). It simplifies the Open-VidVRD process by adopting a unified modeling strategy, thereby mitigating the error propagation inherent in cascaded pipelines, and emphasizes the role of relationship context in enriching object representations, effectively exploiting the interdependence between objects and relationships to promote their mutual enhancement. As illustrated in Fig. 1 (c), our method reduces both object errors and relationship errors, highlighting the benefits of the proposed framework.

To enhance the generalization ability to novel categories, we design a CLIP-based contextual refinement encoding module that captures the contextual information of objects and relationships to refine the encoding process. Specifically, we incorporate learnable object and relationship tokens into the CLIP visual encoder to capture their respective contexts. These contexts are then used to refine the CLIP-encoded text features and object queries, delivering instance-specific semantic knowledge to improve the adaptability in open-vocabulary scenarios.

To fully exploit the interdependence between objects and relationships for recognition, we propose an iterative enhancement module to alternately enhance the representations of objects and relationships. Specifically, this module consists of multiple iterative enhancement layers, where each layer first uses object features to extract relationship features through spatio-temporal modeling and then uses the extracted relationship features to refine the object features, promoting a continuous mutual enhancement process that enables objects and relationships to iteratively improve each other's representations.

To summarize, the main contributions are as follows:

- We propose METOR, a query-based unified framework that jointly models object detection and relationship classification to effectively exploit their interdependence to promote mutual enhancement, simplifying the process of Open-VidVRD.
- We propose an iterative enhancement module that alternately enhances the representations of objects and relationships by using each other's representations for more accurate recognition.
- We design a CLIP-based contextual refinement encoding module that extracts contexts for objects and relationships to refine the encoding of text features and object queries for better open-vocabulary generalization.

2 Related Work

Video visual relationship detection (VidVRD) [Shang *et al.*, 2017] aims to detect relationships between objects over time in a given video, which has been widely applied to various visual understanding tasks [Zhao *et al.*, 2023; Nguyen *et al.*, 2024; Rodin *et al.*, 2024]. Numerous studies have explored various VidVRD methods, which can be broadly categorized into spatio-temporal modeling [Qian *et al.*, 2019; Tsai *et al.*, 2019; Liu *et al.*, 2020; Cong *et al.*, 2021], relationship refinement [Shang *et al.*, 2021; Chen *et al.*, 2021], video relationship debiasing [Xu *et al.*, 2022; Dong *et al.*, 2022; Lin *et al.*, 2024], and end-to-end video relationship detection [Zheng *et al.*, 2022; Zhang *et al.*, 2023; Jiang *et al.*, 2024]. However, these methods are designed for close-set scenarios and struggle to generalize to open-vocabulary settings, thus being limited in pratical applications.

Open-vocabulary VidVRD (Open-VidVRD) [Gao *et al.*, 2023] has emerged in recent years to extend VidVRD by testing novel categories, thereby enhancing its applicability to real-world scenarios. RePro [Gao *et al.*, 2023], OV-MMP [Yang *et al.*, 2024] and UASAN [Wu *et al.*, 2024a] design prompt learning or semantic alignment modules to better align visual and textual modalities. However, these methods rely on a close-set pre-trained trajectory detector for trajectory detection. Incorrect trajectories can lead to relationship classification errors, causing severe error propagation and hurting the overall performance.

The method most related to ours is EOV-MMP [Wang *et al.*, 2025], which extends OV-MMP into an end-to-end model, eliminating the need for the close-set pre-trained trajectory detector used in previous Open-VidVRD methods. It jointly optimizes the object detection and relationship classification modules, introducing an auxiliary loss to capture relationship context during object detection. However, EOV-MMP only perceives the relationship context without explicitly leveraging it to enhance object representations. In addition, it still follows a cascaded pipeline that first detects objects and then classifies relationships. In contrast, our METOR jointly models objects and relationships, and fully exploits their interdependence to promote mutual enhancement.

3 Our Framework

3.1 Problem Definition

Video Visual Relationship Detection (VidVRD) involves identifying visual relationship instances in a video sequence $\mathcal{V} = \{f_t\}_{t=1}^{T}$, where f_t represents the frame at timestamp t, and T denotes the total number of frames. Each visual relationship instance is represented as a tuple $(s, r, o, \tau_s, \tau_o)$, where s, r, and o indicate the subject, relationship, and object categories, respectively. τ_s and τ_o represent the trajectories of subject and object, respectively, defined as sequences of bounding boxes $\{b_t^s\}$ and $\{b_t^o\}$ over a temporal span, respectively, where t ranges from t_{start} to t_{end}, indicating the start and end times of each trajectory. In Open-VidVRD, categories are divided into base and novel splits, including base object categories $\mathbb{C}_o^b$, novel object categories $\mathbb{C}_o^n$, base relationship categories $\mathbb{C}_r^b$, and novel relationship categories $\mathbb{C}_r^n$. Training is only performed on base categories, while evaluation includes both base and novel categories to evaluate the

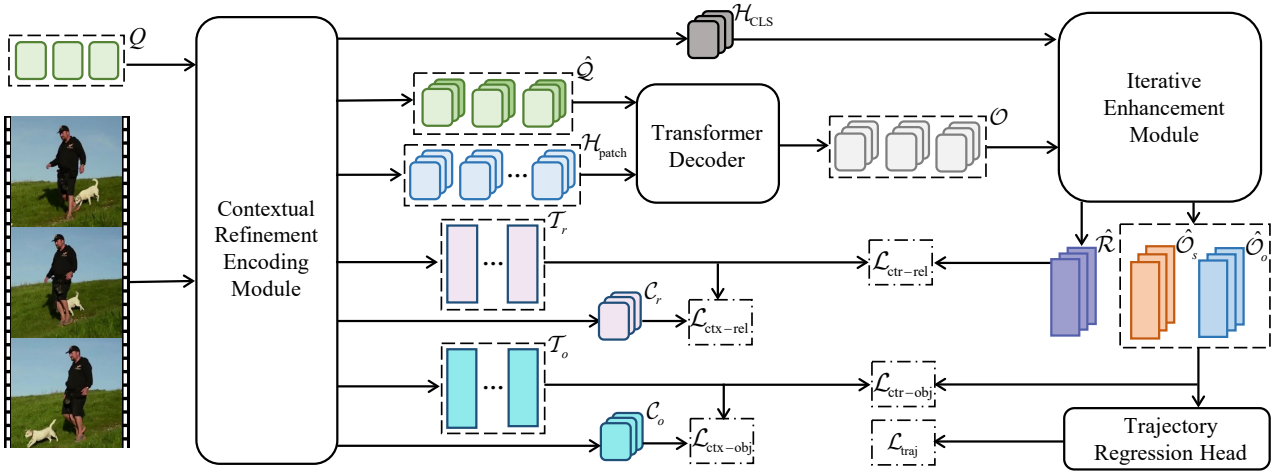

Figure 2: Overview of the proposed framework.

model's generalization.

3.2 Overview

In this paper, we propose METOR, a query-based unified framework that jointly models object detection and relationship classification to effectively exploit their interdependence to promote mutual enhancement, streamlining the process of Open-VidVRD. The framework comprises two key modules: a contextual refinement encoding module (Sec. 3.3) and an iterative enhancement module (Sec. 3.4). An overview of METOR is illustrated in Fig. 2.

For a given video $\mathcal{V}$, we input it into the contextual refinement encoding module along with N_q learnable object queries $\mathcal{Q}$. This module outputs CLS embeddings $\mathcal{H}_{\text{CLS}}$, patch embeddings $\mathcal{H}_{\text{patch}}$, refined text features $\mathcal{T}_o$ for objects, refined text features $\mathcal{T}_r$ for relationships, refined object queries $\hat{\mathcal{Q}}$, object context embeddings $\mathcal{C}_o$ and relationship context embeddings $\mathcal{C}_r$, formulated by

$$(\mathcal{H}_{\text{CLS}}, \mathcal{H}_{\text{patch}}, \mathcal{T}_o, \mathcal{T}_r, \hat{\mathcal{Q}}, \mathcal{C}_o, \mathcal{C}_r) = \Phi(\mathcal{V}, \mathcal{Q}), \tag{1}$$

where $\Phi(\cdot)$ denotes the contextual refinement encoding module.

Then, the refined object queries and patch embeddings are then passed through a Transformer decoder to generate visual object features:

$$\mathcal{O} = Decoder(\hat{\mathcal{Q}}, \mathcal{H}_{\text{patch}}), \tag{2}$$

where $Decoder(\cdot)$ is the Transformer decoder, and $\mathcal{O}$ denotes the visual features of N_q objects in the video.

Next, the visual object features and the CLS embeddings containing global semantic information are fed into the iterative enhancement module to generate mutually enhanced visual subject feature $\hat{\mathcal{O}}_s$, visual object feature $\hat{\mathcal{O}}_o$, and visual relationship feature $\hat{\mathcal{R}}$:

$$(\hat{\mathcal{O}}_s, \hat{\mathcal{O}}_o, \hat{\mathcal{R}}) = \Psi(\mathcal{O}, \mathcal{H}_{\text{CLS}}), \tag{3}$$

where $\Psi(\cdot)$ denotes the iterative enhancement module.

Finally, the mutually enhanced features are matched with the corresponding textual features to predict subject score $\mathcal{S}_s$, object score $\mathcal{S}_o$, and relationship score $\mathcal{S}_r$:

$$\begin{aligned} \mathcal{S}_s &= \sigma(\cos(\hat{\mathcal{O}}_s, \mathcal{T}_o)), \\ \mathcal{S}_o &= \sigma(\cos(\hat{\mathcal{O}}_o, \mathcal{T}_o)), \\ \mathcal{S}_r &= \sigma(\cos(\hat{\mathcal{R}}, \mathcal{T}_r)), \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, and $\cos(\cdot,\cdot)$ represents the cosine similarity. The trajectories for the subject and object are predicted as

$$\begin{aligned} \tau_s &= M_b(\hat{\mathcal{O}}_s), \\ \tau_o &= M_b(\hat{\mathcal{O}}_o), \end{aligned} \tag{5}$$

where $M_b(\cdot)$ denotes a trajectory regression head implemented as a multi-layer perceptron.

3.3 Contextual Refinement Encoding Module

The main goal of Open-VidVRD is to discover novel categories. To improve the generalization ability to novel categories, we propose a contextual refinement encoding module that extracts contexts for objects and relationships to refine the encoding of text features and object queries. An illustration of this module is shown in Fig. 3.

The video $\mathcal{V}$ is first divided into fixed-sized and non-overlapping patches, which are then linearly projected into 1D tokens:

$$h_{\text{patch}} = L(P(\mathcal{V})), \tag{6}$$

where $P(\cdot)$ represents the patchification operation, and $L(\cdot)$ is the linear projection layer. $h_{\text{patch}} \in \mathbb{R}^{T \times N_p \times d}$ represents the patch tokens, where T is the number of video frames, and N_p is the number of patches per frame.

The patch tokens, concatenated with learnable CLS tokens h_{CLS}, learnable object context tokens c_o and learnable relationship context tokens c_r, are fed into the CLIP ViT visual encoder to generate CLS embeddings $\mathcal{H}_{\text{CLS}}$, patch embeddings $\mathcal{H}_{\text{patch}}$, object context embeddings $\mathcal{C}_o$ and relationship context embeddings $\mathcal{C}_r$:

$$(\mathcal{H}_{\text{CLS}}, \mathcal{H}_{\text{patch}}, \mathcal{C}_o, \mathcal{C}_r) = V([h_{\text{CLS}}; h_{\text{patch}}; c_o; c_r]), \tag{7}$$

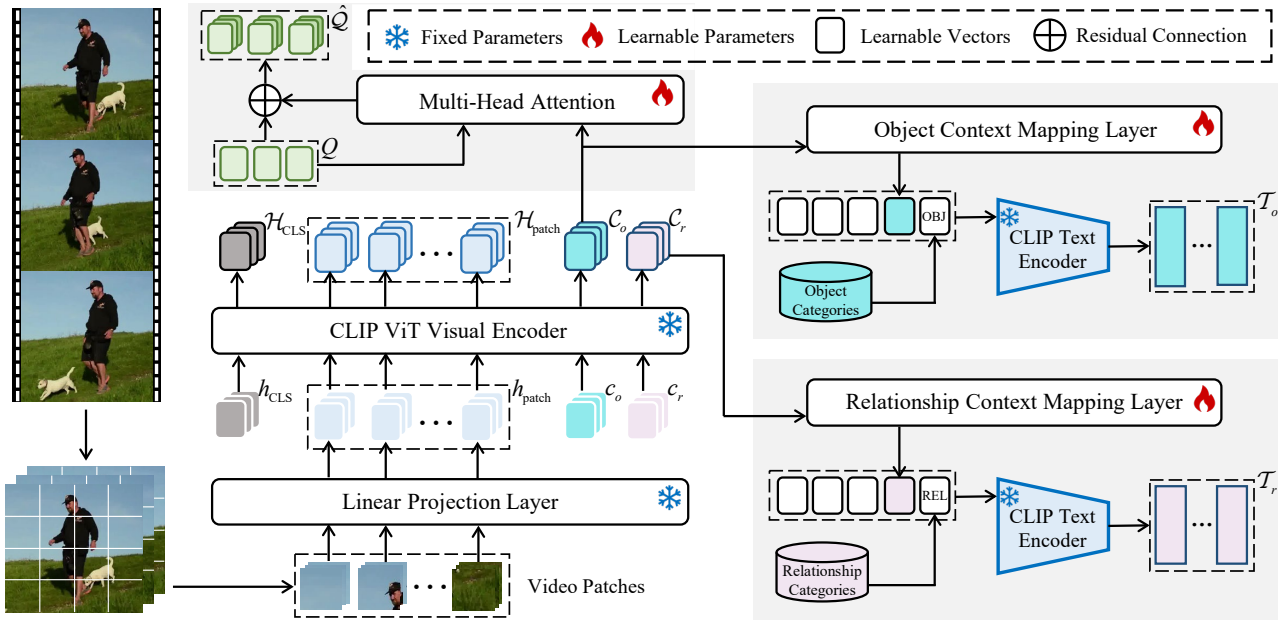

Figure 3: An overview of the proposed contextual refinement encoding module.

where $V(\cdot)$ represents the CLIP ViT visual encoder, and $[\cdot;\cdot]$ denotes the concatenation operation.

To capture contextual dependencies to refine the object queries in each frame, the object context embedding are used as keys and values, and the object queries serve as queries in a multi-head attention mechanism:

$$\hat{\mathcal{Q}} = \text{MHA}(\mathcal{Q}, \mathcal{C}_o, \mathcal{C}_o) + \mathcal{Q}, \tag{8}$$

where $\text{MHA}(\cdot,\cdot,\cdot)$ is the multi-head attention mechanism. The residual connection preserves information in the original object queries while incorporating contextual knowledge from the object context embeddings.

To refine the textual object features, the object context embeddings are first processed through an object context mapping layer, then concatenated with a set of learnable vectors $\mathbf{v}_o$ and an object category embedding OBJ, and finally passed through the CLIP text encoder:

$$\mathcal{T}_o = G([\mathbf{v}_o; \mathcal{M}_o(\mathcal{C}_o); \text{OBJ}]), \tag{9}$$

where $\mathcal{T}_o$ denotes the refined textual object features, $G(\cdot)$ represents the CLIP text encoder, and $\mathcal{M}_o$ denotes the object context mapping layer, implemented as a multi-layer perceptron.

In a similar way, the refined textual relationship features $\mathcal{T}_r$ are learned from the relationship context embeddings $\mathcal{C}_r$:

$$\mathcal{T}_r = G([\mathbf{v}_r; \mathcal{M}_r(\mathcal{C}_r); \text{REL}]), \tag{10}$$

where $\mathbf{v}_r$ denotes a set of learnable vectors, $\mathcal{M}_r$ represents the relationship context mapping layer, implemented as a multi-layer perceptron, and REL corresponds to the relationship category embedding.

After encoding, the refined object queries $\hat{\mathcal{Q}}$ and patch embeddings $\mathcal{H}_{\text{patch}}$ are fed into a Transformer decoder to generate visual object features $\mathcal{O} \in \mathbb{R}^{N_q \times T \times d}$, which represent the visual features corresponding to N_q candidate objects across T video frames. From the visual object features $\mathcal{O}$, we derive subject-object pairs for subsequent recognition, denoted as $(\mathcal{O}_s, \mathcal{O}_o)$, where $\mathcal{O}_s$ and $\mathcal{O}_o$ represent the visual subject and object features, respectively.

3.4 Iterative Enhancement Module

We propose an iterative enhancement module to alternatively enhance the representations of objects and relationships by fully exploiting their interdependence, which consists of N_i iterative enhancement layers, as illustrated in Fig. 4. In each layer, the visual features of the paired subject and object are first concatenated with the CLS embeddings, and then is fed into a spatio-temporal Transformer to generate visual relationship features. These visual relationship features are processed through a relationship feature mapping layer, which in turn enhances the visual subject and object features. This process is expressed as

$$\begin{aligned}
\hat{\mathcal{R}}^{(k)} &= STTrans^{(k)}([\hat{\mathcal{O}}_s^{(k-1)}; \hat{\mathcal{O}}_o^{(k-1)}; \mathcal{H}_{\text{CLS}}]), \\
\hat{\mathcal{O}}_s^{(k)} &= \alpha \hat{\mathcal{O}}_s^{(k-1)} + (1-\alpha) M_f^{(k)}(\hat{\mathcal{R}}^{(k)}), \\
\hat{\mathcal{O}}_o^{(k)} &= \alpha \hat{\mathcal{O}}_o^{(k-1)} + (1-\alpha) M_f^{(k)}(\hat{\mathcal{R}}^{(k)}),
\end{aligned} \tag{11}$$

where $STTrans^{(k)}(\cdot)$ and $M_f^{(k)}(\cdot)$ denote the spatio-temporal Transformer and the relationship feature mapping layer of the k-th iterative enhancement layer, respectively. $\hat{\mathcal{R}}^{(k)}$ represents the visual relationship features, while $\hat{\mathcal{O}}_s^{(k)}$ and $\hat{\mathcal{O}}_o^{(k)}$ are the enhanced visual subject and object features in the k-th layer. α is a balance parameter.

For a subject-object feature pair $(\mathcal{O}_s, \mathcal{O}_o)$, the iterative enhancement module initializes the inputs to the first layer as $\hat{\mathcal{O}}_s^{(0)} = \mathcal{O}_s$ and $\hat{\mathcal{O}}_o^{(0)} = \mathcal{O}_o$, and outputs the the mutually

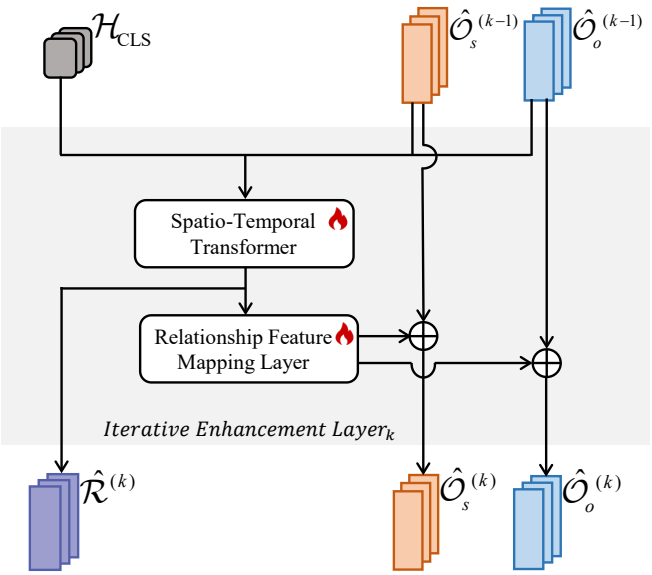

Figure 4: An overview of the proposed iterative enhancement module.

enhanced visual subject, object and relationship features as $\hat{\mathcal{O}}_s = \hat{\mathcal{O}}_s^{(N_i)}$, $\hat{\mathcal{O}}_o = \hat{\mathcal{O}}_o^{(N_i)}$, and $\hat{\mathcal{R}} = \hat{\mathcal{R}}^{(N_i)}$, respectively.

3.5 Training Objective

We train the entire framework in an end-to-end manner. The overall objective function consists of five losses: a relationship contrastive loss $\mathcal{L}_{\text{rel-ctr}}$, an object contrastive loss $\mathcal{L}_{\text{obj-ctr}}$, a trajectory loss $\mathcal{L}_{\text{traj}}$, a relationship contextual loss $\mathcal{L}_{\text{rel-ctx}}$, and an object contextual loss $\mathcal{L}_{\text{obj-ctx}}$, formulated by

$$\mathcal{L} = \mathcal{L}_{\text{rel-ctr}} + \mathcal{L}_{\text{obj-ctr}} + \theta_{\text{traj}}\mathcal{L}_{\text{traj}} + \theta_{\text{ctx}}(\mathcal{L}_{\text{rel-ctx}} + \mathcal{L}_{\text{obj-ctx}}), \quad (12)$$

where θ_{traj} and θ_{ctx} are balance factors.

Contrastive Loss. The relationship contrastive loss is formulated using the binary cross-entropy loss (BCE):

$$\mathcal{L}_{\text{rel-ctr}} = \frac{1}{|\mathbb{C}_r^b|}\text{BCE}(\mathcal{S}_r, \tilde{r}), \quad (13)$$

where $\mathcal{S}_r$ represents the predicted relationship score, $\tilde{r}$ is the ground-truth relationship labels for the subject-object pair, and $\mathbb{C}_r^b$ denotes the set of base relationship categories used during training.

The object contrastive loss is computed using the cross-entropy loss (CE):

$$\mathcal{L}_{\text{obj-ctr}} = \text{CE}(\mathcal{S}_s, \tilde{s}) + \text{CE}(\mathcal{S}_o, \tilde{o}), \quad (14)$$

where $\mathcal{S}_s$ and $\mathcal{S}_o$ are the predicted scores for the subject and object, respectively, and $\tilde{s}$ and $\tilde{o}$ are the ground-truth labels for the subject and object categories.

Trajectory Loss. The trajectory loss consists of a bounding box regression loss and a trajectory consistency loss, given by

$$\begin{aligned} \mathcal{L}_{\text{traj}} &= \mathcal{L}_{\text{box}} + \theta_{\text{cst}}\mathcal{L}_{\text{cst}}, \\ \mathcal{L}_{\text{box}} &= \frac{1}{|T|}\sum_{t=1}^{T}\sum_{e\in\{s,o\}} \text{SL1}(\mathcal{B}_e^{(t)}, \tilde{\mathcal{B}}_e^{(t)}), \\ \mathcal{L}_{\text{cst}} &= \frac{1}{|T-1|}\sum_{t=1}^{T-1}\sum_{e\in\{s,o\}} \|\mathcal{B}_e^{(t+1)} - \mathcal{B}_e^{(t)}\|_1, \end{aligned} \quad (15)$$

where $\mathcal{B}_e^{(t)}$ and $\tilde{\mathcal{B}}_e^{(t)}$ denote the predicted and ground truth bounding boxes for entity $e \in \{s, o\}$ at frame t, derived from the predicted trajectory τ_e and ground truth trajectory $\tilde{\tau}_e$. $\text{SL1}(\cdot)$ represents the Smooth L1 loss for bounding box regression. $\|\cdot\|_1$ is the L1 norm enforcing temporal smoothness by penalizing large deviations between consecutive bounding boxes. θ_{cst} balances the contributions of the bounding box regression loss and the trajectory consistency loss.

Contextual Loss. To capture more effective contextual information for relationships and objects, we define the contextual losses for them using the binary cross-entropy loss (BCE):

$$\begin{aligned} \mathcal{L}_{\text{rel-ctx}} &= \text{BCE}(\cos(\mathcal{C}_r, \mathcal{T}_r), \tilde{\mathcal{R}}), \\ \mathcal{L}_{\text{obj-ctx}} &= \text{BCE}(\cos(\mathcal{C}_o, \mathcal{T}_o), \tilde{\mathcal{O}}), \end{aligned} \quad (16)$$

where $\tilde{\mathcal{R}}$ and $\tilde{\mathcal{O}}$ represent the sets of relationship and object categories present in each frame, respectively.

4 Experiment

4.1 Dataset and Evaluation

Datasets. We evaluate our framework on the VidVRD [Shang *et al.*, 2017] and VidOR [Shang *et al.*, 2019] datasets. VidVRD consists of 1,000 videos, with 800 videos for training and 200 for testing, covering 35 object categories and 132 relationship categories. VidOR contains 10,000 videos, including 7,000 for training, 835 for validation, and 2,165 for testing, covering 80 object categories and 50 relationship categories.

Evaluation Settings. Following Repro [Gao *et al.*, 2023], we designate common object and relationship categories as base categories and rarer ones as novel categories. Training is performed on the base categories, and testing is conducted under two settings: (1) Novel-split evaluation, which includes all object categories and novel relationship categories; (2) All-split evaluation, covering all object and relationship categories as the standard evaluation. Testing is carried out on both the VidVRD test set and the VidOR validation set, as the annotations for the VidOR test set are not available.

Evaluation Tasks. Three evaluation tasks are usually used for VidVRD: scene graph detection (SGDet), scene graph classification (SGCls), and predicate classification (PredCls). SGCls and PredCls are often used to evaluate methods that depend on pre-detected trajectories and are not suitable for our framework which jointly models objects and relationships. So we use SGDet that do not rely on pre-detected trajectories for evaluation.

Evaluation Metrics. 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.*, 2025], we also use mean Average Precision of object trajectory (mAP_o) to evaluate the quality of object detection.

4.2 Implementation Details

In all experiments, key frames are sampled every 30 video frames to form 30-frame video segments. Following [Shang

Split	Method	VidVRD				VidOR			
		mAP	R@50	R@100	mAP$_o$	mAP	R@50	R@100	mAP$_o$
Novel	RePro [Gao *et al.*, 2023]	5.87	12.75	16.23	10.36	-	-	-	-
	UASAN [Liu *et al.*, 2024]	9.88	12.80	17.68	12.15	-	-	-	-
	OV-MMP [Yang *et al.*, 2024]	12.15	13.72	15.21	14.37	0.84	1.44	1.44	1.11
	EOV-MMP [Wang *et al.*, 2025]	15.04	16.03	18.18	36.31	2.45	4.79	4.79	2.33
	METOR (Ours)	**16.74**	**16.72**	**19.43**	**38.91**	**3.75**	**4.86**	**5.32**	**3.37**
All	RePro [Gao *et al.*, 2023]	21.12	12.63	15.42	18.18	-	-	-	-
	UASAN [Liu *et al.*, 2024]	22.93	15.74	18.89	23.74	-	-	-	-
	OV-MMP [Yang *et al.*, 2024]	22.10	13.26	16.08	34.61	7.15	6.54	8.29	3.38
	EOV-MMP [Wang *et al.*, 2025]	26.34	16.48	19.54	52.72	11.08	8.43	**9.82**	12.99
	METOR (Ours)	**27.52**	**16.69**	**19.58**	**55.09**	**12.32**	**8.54**	9.72	**14.02**

Table 1: Comparison with existing methods on VidVRD and VidOR datasets.

et al., 2017; Gao *et al.*, 2023; Liu *et al.*, 2024], visual relationship triplets are generated for video segments and merged using the greedy relation association algorithm proposed in [Shang *et al.*, 2017]. We adopt the ViT-L/14 variant of CLIP with fixed parameters. The number of iterative enhancement layers N_i is set to two for VidVRD and three for VidOR. The hyperparameter α in Eq. 11 is set to 0.9. The loss balance factors θ_{traj}, θ_{ctx}, and θ_{cst} in Eq. 12 and Eq. 15 are set to 1.0, 0.2, and 0.1, respectively. The number of object queries N_q is set to 100. The Transformer decoder is initialized with the parameters of an object detector pre-trained on the MS-COCO dataset [Lin *et al.*, 2014], excluding novel object categories. For object detection results, we retain object trajectories with an average classification score greater than 0.2 and filter bounding boxes using a threshold of 0.35. The optimization process employs the AdamW algorithm [Loshchilov and Hutter, 2019] with an initial learning rate of 1e-4. A multi-step decay schedule is applied at epochs 15, 20, and 25, reducing the learning rate by a factor of 0.1 at each step, and the model is trained for a total of 30 epochs. The batch size is set to 1, which means that only one video is processed at a time. All experiments are conducted using a single NVIDIA GeForce RTX 4090 GPU.

4.3 Comparison Results

We compare our method with existing Open-VidVRD methods, including RePro [Gao *et al.*, 2023], UASAN [Wu *et al.*, 2024a], OV-MMP [Yang *et al.*, 2024] and EOV-MMP [Wang *et al.*, 2025]. Notably, RePro, UASAN, and OV-MMP rely on trajectory detectors pre-trained on a close set. For a fair comparison, we exclude data from novel categories and retrain the trajectory detector to reproduce these methods. Due to the lack of publicly available models or codes of RePro and UASAN on the SGDet task on the VidOR dataset, we does not show their results on the VidOR dataset.

Tab. 1 reports the evaluation results of our method and existing Open-VidVRD methods on the SGDet task on the VidVRD and VidOR datasets under both novel-split and all-split settings. From Tab. 1, we can draw the following observations: (1) METOR achieves improvements over contemporary models on almost all metrics on both datasets, especially substancial gains in mAP and mAP$_o$. This indicates that our method helps a lot improving the performance of object detection and relationship classification by mutually enhancing the representations of objects and relationships. (2) Compared with the all-split, METOR achieves more significant improvements on the novel-split. For instance, in terms of mAP metric on VidOR, our method surpasses the best competing approach by an absolute margin of 1.30% (3.75% vs. 2.45%) and a relative margin of 53.06% on the novel-split, while achieving an absolute margin of 1.24% (12.32% vs. 11.08%) and a relative margin of 11.19% on the all-split. This demonstrates that by leveraging the rich semantic knowledge in CLIP to capture object and relationship contexts to enhance feature representations, our method improves the generalization ability to novel categories.

We also compare METOR with state-of-the-art visual language pre-trained models such as Video-LLMs, and the experimental results provided in Supplementary Materials.

Enc	Itr	Novel		All	
		mAP	mAP$_o$	mAP	mAP$_o$
		11.64	29.38	24.33	48.27
✓		15.16	35.11	25.67	50.30
	✓	13.49	29.38	25.97	52.17
✓	✓	**16.43**	**38.56**	**27.45**	**55.09**

Table 2: Performance of ablation study for the two modules in METOR on the VidVRD dataset. "Enc" and "Itr" denote the contextual refinement encoding module and the iterative enhancement module, respectively.

4.4 Ablation Studies

We conduct comprehensive ablation studies on the VidVRD dataset to assess the contribution of each component.

Effectiveness of Different Modules. To evaluate the effectiveness of the contextual refinement encoding module (denoted as "Enc"), we remove it and use the original CLIP encoder as its replacement for comparison. To evaluate the iterative enhancement module (denoted as "Itr"), we remove it and design a non-mutual enhancement module as replacement. Tab. 2 shows the evaluation results where the consistent improvements in all metrics highlight the effectiveness of the proposed modules in our method. Specifically, incorporating the contextual refinement module boosts performance,

Variant	Novel		All	
	mAP	mAP_o	mAP	mAP_o
w/o CRE	13.49	29.38	25.97	52.17
w/o CRQ	15.32	35.19	26.25	53.30
w/o CRT	14.91	33.26	26.73	54.08
METOR	**16.43**	**38.56**	**27.45**	**55.09**

Table 3: Performance of ablation study for the contextual refinement encoding on the VidVRD dataset.

Iteration Number	Novel		All	
	mAP	mAP_o	mAP	mAP_o
0	15.16	35.11	25.67	50.30
1	16.04	37.82	26.88	53.76
2	**16.43**	**38.56**	**27.45**	55.09
3	16.25	38.28	27.39	**55.16**

Table 4: Performance of model with different iteration numbers on the VidVRD dataset.

particularly for novel categories, suggesting that contextual refinement improves generalization to real-world scenarios. Additionally, adding the iterative enhancement module improves both object and relationship detection, validating the benefits of mutual enhancement.

Effectiveness of Contextual Embeddings. To evaluate the effectiveness of the contextual embeddings in the contextual refinement encoding module, we design several variants of METOR for comparison: (1) "w/o CRE", removing contextual refinement encoding; (2) "w/o CRQ", removing contextual refinement of object queries; (3) "w/o CRT", removing contextual refinement of text features. As shown in Tab. 3, the results demonstrate that METOR outperforms all variants. Removing contextual refinement (whether applied to object queries, text features, or both) leads to significant drops in performance, especially in the novel categories. These comparisons underscore the critical contribution of contextual refinement encoding to relationship detection.

Evaluation of the Number of Iteration. To evaluate the impact of the number of iterations in the iterative enhancement module, we start with a baseline model without the mutual enhancement, that is, the iteration number is set to zero. We then gradually increase the number of iterations by adding more iterative enhancement layers and analyze their impact on the performance. As shown in Tab. 4, the results demonstrate that as the number of iterations increases, the performance improves significantly, peak at two layers, and then decreases slightly.

4.5 Qualitative Analysis

To further evaluate the impact of the iterative mutual enhancement module on feature representation, we visualize the feature distributions of object and relationship categories before and after mutual enhancement. Specifically, we project the features onto a 2D plane using T-SNE [Hinton and Roweis, 2002]. As illustrated in Fig. 5, the features after mutual enhancement exhibit better clustering. Features within the same

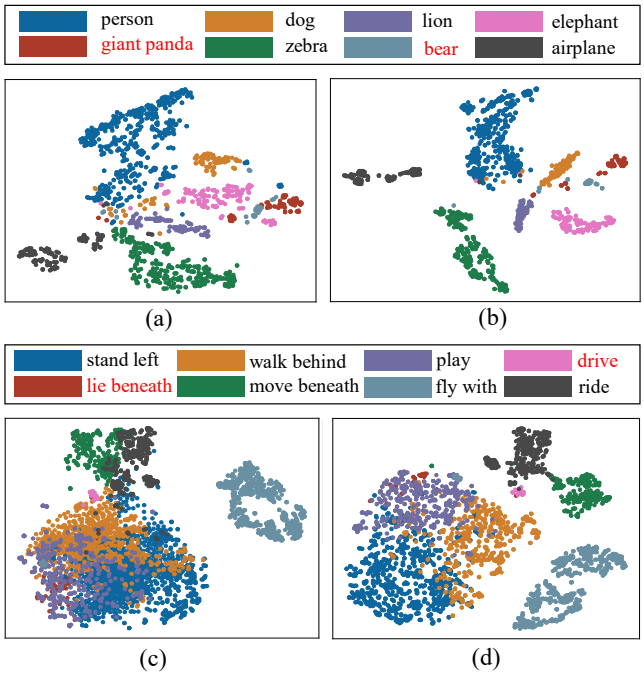

Figure 5: Qualitative results of feature distributions via T-SNE. (a) and (b) display the feature distribution of objects before and after mutual enhancement, while (c) and (d) show the feature distribution of relationships. The categories labeled in red font represent the novel categories.

category become more compact, while those between different categories are more clearly separated. This indicates that the mutual refinement process effectively enhances the feature discrimination of objects and relationships. More qualitative analysis are provided in the Supplementary Materials.

5 Conclusion

In this paper, we propose METOR, a query-based unified framework that jointly models object detection and relationship classification. It is simple yet effective and can mutually enhance object detection and relationship classification by effectively exploiting their interdependence. Under this framework, we design an iterative enhancement module that alternately enhances the representations of objects and relationships by using each other's representation. Additionally, we design a contextual refinement encoding module that extracts contexts for objects and relationships to refine the encoding of text features and object queries. Extensive experimental results on the VidVRD and VidOR datasets demonstrate that our method achieves state-of-the-art performance. In the future, we will explore leveraging additional modalities such as audio and 3D information to support Open-VidVRD in a wider range of applications to enhance scene understanding in dynamic environments.

Acknowledgments

This work was supported by the Shenzhen Science and Technology Program under Grant No. JCYJ20241202130548062,

the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006, and the Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No.2023ZDZX1034.

References

[Chen *et al.*, 2021] Shuo Chen, Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Social fabric: Tubelet compositions for video relation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13485–13494, 2021.

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

[Dong *et al.*, 2022] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022.

[Fan *et al.*, 2024] Lei Fan, Jianxiong Zhou, Xiaoying Xing, and Ying Wu. Active open-vocabulary recognition: Let intelligent moving mitigate clip limitations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16394–16403, 2024.

[Fang *et al.*, 2024] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1716–1725, 2024.

[Gao *et al.*, 2023] 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.

[Hinton and Roweis, 2002] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 15, pages 833–840, 2002.

[Jiang *et al.*, 2024] Xinjie Jiang, Chenxi Zheng, Xuemiao Xu, Bangzhen Liu, Weiying Zheng, Huaidong Zhang, and Shengfeng He. Vrdone: One-stage video visual relation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1437–1446, 2024.

[Li *et al.*, 2022] 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*, pages 4953–4963, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[Liang *et al.*, 2023] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2024] Xin Lin, Chong Shi, Yibing Zhan, Zuopeng Yang, Yaqi Wu, and Dacheng Tao. Td^2-net: Toward denoising and debiasing for video scene graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3495–3503, 2024.

[Liu *et al.*, 2020] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 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*, pages 10840–10849, 2020.

[Liu *et al.*, 2024] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019.

[Nguyen *et al.*, 2024] Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18384–18394, 2024.

[Qian *et al.*, 2019] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. Video relation detection with spatio-temporal graph. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 84–93, 2019.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Rodin *et al.*, 2024] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition, pages 18622–18632, 2024.

[Shang *et al.*, 2017] 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*, pages 1300–1308, 2017.

[Shang *et al.*, 2019] Xindi Shang, Junbin Xiao, Donglin Di, and Tat-Seng Chua. Relation understanding in videos: A grand challenge overview. In *Proceedings of the ACM International Conference on Multimedia*, pages 2652–2656, 2019.

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

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10433, 2019.

[Wang *et al.*, 2021] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[Wang *et al.*, 2025] 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*, pages 1–17, 2025.

[Wu *et al.*, 2024a] 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*, pages 8566–8575, 2024.

[Wu *et al.*, 2024b] Zuxuan Wu, Zejia Weng, Wujian Peng, Xitong Yang, Ang Li, Larry S Davis, and Yu-Gang Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Xu *et al.*, 2022] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In *European Conference on Computer Vision*, pages 374–390. Springer, 2022.

[Yang *et al.*, 2024] 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, pages 6513–6521, 2024.

[Zhang *et al.*, 2023] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. End-to-end video scene graph generation with temporal propagation transformer. *IEEE Transactions on Multimedia*, 26:1613–1625, 2023.

[Zhao *et al.*, 2023] Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5281–5291, 2023.

[Zheng *et al.*, 2022] Sipeng Zheng, Shizhe Chen, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18836–18846, 2022.