Multi-Modal Prompting for Open-Vocabulary Video Visual Relationship Detection

Shuo Yang[1,2*], **Yongqi Wang**[2*], **Xiaofeng Ji**[2], **Xinxiao Wu**[2,1†]

[1]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
[2]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology
Beijing Institute of Technology, China
{shuoyang,3120230916,jixf,wuxinxiao}@bit.edu.cn

Abstract

Open-vocabulary video visual relationship detection aims to extend video visual relationship detection beyond annotated categories by detecting unseen relationships between objects in videos. Recent progresses in open-vocabulary perception, primarily driven by large-scale image-text pre-trained models like CLIP, have shown remarkable success in recognizing novel objects and semantic categories. However, directly applying CLIP-like models to video visual relationship detection encounters significant challenges due to the substantial gap between images and video object relationships. To address this challenge, we propose a multi-modal prompting method that adapts CLIP well to open-vocabulary video visual relationship detection by prompt-tuning on both visual representation and language input. Specifically, we enhance the image encoder of CLIP by using spatio-temporal visual prompting to capture spatio-temporal contexts, thereby making it suitable for object-level relationship representation in videos. Furthermore, we propose vision-guided language prompting to leverage CLIP's comprehensive semantic knowledge for discovering unseen relationship categories, thus facilitating recognizing novel video relationships. Extensive experiments on two public datasets, VidVRD and VidOR, demonstrate the effectiveness of our method, especially achieving a significant gain of nearly 10% in mAP on novel relationship categories on the VidVRD dataset.

Introduction

The task of Open-vocabulary Video Visual Relationship Detection (Open-VidVRD) (Gao et al. 2023) aims to detect video relationships between two objects as triplet format $\langle subject, predicate, object \rangle$ following an open-vocabulary setting, where the model is learned on base relationship categories during training and is applied to infer both base and novel relationship categories during testing. As shown in Figure 1, the novel categories of $\langle person, wear, hat \rangle$ and $\langle person, pull, dog \rangle$, which are absent during training, can be recognized during testing. Different from the closed-set setting that does not involve novel categories, the Open-VidVRD task is more essential and practical for real-world scenarios characterized by complex and diverse object relationships.

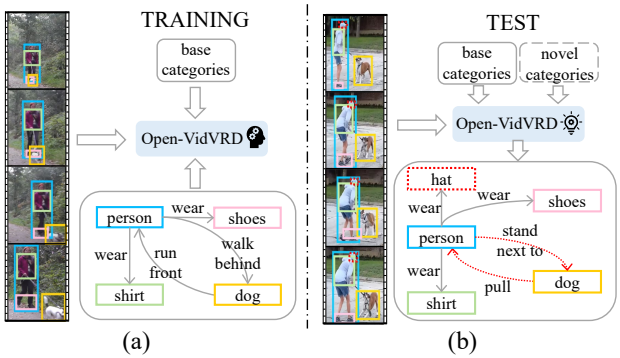


Figure 1: Illustration of the Open-VidVRD task. Training is performed on base categories. Testing is performed on both base categories and novel categories that are absent in training.

The recent emergence and advancement of large-scale pre-trained image-text models (Jia et al. 2021; Pham et al. 2021; Radford et al. 2021) present a promising avenue for Open-VidVRD, leveraging their learned vision-language joint embedding space, which contains rich semantic knowledge about objects, scenes, actions, and interactions (Gao et al. 2022b; Gu et al. 2022; Kuo et al. 2023; Ni et al. 2022; Weng et al. 2023; Xu et al. 2023). However, directly applying these models to open-vocabulary video visual relationship detection entails confronting two challenges. The first challenge pertains to handling the domain gap between images and video object relationships. Most image-text pre-trained models are learned on images and their corresponding textual descriptions, which typically encompass the entire image content rather than specific local objects. Consequently, such models lack the ability to capture both temporal context information and object-level details, limiting their applicability in video object relationship analysis. The second challenge revolves around leveraging comprehensive semantic knowledge learned in these pre-trained models to discover novel video visual relationships.

In this paper, we propose a multi-modal prompting

*These authors contributed equally.
†Corresponding author.

method for Open-VidVRD[1], which prompts the pre-trained image-text model, *i.e.*, CLIP (Radford et al. 2021), on both visual and language sides, to improve the alignment between visual and language representations of relationships. To address the first challenge, we propose spatio-temporal visual prompting to improve CLIP's ability to capture spatial and temporal relationships between objects. Starting with using CLIP to identify and detect novel objects, we then introduce sequential Transformer blocks (Vaswani et al. 2017) to model the spatial and temporal context among these objects. Therefore, we adapt CLIP well to the spatio-temporal object-level relationship domain, ultimately enhancing its applicability to video visual relationship detection. To address the second challenge, we propose vision-guided language prompting to exploit CLIP's comprehensive semantic knowledge for discovering novel relationships. We introduce two kinds of prompts, learnable continuous prompts and learnable conditional prompts, where the former imparts task-specific prior knowledge and the latter dynamically adapts to visual cues. Through the integration of these two prompts, we not only obtain shared task-specific priors, but also retain the ability to effectively incorporate novel categories.

By integrating the spatio-temporal visual prompting and the vision-guided language prompting, our method successfully bridges the domain gap and exploits the semantic knowledge embedded in CLIP for video visual relationship detection in the open-vocabulary scenario. Extensive experiments show that our method achieves significant performance improvements over existing methods, especially achieving nearly 10% mAP gains on the novel relationship categories on the VidVRD dataset.

To summarize, our main contributions are three-fold: (1) We propose a multi-modal prompting method for detecting video visual relationships in an open-vocabulary setting, which prompts CLIP on both the visual and language sides. (2) We propose spatio-temporal visual prompting to imbue CLIP with the capabilities of spatial and temporal modeling, effectively bridging the gap between images and video visual relationships. (3) We propose vision-guided language prompting, which exploits semantic knowledge from CLIP to discover novel visual relationships in videos.

Related Work

Open-vocabulary Visual Relationship Detection. The task of visual relationship detection in images (Lu et al. 2016) or videos (Shang et al. 2017), involving the classification and localization of relationship triplets, has become a hot topic in the field of computer vision (Tang et al. 2020; Li et al. 2022b; Cong, Yang, and Rosenhahn 2023; Zheng, Chen, and Jin 2022; Xu et al. 2022; Chen, Xiao, and Chen 2023). This field has also explored the concept of zero-shot detection (Shang et al. 2021), where all object and relationship categories are seen during training, but some certain triplet combinations remain unseen during test.

[1]Codes are at https://github.com/wangyongqi558/MMP_OV_VidVRD

In recent years, open-vocabulary visual relationship detection (He et al. 2022; Gao et al. 2023) has emerged, aiming to recognize visual relationships involving objects or predicates that are completely unseen in the training data. SVRP (He et al. 2022) adopts a two-step method for open-vocabulary visual relationship detection, including visual-relationship pre-training and prompt-based fine-tuning. However, this method caters to relationships within static images, but not within videos. In contrast, Repro (Gao et al. 2023) pioneers the open-vocabulary video visual relationship detection by leveraging the video-language pre-training model ALPro (Li et al. 2022a), sidestepping the necessity of training from scratch. However, Repro predominantly aligns the embedding space of video relationships from a linguistic perspective, ignoring the inherent spatial and temporal contextual dependencies in object relationships. The multi-modal prompting method proposed in this paper adapts the more broadly applicable image-text pre-trained CLIP on both visual and language sides to capture the spatio-temporal contexts of relationships and exploit prior semantic knowledge of novel relationships.

Prompting CLIP. Recently, visual-language pre-trained models (Radford et al. 2021; Alayrac et al. 2022; Li et al. 2020; Luo et al. 2020) have demonstrated significant progress in many downstream vision-language tasks. As one of the most successful visual-language pre-training models, CLIP (Radford et al. 2021), is extensively pre-trained using 400 million image-text pairs from the Internet, resulting in a visual-language embedding space with comprehensive semantic knowledge.

Various techniques for prompt learning (Gu et al. 2022; Ding, Wang, and Tu 2022; Xu et al. 2023; Kuo et al. 2023; Gao et al. 2022b; Wang, Xing, and Liu 2021) have emerged to facilitate the efficient transfer of knowledge from CLIP to the downstream tasks. These tasks range from few/zero-shot classification (Pham et al. 2021; Zhou et al. 2022), open-vocabulary recognition and detection (Gu et al. 2022; Du et al. 2022; Ding, Wang, and Tu 2022; Ma et al. 2022; Wang et al. 2022; Kuo et al. 2023), to video-related applications (Xu et al. 2021; Ju et al. 2022; Wang, Xing, and Liu 2021; Ni et al. 2022; Weng et al. 2023). However, these approaches prompt CLIP in a single modality, either visual or language, resulting in sub-optimal performance.

A method closely related to our method is MaPLe (Khattak et al. 2023), which prompts CLIP in both vision and language to enhance the alignment between visual and language representations. It is worth noting that MaPLe primarily focuses on image recognition tasks. In contrast, our method is tailored for video visual relationship detection, which is more challenging than image domain tasks.

Our Method

Overview

Video Visual Relationship Detection (VidVRD) aims to detect instances of visual relationships of interest in a video, where a visual relationship instance is represented by a triplet $\langle subject, predicate, object \rangle$ with the trajectories of subject and object. For the Open-vocabulary Video Visual

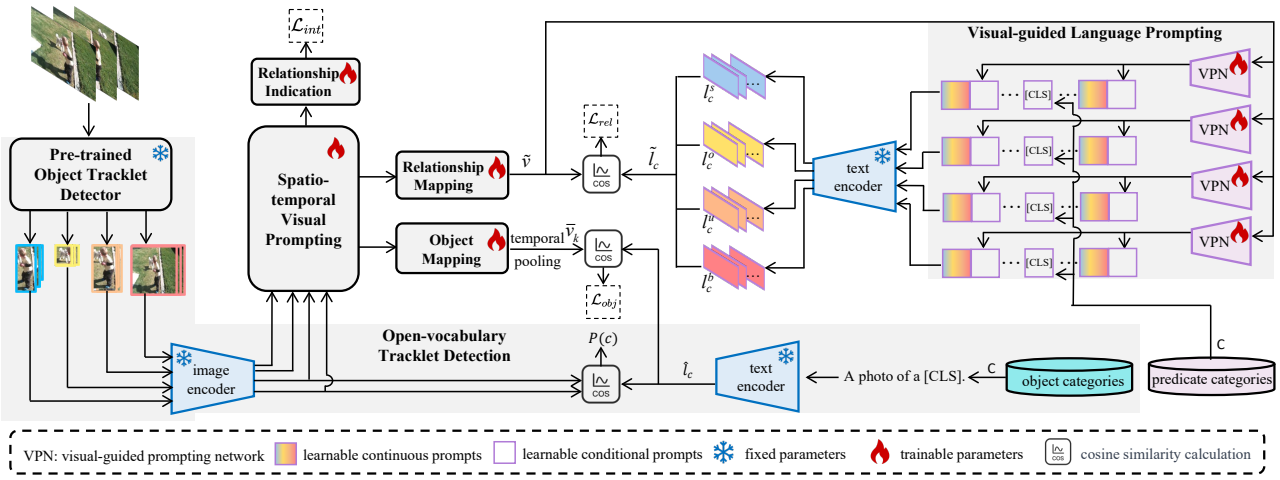


Figure 2: Overview of proposed method.

Relationship Detection (Open-VidVRD), the categories of objects and predicates are divided into base and novel splits, *i.e.,* base objects ($\mathcal{C}_b^O$), novel objects ($\mathcal{C}_n^O$), base predicate ($\mathcal{C}_b^P$), and novel predicate ($\mathcal{C}_n^P$). Only base object and predicate categories are used in the training stage, and all categories are used in the testing stage.

To address the Open-VidVRD task, we propose a multi-modal prompting method to prompt the well-known CLIP model on both visual representation and language input. Our method consists of three main components: an open-vocabulary object tracklet detection module for both base and novel objects, a spatio-temporal prompting module to enable the spatio-temporal modeling of object relationships, and a language prompting module with both learnable continuous prompts and learnable conditional prompts. The overview of our method is shown in Figure 2.

Open-vocabulary Object Tracklet Detection

We use a pre-trained tracklet detector (Gao et al. 2022a) to obtain N class-agnostic visual object tracklets $\boldsymbol{T} = \{\boldsymbol{T}_i\}_{i=1}^N$, $\boldsymbol{T}_i = \{\boldsymbol{B}_t\}_{t=1}^T$, where $\boldsymbol{T}_i$ are visual object trajectories with T frames and $\boldsymbol{B}_t$ is the bounding box of object in frame t. We classify each tracklet using CLIP. Specifically, we extract the visual representations of cropped object regions corresponding to detected bounding boxes using CLIP's image encoder and average the features along temporal axis of the tracklet. Meanwhile, we extract text embeddings for all object categories by feeding handcrafted prompts, *i.e.*, "a photo of [CLS]" into the text encoder of CLIP, where [CLS] can be replaced with the names of objects. Then, we assign each tracklet an object category label c by its maximum scores within all objects:

$$p(c) = \frac{\exp(\cos(\boldsymbol{v}_i, \hat{\boldsymbol{l}}_c)/\tau)}{\sum_{c' \in \mathcal{C}_b^o \cup \mathcal{C}_n^o} \exp(\cos(\boldsymbol{v}_i, \hat{\boldsymbol{l}}_c)/\tau)}, \tag{1}$$

where $c \in \mathcal{C}_b^o \cup \mathcal{C}_n^o$, τ is a temperature parameter, and $\cos(\boldsymbol{v}_i, \hat{\boldsymbol{l}}_c)$ is the cosine similarity between the visual features $\boldsymbol{v}_i$ of the i-th object trajectory and the text embedding features $\hat{\boldsymbol{l}}_c$ of the object category c.

Spatio-temporal Visual Prompting

Pre-trained image-text models such as CLIP are typically trained using images and their associated textual descriptions. These descriptions usually encapsulate the entire content of the image, rather than focusing on individual objects. Therefore, these models lack detailed object-level information and temporal context. To address the disparity in relationships between the image and video domains, we introduce the spatio-temporal visual prompting, which is integrated into the image encoding process of CLIP, to bridge the domain gap and enhance the CLIP's ability to capture both spatial and temporal visual relationships in videos.

For each pair of tracklet, *i.e.*, a subject tracklet and an object tracklet, we first set the regions outside the bounding boxes of the tracklets to 0, resulting in four masked frames: frames corresponding to the subject, object, their union, and background (whole image). And then, we extract their features using CLIP and capture their spatio-temporal relationships by standard Transformer Blocks. To reduce the computational complexity, we decouple the spatio-temporal modeling into separate and successive modules, namely spatial modeling and temporal modeling, as illustrated in Figure 3.

Spatial Modeling. Spatial relationships between objects are typically defined by their positional orientations, such as being in front of or above each other. Therefore, spatial modeling requires combining four key elements: features of subject region, features of object region, features encompassing their union region (*i.e.*, the smallest area covering subject and object), and features representing the backgrounds (*i.e.*, the whole image). This process involves modeling interactions between objects and their backgrounds to enhance object features, thus capturing the spatial context.

Specifically, given the masked frames of subject, object, their union and background, we first extract their features by the image encoder of CLIP, denoted by $\boldsymbol{v}^k, k \in \{s, o, u, b\}$.

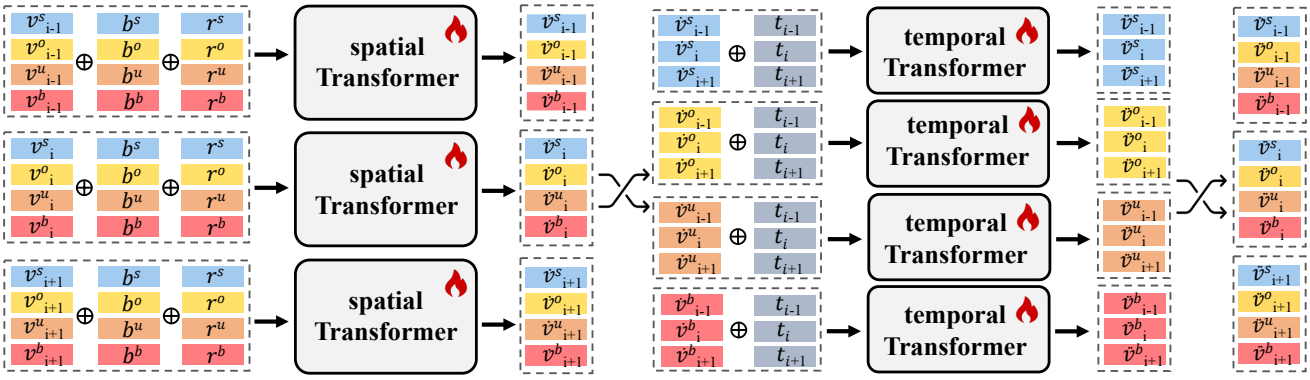


Figure 3: Overview of spatial modeling and temporal modeling. $\boldsymbol{b}^k, \boldsymbol{r}^k, k \in \{s, o, u, b\}$ denotes the spatial position embedding and role embedding corresponding to subject, object, their union, and background, respectively. $\boldsymbol{t}_t$ is the temporal embedding of t-th frame.

Note that the same spatial modeling is used across different frames, thus we omit the frame index here for simplify. And then we add two types of learnable embedding: positional embedding $\boldsymbol{b}^k, k \in \{s, o, u, b\}$, which are learned related to the normalized bounding box, and role embedding $\boldsymbol{r}^k, k \in \{s, o, u, b\}$. These two types of embeddings are learned and shared over all video frames, respectively. Finally, we update the visual features by

$$(\dot{\boldsymbol{v}}^s, \dot{\boldsymbol{v}}^o, \dot{\boldsymbol{v}}^u, \dot{\boldsymbol{v}}^b) = \mathrm{STrans}(\boldsymbol{I}^s, \boldsymbol{I}^o, \boldsymbol{I}^u, \boldsymbol{I}^b), \tag{2}$$

where $\boldsymbol{I}^k = \boldsymbol{v}^k + \boldsymbol{b}^k + \boldsymbol{r}^k, k \in \{s, o, u, b\}$ and $\mathrm{STrans}(\cdot)$ denotes the spatial Transformer blocks.

Temporal Modeling. Temporal relationships of objects are time-dependent, such as toward or away, so the inputs of temporal modeling contain two items: visual features and temporal embeddings. Note that the same temporal modeling is used for different roles, *i.e.*, subject, object, their union, and background. Through the exploration of dynamic state transformations, visual features are systematically updated.

Specifically, given the spatial encoded visual features $\dot{\boldsymbol{v}} = \{\dot{\boldsymbol{v}}_t^s, \dot{\boldsymbol{v}}_t^o, \dot{\boldsymbol{v}}_t^u, \dot{\boldsymbol{v}}_t^b\}_{t=0}^T$, we collect each role features separately from all frames as $\dot{\boldsymbol{v}}^k = \{\dot{\boldsymbol{v}}_t^k\}_{t=0}^T, k \in \{s, o, u, b\}$, and add temporal embedding $\boldsymbol{t}_t$, which are learned related to frame t and shared over all video frames. For each region, the visual features are then updated by

$$\ddot{\boldsymbol{v}}^k = \{\ddot{\boldsymbol{v}}_t^k\}_{t=0}^T = \mathrm{TTrans}(\dot{\boldsymbol{I}}_0^k, \dot{\boldsymbol{I}}_1^k, \cdots, \dot{\boldsymbol{I}}_T^k), \tag{3}$$

where $\dot{\boldsymbol{I}}_t^k = \dot{\boldsymbol{v}}_t^k + \boldsymbol{t}_t$, and $\mathrm{TTrans}(\cdot)$ denotes the temporal Transformer blocks. The updated features are then reorganized by their frame indexes for the next layer.

Vision-guided Language Prompting

The main goal of the Open-VidVRD task is to discover novel video visual relationships. To achieve this goal, considering both task-related prior knowledge and visual-related prior knowledge, we propose vision-guided language prompting to leverage the rich semantic knowledge stored in CLIP by combining learnable continuous language prompts and learnable conditional language prompts.

Learnable Continuous Language Prompts. For each predicate category [CLS], [CLS] $\in \mathcal{C}_b^P$ when training and [CLS] $\in \mathcal{C}_n^P \cup \mathcal{C}_b^P$ when testing, N_ς-token language prompts corresponding to each of the four roles (*i.e.*, subject, object, union, and background) are initialized as $\varsigma^k = [\boldsymbol{\varsigma}_1^k, \boldsymbol{\varsigma}_2^k, \cdots, \boldsymbol{\varsigma}_{N_\varsigma}^k], k \in \{s, o, u, b\}$ and are learned by gradient backpropagation.

Learnable Conditional Language Prompts. For each predicate category [CLS], [CLS] $\in \mathcal{C}_b^P$ when training and [CLS] $\in \mathcal{C}_n^P \cup \mathcal{C}_b^P$ when testing, N_ζ-token learnable conditional language prompts corresponding to each of the four roles (subject, object, union, and background) are learned by taking into account their visual features:

$$\boldsymbol{\zeta}^k = [\boldsymbol{\zeta}_1^k, \boldsymbol{\zeta}_2^k, \cdots, \boldsymbol{\zeta}_{N_\zeta}^k] = \varphi_k(\ddot{\boldsymbol{v}}^{\boldsymbol{k}}), \tag{4}$$

where $\varphi_k(\cdot)$ denotes the vision-guided prompting network, consisting of two linear layers.

Learnable Vision-guided Language Prompts. We concatenate the token of learnable continuous language prompts and token of learnable conditional language prompts interlaced, and then insert the [CLS] token of predicate categories into the later middle (75% of the length) of token sequence, resulting in the final language prompts $\boldsymbol{\ell}_{\mathrm{CLS}}^k = \left[\boldsymbol{\varsigma}_1^k, \boldsymbol{\zeta}_1^k, \boldsymbol{\varsigma}_2^k, \boldsymbol{\zeta}_2^k, \cdots, \mathrm{CLS}, \cdots, \boldsymbol{\varsigma}_{N_\varsigma}^k, \boldsymbol{\zeta}_{N_\zeta}^k\right], k \in \{s, o, u, b\}$.

The final language features of category c corresponding to each visual region are generated by

$$\boldsymbol{l}_c^k = \pi(\boldsymbol{\ell}_c^k), \tag{5}$$

where $\pi(\cdot)$ is the text encoder of CLIP.

Training Loss

The training loss of our method consists of three parts: a relationship contrastive loss $\mathcal{L}_{rel}$, an object contrastive loss $\mathcal{L}_{obj}$, and an interaction loss $\mathcal{L}_{int}$, as shown in Figure 2. The overall objective loss function is given by

$$\mathcal{L} = \mathcal{L}_{rel} + \alpha \mathcal{L}_{obj} + \beta \mathcal{L}_{int}. \tag{6}$$

Split	Method	DET Trajectory			GT Trajectory		
		mAP	R@50	R@100	mAP	R@50	R@100
Novel	ALPro (Li et al. 2022a)	1.05	3.14	4.62	4.09	9.42	10.41
	CLIP (Radford et al. 2021)	2.88	3.80	4.96	4.54	7.27	11.74
	VidVRD-II (Shang et al. 2021)	3.57	8.59	12.39	7.35	18.84	26.44
	RePro (Gao et al. 2023)	6.10	13.38	16.52	12.74	25.12	33.88
	Ours	**15.99**	**16.69**	**18.68**	**21.14**	**30.41**	**37.85**
All	ALPro (Li et al. 2022a)	3.20	2.62	3.18	4.97	4.50	5.79
	CLIP (Radford et al. 2021)	5.03	3.04	3.68	6.49	5.21	6.54
	VidVRD-II (Shang et al. 2021)	12.74	9.90	12.59	19.73	18.17	24.90
	RePro (Gao et al. 2023)	21.33	12.92	15.94	34.90	25.50	32.49
	Ours	**27.06**	**16.71**	**19.61**	**38.08**	**30.47**	**37.46**

Table 1: Results of different methods on the VidVRD dataset."DET" and "GT" denote using detected trajectory and ground-truth trajectory, respectively.

where α and β represent balance factors.

Relationship Contrastive Loss. Given the visual representations $\ddot{\boldsymbol{v}}^k$ and the language representations $\boldsymbol{l}_c^k$, the prediction score of the relationship category c is calculated by

$$\hat{y}_c^{rel} = \sigma(\cos(\psi(\tilde{\boldsymbol{v}}), \tilde{\boldsymbol{l}}_{\boldsymbol{c}})) \tag{7}$$

where $\tilde{\boldsymbol{v}} = ||\ddot{\boldsymbol{v}}^s, \ddot{\boldsymbol{v}}^o, \ddot{\boldsymbol{v}}^u, \ddot{\boldsymbol{v}}^b||$; $\tilde{\boldsymbol{l}}_{\boldsymbol{c}} = ||\boldsymbol{l}_c^s, \boldsymbol{l}_c^o, \boldsymbol{l}_c^u, \boldsymbol{l}_c^b||$; $||\cdot||$ denotes concatenation; $\sigma(\cdot)$ is the sigmoid function; $\cos(\cdot,\cdot)$ is the cosine similarity and $\psi(\cdot)$ denotes the relationship mapping layer in Figure 2. Then the relationship contrastive loss is formulated by using the binary cross-entropy loss (BCE):

$$\mathcal{L}_{rel} = 1/|\mathcal{C}_b^P| \cdot \sum\nolimits_{c \in \mathcal{C}_b^P} \mathrm{BCE}(\hat{y}_c^{rel}, y_c^{rel}), \tag{8}$$

where $y_c^{rel} = 1$ when the class c is the ground-truth predicate category, otherwise $y_c^{rel} = 0$.

Object Contrastive Loss. To avoid the visual features drift caused by the proposed sptio-temporal prompting, we introduce an object contrastive loss to enforce the spatial encoded object features to have the ability to distinguish each other as the original CLIP. Specifically, after the spatial modeling, we collect subject and object features from all frames and average them as $\bar{\boldsymbol{v}}^k = \mathrm{avg}(\{\phi(\dot{\boldsymbol{v}}_t^k)\}_{t=0}^T)$, $k \in \{s, o\}$ and $\phi(\cdot)$ denotes the object mapping layer in Figure 2. Meanwhile, we extract the text embeddings for all subject or object categories by feeding the handcrafted prompts (i.e.,"a photo of [CLS]") into the text encoder of CLIP, where [CLS] can be replaced with the names of subjects or objects. The similarity between the visual feature and the language features of category c is calculated by $\hat{y}_c^k = \cos(\bar{\boldsymbol{v}}^k, \hat{\boldsymbol{l}}_c), k \in \{s, o\}$. Finally, the object contrastive loss is computed over all object categories using the cross-entropy loss (CE):

$$\mathcal{L}_{obj} = \mathrm{CE}(\hat{\boldsymbol{y}}^s, y^s) + \mathrm{CE}(\hat{\boldsymbol{y}}^o, y^o), \tag{9}$$

where $\hat{\boldsymbol{y}}^s$ is the predicted subject similarity between visual features and language features of base object categories ($\mathcal{C}_b^O$), and $\hat{\boldsymbol{y}}^o$ is the corresponding predicted object similarity. y^s and y^o denote the ground-truth category labels of the subject and object, respectively.

Interaction 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 in video frame t, we set the ground-truth interaction by $y_t^{int} = 1$, otherwise $y_t^{int} = 0$. To learn this weak interaction, we concatenate all the features in frame t and predict the interaction probability by $\hat{y}_t^{int} = \psi(||\ddot{\boldsymbol{v}}_t^s, \ddot{\boldsymbol{v}}_t^o, \ddot{\boldsymbol{v}}_t^u, \ddot{\boldsymbol{v}}_t^b||)$, where $\psi(\cdot)$ denotes the relationship indication layer in Figure 2. The interaction loss is then computed using the binary cross-entropy loss (BCE):

$$\mathcal{L}_{int} = 1/T \cdot \sum\nolimits_{t=1}^{T} \mathrm{BCE}(\hat{y}_t^{int}, y_t^{int}). \tag{10}$$

Experiment

Datasets and Evaluation Metrics

Datasets. We evaluate our method on the **VidVRD** (Shang et al. 2017) and **VidOR** (Shang et al. 2019) datasets. The VidVRD dataset contains 1000 videos, 800 videos for training and 200 for testing, covering 35 object categories and 132 predicate categories. The VidOR dataset contains 10000 videos, 7000 videos for training, 835 videos for validation, and 2165 videos for testing, covering 80 object categories and 50 predicate categories.

Evaluation Settings. For the open-vocabulary evaluation, the base and novel categories are selected based on frequency. Following Repro (Gao et al. 2023), we choose the common object and predicate categories as base categories and the rare ones as novel categories. Training is performed on base categories. Test is performed under two settings: (1) **Novel**-split evaluation involves all object categories and novel predicate categories; (2) **All**-split evaluation involves all object categories and all predicate categories, which is a standard evaluation. Note that the test is performed on both the test set of VidVRD and the validation set of VidOR (the annotations of test set of VidOR are not available). To remove the impact of inaccurately detected trajectories, we also evaluate the methods using ground-truth trajectories, focusing on relationship detection with accurately detected objects.

Split	Method	mAP	R@50	R@100
Novel	ALPro	-	8.35	9.79
	CLIP	1.08	5.48	7.20
	VidVRD-II	-	4.32	4.89
	RePro	-	7.20	8.35
	Ours	**3.58**	**9.22**	**11.53**
All	ALPro	-	2.61	3.66
	CLIP	1.29	1.71	3.13
	VidVRD-II	-	24.81	34.11
	RePro	-	27.11	35.76
	Ours	**38.52**	**33.44**	**43.80**

Table 2: Results of different methods using ground-truth trajectory on the VidOR dataset.

Vis	Lan	Novel		All		AVG
		DET	GT	DET	GT	
		5.03	5.23	20.61	27.14	14.50
✓		11.04	16.83	26.35	37.10	22.83
	✓	12.33	17.98	26.94	36.79	23.51
✓	✓	**15.99**	**21.14**	**27.06**	**38.08**	**25.57**

Table 3: Performance (mAP) of ablation study for multi-modal prompting on the VidVRD dataset. "Vis" and "Lan" denote visual prompting and language prompting, respectively.

Metrics. We use mean Average Precision (mAP) and Recall@K (R@K) with K=50,100 for evaluation. The detected triplet is considered correct if it matches a triplet in the ground-truth and the IoU between them is greater than a threshold (*i.e.,* 0.5).

Implementation Details

For all experiments, video frames are sampled every 30 frames. We adopt the ViT-B/16 version of CLIP while keeping the parameters fixed. The number of Transformer blocks of spatio-temporal visual prompting is set to 1 and 2 for the VidVRD dataset and the VidOR dataset, respectively. The head number of multi-head self-attention of Transformer blocks is set to 8, and the dropout rate is set to 0.1. For language prompting, we set the number of tokens for both learnable continuous prompts and conditional prompts to 8. The [CLS] token is positioned at 75% of the token length. For optimization, we use the AdamW (Loshchilov and Hutter 2019) algorithm with an initial learning rate of 0.001. A multi-step decay schedule is applied at epochs 15, 20, and 25, reducing the learning rate by a factor of 0.1 each time. The batch size is set to 32.

Comparison with Existing Methods

We conduct a comprehensive comparison of our method with the state-of-the-art methods, including RePro (Gao et al. 2023), VidVRD-II (Shang et al. 2021), and the pre-trained models ALPro (Li et al. 2022a) and CLIP.

The comparison results on the VidVRD dataset are shown in Table 1. It is interesting to observe that our method achieves the best performance in terms of all evaluation metrics, and specifically achieves nearly 10% improvement in mAP using both detected trajectories (denoted as "DET") and ground-truth trajectories (denoted as "GT") in the "Novel" split. This clearly validates the superiority of the proposed multi-modal prompting method in Open-VidVRD.

The comparison results on the VidOR dataset are shown in Table 2. Since the existing methods only provide the results using ground-truth trajectory on the VidOR dataset, we only report the results using ground-truth trajectory for comparison. Our method outperforms the other methods by 0.87% and 1.74% on R@50 and R@100, respectively, in the "Novel" split. Moreover, our method achieves gains of 6.8% and 8.04% on R@50 and R@100, respectively, in the "All" split.

Spa	Tem	Novel		All		AVG
		DET	GT	DET	GT	
		12.33	17.98	26.94	36.79	23.51
✓		13.92	17.54	26.79	36.73	23.75
	✓	9.10	12.41	23.27	31.39	19.07
✓	✓	**15.99**	**21.14**	**27.06**	**38.08**	**25.57**

Table 4: Performance (mAP) of ablation study for the spatio-temporal visual prompting on the VidVRD dataset. Note that we add linear layers to keep similar amount parameters when a module is absent. "Spa" and "Tem" denote spatial modeling and temporal modeling, respectively.

Variants	Novel		All		AVG
	DET	GT	DET	GT	
Manual	11.04	16.83	26.35	37.10	22.83
Continuous	14.21	18.21	**27.27**	36.94	24.16
Conditional	15.29	**21.88**	25.85	36.44	24.87
Ours	**15.99**	21.14	27.06	**38.08**	**25.57**

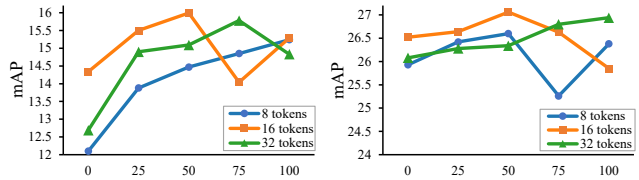
Table 5: Performance (mAP) of ablation study for the language prompting on the VidVRD dataset.

Abaltion Studies

We perform in-depth ablation studies on the VidVRD dataset to evaluate each component of our method.

Effectiveness of multi-modal prompting. To evaluate the multi-modal prompting, we replace the visual prompting (denoted as "Vis") with linear layers and replace the language prompting (denoted as "Lan") with handcraft language prompting. The results are shown in Table 3. The consistent improvements in the "Novel" split and the "All" split demonstrate the effectiveness of the proposed visual promoting and language prompting.

Effectiveness of spatio-temporal visual prompting. To evaluate the components of the spatio-temporal visual prompting, we replace the spatial modeling module (denoted as "Spa") or the temporal modeling module (denoted as "Tem") with linear layers. From the results shown in Table 4, the average gain in mAP exceeds 2% when perform-



(a) Novel (b) All

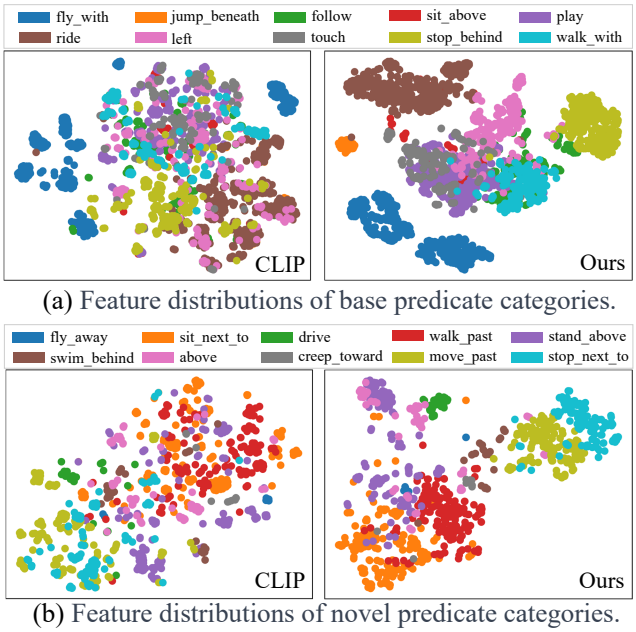
Figure 4: Effectiveness of tokens of vision-guided languge prompts on the VidVRD dataset. The x-axis represents the percentage of tokens from conditional prompts, from 0 (all tokens are from learnable continuous prompts) to 100% (all tokens are from learnable conditional prompts), and different colors denote different lengths.

ing both spatial and temporal modeling. Furthermore, we observe that the performance drops significantly when only temporal modeling is performed without incorporating spatial modeling. This is in line with expectations, as it is difficult to recognize object relationships based only on the dynamic state changes of individual objects.

Effectiveness of vision-guided language prompting. To evaluate the effectiveness of the vision-guided language prompting, we design three variants of our method for comparison: (1) "Manual" involves pre-defined templates for subjects (*i.e.*, "An image of a person or object [CLS] something"), objects (*i.e.*, "An image of something [CLS] a person or object"), unions and backgrounds (*i.e.*, "An image of the visual relationship [CLS] between two entities"); (2) "Continuous" involves learnable continuous prompts; (3) "Conditional" tailors all prompts to input visual features.

From the results shown in Table 5, it is obvious that the proposed viual-guided language prompting ("Ours") outperforms the other variants on average. Specifically, an impressive improvement of nearly 5% is achieved in the "Novel" split using the detected trajectories. We also observe that the learnable prompts, including the learnable continuous prompts and the learnable conditional prompts, generally achieve better results than the manually designed prompts.

Effectiveness of tokens of languge prompting. We analyze the effectiveness of tokens of vision-guided language prompts by alternately placing tokens of learnable continuous prompts and conditional prompts and inserting the [CLS] token at the 75% position of the length, as shown in Figure 4. We observe that the performance initially ascends and then declines with the increasing token number. And along with the increasing percentages of tokens from learnable conditional prompts, the results first increase and then become unstable. The best performance surfaces when the token number is 16, and half of the tokens are from learnable conditional prompts. These observations highlight the importance of combining task-specific knowledge and visual cues, further validating the effectiveness of our vision-guided prompting on combining learnable continuous prompts and learnable conditional prompts.



(a) Feature distributions of base predicate categories.

(b) Feature distributions of novel predicate categories.

Figure 5: Qualitative results of visual feature (union region of subject and object) distributions by T-SNE.

Qualitative Analysis

We visualize the feature distributions of randomly selected 10 predicate categories by projecting the features of the union regions onto a 2D plane using T-SNE (Hinton and Roweis 2002), to demonstrate how well our spatio-temporal visual prompting method adapts the image encoder of CLIP. As shown in Figure 5, features of our method (the right parts of Figure 5 (a) , (b)) within the same categories are pulled closer while features across different categories are pushed further apart, improving the discrimination on both base and novel categories. These qualitative results further verify the effectiveness of our spatio-temporal visual prompting method.

Conclusion

We have presented a multi-modal prompting method for open-vocabulary video visual relationship detection. By introducing spatio-temporal visual prompting and vision-guided language prompting to leverage the large-scale pre-trained image-text model, our method demonstrates remarkable potential in bridging the domain gap between image and video relationships and discovering novel objects and relationships. Extensive experiments conducted on public datasets show the superiority of our method, substantiated by a notable gain of performance on novel relationship categories while keeping the performance of base categories. A limitation of our method is the dependence on a pre-trained trajectory detector, so investigating an end-to-end pipeline to alleviate this dependency is an interesting avenue for future research.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006 and the Natural Science Foundation of China (NSFC) under Grant No 62072041.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 23716–23736.

Chen, S.; Xiao, J.; and Chen, L. 2023. Video scene graph generation from single-frame weak supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Cong, Y.; Yang, M. Y.; and Rosenhahn, B. 2023. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Ding, Z.; Wang, J.; and Tu, Z. 2022. Open-Vocabulary panoptic segmentation with MaskCLIP. *arXiv*, abs/2208.08984.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14084–14093.

Gao, K.; Chen, L.; Niu, Y.; Shao, J.; and 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)*, 19497–19506.

Gao, K.; Chen, L.; Zhang, H.; Xiao, J.; and Sun, Q. 2023. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

He, T.; Gao, L.; Song, J.; and Li, Y.-F. 2022. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 56–73. Springer.

Hinton, G. E.; and Roweis, S. 2002. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 15, 833–840.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and 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)*, 4904–4916. PMLR.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 105–124. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19113–19122.

Kuo, W.; Cui, Y.; Gu, X.; Piergiovanni, A.; and Angelova, A. 2023. Open-vocabulary object detection upon frozen vision and language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, D.; Li, J.; Li, H.; Niebles, J. C.; and Hoi, S. C. 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)*, 4953–4963.

Li, L.; Chen, L.; Huang, Y.; Zhang, Z.; Zhang, S.; and Xiao, J. 2022b. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18869–18878.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2046–2065.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 852–869.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv*, abs/2002.06353.

Ma, C.; Yang, Y.; Wang, Y.; Zhang, Y.; and Xie, W. 2022. Open-vocabulary semantic segmentation with frozen vision-Language models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 45.

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

Pham, H.; Dai, Z.; Ghiasi, G.; Liu, H.; Yu, A. W.; Luong, M.-T.; Tan, M.; and Le, Q. V. 2021. Combined scaling for zero-shot transfer learning. *arXiv*, abs/2111.10050.

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)*, 8748–8763. PMLR.

Shang, X.; Li, Y.; Xiao, J.; Ji, W.; and Chua, T.-S. 2021. Video visual relation detection via iterative inference. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 3654–3663.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017. Video visual relation detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1300–1308.

Shang, X.; Xiao, J.; Di, D.; and Chua, T.-S. 2019. Relation understanding in videos: A grand challenge overview. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2652–2656.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3716–3725.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 5998–6008.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv*, abs/2109.08472.

Wang, S.; Duan, Y.; Ding, H.; Tan, Y.-P.; Yap, K.-H.; and Yuan, J. 2022. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 939–948.

Weng, Z.; Yang, X.; Li, A.; Wu, Z.; and Jiang, Y.-G. 2023. Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. *arXiv*, abs/2302.00624.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6787–6800.

Xu, L.; Qu, H.; Kuen, J.; Gu, J.; and Liu, J. 2022. Meta spatio-temporal debiasing for video scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 374–390. Springer.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. *arXiv*, abs/2302.12242.

Zheng, S.; Chen, S.; and 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)*, 18836–18846.

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