Improving Open-vocabulary Video Visual Relation Detection with Decomposed Prompt Learning and Relation Adjustment

Ming Pei[1*], Yi Tan[1*], Yanbin Hao[1†], Hao Zhang[2], Jinmeng Wu[3], Basura Fernando[2], Xun Yang[1]
[1] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[2] Centre for Frontier AI Research, Agency for Science, Technology and Research(A*STAR), Singapore
[3] Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, China
{pb14210,ty133}@mail.ustc.edu.cn, haoyanbin@hotmail.com, {zhang_hao, fernando_basura}@cfar.a-star.edu.sg,
Jinmeng2004910@outlook.com, xyang21@ustc.edu.cn

***Abstract*—Open-vocabulary video visual relation detection (VidVRD) expands the scope of detecting object relations in videos to include unseen categories. It marks considerable advancement in recognizing novel relations solely by training on a base set, thus extending the frontiers of automated video understanding. However, the performance of current methods on novel predicates remains significantly inferior to that on base categories. We attribute this discrepancy to two primary factors: (1) A significant task misalignment between the Visual Relation Detection (VRD) task and the pre-trained models' visual feature extractors, which are often designed for tasks like video-text retrieval and image-text retrieval, resulting in poor generalization to the novel set. (2) The relatively small size and limited vocabulary of open-vocabulary datasets, which create a substantial gap between base and novel predicates. Consequently, text prompts trained on the base set fail to generalize effectively to the novel set. To address these issues, we propose two improvement measures: (1) We decompose base and novel relations into actional and spatial patterns and introduce an innovative text prompt learning method that leverages the shared patterns between base and novel relations. (2) We develop a relation probability adjustment mechanism that utilizes reliable base relation predictions to adjust the probabilities of relations in novel classes by considering their overlaps in either actional or spatial contents. Experimental results on the benchmark dataset demonstrate significant performance improvements.**

***Index Terms*—Visual relation detection, Video understanding**

I. Introduction

Video visual relation detection (VidVRD) aims at recognizing visual relations between *subject* and *object* as a triplet form: *[subject* $\xrightarrow{predicate}$ *object]*, e.g., *[dog* $\xrightarrow{sit\,left}$ *person]*, *[person* $\xrightarrow{stand\,right}$ *dog]* in Fig. 1(A). It offers a high-level understanding and summarization of video knowledge for downstream tasks such as visual question answering [1]–[5], scene understanding [6]–[10], and video retrieval [11]–[15].

Currently, with the development of vision-language pre-trained models, e.g., CLIP [16] and Alpro [17], open-vocabulary VidVRD has garnered attention due to its ability to infer the unseen relations (novel relations) from the training set (base relations). For instance, [18], [19] achieve reasoning on novel video relations by calculating the feature similarity between visual object pairs and textual descriptions of candidate relations. Although the aforementioned methods customize CLIP to address the open-vocabulary VidVRD scenario, they empirically demonstrate poor generalization: as experiments reveal a significant performance gap between base relations and novel relations. We identify the challenges in two aspects: *first*, the VidVRD task differs significantly from CLIP's pre-trained image-text matching task. The former requires inferring relations between objects in complex video scenarios, while CLIP only considers the correspondence between static images and text. This difference prevents CLIP's strong generalization capability from fully manifesting, especially on the novel set of open-vocabulary VidVRD. *Second*, another

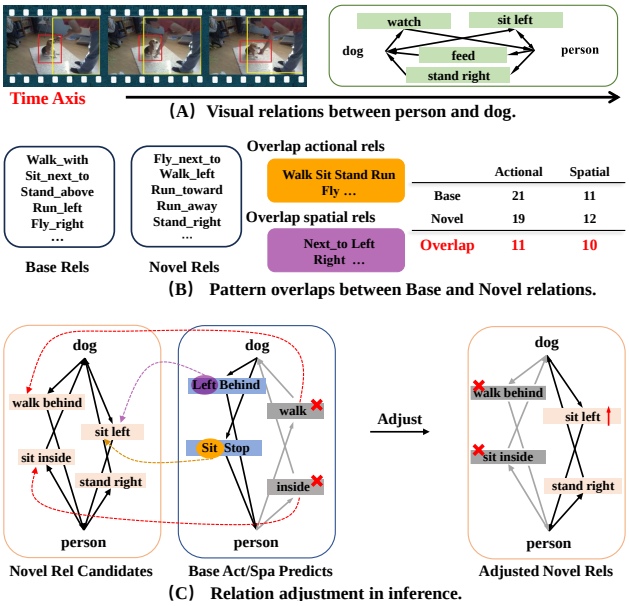

Fig. 1. (A) An example of VidVRD. (B) Demonstration and statistical analysis of the actional/spatial pattern overlap in the base and novel relations. (C) An illustration of the relation adjustment procedure. Probabilities of actional pattern *"sit"* and spatial pattern *"left"* are above the threshold resulting in further activation of novel relation *"sit_left"* while probabilities of actional pattern *"walk"* and spatial pattern *"inside"* are below threshold leading to a repression on novel relations *"walk_behind"* and *"sit_inside"*.

* Co-first author † Corresponding author

challenge in open-vocabulary VRD tasks is the difference between the novel and the base relations, which makes it difficult to directly transfer reliable knowledge from the base relation training to novel relation inference.

To address these two challenges, we propose two targeted solutions: *First*, to improve CLIP's generalization in VidVRD scenarios, we decouple the visual relations into actional and spatial relations. This approach is inspired by our observation of visual relations, which typically consist of both an actional pattern and a spatial pattern. After decoupling, we found that base and novel relations actually share a significant portion of these actional and spatial patterns, as shown in Fig. 1(B). We view this as a strategy to build a bridge between base and novel sets. Specifically, during training, we decouple visual relations into actional and spatial relations, enabling the model to generalize better to unseen novel relations by leveraging the underlying commonalities between novel and base relations in terms of actional and spatial patterns. *Second*, during testing, we aim to leverage the model's reliable prediction on the base set to enhance the inference of novel relations, as shown in Fig. 1(C). Technically, we first rank all candidate novel relations based on their relevance to the current pair of visual objects, generating an initial relation rank. We then adjust the rank by utilizing the reliable prediction on actional and spatial relations in the base set. Specifically, novel relations containing actional/spatial relations with a high likelihood of occurrence are further activated, while those containing actional/spatial relations with a low probability of occurrence are repressed.

Our main contributions are three-fold: (1) we propose a novel open-vocabulary VidVRD method that leverages shared patterns in base and novel relations by decomposed prompt learning; (2) we devise a relation adjustment mechanism to further transfer the knowledge on base visual relations to improve performance on novel relations; (3) our method achieves significant improvement in Open-vocabulary VidVRD.

II. Related work

Video Visual Relation Detection (VidVRD) was proposed in [20] together with ImageNet-VidVRD benchmark. This task spatiotemporally localizes pairwise visual relations. Existing methods mainly focus on modeling better visual or spatiotemporal contexts [21]–[24], and detecting visual relations with more granularity either by sliding windows [25] or temporal grounding [26]. They mainly worked on the pre-defined (closed) sets of object and visual relation categories. Reference [18], [19] concentrate on open-vocabulary VidVRD and propose prompt engineering methods to customize pre-trained VLMs. However, their methods neglect the characteristics of the VidVRD task and the correlation between base and novel relations in the dataset. In this work, based on the shared actional and spatial patterns between base and novel relations, we propose a decomposed prompt learning paradigm and relation adjustment mechanism for open-vocabulary VidVRD.

III. Approach

A. Preliminary

Video Visual Relation Detection (VidVRD) aims at detecting visual relation instances from the given untrimmed video. Each relation instance is represented by the triplet *[subject* $\xrightarrow{predicate}$ *object]* from a set of predefined predicate categories $\mathcal{C}^P$ and object categories $\mathcal{C}^O$. In open-vocabulary setting, the categories of predicate and object are divided into base and novel splits. We denote $\mathcal{C}_b^{O/P}$ and $\mathcal{C}_n^{O/P}$ as the set of base and novel object/predicate categories, respectively. Only base split is available to the model during training. Both base and novel splits are used for evaluation in the test stage.

As shown in Fig. 2, VidVRD begins with detecting N class-agnostic object tracklets in videos, i.e., $\mathcal{T} = \{T_i\}_{i=1}^N$, where each tracklet T_i is characterized with a temporal related sequence of bounding boxes. For any two potentially interacting tracklets (T_i and T_j), a new interaction tracklet T_{ij} is conducted, where the bounding box in each frame is the bounding rectangle that encloses the bounding boxes of both tracklets for interaction detection enhancement. Subsequently, CLIP encodes T_i, T_j, T_{ij}, as well as the corresponding frame into a group of visual features f_s, f_o, f_i, and f_b, where s, o, i, and b represent subject, object, interaction, and background, respectively. A light-weight transformer is then used to model the interactions within and cross each frame, the output features are denoted as [1]:

$$[\widetilde{f}_s, \widetilde{f}_o, \widetilde{f}_i, \widetilde{f}_b] = \mathrm{Trans}([f_s, f_o, f_i, f_b]). \tag{1}$$

Upon these features, a cross-entropy loss $\mathcal{L}_{int}$ equipped with a binary FC classifier is adopted to train the model to figure out whether two tracklets interact or not. Besides, L_{rel} and L_{obj} align these features to the relation descriptions features and object descriptions features (produced by the frozen CLIP text encoder), following the paradigm of image-text contrastive learning. The overall training objective is given by:

$$\mathcal{L} = \mathcal{L}_{int} + \mathcal{L}_{rel} + \mathcal{L}_{obj}. \tag{2}$$

B. Decomposed Prompt Learning

Relations such as *"sit above"* incorporate actional patterns (*"sit"*) and spatial patterns (*"above"*) as observed in [27]. We define functions ACT and SPA to map a visual relation to corresponding actional relation and spatial relation, (e.g., $\mathrm{ACT}(\text{“}sit\_above\text{”}) = \text{“}sit\text{”}$, $\mathrm{SPA}(\text{“}sit\_above\text{”}) = \text{“}above\text{”}$). For some visual relations like *"beneath"* which have no actional pattern, the mapping function ACT maps it to $\emptyset$. We define base actional relations set as $\mathcal{C}_b^A = \{\mathrm{ACT}(c^p) | c^p \in \mathcal{C}_b^P\}$. Similarly, the mapping function SPA maps visual relations like *"feed"* to $\emptyset$ and we define base spatial relation set as $\mathcal{C}_b^S = \{\mathrm{SPA}(c^p) | c^p \in \mathcal{C}_b^P\}$.

Through this design, we transform the relation recognition component ($\mathcal{L}_{rel}$) in the preliminary phase from recognizing

[1] VidVRD infers relationships on each frame and ultimately outputs video-level relations based on the relations across consecutive frames. Since the contributions of this paper do not involve operations in the temporal domain, and for simplicity of expression, we omit the frame indices in our description.

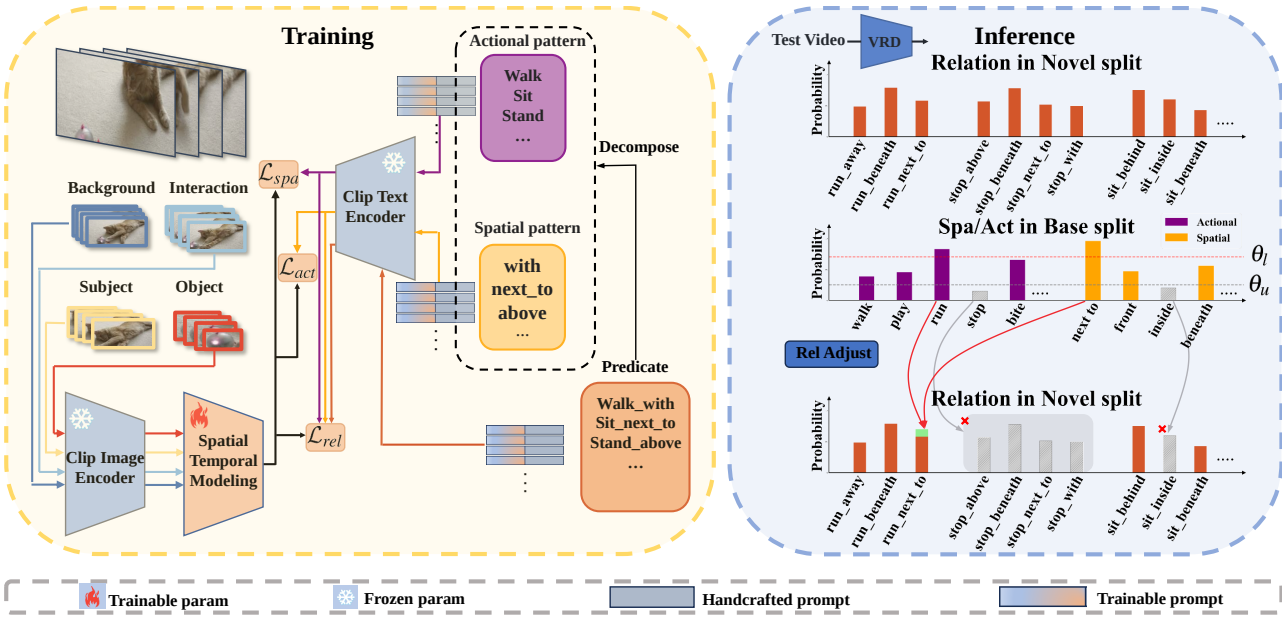

Fig. 2. The overall framework of our method. In the training phase (left part), candidate object pairs are encoded by CLIP visual encoder and a light-weight transformer. To leverage the shared patterns, we decouple base visual relations and align visual features to both actional/spatial/predicate embedding through prompt learning. In the inference stage (right part), we further activate novel relations that have actional and spatial patterns with high probabilities of occurrence like "run next to" and repress relations that contain patterns with low probabilities of occurrence (gray bars).

the predicate to recognizing the actional pattern, spatial pattern, and corresponding predicate. By leveraging the shared actional patterns and spatial patterns between the base split and novel split, we enhance the model's generalization capability from the base split to the novel split. Specifically, we first insert the decoupled actional pattern descriptions, spatial pattern descriptions, and the complete predicate descriptions into a pre-designed prompt template. Then, we concatenate them with learnable prompt tokens and feed them into the CLIP text encoder to obtain the actional pattern embedding, spatial pattern embedding, and predicate embedding. These embeddings are then used in CLIP-style contrastive learning with the visual tracklet features. Ultimately, our $\mathcal{L}_{rel}$ is rewritten as the sum of three components:

$$\mathcal{L}_{rel} = \mathcal{L}_{pre} + \mathcal{L}_{act} + \mathcal{L}_{spa}. \tag{3}$$

where, $\mathcal{L}_{pre}$, $\mathcal{L}_{act}$ and $\mathcal{L}_{spa}$ denotes the contrastive learning loss between predicate/actional/spatial pattern embeddings and tracklet features.

C. Relation adjustment

As mentioned above, base and novel relations share a significant proportion of actional and spatial patterns. Analyzing actional and spatial relation predictions in the base set will help in distinguishing between candidate relations in the novel set. Besides, the model has a better understanding and classification ability on relations appearing in training data which makes the prediction of these relations more reliable. In this section, we aim to use the prediction results on actional and spatial relations from the base split to adjust the prediction results of visual relations in the novel split to enhance the overall performance. As this procedure essentially changes the priority of visual relations, we call it a relation adjustment procedure [2]. Given a pair of visual objects to be inferred, we obtain a prediction score for each novel predicate $p(c_n^P)$, as well as for each base actional relation $p(c_b^A)$ and spatial relation $p(c_b^S)$, where c_b^A is a base split actional relation instance $\in \mathcal{C}_b^A$, c_b^S is a base split spatial relation instance $\in \mathcal{C}_b^S$ and c_n^P is a novel split predicate instance $\in \mathcal{C}_n^P$. We then deliberate on how to use $p(c_b^A)$ and $p(c_b^S)$ for refinement of $p(c_n^P)$. We split candidate $c_n^P \in \mathcal{C}_n^P$ according to: *(1)* if its corresponding actional/spatial relation $\mathrm{ACT}(c_n^P)$/$\mathrm{SPA}(c_n^P)$ appears in the base split as:

$$\mathcal{C}_1 = \{c_n^P \,|\, \mathrm{ACT}(c_n^P) \, or \, \mathrm{SPA}(c_n^P) \in \mathcal{C}_b^A \cup \mathcal{C}_b^S\}, \tag{4}$$

$$\mathcal{C}_2 = \mathcal{C}_n^P \setminus \mathcal{C}_1. \tag{5}$$

and *(2)* if the prediction score of $\mathrm{ACT}(c_n^P)$ and $\mathrm{SPA}(c_n^P)$ ($c_n^P \in \mathcal{C}_1$) both exceed a pre-defined upper threshold θ_u, and whether one of them fall below the lower threshold θ_l:

$$\mathcal{C}_1^+ = \{c_n^P | p(\mathrm{ACT}(c_n^P)) > \theta_u \,\&\, p(\mathrm{SPA}(c_n^P)) > \theta_u\}, \tag{6}$$

$$\mathcal{C}_1^- = \{c_n^P | p(\mathrm{ACT}(c_n^P)) < \theta_l \,||\, p(\mathrm{SPA}(c_n^P)) < \theta_l\}, \tag{7}$$

$$\mathcal{C}_1^0 = \mathcal{C}_1 \setminus (\mathcal{C}_1^+ \cup \mathcal{C}_1^-). \tag{8}$$

For predicate $c_n^P \in \mathcal{C}_2 \cup \mathcal{C}_1^0$, we leave the predicted $p(c_n^P)$ unchanged because no significant additional knowledge can be transferred from the base split; for $c_n^P \in \mathcal{C}_1^+$, we further activate the predict $p(c_n^P)$ as:

$$p(c_n^P) = 1 - (1 - p(\mathrm{ACT}(c_n^P)) \cdot (1 - p(\mathrm{SPA}(c_n^P)); \tag{9}$$

[2]The main challenge of open-vocabulary VidVRD lies in the detection of novel relations, while we keep the detection process for base relations unchanged.

TABLE I
PERFORMANCE(%) COMPARISON TO EXISTING METHODS. THE **BOLD** ITEMS ARE THE BEST RESULTS.

	Methods	RelDet			RelTag		
		mAP	R@50	R@100	P@1	P@10	P@50
Novel	RePro [18]	6.10	13.38	16.52	13.97	9.55	7.42
	MMP [19]	16.56	16.03	**18.68**	22.05	12.94	10.73
	Ours	**19.90**	**16.86**	18.35	**32.35**	17.50	11.76
	Ours+ReAd	19.84	16.69	18.51	**32.35**	**17.64**	**11.91**
All	RePro [18]	21.33	12.92	15.94	59.00	41.09	28.87
	MMP [19]	26.80	16.26	19.46	72.5	51.25	37.10
	Ours	27.87	16.53	19.54	72.5	53.10	39.81
	Ours+ReAd	**27.88**	**16.59**	**19.61**	**72.5**	**53.10**	**39.91**

for $c_n^P \in \mathcal{C}_1^-$, its actional pattern or spatial pattern is assigned an unacceptable low score, we then repress the predict $p(c_n^P)$ to zero:

$$p(c_n^P) = 0. \tag{10}$$

Through this refinement, we recalculate the prediction score for each candidate relation, resulting in an updated rank of the candidate relations. Predicates containing actional/spatial patterns with high occurrence probability are strengthened, while those with unlikely actional/spatial patterns have their probabilities reduced.

IV. EXPERIMENT

A. Dataset, implementation details and evaluation metrics

Dataset. We adopt the widely used benchmark ImageNet-VidVRD proposed in [20] for our experiments. It contains 1000 videos selected from ILSVRC2016-VID with 800 in the train set and 200 in the test set. The dataset is densely annotated covering 35 categories of objects, e.g., *"airplane"*, *"frisbee"*, etc., as well as 132 categories of visual relations, e.g., *"run left"*, *"walk with"*, etc.

Implementation Details. We use the object trajectories provided in [19]. Object pairs that co-occur in the video are enumerated. We then extract overlapping portions for each pair from the entire video and divide them into multiple clips of 30 frames, from which one frame is sampled. We first obtain prediction scores of visual relations in a single clip then a greedy association algorithm is used to merge relations and scores across segments. Finally, pairs are sorted according to prediction scores, and the top 200 are kept. We set θ_u and θ_l to 0.9 and 0.1 respectively by default unless otherwise specified. We follow [18], [19] in base and novel splitting.

Evaluation Metrics. We adopt standard evaluation metrics as in [18], [19], i.e., *RelDet*: it requires that not only the visual relation should be correct, but the detected object trajectories must have a vIoU greater than 0.5 with the ground truth, with using Recall@K (R@K, K=50,100) and mAP; and *RelTag*: it focuses on object and relation categories measured by Precision@K (P@K, K=1,5,10). We report model performance on (1) Novel split, and (2) All split which involves all visual relations in base and novel sets.

TABLE II
ABLATION STUDY FOR EVALUATING OUR PROPOSED DECOMPOSED PROMPT LEARNING.

Methods	Novel			All		
	mAP	R@50	R@100	mAP	R@50	R@100
ours	**19.90**	**16.86**	18.35	**27.87**	**16.53**	**19.54**
w/o loss	14.50	16.20	**19.34**	26.56	16.17	19.36
w/o prompt	16.70	15.21	17.02	27.26	15.68	18.72

B. Comparison with Existing Methods

We compare our method with previous SOTA methods RePro [18] and MMP [19] in Table I. *"ReAd"* denotes the relation adjustment procedure. Under all split evaluation, our method consistently outperforms across all metrics. In particular, our approach delivered a 1.08% improvement in mAP(RelDet) and a 2.81% gain in P@50(RelTag). Under novel split evaluation, our method achieves significant improvements in all precision-related metrics. Notably, we observed a 10.3% performance boost on P@1(RelTag), and a 3.34% improvement in the mAP(RelDet). These substantial improvements strongly validate the effectiveness of our proposed method.

C. Ablation Study

Decomposed Prompt Learning. Our proposed decomposed prompt learning paradigm leverages the correlation between novel and base visual relations by decoupling visual relations into actional and spatial patterns. Accordingly, we introduce actional and spatial relation prompts on top of the original predicate prompt then we incorporate actional/spatial relation loss to explicitly ensure that the model captures these patterns. To verify our approach, we first remove the actional/spatial relation loss, which impairs the model's ability to grasp and understand actional and spatial patterns. Secondly, we remove the decomposed prompt learning using only the simplest predicate prompt. The results are presented in Table II. A marked decline in performance is evident in either setting, which proves the rationality and effectiveness of our design.

V. CONCLUSION

In this work, we propose a novel method in open-vocabulary video visual relation detection task. We mine the shared actional and spatial patterns between base and novel visual relations and propose a decomposed prompt learning paradigm as well as a relation adjustment mechanism. We conduct a series of experiments on the ImageNet-VidVRD dataset. Through comparison with previous works and ablation studies, we observe substantial improvements and thoroughly demonstrate the effectiveness of our method.

VI. ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant U22A2094 and Grant 62272435, and also supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016), and also supported by the advanced computing resources provided by the Supercomputing Center of the USTC.

REFERENCES

[1] Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67(10):1–16, 2024.

[2] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[3] Yuzhu Wang, Lechao Cheng, Chaowei Fang, Dingwen Zhang, Manni Duan, and Meng Wang. Revisiting the power of prompt for visual tuning. *arXiv preprint arXiv:2402.02382*, 2024.

[4] Hao Zhang, Yeo Keat Ee, and Basura Fernando. Rca: Region conditioned adaptation for visual abductive reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9455–9464, 2024.

[5] Yicong Li, Xun Yang, An Zhang, Chun Feng, Xiang Wang, and Tat-Seng Chua. Redundancy-aware transformer for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3172–3180, 2023.

[6] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *Advances in neural information processing systems*, 33:20612–20623, 2020.

[7] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3940–3956, 2021.

[8] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 928–938, 2022.

[9] Yanbin Hao, Diansong Zhou, Zhicai Wang, Chong-Wah Ngo, and Meng Wang. Posmlp-video: Spatial and temporal relative position encoding for efficient video recognition. *International Journal of Computer Vision*, pages 1–21, 2024.

[10] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. Benchmarking micro-action recognition: Dataset, method, and application. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[11] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1339–1348, 2020.

[12] Cees GM Snoek, Marcel Worring, et al. Concept-based video retrieval. *Foundations and Trends® in Information Retrieval*, 2(4):215–322, 2009.

[13] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 917–925, 2021.

[14] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021.

[15] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Kaifeng Gao, Long Chen, Yulei Niu, Jian Shao, and Jun Xiao. Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19497–19506, 2022.

[27] Wenqing Wang, Yawei Luo, Zhiqing Chen, Tao Jiang, Yi Yang, and Jun Xiao. Taking a closer look at visual relation: Unbiased video scene graph generation with decoupled label learning. *IEEE Transactions on Multimedia*, 2023.