Compositional Prompt Tuning with Motion Cues for Open-Vocabulary Video Relation Detection

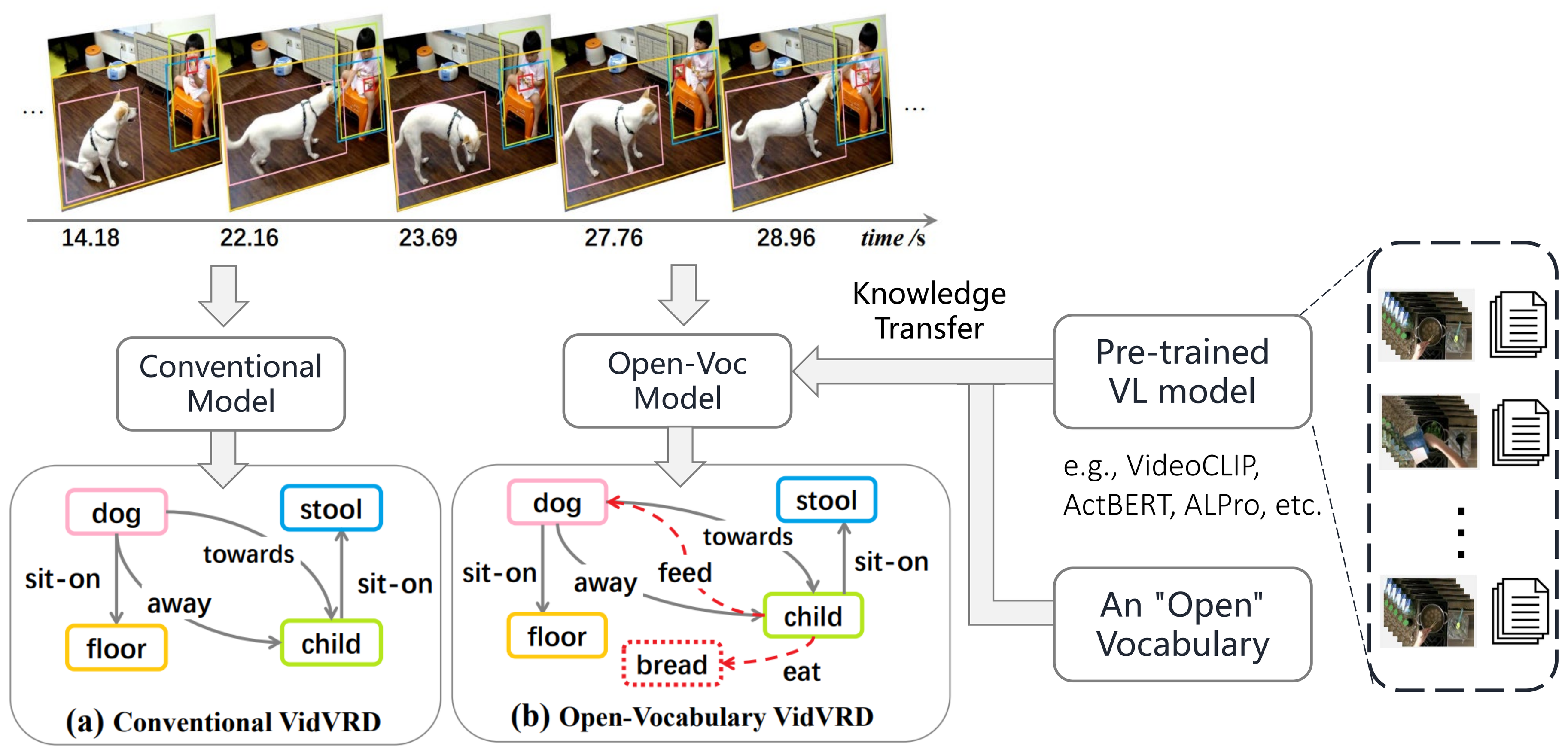
Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, Qianru Sun

Zhejiang University · The Hong Kong University of Science and Technology · Nanyang Technological University Singapore · SMU Singapore Management University

Paper Code

ICLR International Conference On Learning Representations

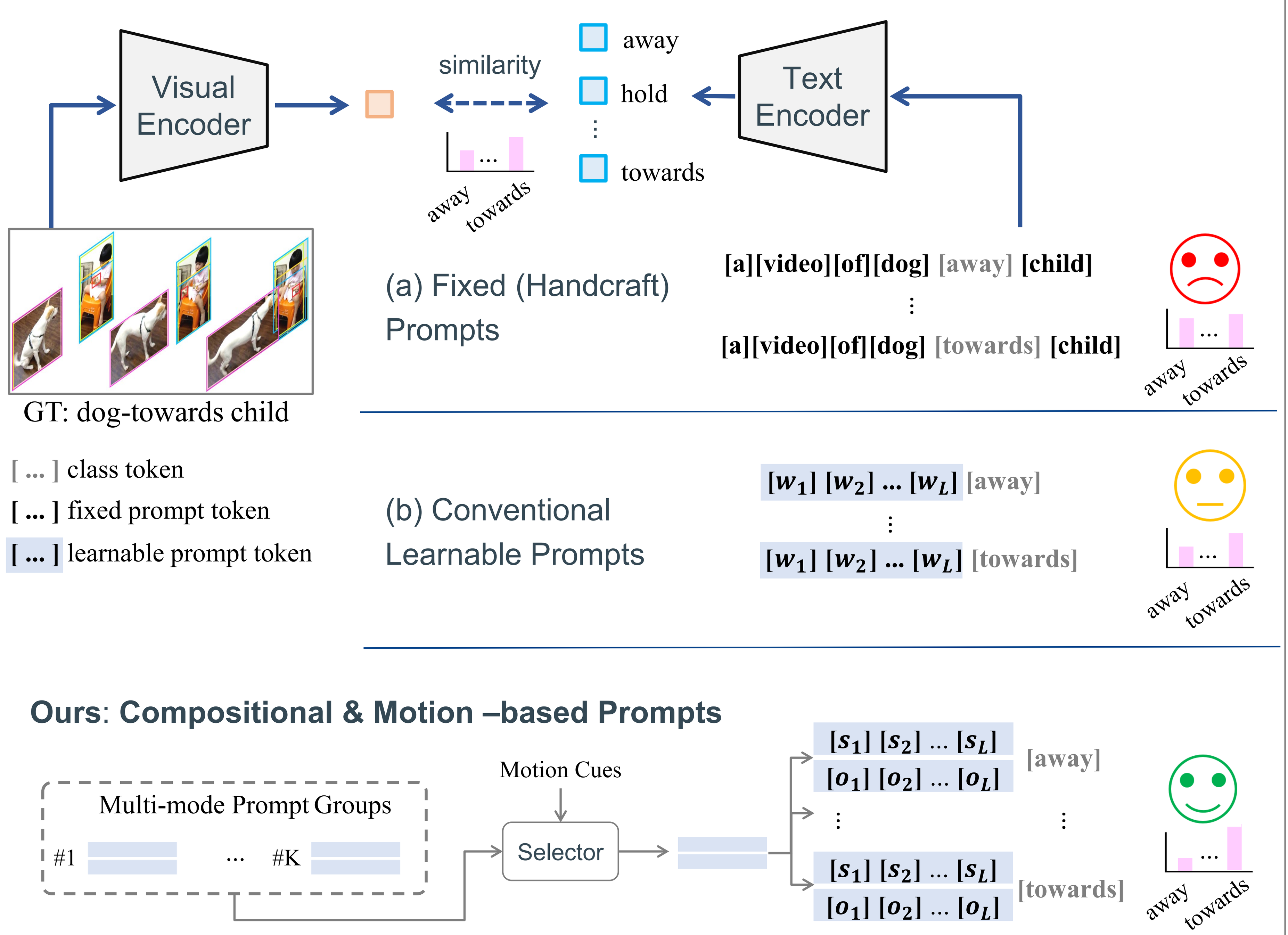
Open-Vocabulary Video Visual Relation Detection



Our contributions:

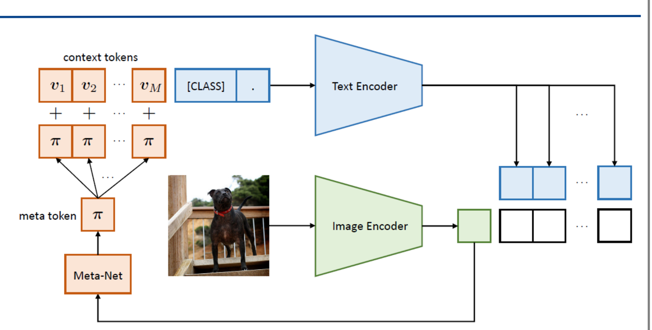
- a new paradigm for Open vocabulary Video VRD.
- a compositional & motion-based prompt tuning/selection approach, which is tailored for **Re**lation **Pro**mpt tuning (RePro)

Compositional & Motion –based Prompting



Advantages:

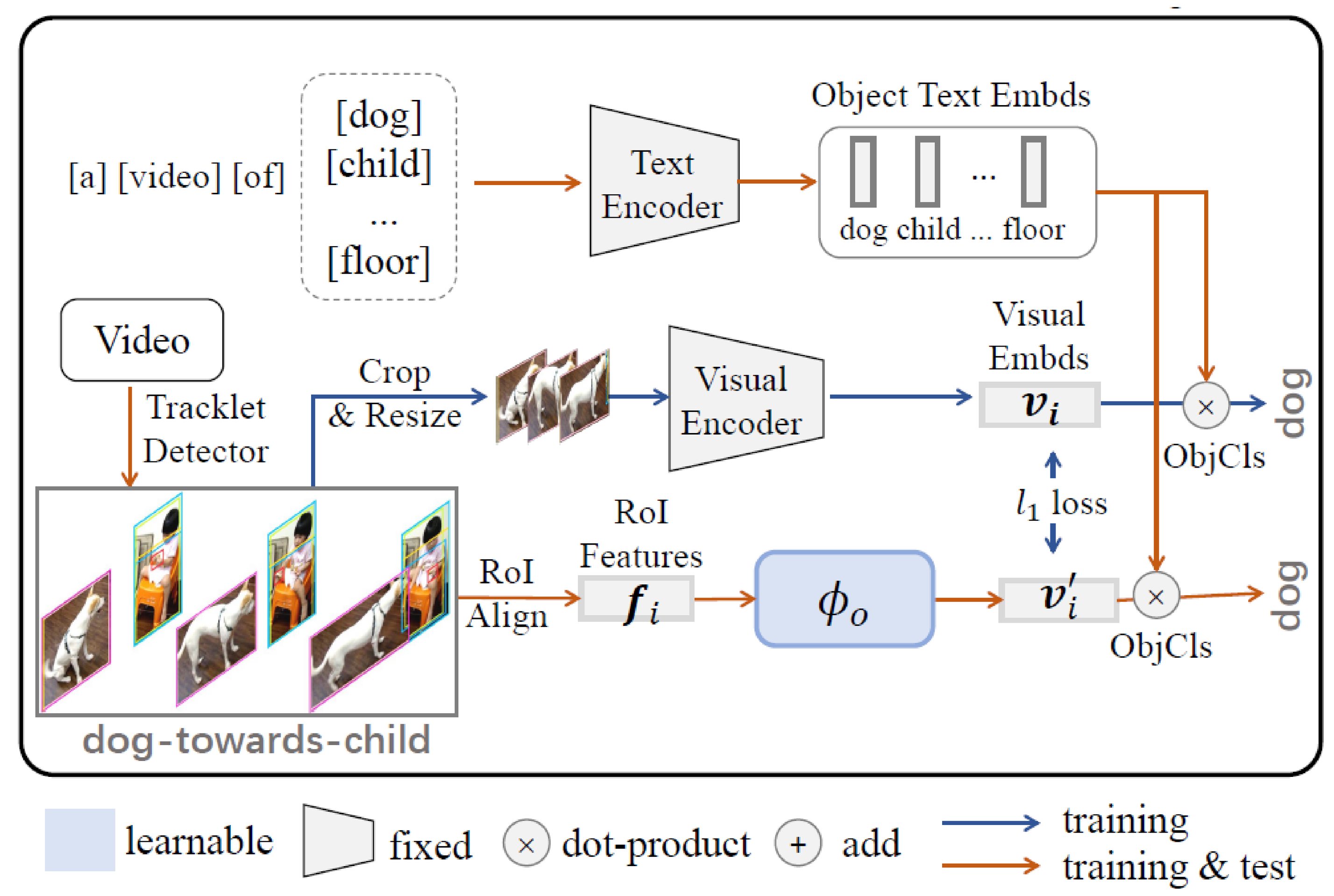
- It considers spatial-temporal motions of different relations;
- It considers different semantic contexts of subject & object;
- Compared to category or instance -conditioned context(*), it achieves better cross-category generalization



* Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." In CVPR 2022.

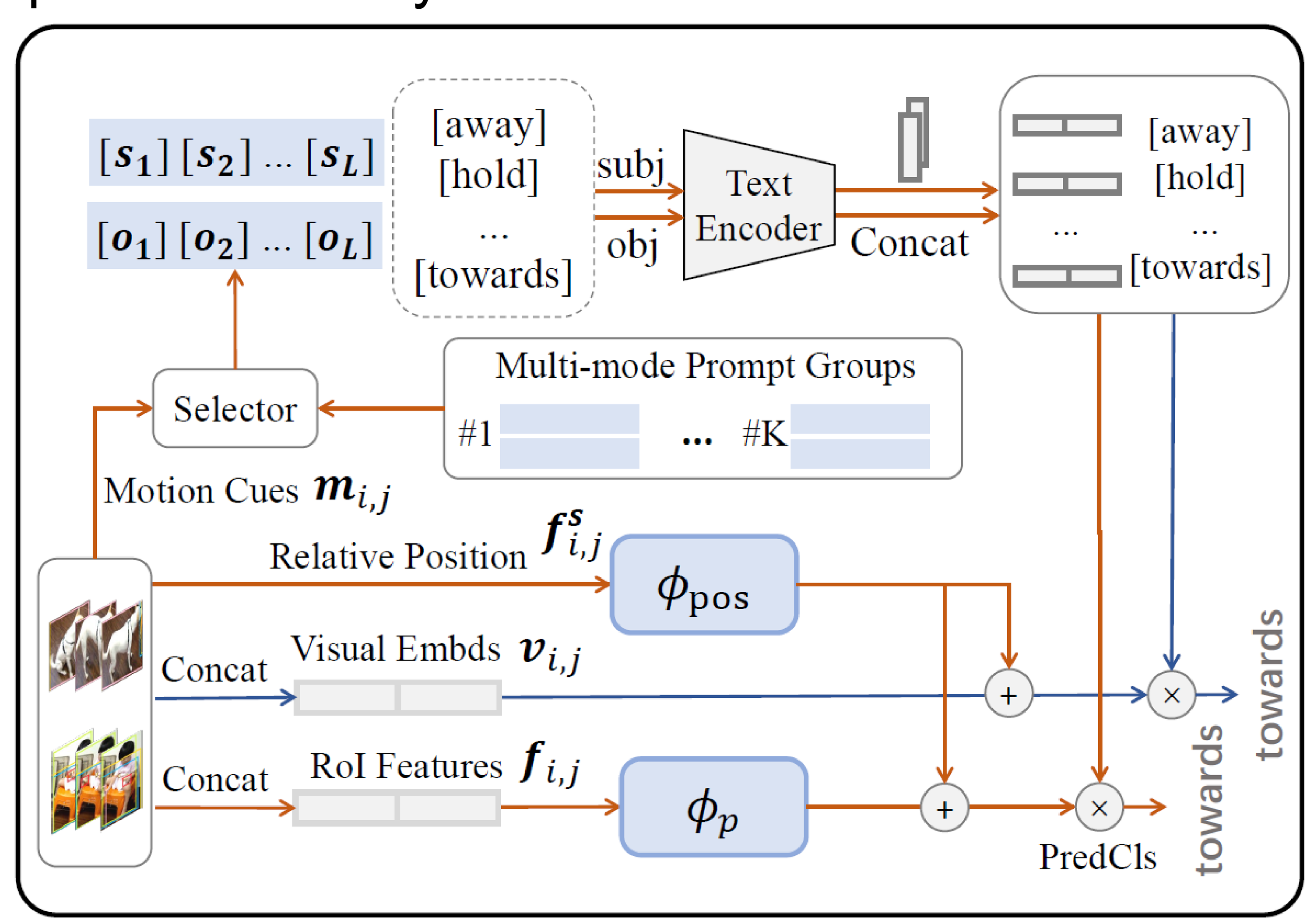
Distillation vs. Prompt –based Knowledge Transfer

1. Open-Vocabulary Tracklet Detection



- Train a visual-to-language (V2L) projection module $\phi_o(\cdot)$ (on base classes);
- Transfer knowledges from VLM via distillation (i.e., l_1 loss);
- Avoid directly inference the heavy pipeline of VLM's visual encoder (test time).

2. Open-Vocabulary Relation Detection



Motion Pattern: $m_{i,j} = \text{sign}([G^s_{i,j} - \gamma, G^e_{i,j} - \gamma, G^e_{i,j} - G^s_{i,j}])$, where $m_{i,j} \in \{+,-\}^3$,

$G^*_{i,j} = \text{GIoU}(T_i, T_j)$ (G^s : start frame, G^e : end frame).

Transfer knowledge via prompt, instead of distillation

- stage-1: train the prompt representations in the comp. & motion –based manner
- stage-2: train V2L module, i.e., $\phi_p(\cdot)$ (on base classes) based on the learned prompt representations.

Experiments Results

➢ Compare with SOTA in conventional setting

Methods	Training Data	SGDet mAP	SGDet R@50	SGDet R@100	RelTag P@1	RelTag P@5	RelTag P@10
Su et al. (2020)	base+novel	19.03	9.53	10.38	57.50	41.40	29.45
Liu et al. (2020)	base+novel	18.38	11.21	13.69	60.00	43.10	32.24
Li et al. (2021)	base+novel	**22.97**	12.40	14.46	**68.83**	**49.87**	**35.57**
Gao et al. (2022)	base+novel	17.67	9.63	11.29	56.00	43.80	**32.85**
RePro (Ours)	base	21.33	**12.92**	**15.94**	59.00	41.09	28.87
RePro (Ours)	base+novel	**25.55**	**13.83**	**17.33**	**62.50**	**45.80**	32.05

- When trained with only ***base*** category samples, our RePro still achieves comparable performance with SOTA.

➢ Comparison in the Open-Vocabulary setting

Split	Methods	SGDet mAP	SGDet R@50	SGDet R@100	SGCls mAP	SGCls R@50	SGCls R@100	PredCls mAP	PredCls R@50	PredCls R@100
Novel	★ ALPro	1.05	3.14	4.62	3.69	7.27	8.92	4.09	9.42	10.41
	✦ VidVRD-II	3.57	8.59	12.39	5.70	13.22	18.34	7.35	18.84	26.44
	RePro†	2.56	8.26	11.73	8.63	15.04	18.84	9.34	18.67	24.13
	RePro	**6.10**	**13.38**	**16.52**	**10.32**	**19.17**	**25.28**	**12.74**	**25.12**	**33.88**
All	★ ALPro	3.20	2.62	3.18	3.92	3.88	4.75	4.97	4.50	5.79
	✦ VidVRD-II	12.74	9.90	12.59	17.26	14.93	19.68	19.73	18.17	24.90
	RePro†	16.21	11.14	14.56	22.37	16.83	21.71	25.43	21.36	28.04
	RePro	**21.33**	**12.92**	**15.94**	**30.15**	**19.75**	**25.00**	**34.90**	**25.50**	**32.49**

★ Pre-trained VLM zero-shot inference

- Li, Dongxu, et al. "Align and prompt: Video-and-language pre-training with entity prompts." In CVPR 2022.

✦ Baseline VidVRD model

- Shang, Xindi, et al. "Video visual relation detection via iterative inference." ACM Multimedia. 2021.

➢ Ablation Studies for Comp. & Motion Prompting

		C	M	SGDet mAP	SGDet R@50	SGDet R@100	SGCls mAP	SGCls R@50	SGCls R@100	PredCls mAP	PredCls R@50	PredCls R@100
Novel-split	#1	×	×	3.50	9.91	13.88	7.21	14.54	19.83	8.63	20.33	27.43
	#2	✓	×	5.57	11.40	14.87	10.31	16.52	21.81	11.83	22.31	30.90
	#3	✓	Ens	6.24	11.57	15.20	10.77	16.03	21.98	12.36	21.32	29.91
	#4	✓	Rand	**7.14**	11.90	14.87	**10.85**	16.52	23.30	12.42	22.64	30.90
	#5	✓	✓	6.10	**13.38**	**16.52**	10.32	**19.17**	**25.28**	**12.74**	**25.12**	**33.88**
All-splits	#1	×	×	19.73	12.26	15.36	26.80	18.24	23.06	30.80	23.70	30.42
	#2	✓	×	18.47	11.95	15.28	25.52	18.13	23.12	29.45	23.39	30.17
	#3	✓	Ens	20.15	12.38	15.61	27.93	18.61	23.55	31.68	23.61	30.29
	#4	✓	Rand	**21.72**	12.71	15.78	29.15	19.15	24.13	33.11	24.38	31.49
	#5	✓	✓	21.33	**12.92**	**15.94**	**30.15**	**19.75**	**25.00**	**34.90**	**25.50**	**32.49**

C: Compositional; **M**: Motion cues;

Ens: ensemble all the learned prompts by averaging their representations.

Rand: randomly select a prompt without considering motion cues

➢ Ablation Studies for different predicate groups

Methods	move	sit	run	walk	stop	stand	fly	swim
Ens	34.48	50.92	12.90	18.30	37.03	35.51	37.50	**15.38**
Rand	37.93	51.85	16.12	18.30	**44.44**	36.44	**50.00**	**15.38**
RePro	**44.82**	**55.55**	**25.80**	**18.95**	40.47	**41.12**	**50.00**	12.82

- Performance reported as Recall@100 (%) of PredCls
- Predicates are grouped by the prefix their words, e.g. "run past", "run next to"
- It indicates that the performance improvements of RePro are largely attributed to motion cues