

Fig. R 1: Illustration of different zero-shot settings. O^{train} (O^{test}), R^{train} (R^{test}), and T^{train} (T^{test}) represent the set of object categories, relation categories, and triplet categories during the training (test) stage, respectively.

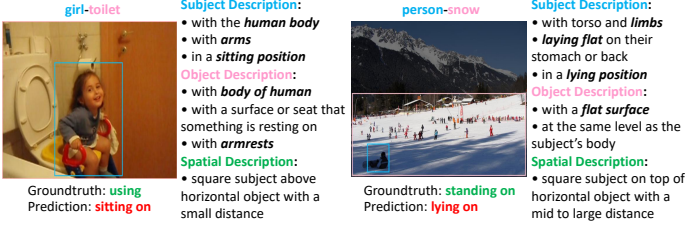


Fig. R 2: Failure cases analysis on VG dataset. The descriptions of prediction errors for each category are provided on the right.

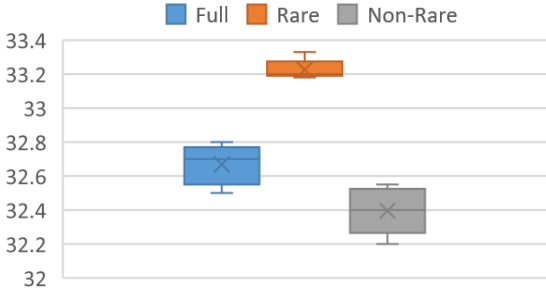


Fig. R 3: The error bar of HICO-DET. It comes from the results obtained by generating descriptions through five invocations of the LLM.

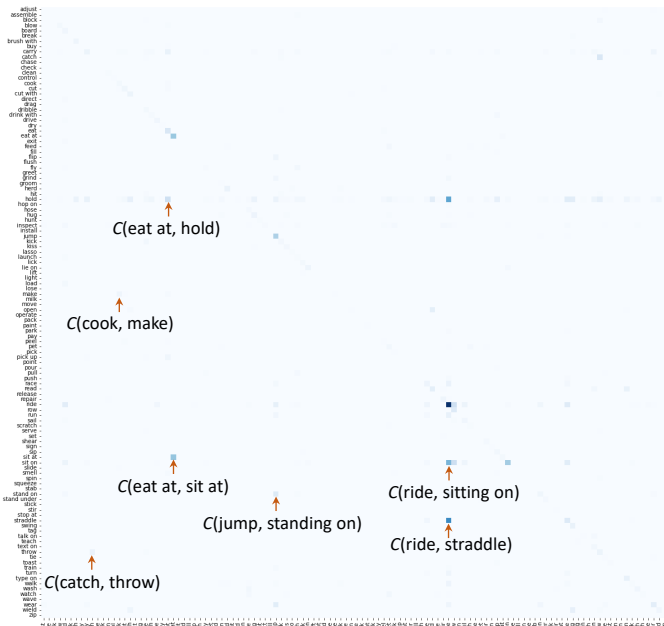


Fig. R 4: The confusion matrix on the HICO-DET dataset.

References

[1] Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, 2022. 1

Table R 1: Performance (%) of combining with different SOTA pre-trained visual-language models on VG dataset. CLS* denotes the model uses class-based prompts to compute the training-free zero-shot similarity between the image and text.

Backbone	Method	Predicate Classification					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
MS-CLIP [1]	Baseline(CLS*)	8.2	15.1	21.5	7.9	16.4	22.4
	RECODE*	9.2	17.3	24.7	8.3	15.4	22.6
DECLIP [2]	Baseline(CLS*)	11.0	18.3	24.4	11.0	19.0	27.1
	RECODE*	11.4	19.3	25.9	10.5	19.5	27.8

Table R 2: Comparison with SOTA VRD methods on the VG dataset. Note that none of these methods can be applied in the training-free zero-shot setting.

Model	No Training	Unseen Relation	Training Data Source	Predicate Classification		
				zR@20	zR@50	zR@100
Motifs [3]	×	×	VG	8.9	15.2	18.5
COACHER [4]	×	×	VG& ConceptNet	28.2	34.1	37.2
DPL [5]	×	×	VG	6.0	7.7	9.3
CaCao [6]	×	✓	VG&CC3M&COCO	17.2	21.3	23.1
RECODE	✓	✓	-	8.2	16.1	23.2

Table R 3: Ablation studies on the HICO-DET and VCOCO datasets.

Cue	Spatial	Weight	HICO-DET			VCOCO	
			Full	Rare	Non-Rare	Scenario 1	Scenario 2
			30.9	30.7	31.0	25.5	28.6
✓			32.5	33.0	32.2	25.8	28.9
✓	✓		32.6	33.0	32.4	25.7	28.8
✓		✓	32.7	33.1	32.5	25.9	29.0
✓	✓	✓	32.7	33.2	32.5	26.0	29.0

Table R 4: Comparison with or without CoT on the VG dataset.

CoT	Predicate Classification					
	R@20	R@50	R@100	mR@20	mR@50	mR@100
×	9.5	17.3	24.6	10.2	18.0	25.6
✓	10.6	18.3	25.0	10.7	18.7	27.8

Table R 5: Analysis of key components on the VG dataset. Time (ms) represents the computation time of each triplet. The (·) represents the time when spatial component is retrieved offline.

Cue	Spatial	Weight	Predicate Classification					
			R@20	R@50	R@100	mR@20	mR@50	mR@100
			7.2	10.9	13.2	9.4	14.0	17.6
✓			7.4	12.3	16.6	9.0	14.0	19.5
✓	✓		9.1	13.4	17.4	9.3	15.0	20.3
✓		✓	7.9	13.4	17.7	9.3	14.7	20.5
✓	✓	✓	9.7	14.9	19.3	10.2	16.4	22.7

[2] Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 1

[3] Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1

[4] Zero-shot scene graph relation prediction through common-sense knowledge integration. In *ECML PKDD*, 2021. 1

[5] Decomposed prototype learning for few-shot scene graph generation. *arXiv preprint arXiv:2303.10863*, 2023. 1

[6] Visually-prompted language model for fine-grained scene graph generation in an open world. In *ICCV*, 2023. 1