Scene Graph Generation with Role-Playing Large Language Models

Guikun Chen[1], Jin Li[2], Wenguan Wang[1]

[1]ZJU

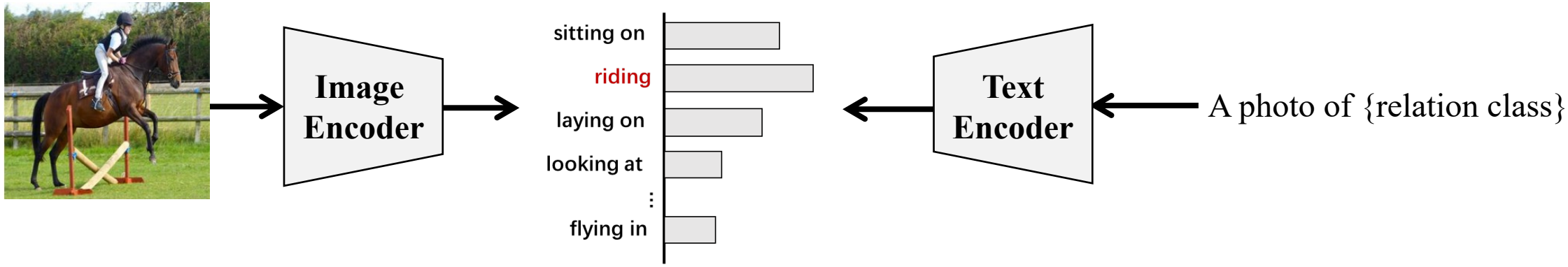
[2]CSUST

1 Definition of OVSGG

Background:

Given an input image, open-vocabulary SGG (OVSGG) aims to detect all objects and their pairwise relationships beyond pre-defined categories. This work focuses on the predicate classification task of SGG (*i.e.,* given the ground-truth object boxes and categories).

A simple baseline using CLIP：

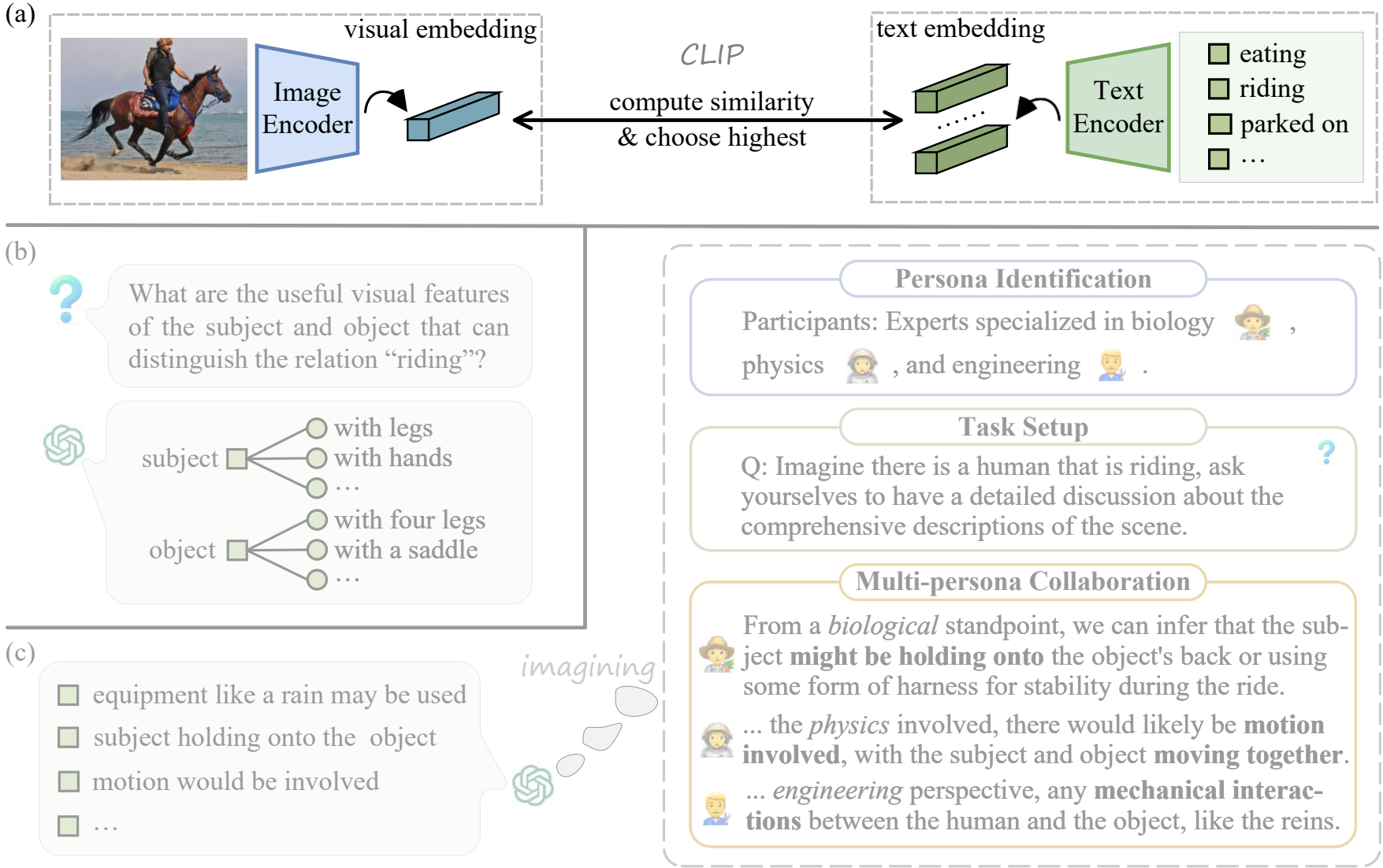


2 Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

Challenges:

- struggle to model the large variance in visual relations.
- overlook the possibility that some text classifiers might be contrary to specific contexts.

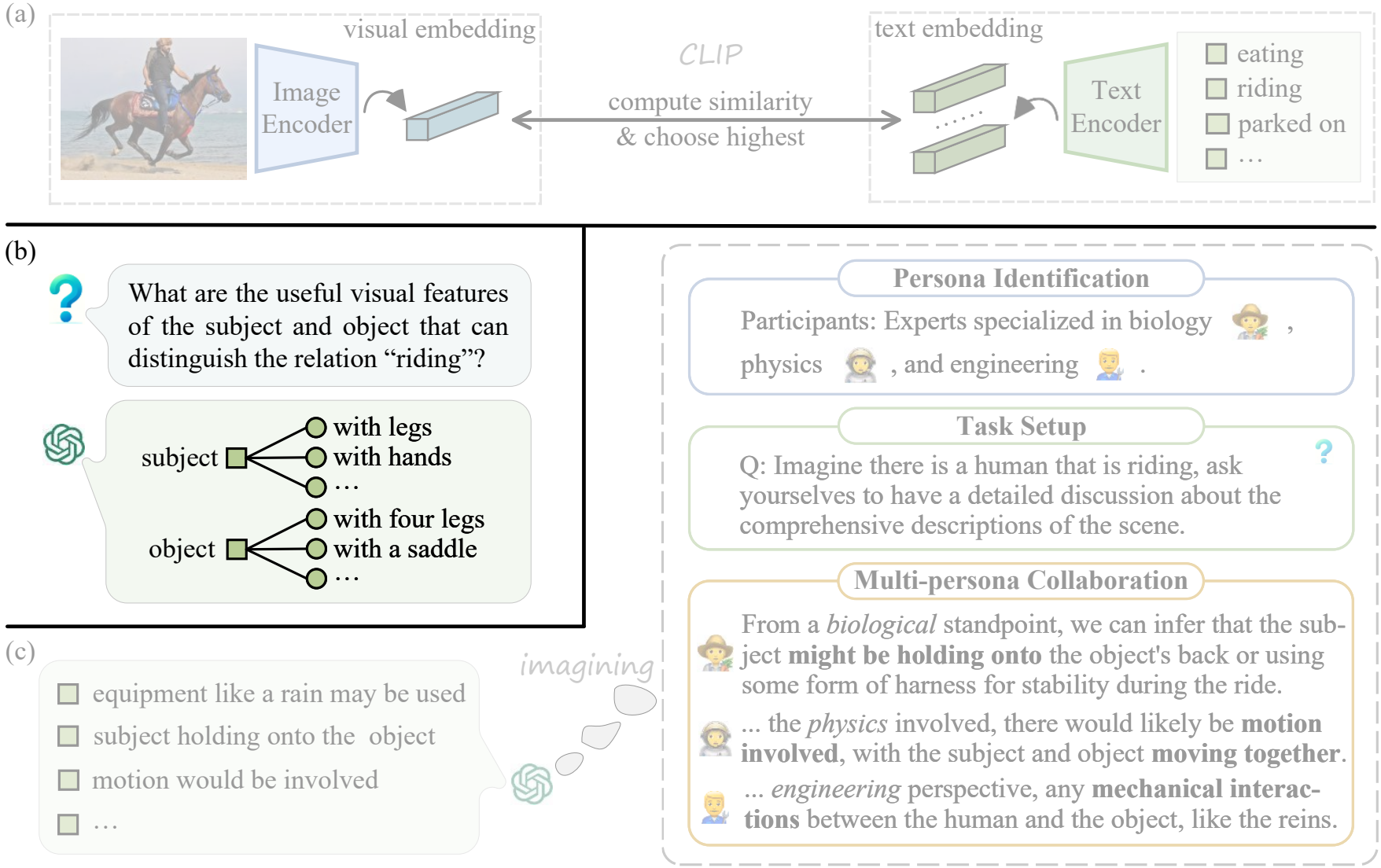


2 Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

Challenges:

- struggle to model the large variance in visual relations.
- overlook the possibility that some text classifiers might be contrary to specific contexts.

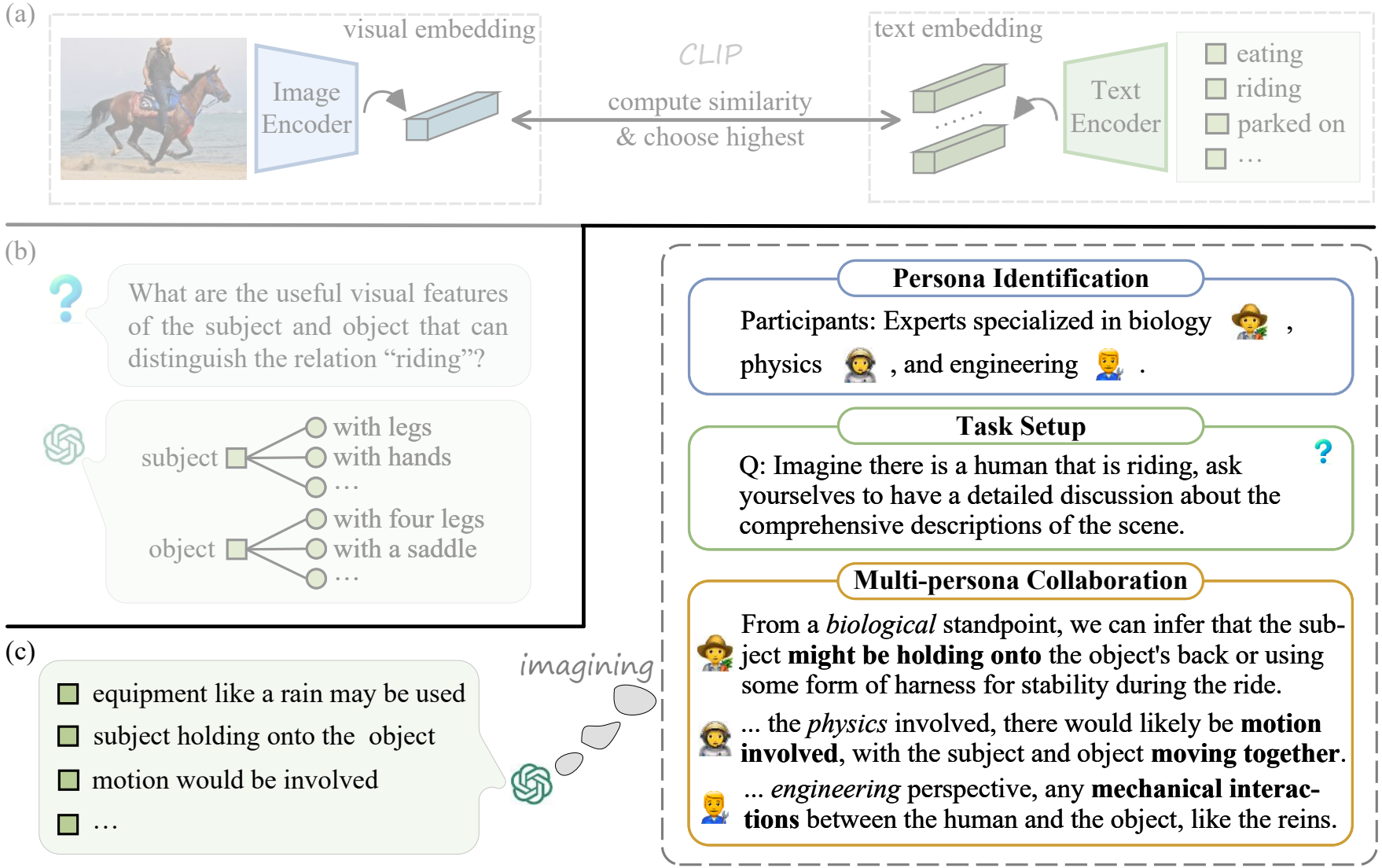


2 Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

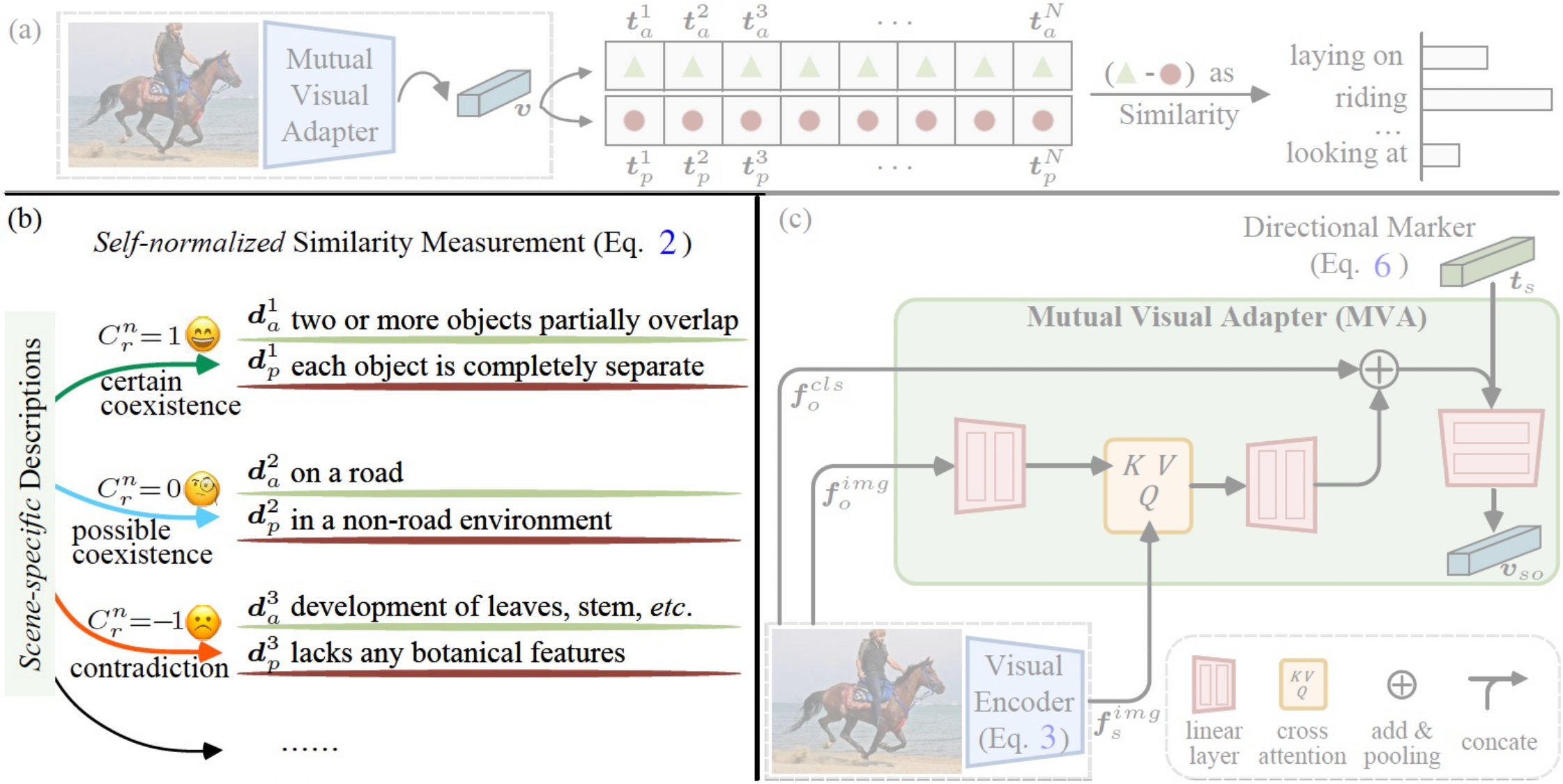
Core ideas:

- generate comprehensive and diverse scene descriptions via LLMs' role-playing.
- the weights of text classifiers are adaptively adjusted according to the visual content.



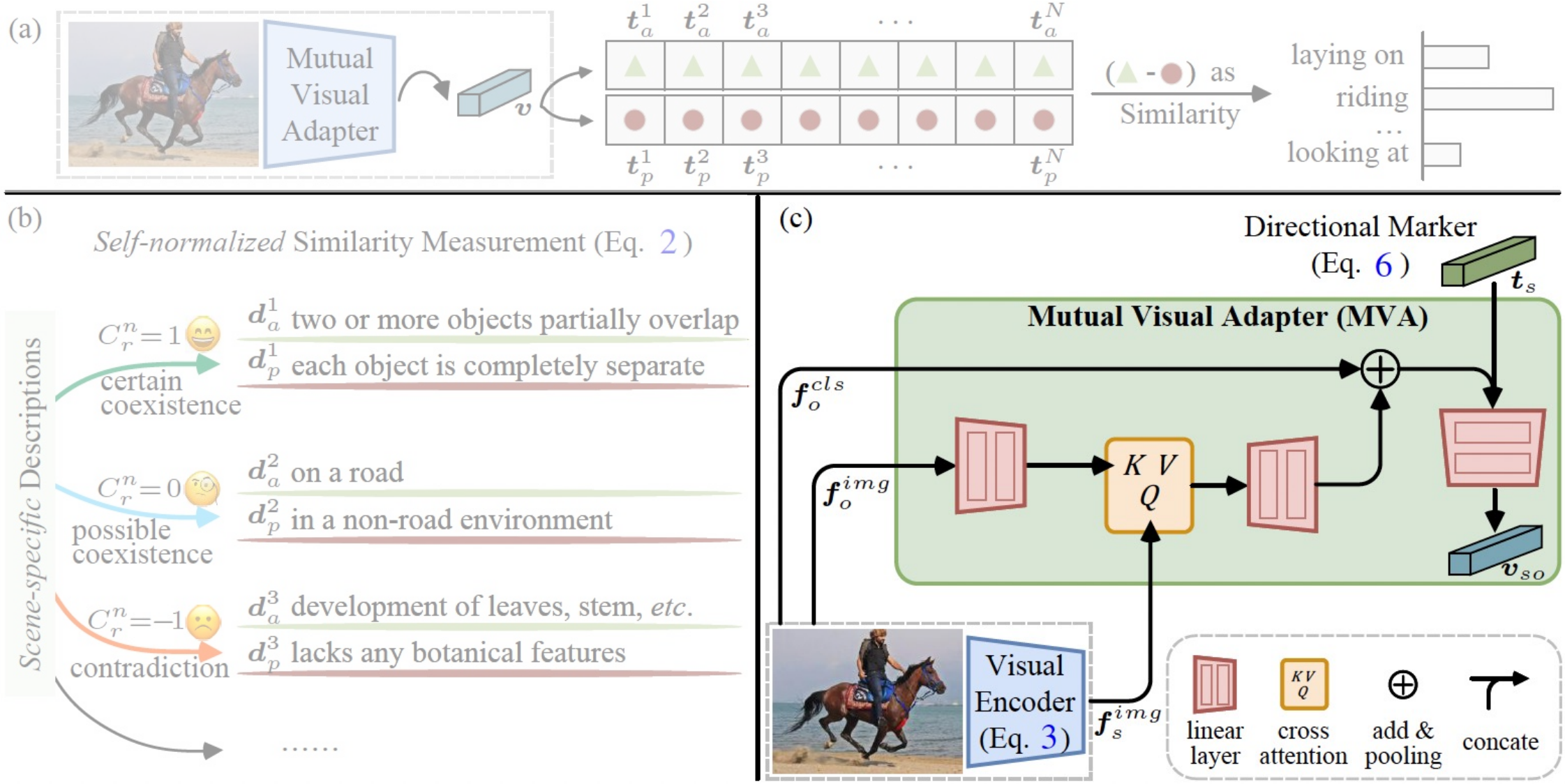
Methodology

NEURAL INFORMATION PROCESSING SYSTEMS



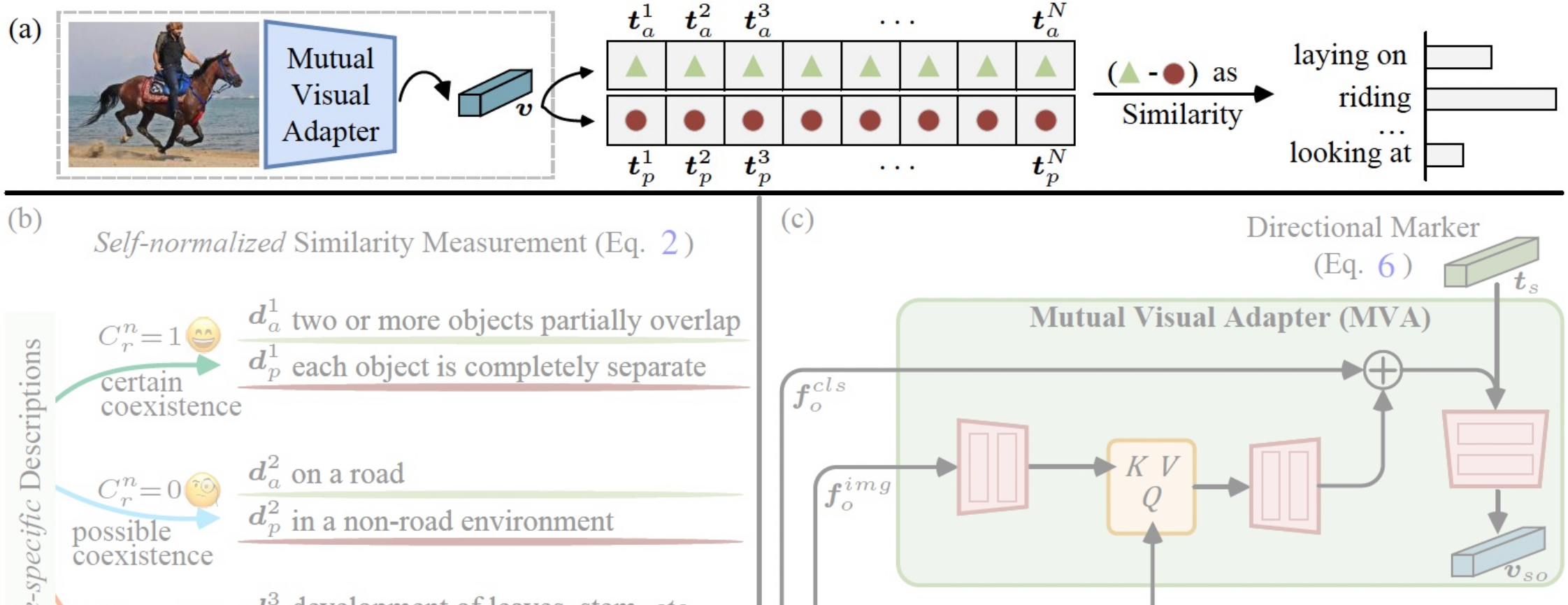
Methodology

NEURAL INFORMATION PROCESSING SYSTEMS



Methodology

NEURAL INFORMATION PROCESSING SYSTEMS



4 Experiment

NEURAL INFORMATION PROCESSING SYSTEMS

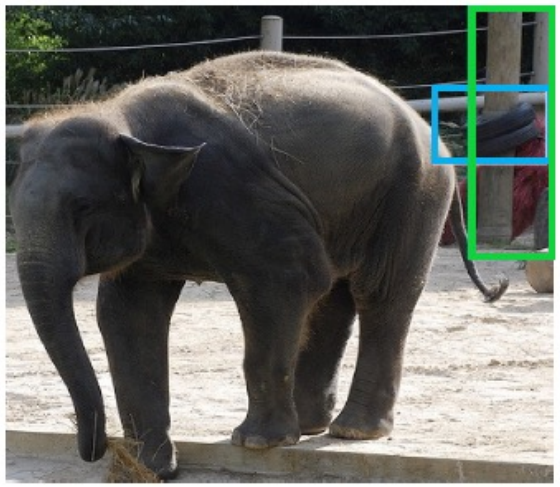
Table 1: Quantitative results (§4.2) on VG [14] `base` and `novel`.

Method	Split	R@20↑	R@50↑	R@100↑	mR@20↑	mR@50↑	mR@100↑
CLS[ICML21][4]	base	2.1	3.2	3.9	7.0	9.0	10.9
Epic[ICCV23][3]		-	22.6	27.2	-	-	-
Ours		**18.7**$_{\pm 0.69}$	**26.5**$_{\pm 0.92}$	**31.6**$_{\pm 1.00}$	**9.2**$_{\pm 0.14}$	**12.4**$_{\pm 0.12}$	**14.8**$_{\pm 0.10}$
CLS[ICML21][4]	novel	13.2	18.1	22.2	11.5	17.9	23.8
Epic[ICCV23][3]		-	7.4	9.7	-	-	-
Ours		**18.4**$_{\pm 0.53}$	**25.4**$_{\pm 0.48}$	**29.6**$_{\pm 0.42}$	**17.1**$_{\pm 0.42}$	**25.2**$_{\pm 0.95}$	**31.2**$_{\pm 1.09}$

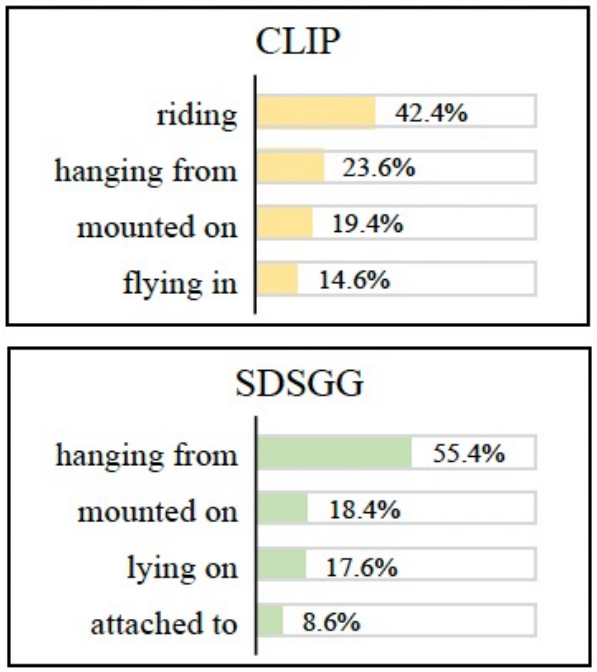
Table 2: Quantitative results (§4.2) on VG [14] `semantic`.

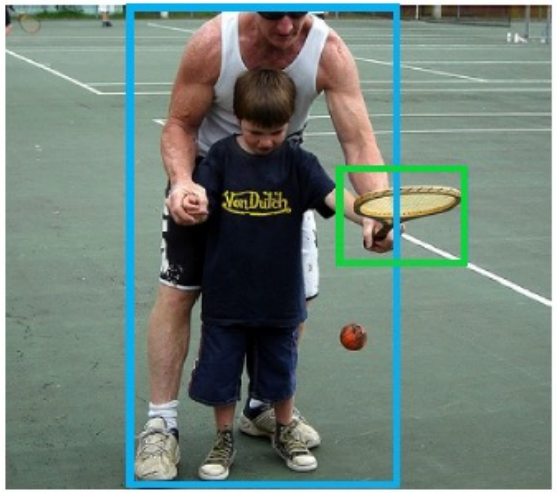
Method	R@20↑	R@50↑	R@100↑	mR@20↑	mR@50↑	mR@100↑
CLS[ICML21][4]	7.2	10.9	13.2	9.4	14.0	17.6
CLSDE[NeurIPS23][12]	7.0	10.6	12.9	8.5	13.6	16.9
RECODE†[NeurIPS23][12]	7.3	11.2	15.4	8.2	13.5	18.3
RECODE[NeurIPS23][12]	9.7	14.9	19.3	10.2	16.4	22.7
RECODE⋆[NeurIPS23][12]	10.6	18.3	25.0	10.7	18.7	27.8
Ours	**21.5**$_{\pm 0.47}$	**29.3**$_{\pm 0.53}$	**34.9**$_{\pm 0.66}$	**16.8**$_{\pm 0.08}$	**22.7**$_{\pm 0.41}$	**28.4**$_{\pm 0.67}$

Experiment



<tire, hanging from, pole>





<man, holding, racket>

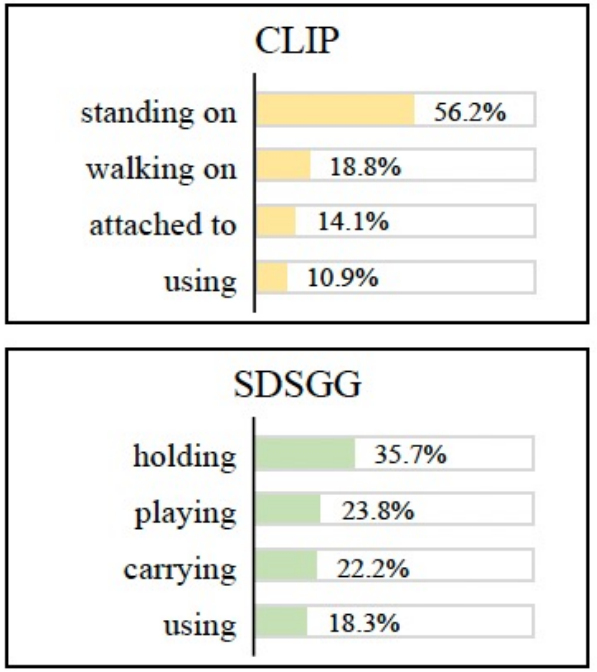


Figure 3: Visual results (§4.3) on VG [14].

Our code will be available at

https://github.com/guikunchen/SDSGG

ACCEPT MY ENDLESS GRATITUDE