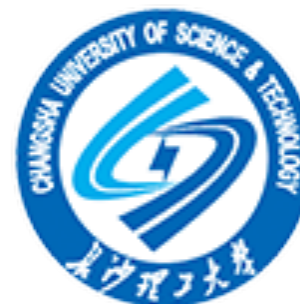


Scene Graph Generation with Role-Playing Large Language Models

Guikun Chen¹, Jin Li², Wenguan Wang¹



¹ZJU



²CSUST



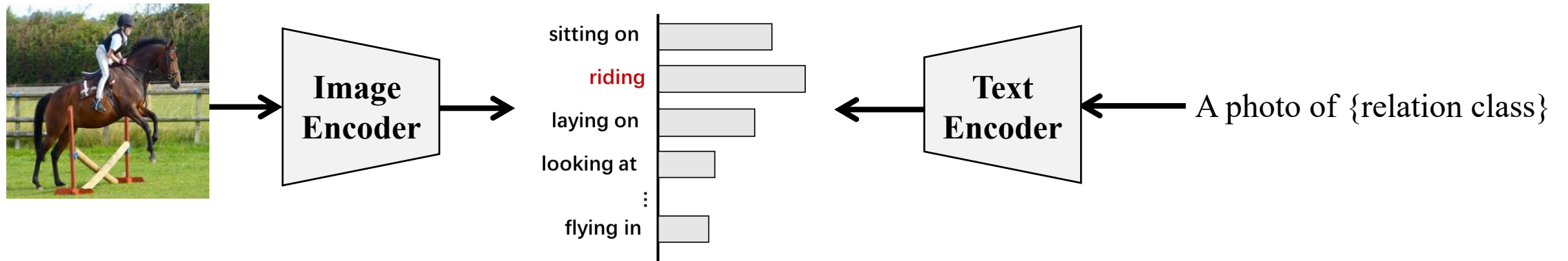
Definition of OVSGG



Background:

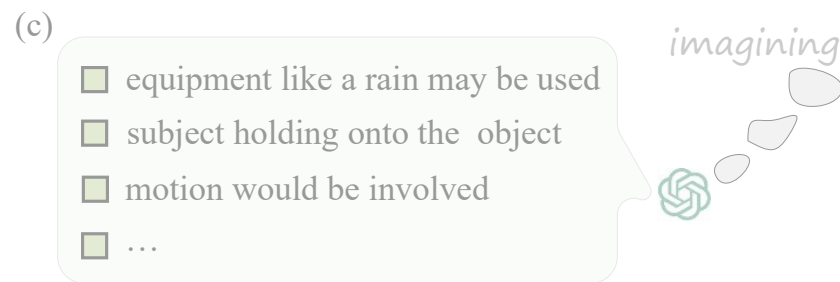
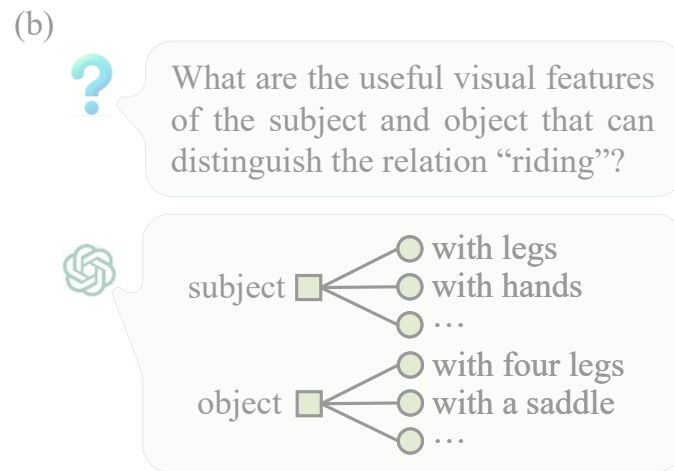
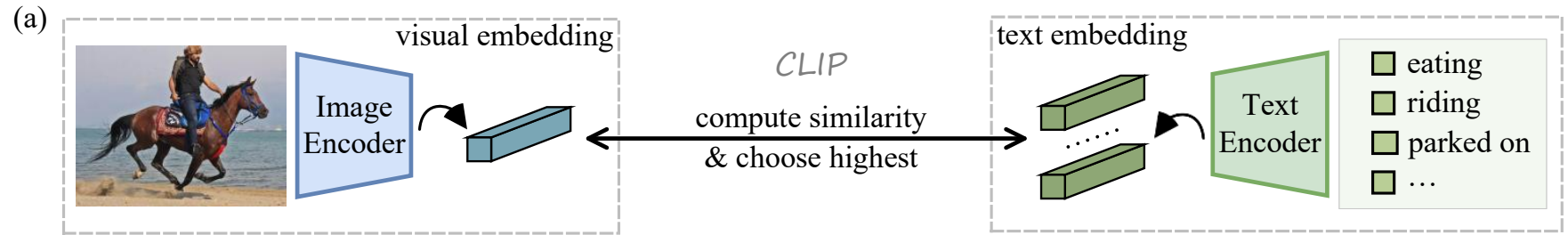
Given an input image, open-vocabulary SGG (OVSGG) aims to detect all objects and their pairwise relationships **beyond pre-defined categories**. This work focuses on the predicate classification task of SGG (*i.e.*, given the ground-truth object boxes and categories).

A simple baseline using CLIP :



Challenges:

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



Persona Identification

Participants: Experts specialized in biology 🧑‍🔬, physics 🧑‍🔧, and engineering 🧑‍🎓.

Task Setup

Q: Imagine there is a human that is riding, ask yourselves to have a detailed discussion about the comprehensive descriptions of the scene. ?

Multi-persona Collaboration

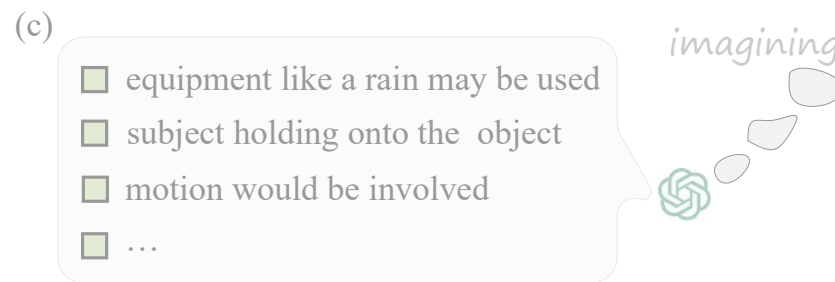
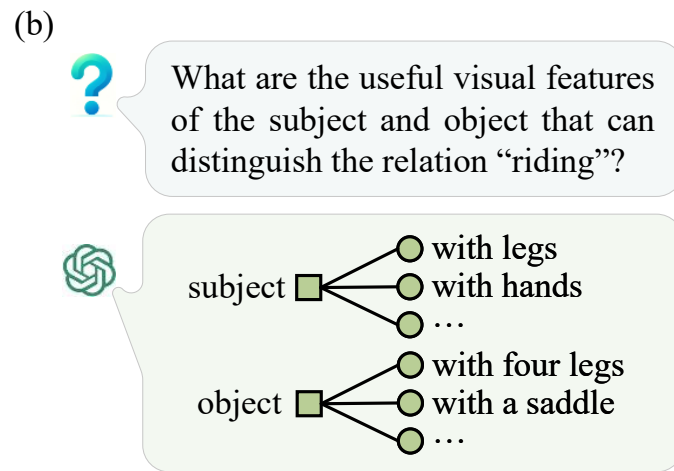
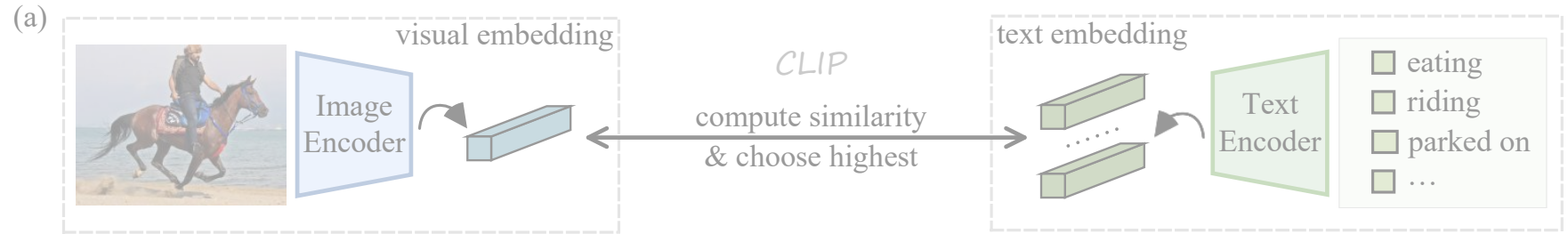
From a *biological* standpoint, we can infer that the subject **might be holding onto** the object's back or using some form of harness for stability during the ride.

... the *physics* involved, there would likely be **motion involved**, with the subject and object **moving together**.

... *engineering* perspective, any **mechanical interactions** between the human and the object, like the reins.

Challenges:

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



Persona Identification

Participants: Experts specialized in biology 🧑‍🔬, physics 🧑‍🔧, and engineering 🧑‍🚀.

Task Setup

Q: Imagine there is a human that is riding, ask yourselves to have a detailed discussion about the comprehensive descriptions of the scene. ?

Multi-persona Collaboration

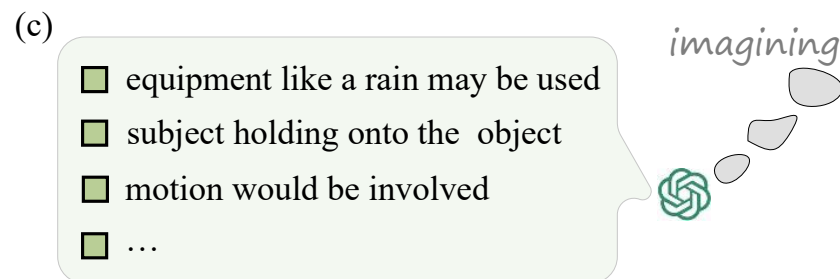
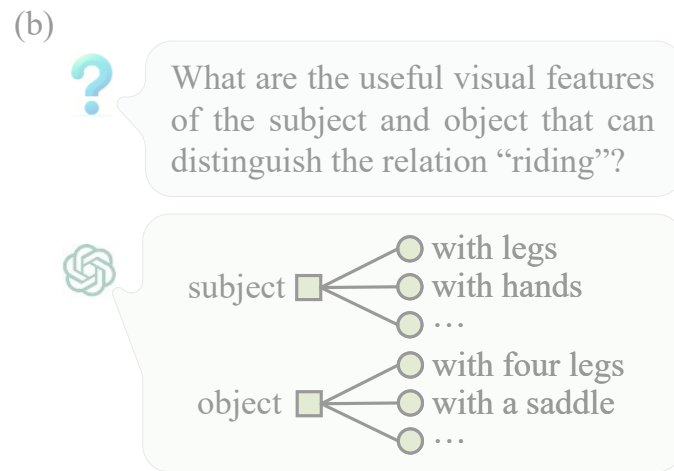
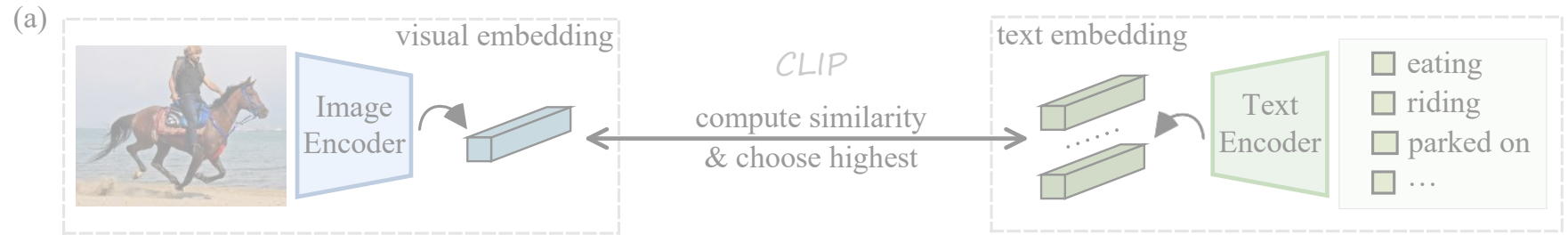
From a *biological* standpoint, we can infer that the subject **might be holding onto** the object's back or using some form of harness for stability during the ride.

... the *physics* involved, there would likely be **motion involved**, with the subject and object **moving together**.

... *engineering* perspective, any **mechanical interactions** between the human and the object, like the reins.

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.



Persona Identification

Participants: Experts specialized in biology 🧑‍🌾 , physics 🧑‍🔬 , and engineering 🧑‍🏭 .

Task Setup

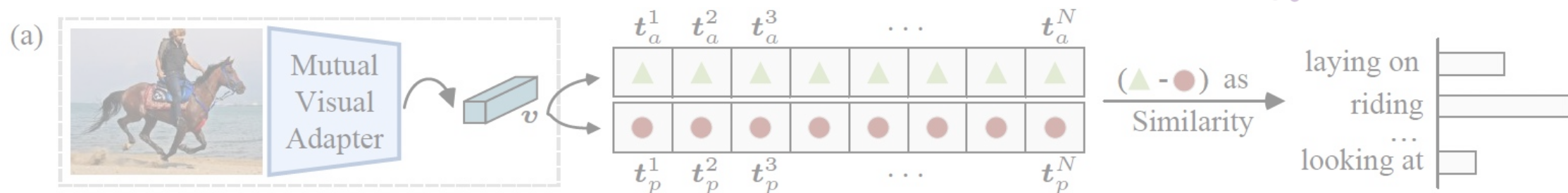
Q: Imagine there is a human that is riding, ask yourselves to have a detailed discussion about the comprehensive descriptions of the scene. ?

Multi-persona Collaboration

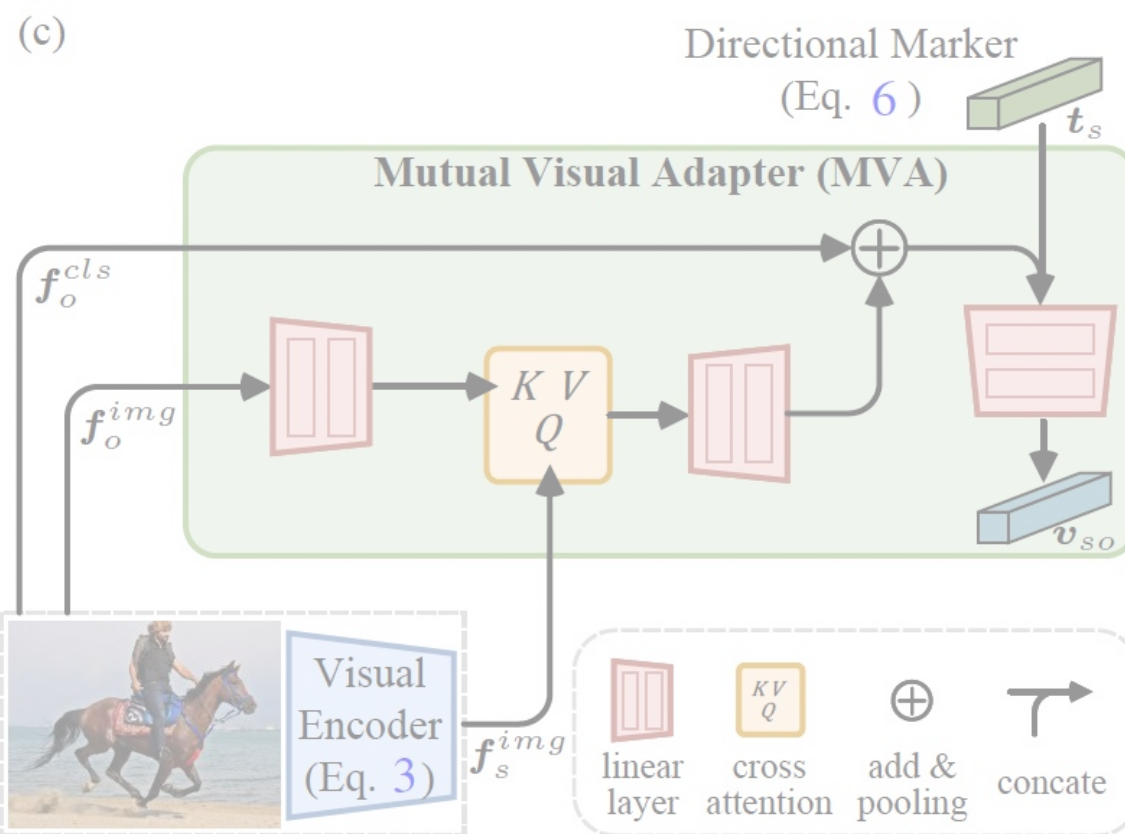
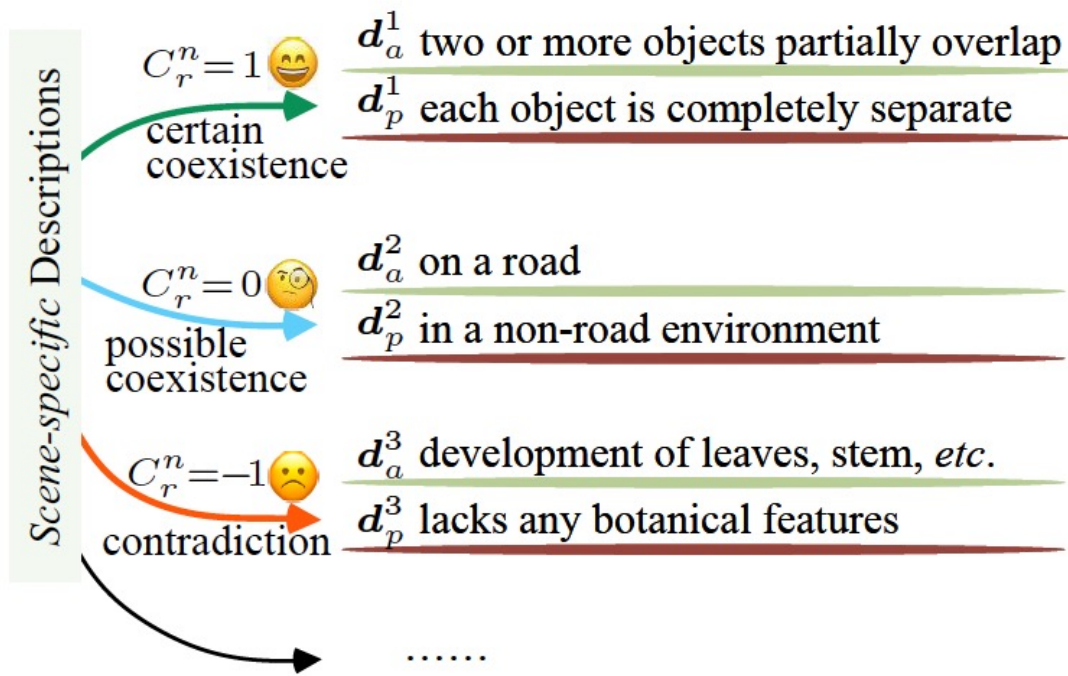
From a *biological* standpoint, we can infer that the subject **might be holding onto** the object's back or using some form of harness for stability during the ride.

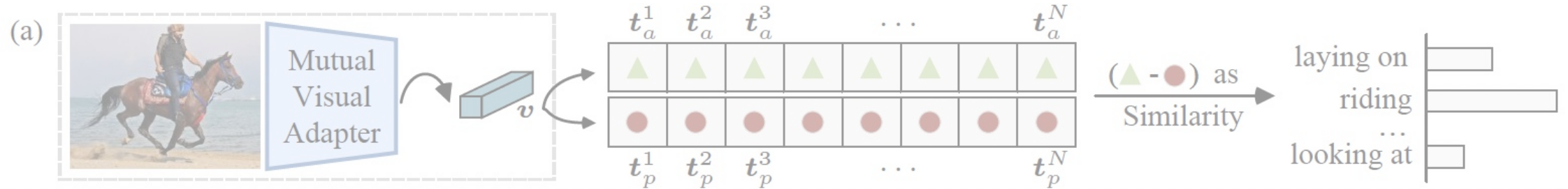
... the *physics* involved, there would likely be **motion involved**, with the subject and object **moving together**.

... *engineering* perspective, any **mechanical interactions** between the human and the object, like the reins.

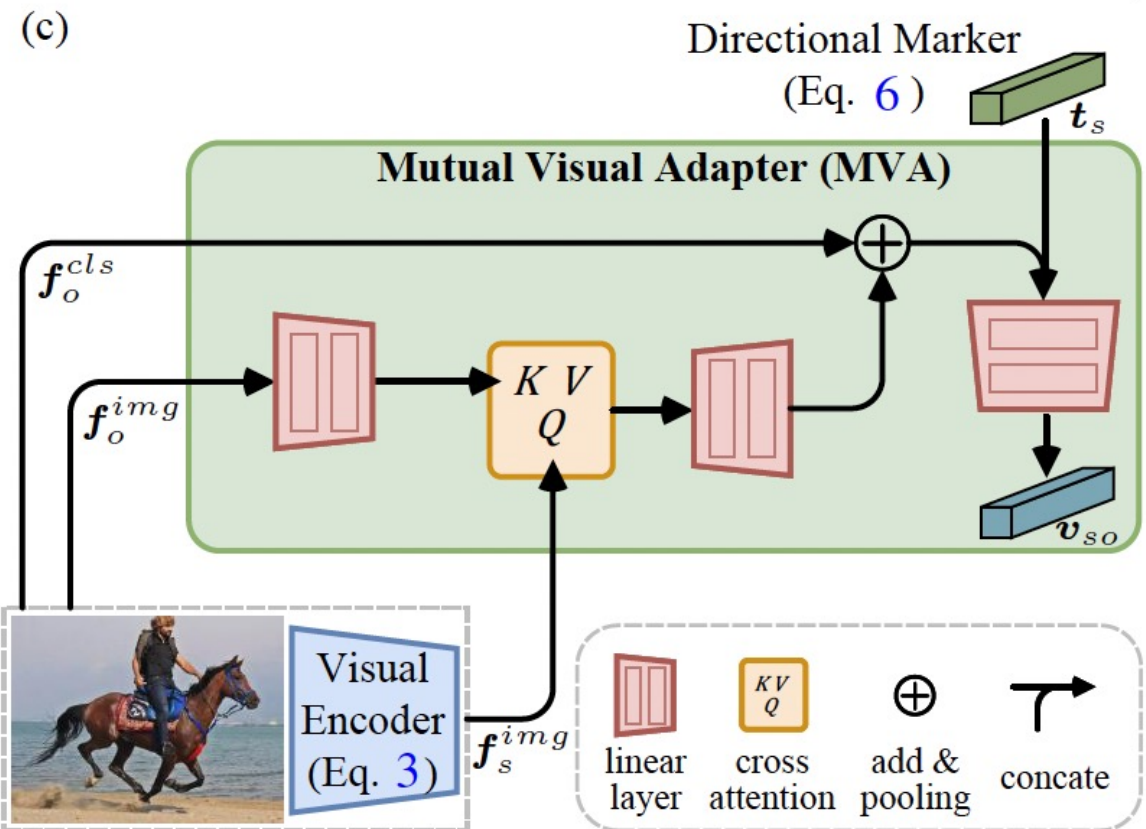
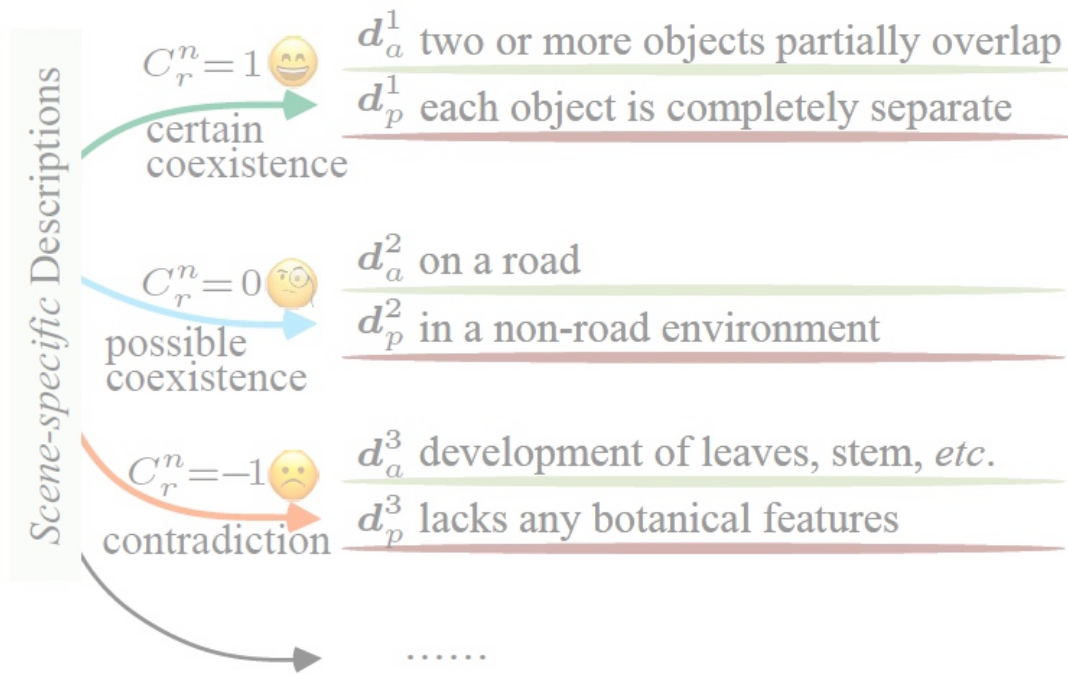


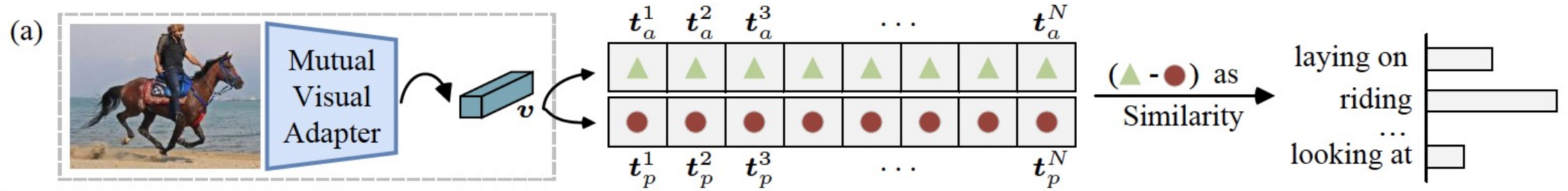
(b) *Self-normalized Similarity Measurement* (Eq. 2)





(b) *Self-normalized Similarity Measurement (Eq. 2)*





(b) *Self-normalized Similarity Measurement (Eq. 2)*

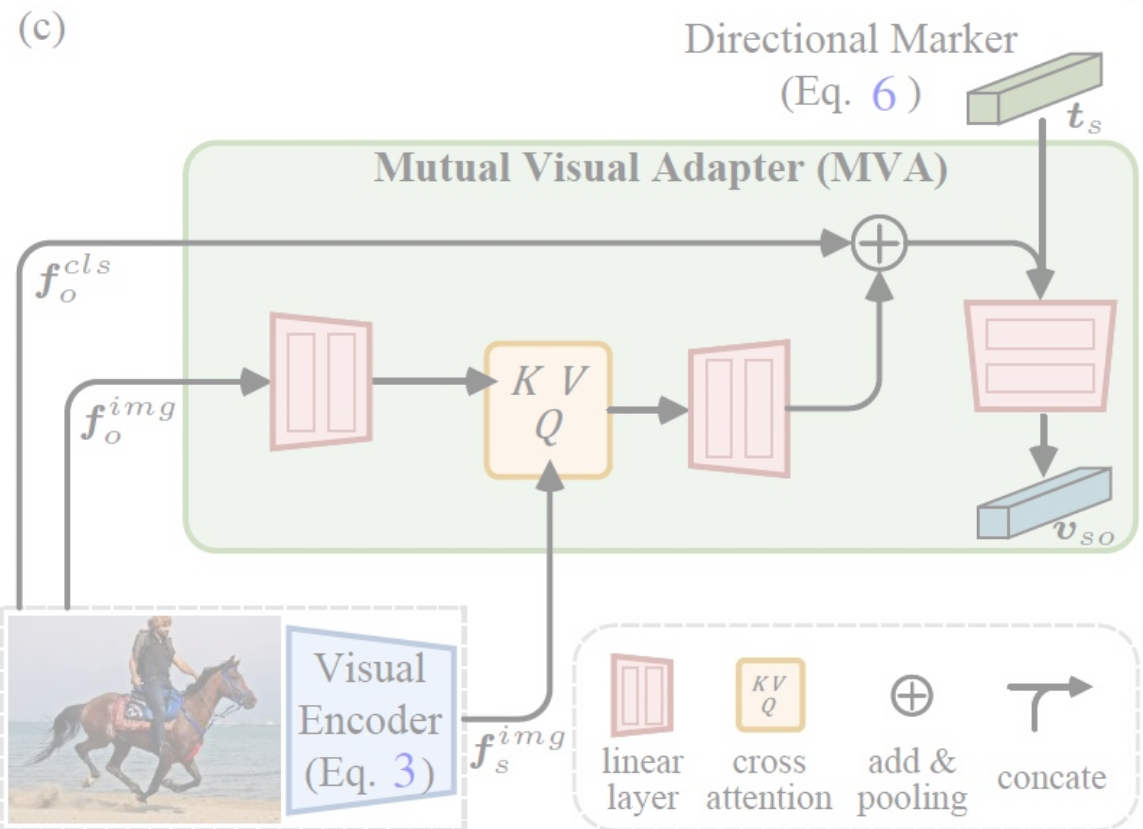
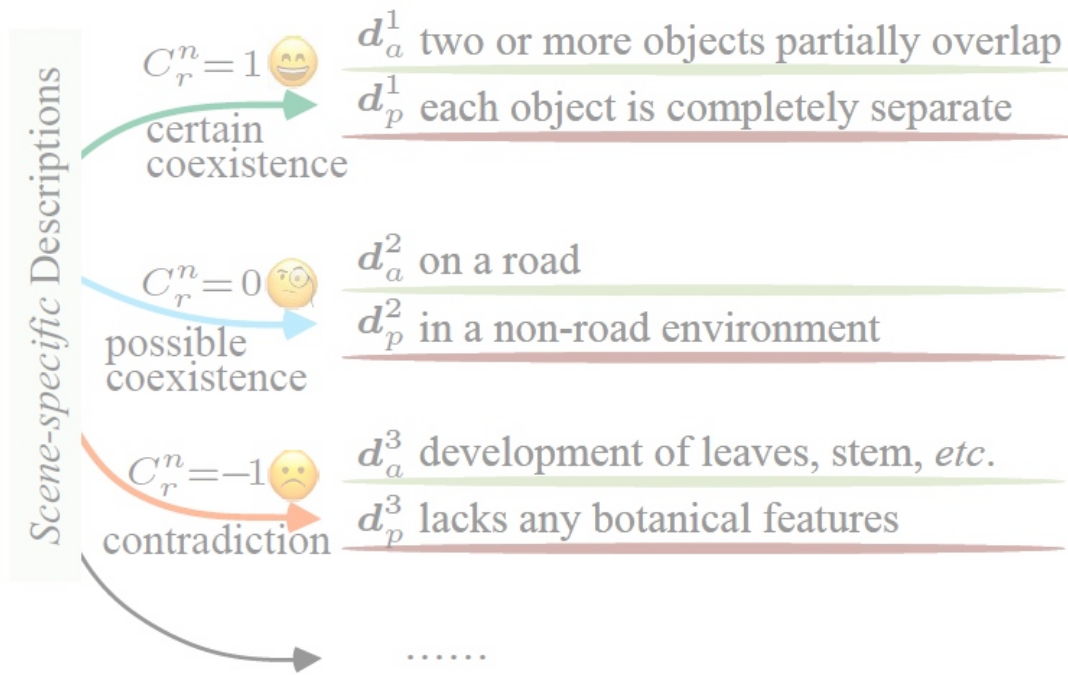
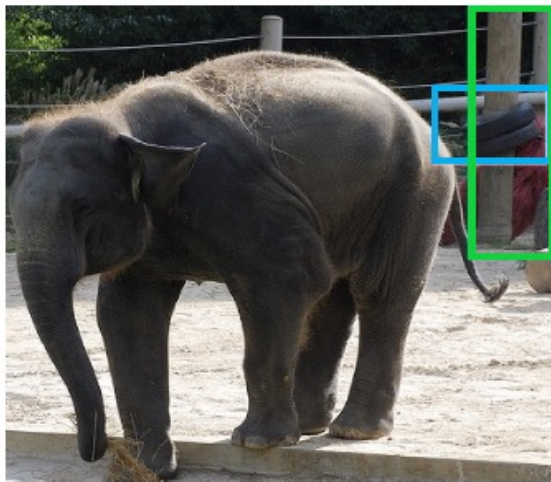


Table 1: Quantitative results (§4.2) on VG [14] base and novel.

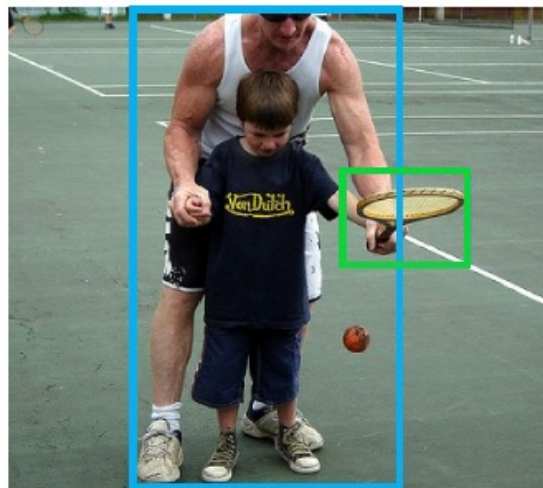
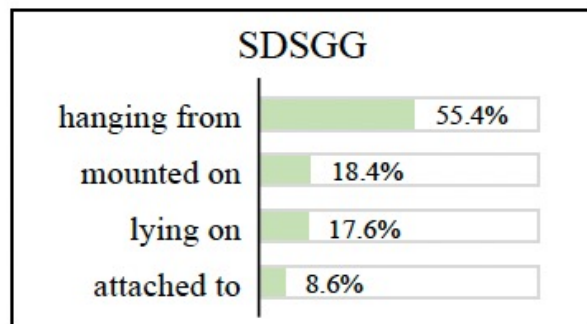
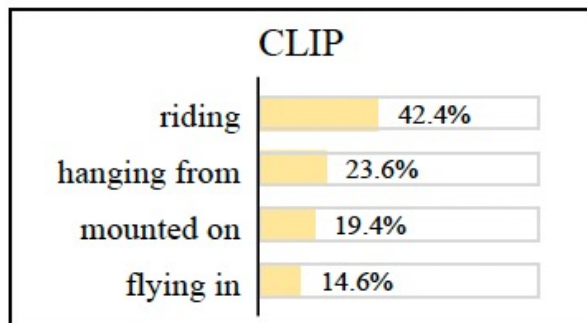
Method	Split	R@20↑	R@50↑	R@100↑	mR@20↑	mR@50↑	mR@100↑
CLS _[ICML21] [4]	base	2.1	3.2	3.9	7.0	9.0	10.9
Epic _[ICCV23] [3]		-	22.6	27.2	-	-	-
Ours		18.7 _{±0.69}	26.5 _{±0.92}	31.6 _{±1.00}	9.2 _{±0.14}	12.4 _{±0.12}	14.8 _{±0.10}
CLS _[ICML21] [4]	novel	13.2	18.1	22.2	11.5	17.9	23.8
Epic _[ICCV23] [3]		-	7.4	9.7	-	-	-
Ours		18.4 _{±0.53}	25.4 _{±0.48}	29.6 _{±0.42}	17.1 _{±0.42}	25.2 _{±0.95}	31.2 _{±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 _{±0.47}	29.3 _{±0.53}	34.9 _{±0.66}	16.8 _{±0.08}	22.7 _{±0.41}	28.4 _{±0.67}



<tire, hanging from, pole>



<man, holding, racket>

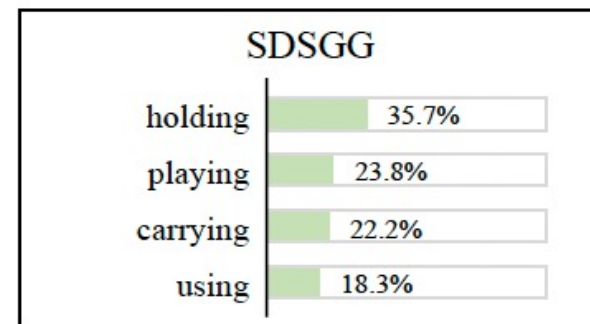
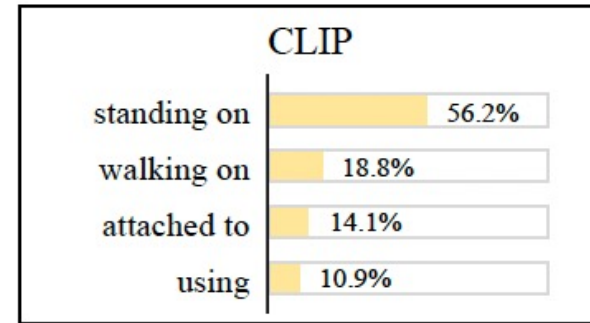


Figure 3: Visual results (§4.3) on VG [14].

Our code will be available at
<https://github.com/guikunchen/SDSGG>

A C C E P T M Y E N D L E S S G R A T I T U D E