Semi-supervised Grounding Alignment for Multi-modal Feature Learning

Shih-Han Chou[1,2], Zicong Fan[3], James J. Little[1], Leonid Sigal[1,2,4]
[1]*Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada*
[2]*Vector Institute for AI, Toronto, Ontario, Canada*
[3]*Department of Computer Science, ETH Zürich, Rämistrasse, Zürich, Switzerland*
[4]*Canada CIFAR AI Chair*
{*shchou75, little, lsigal*}*@cs.ubc.ca; zicong.fan@inf.ethz.ch*

In this supplementary material, we present additional results to complement the main paper.

I. Grounding Alignment

We show examples of the pseudo ground truth of grounding alignment on the Conceptual Captions dataset in Figure 1.

II. Details of the grounding alignment module

Figure 2 summarizes the details of the grounding alignment module.

III. Hyper-parameter Tuning

We show hyper-parameter tuning in Table I. The settings with SPE or at phrase level achieve higher accuracy than the settings at token level which indicates that SPE and phrase level grounding are more important.

IV. Adjusting Learning Rates

The adjustment of learning rates is shown in Table II. We train our model with 0.25x, 0.5x, and 2x of the original setting.

V. Additional Qualitative Results

A. Visual Grounding

We show some qualitative results of the visual grounding task in Figure 3. The orange boxes denote the ground truth annotations. The red boxes are the baseline [1] and the blue boxes are ours. The top two rows are the results where we perform better than the baseline. As shown in the first example, the 'silver truck', our predicted region is more aligned with the ground truth than the baseline. We also show some failure cases in the bottom row. Although our method does not exactly match the ground truth, the predictions are more reasonable than the baseline, we predict the oven but not the table in the first example in the bottom row. We also show some examples in Figure 4 for the model pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 RefCOCO+ dataset.

B. Visual Question Answering

We show some qualitative results of the VQA task in Figure 5. The green color means the prediction is the same as the ground truth and the red color represents the wrong answer. The top row shows the positive examples where our method predicts the correct answers but the baseline doesn't. We also show some failure cases in the bottom row. In the failure case, 'How many vegetables are there?', our model is confused by the beans so that the prediction is '3' but not '2'. Similarly, in the last example, the color of the curtains is too dark to distinguish whether they are blue or black. We also show some examples in Figure 6 for the model pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 VQA dataset.

C. Visual Commonsense Reasoning

We show some qualitative results of the VCR task in Figure 7. The green color denotes the correct answers and the red color denotes the wrong answers. We show the probabilities among the answer choices. The examples show that our method is able to put the right attention on the correct answers. For the last example, although the answer is (b) in the QA $\rightarrow$ R task, we argue that the answer choices are somewhat ambiguous. Answers (a) and (b) are both suitable for this question. We also show some examples in Figure 8 for the model pre-trained on 1/16 Conceptual Captions dataset and fine-tuned on 1/4 VCR dataset.

References

[1] Lu, Jiasen and Batra, Dhruv and Parikh, Devi and Lee, Stefan, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019. 1, 3, 4, 5

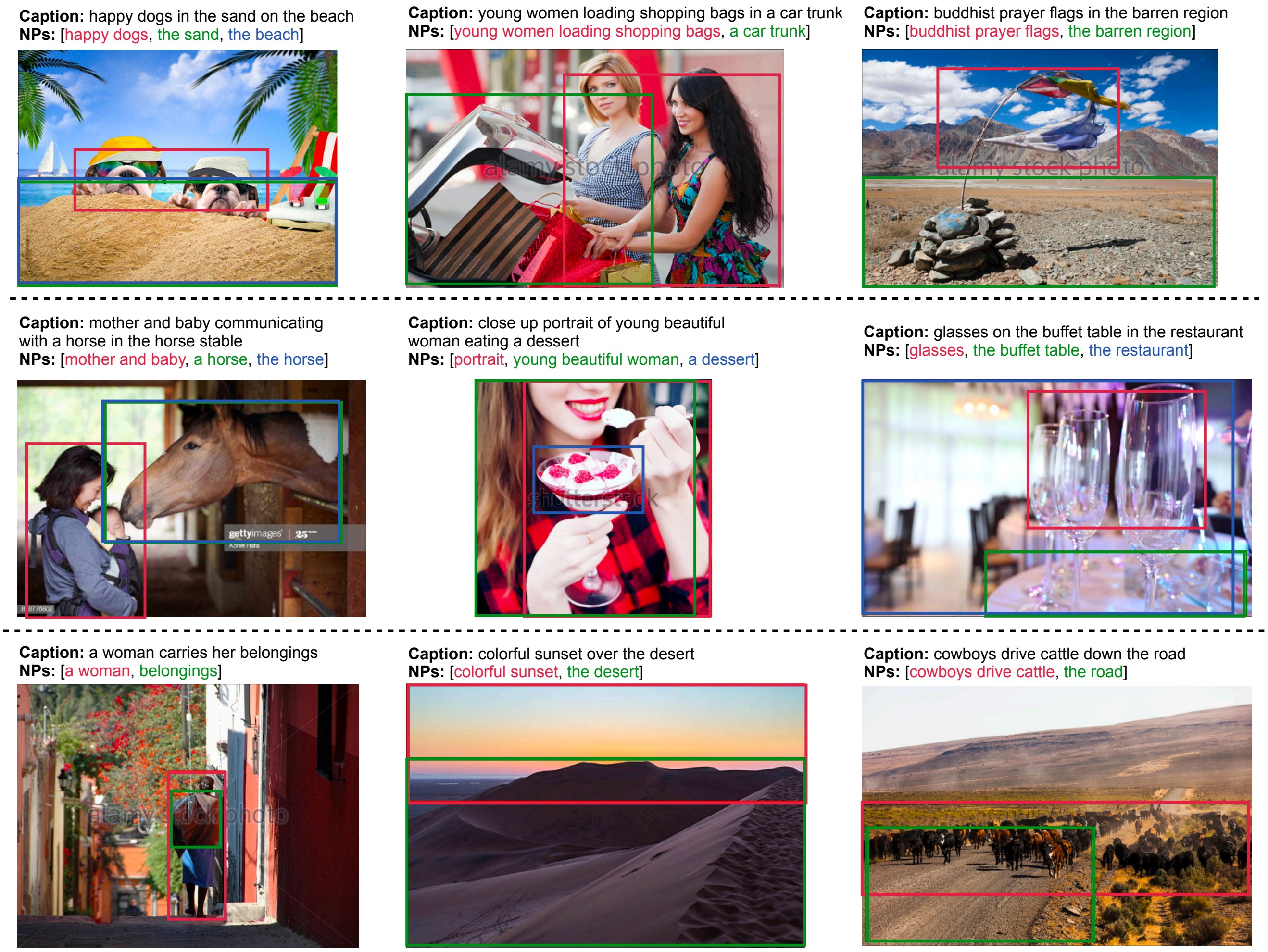


Figure 1: **Pseudo Ground Truth of Grounding Alignment.**

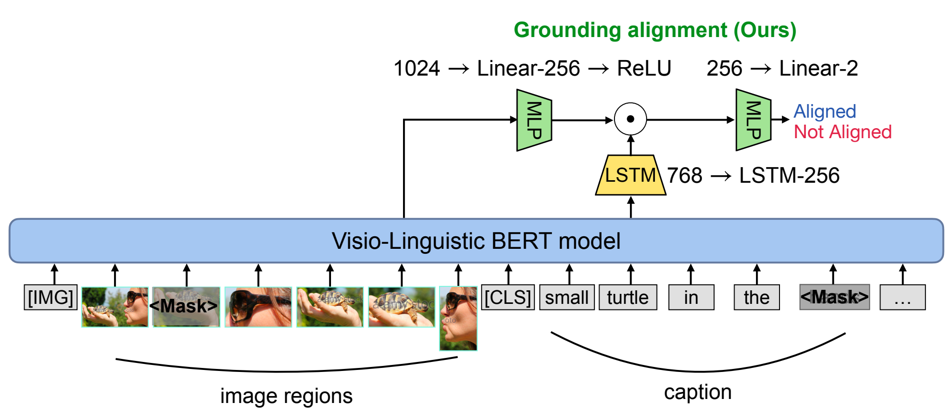


Figure 2: **Detailed architecture of Grounding Alignment.**

Table I: **Hyper-parameter Tuning.** We do hyper-parameter tuning using $1/8$ amount of the Conceptual Caption Dataset. SPE denotes spatial positional encoding and SS Ground represents semi-supervised grounding alignment. Best results are highlighted in bold.

Settings	SPE	λ_{align}	λ_{gnd}	token	phrase	Visual Grounding	VQA	VCR (Q→A)	VCR (QA→R)
ViLBERT [1]	-	1	-	-	-	70.92	67.85	70.83	72.47
+ SS Ground (token)		0	1	✓		-	67.38	71.28	72.73
+ SS Ground (token)		1	1	✓		-	67.79	71.77	73.12
+ SS Ground (token)		1	20	✓		70.98	67.84	**71.91**	73.36
+ SS Ground (phrase)		1	20		✓	-	67.39	71.52	72.55
+ SPE	✓	1	0			71.19	68.12	71.67	73.58
+ SS Ground (token)	✓	1	20	✓		69.02	67.64	71.72	73.06
+ SS Ground (phrase)	✓	1	20		✓	**72.23**	**68.98**	71.88	**73.62**

Table II: **Adjustment of Learning Rates.** We conduct the learning adjustment on the full dataset and fine-tune to visual grounding and VQA tasks. The setting is using our final model which includes SPE and semi-supervised grounding at the phrase level. Best results for each task are bold.

Tasks / Learning Rate	lr=2.5e-05	lr=5e-05	lr=1e-04 (original)	lr=2e-04
Visual Grounding	71.74	71.83	**72.47**	71.45
VQA	68.82	69.19	**69.63**	68.02

Figure 3: **Qualitative Results for visual grounding task pre-trained on full Conceptual Captions dataset and fine-tuned on full RefCOCO+ dataset.** Orange box: Ground truth. Red box: Baseline (ViLBERT [1]). Blue box: Ours. The top two rows are the positive examples and the bottom row are the failure examples.

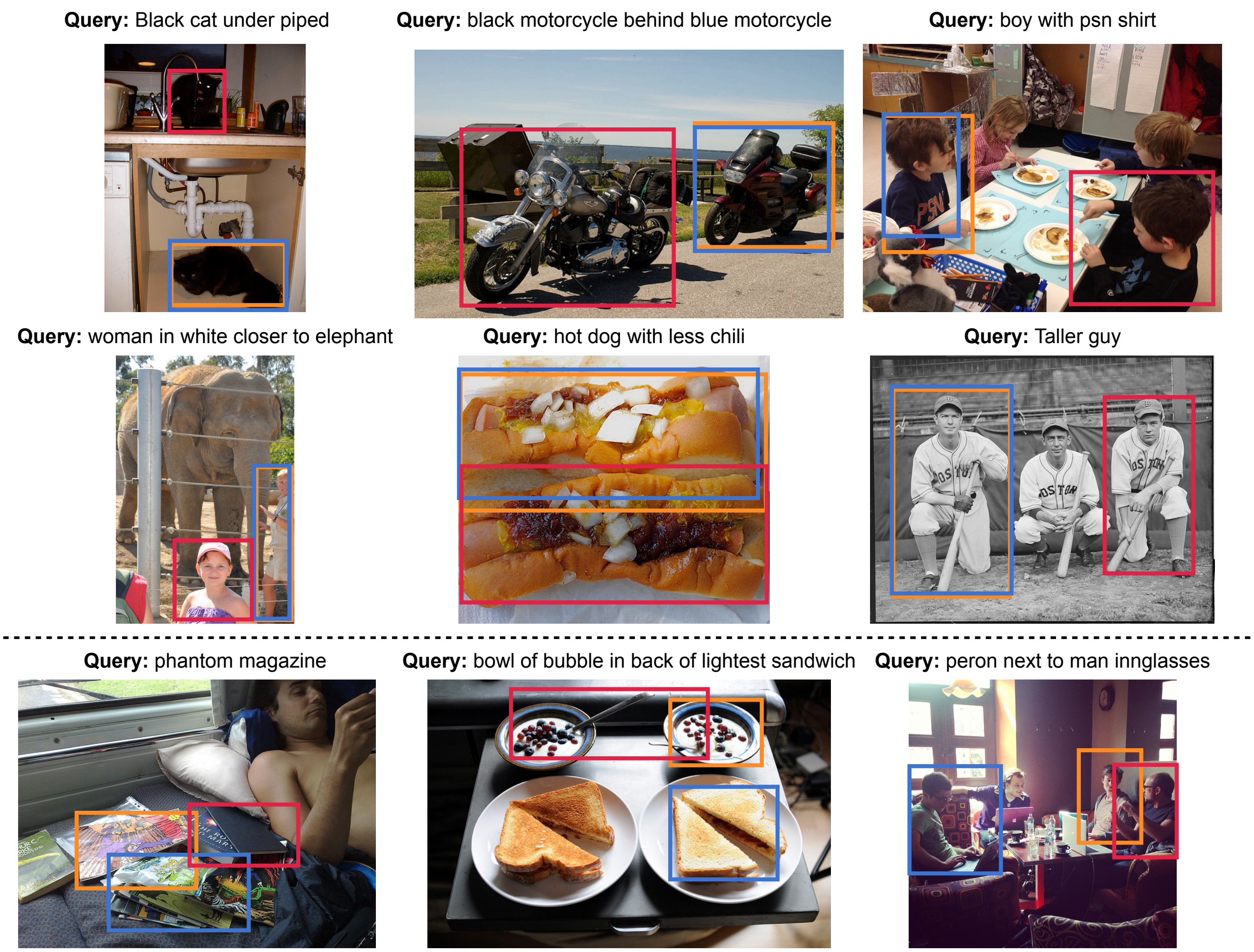


Figure 4: **Qualitative Results for visual grounding task pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 RefCOCO+ dataset.** Orange box: Ground truth. Red box: Baseline (ViLBERT [1]). Blue box: Ours. The top two rows are the positive examples and the bottom row are the failure examples.

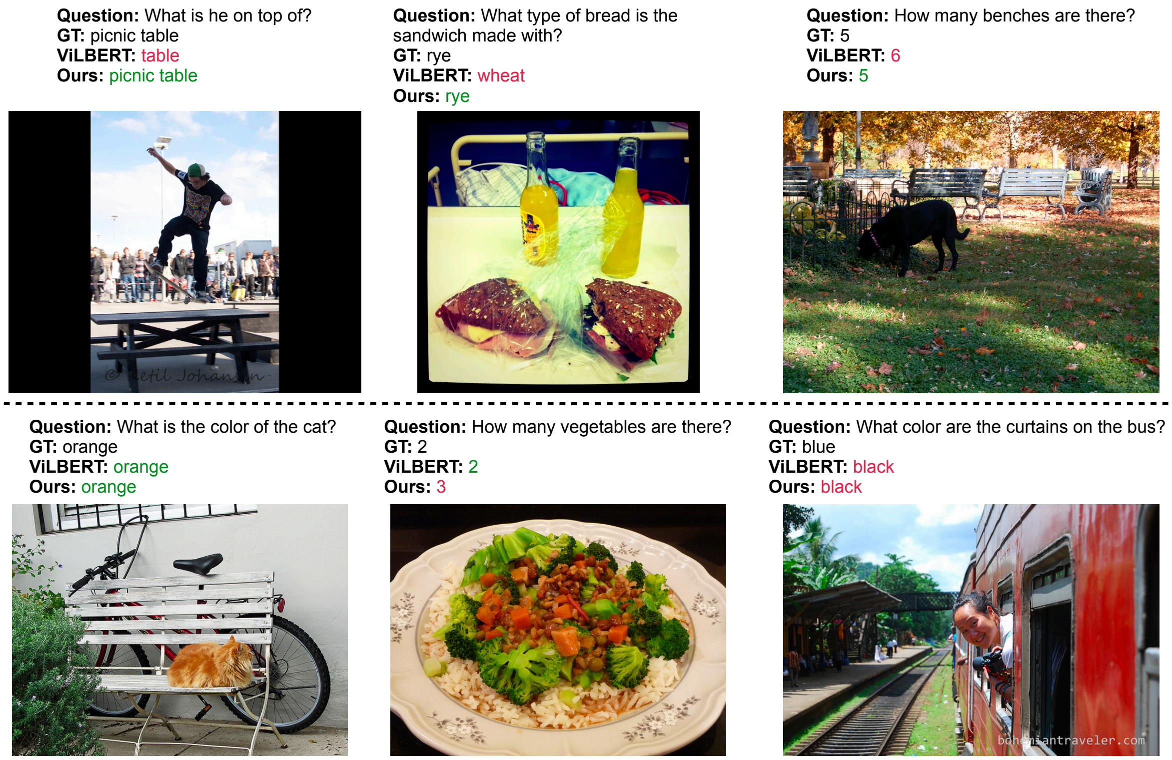


Figure 5: **Qualitative Results for VQA task pre-trained on full Conceptual Captions dataset and fine-tuned on full VQA dataset.** The top row is the positive examples and the two on the right of the bottom row are failure cases.

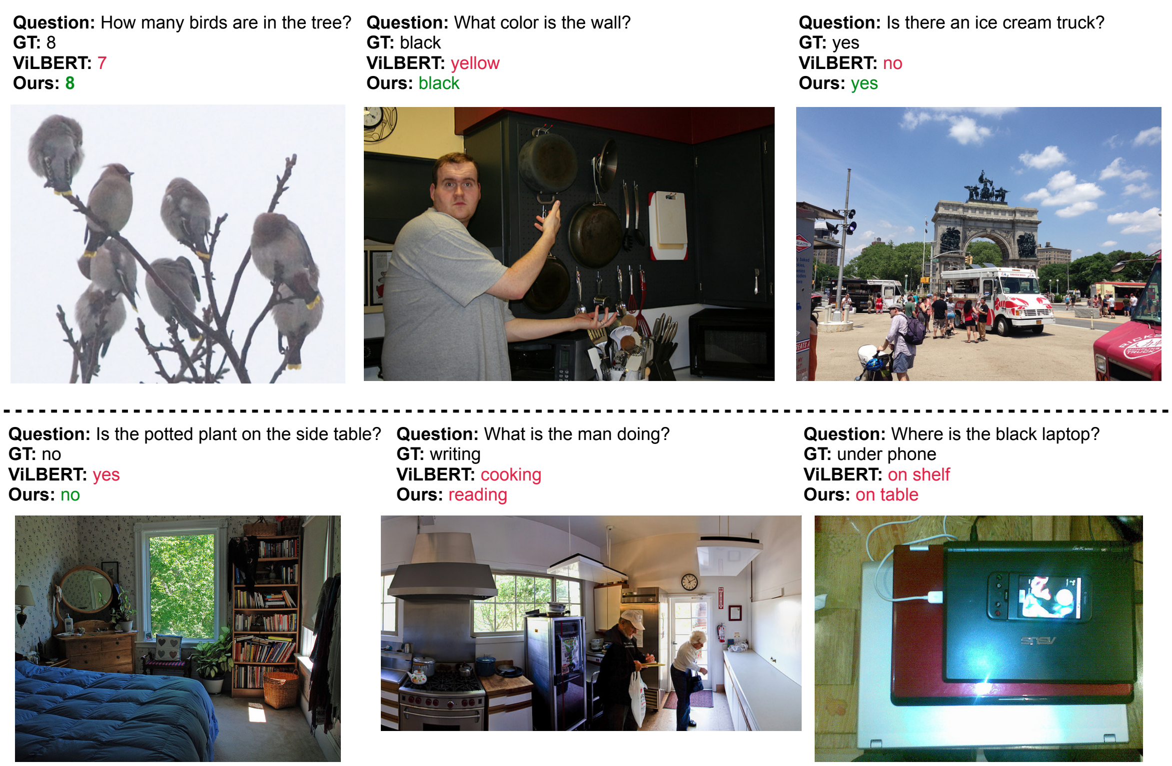
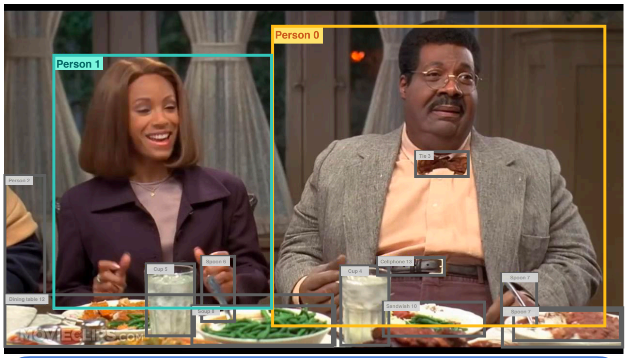


Figure 6: **Qualitative Results for VQA task pre-trained on 1/8 Conceptual Captions dataset and fine-tuned on 1/8 VQA dataset.** The top row is the positive examples and the two on the right of the bottom row are failure cases.

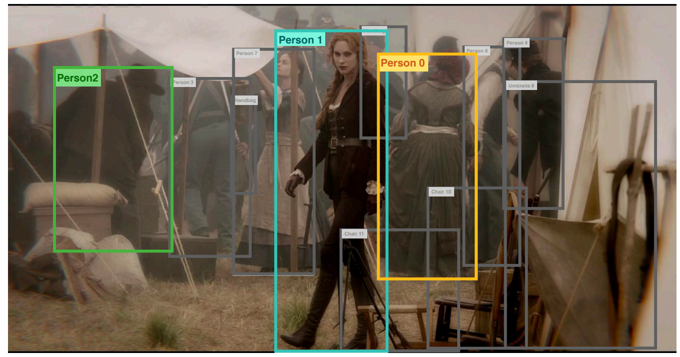


Q → A task

What are [person0, person1], looking at?	Baseline	Ours
(a) They are looking at something at the bottom of the cliff.	~=0	~=0
(b) They are looking at the structure in front of them.	~=0	~=0
(c) They are watching something out of the window of the plane.	~=0	~=0
(d) [person0, person1] are looking at someone who is joking about [person0].	**~=1**	**~=1**

QA → R task

I think so because…	Baseline	Ours
(a) [person1] is smiling and opening her mouth as if about to speak to someone she is looking at.	~=0	~=0
(b) [person1] is amused by what she is seeing while [person0], is embarrassed and uncomfortable.	0.4397	**0.999**
(c) [person2] is wearing a silly outfit and has her hair tied back to avoid getting hair in the food . the table is also looking at her and [person1] is speaking to [person2] like she is giving an order.	~=0	~=0
(d) Looking away from [person0]'s eyes are closed and her mouth is pursed as if she is trying to contain her laughter.	**0.5602**	~=0



Q → A task

Why is [person1] looking so defiant?	Baseline	Ours
(a) [person1] has tried something in this shop she does not like.	~=0	~=0
(b) [person1] may not think that a teenage girl can do what he can.	~=0	~=0
(c) [person1] is just in a stance that shows she is very excited about [person2, person0]'s dance moves .	0.011	~=0
(d) Someone is threatening her.	**0.9889**	**~=1**

QA → R task

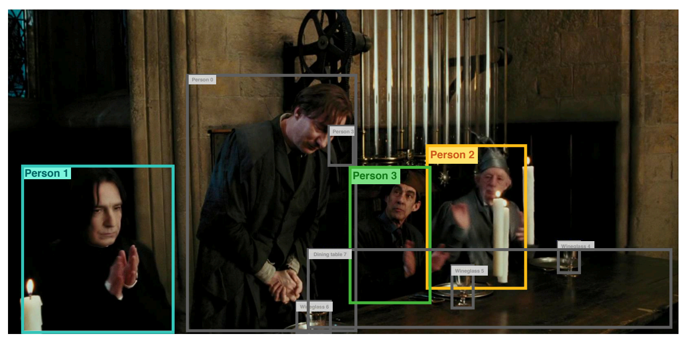
I think so because…	Baseline	Ours
(a) She has a look of fear and apprehension.	0.2665	0.205
(b) Someone could be forcing her to leave and she is looking back at her family.	**0.721**	0.0014
(c) She is gagged and tied up in order not to make any noise.	~=0	~=0
(d) She has a daring expression on her face as if she is saying i dare you to come after me.	0.012	**0.7939**

Q → A task

Why does [person0] keep this safe in a coat closet?	Baseline	Ours
(a) [person0] expects bad weather outside.	**0.945**	0.0001
(b) It is winter outside and [person0] needs to wear warm clothes.	~=0	~=0
(c) This keeps [person0] safe.	0.0001	~=0
(d) So that people will not know where to find it.	0.055	**0.999**

QA → R task

I think so because…	Baseline	Ours
(a) They are in this wooded like they're hiding so they would not want people to discover them.	~=0	~=0
(b) Safes are designed to stay closed if you do not put in the right combination.	~=0	~=0
(c) People ask for the gun to be kicked away so that it can't be immediately picked up again.	~=0	~=0
(d) This is an unusual place to keep a safe, but might help to conceal it.	**0.9999**	**0.9999**



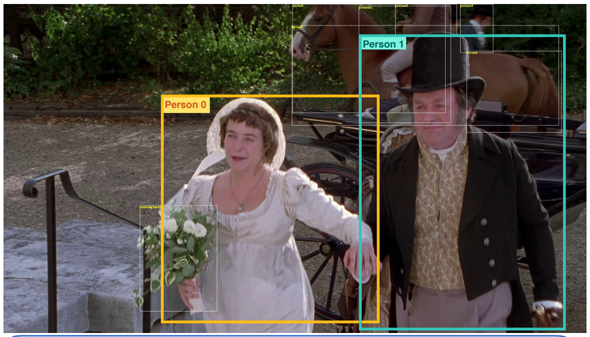
Q → A task

What are [person1, person2], and [person3] doing ?	Baseline	Ours
(a) [person1, person2], and [person3] are having a group meeting.	**0.954**	0.038
(b) [person1, person2], and [person3] are exiting the room with door open.	0.0001	0.001
(c) Clapping their hands.	0.046	**0.96**
(d) [person1, person2], and [person3] are participating in a riot.	~=0	~=0

QA → R task

I think so because…	Baseline	Ours
(a) Clapping one's hands is a common sign of approval or support.	**0.990**	**0.999**
(b) They are holding their hands in front of themselves and moving them back and forth.	0.001	~=0
(c) People clap when they are happy or excited to see something or someone.	0.008	~=0
(d) When you like the beat it is common to clap along.	~=0	~=0

Figure 7: **Qualitative Results for VCR task pre-trained on full Conceptual Captions dataset and fine-tuned on full VCR dataset.** We show the ground truth answers with green color on the multiple choices side. The green/red colors on the probability side mean the correct/incorrect answers. The top row is the positive examples and the one on the right of the bottom row is the failure case.

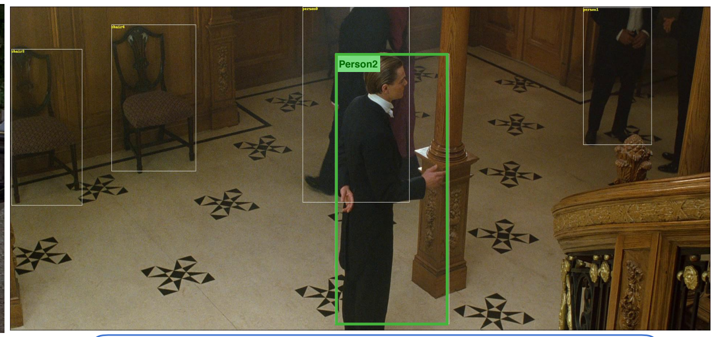


Q → A task

Are [person0] and [person1] happy to get married?	Baseline	Ours
(a) Yes [person0, person1] are in love.	~=0	~=0
(b) No, [person0, person1] are not discussing something happy.	~=0	~=0
(c) No, they are not.	0.999	0.068
(d) Yes, they re both very happy today.	0.0008	0.932

QA → R task

I think so because…	Baseline	Ours
(a) They re facing each other as they take their vows, while dressed in wedding attire.	~=0	~=0
(b) [person0, person1] are dressed formally, [person3] has on a wedding dress and there is draping above them.	~=0	~=0
(c) Both of them raise there arms up and slap hands together in a sign of celebration.	~=0	~=0
(d) They are both smiling and seem delighted.	0.999	0.999

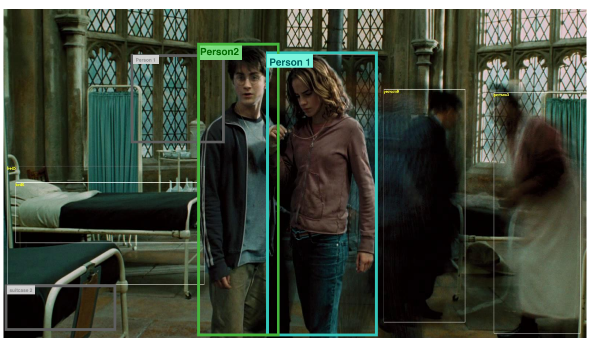


Q → A task

What is [person2] doing next?	Baseline	Ours
(a) [person2] tells [person1] he is sneaking up on the staff to perform an audit.	0.7556	0.00016
(b) He is going to attack [person1].	~=0	~=0
(c) He is going to greet someone.	0.2445	0.9998
(d) [person0] is going to retrieve his hat from the hat rack.	~=0	~=0

QA → R task

I think so because…	Baseline	Ours
(a) He's standing next to the helicopter with an open door and someone is looking at him.	~=0	~=0
(b) [person2] is standing near the door and wearing a store uniform, indicating that he works at the store. store greeters stand near the entrance of their store.	~=0	~=0
(c) He is walking through a door and holding it open.	~=0	~=0
(d) He has his hand extended waiting on someone to arrive.	~=1	0.999



Q → A task

How does [person2] feel about [person1]?	Baseline	Ours
(a) [person2] is romantically interested in [person1].	0.8259	0.0028
(b) [person2] is angry with [person1].	0.0002	0.3915
(c) [person0] is worried and doesn't want [person1] to go, but can not do much else.	~=0	~=0
(d) [person2] trusts [person1] quite a lot.	0.1739	0.6057

QA → R task

I think so because…	Baseline	Ours
(a) He is looking at him keenly.	~=0	~=0
(b) [person2] is making no move to step away from [person1] and he has his eyes closed as he continues to interact with him.	~=0	~=0
(c) [person2] has his arms around [person1].	~=0	~=0
(d) He is standing close to her, and she is sharing something with him.	0.9999	0.9999

Q → A task

What is [person0] looking at?	Baseline	Ours
(a) [person0] is looking at some drawings.	0.00102	0.99964
(b) At something that is just out of view.	~=0	~=0
(c) [person0] is looking down.	0.9981	0.00035
(d) [person1] is looking towards [person0].	~=0	~=0

QA → R task

I think so because…	Baseline	Ours
(a) She is staring intently at something along with everyone else.	~=0	~=0
(b) She seems to have noticed'something but because of [couch5] we can not see it.	~=0	0.00012
(c) She is turned around looking over her shoulder.	0.99993	0.99988
(d) She is looking straight ahead.	~=0	~=0

Figure 8: **Qualitative Results for VCR task pre-trained on 1/16 Conceptual Captions dataset and fine-tuned on 1/4 VCR dataset.** We show the ground truth answers with green color on the multiple choices side. The green/red colors on the probability side mean the correct/incorrect answers. The top row is the positive examples and the one on the right of the bottom row is the failure case.