Embodied Image Captioning: Self-supervised Learning Agents for Spatially Coherent Image Descriptions Supplementary Material

Tommaso Galliena[1,2], Tommaso Apicella[1], Stefano Rosa[1], Pietro Morerio[1], Alessio del Bue[1], Lorenzo Natale[1]
[1]Istituto Italiano di Tecnologia, Genoa, Italy
[2]University of Genoa, Genoa, Italy
name.surname@iit.it

A. Experimental setup

Datasets. Gibson reflects the complexity of 572 real-world environments featuring object categories such as tables, sofas, plants, in different settings (rooms and object instances). HM3D comprises 1,000 reconstructed indoor residential and commercial environments having higher quality than Gibson [4].

Methods under comparison. We compare three exploration policies: *random goals*, *frontier exploration* and the learned policy (*CLA*). The first two policies use the same path planner that first converts the explored map into a visibility graph (nodes represent feasible goals and edges the straight trajectory without obstacles) via skeletonization, then computes the optimal trajectory between the current position and the goal using graph search, considering nodes in the trajectory as sub-goals. Moreover, we delete repeated samples caused by the agent revisiting certain poses during exploration, to train and test on different images. Training images are augmented with 50% probability using horizontal flip, gaussian noise and affine transformations.

Captioners fine-tuning setup. We report the training setup used to fine-tune CoCa [5] and BLIP2 [2] in Table 1. Note that we use LoRA [1] to fine-tune BLIP2 due to the large amount of parameters to reduce instability during the fine-tuning processing and reduce the amount of time needed for the fine-tuning processing.

B. Ablation studies

Pseudo-caption ablation. Results in Tab. 2 show that compared to LLaMa, Mistral achieves -1.85 p.p. in SP and -0.42 p.p. in CI while Qwen achieves -1.41 p.p. in SP and -0.13 p.p. in CI using BLIP2. Also using CoCa, LLaMa outperforms other LLMs. Varying the LLMs, our prompt with caption frequency outperforms ECO and IC3 (Tab.1 main paper).

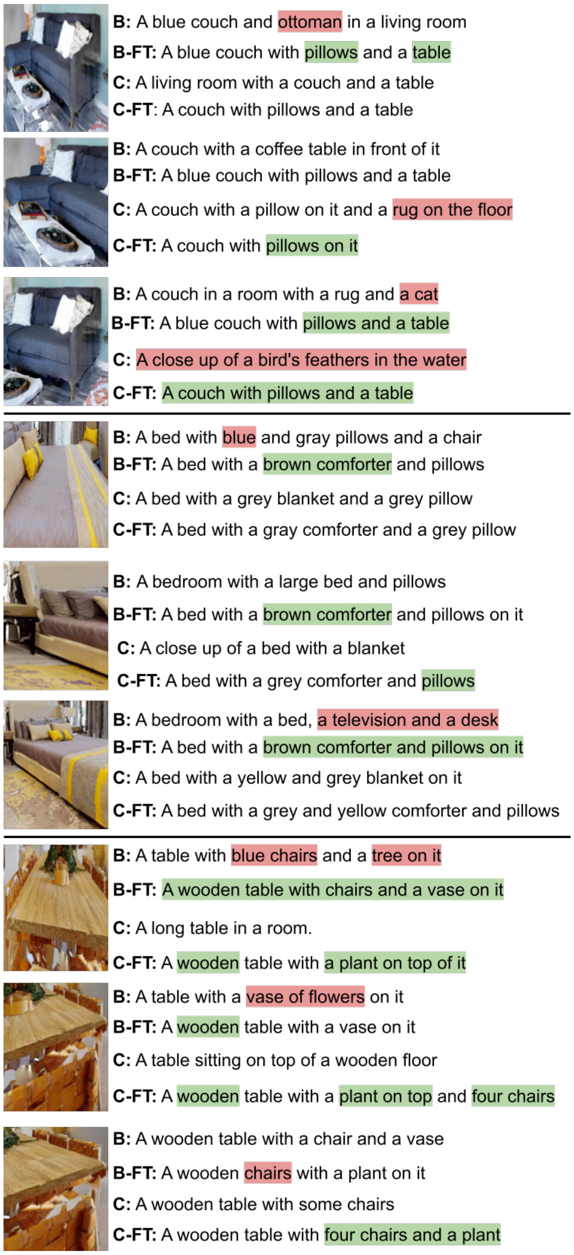
Hyperparameter	CoCa	BLIP2
Learning Rate	0.0005	0.0001
lr scheduler	cosine	-
Batch Size	64	64
Num Workers	4	4
Optimizer	AdamW	AdamW
Weight Decay	0.001	0.001
Epochs	10	10
Patience	3	3
Rotation	[-10.0, 10.0]	[-10.0, 10.0]
Shear	[-10.0, 10.0]	[-10.0, 10.0]
Gaussian noise	0 ± 0.5	0 ± 0.5
Contrastive-loss	0	-
LoRA rank	-	8
LoRA alpha	-	16
LoRA droput	-	0.3
LoRA bias	-	None

Table 1. Finetuning hyperparameters

Captioner	**LLM**	B_4	M	R_L	CI	SP	CS
CoCa	Mistral-7B	17.21	22.07	44.70	1.13	29.29	71.63
	Qwen3-8B	19.95	23.95	49.60	1.31	33.59	73.37
BLIP-2	Mistral-7B	16.46	24.05	48.02	1.20	33.73	72.11
	Qwen3-8B	19.37	24.62	51.43	1.49	34.17	73.02

Table 2. LD-CPS ablation on HM3D with CLA.

Ablation study on triplet loss. Table 3 shows the effect of the λ_{tr} coefficient on the captioners fine-tuning. We experiment $\lambda_{tr} = \{1, 0.5, 0.1\}$. Results show that a choice of $\lambda_{tr} = 0.1$ achieves higher performance values compared to standard fien-tuning and other triplet weights. With $\lambda_{tr} = 1$ the fine-tuned model has worse performance values compared to other triplet loss weights and standard fine-tuned model with all the exploration policies apart from CLA.



B: A blue couch and ottoman in a living room
B-FT: A blue couch with pillows and a table
C: A living room with a couch and a table
C-FT: A couch with pillows and a table

B: A couch with a coffee table in front of it
B-FT: A blue couch with pillows and a table
C: A couch with a pillow on it and a rug on the floor
C-FT: A couch with pillows on it

B: A couch in a room with a rug and a cat
B-FT: A blue couch with pillows and a table
C: A close up of a bird's feathers in the water
C-FT: A couch with pillows and a table

B: A bed with blue and gray pillows and a chair
B-FT: A bed with a brown comforter and pillows
C: A bed with a grey blanket and a grey pillow
C-FT: A bed with a gray comforter and a grey pillow

B: A bedroom with a large bed and pillows
B-FT: A bed with a brown comforter and pillows on it
C: A close up of a bed with a blanket
C-FT: A bed with a grey comforter and pillows

B: A bedroom with a bed, a television and a desk
B-FT: A bed with a brown comforter and pillows on it
C: A bed with a yellow and grey blanket on it
C-FT: A bed with a grey and yellow comforter and pillows

B: A table with blue chairs and a tree on it
B-FT: A wooden table with chairs and a vase on it
C: A long table in a room.
C-FT: A wooden table with a plant on top of it

B: A table with a vase of flowers on it
B-FT: A wooden table with a vase on it
C: A table sitting on top of a wooden floor
C-FT: A wooden table with a plant on top and four chairs

B: A wooden table with a chair and a vase
B-FT: A wooden chairs with a plant on it
C: A wooden table with some chairs
C-FT: A wooden table with four chairs and a plant

Figure 1. Examples of predicted captions before and after fine-tuning for different object views. We highlight mistakes and correct details in generated captions. KEYS – C: off-the-shelf CoCa [5], C-FT: fine-tuned (triplet loss) CoCa, B: BLIP2 [2], B-FT: fine-tuned (triplet loss) BLIP2.

C. Inference time

On an NVIDIA Tesla V100 16GB, the time for an exploration step is 0.266 s for CLA equipped with CoCa, 0.271 s equipped with BLIP2 (averaged over 300 steps); of these timings, the captioning inference time is 6.65 ms for CoCa and 10.2 ms for BLIP2 (averaged over 754 bounding boxes); and the pseudo-captioning time per object instance is 1.02 s (averaged over all 160 testing objects).

	Captioner	**Pseudo-caption**	**Policy**	**Fine-tuning**	B_4	M	R_L	CI	SP	CS
Gibson	CoCa	-	-	NoFT	12.15	21.47	42.83	0.63	28.30	63.30
		Ours	Random	Standard	19.55	23.36	51.10	1.01	33.21	70.11
				$\lambda_{tr} = 1.0$	18.94	23.73	51.95	0.97	34.11	71.18
				$\lambda_{tr} = 0.5$	19.49	24.52	52.28	1.03	35.56	73.87
				$\lambda_{tr} = 0.1$	19.85	24.83	53.10	1.12	35.58	73.26
			Frontier	Standard	17.93	24.92	51.00	1.08	35.65	73.20
				$\lambda_{tr} = 1.0$	16.53	23.58	48.97	0.89	30.22	69.28
				$\lambda_{tr} = 0.5$	16.93	24.72	50.46	0.95	33.20	72.87
				$\lambda_{tr} = 0.1$	17.17	24.82	51.43	0.94	36.10	73.44
			CLA	Standard	13.70	20.50	45.43	0.72	26.48	61.76
				$\lambda_{tr} = 1.0$	14.19	21.12	45.97	0.72	27.00	60.89
				$\lambda_{tr} = 0.5$	17.72	24.36	50.71	0.99	35.53	70.86
				$\lambda_{tr} = 0.1$	14.96	20.79	45.50	0.71	25.30	63.21
HM3D	CoCa	-	-	NoFT	09.93	17.36	38.91	0.44	22.19	62.08
		Ours	Random	Standard	17.34	21.68	47.46	0.82	26.90	72.67
				$\lambda_{tr} = 1.0$	16.61	19.46	44.83	0.74	25.84	70.33
				$\lambda_{tr} = 0.5$	16.51	20.92	46.29	0.81	27.41	72.53
				$\lambda_{tr} = 0.1$	16.57	21.39	46.73	0.83	29.36	72.77
			Frontier	Standard	19.04	21.73	47.12	0.87	28.64	72.47
				$\lambda_{tr} = 1.0$	18.51	21.47	46.23	0.83	27.42	70.28
				$\lambda_{tr} = 0.5$	16.80	21.42	45.44	0.74	27.43	70.32
				$\lambda_{tr} = 0.1$	19.11	22.05	49.02	0.95	30.21	73.99
			CLA	Standard	16.35	20.62	46.19	0.80	28.38	70.19
				$\lambda_{tr} = 1.0$	17.01	21.33	47.01	8.75	28.71	71.02
				$\lambda_{tr} = 0.5$	17.31	21.55	47.19	0.79	28.74	71.74
				$\lambda_{tr} = 0.1$	17.57	20.73	47.81	0.81	29.36	72.23

Table 3. Comparison of captioning performance on Gibson and HM3D testing sets, varying fine-tuning method and policy. KEYS – B_4: BLEU, M: METEOR, R_L: ROUGE-L, CI: CIDER, SP: SPICE, CS: cosine similarity between SBERT embedding of prediction and annotation.

D. Qualitative comparison of captioners prediction

In Fig. 1 we present some qualitative examples of predicted captions before and after fine-tuning (with triplet loss) for both CoCa and BLIP2. Off-the-shelf models predict captions containing wrong attributes (colors) or object categories not present in the images, e.g., cat, bird's feathers, television. On the contrary, the fine-tuned models show higher accuracy in describing details of objects and generate predictions that are more consistent across different views of the same object, showing that learning the unique pseudo-caption for each object instance combined with the triplet loss enhances captions accuracy and consistency.

E. Pseudo-captioning and ChatGPT prompts

Listing 1 shows the text prompt used to generate results with ChatGPT [3]. We use ChatGPT to generate captions from the objects crop.

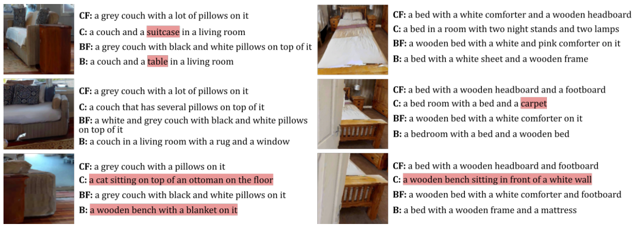
Listing 2 shows the proposed text prompt for the pseudo-caption generation, where `str(captions_freq_list)` is a list of captions for different views of the same object with their frequency. Our approach combines the frequency of the captions with

Listing 1. The prompt for off-the-shelf ChatGPT.

```
Please provide an image caption for the provided picture. Do not
use more than 5−7 words to describe the object in the image. Use
simple words and don’t use more than one adjective per noun.
Some examples: ’A red couch with pillows on it’,’A television
set on top of a table’,’A grey patterned armchair’
```

Env	**Finetuning**	B_4	M	R_L	CI	SP	CS
CoCa	NoFT	16.01	22.19	46.37	0.88	31.07	68.71
	Triplet	**17.95**	**24.10**	**48.20**	**1.05**	**32.98**	**74.03**
BLIP-2	NoFT	20.85	23.18	50.39	1.32	36.48	73.92
	Triplet	**22.70**	**25.29**	**53.12**	**1.38**	**39.90**	**77.58**

Table 4. Comparison of captioning performance on occluded objects. KEYS – B_4: BLEU, M: METEOR, R_L: ROUGE-L, CI: CIDER, SP: SPICE, CS: cosine similarity between SBERT embedding of predicted and annotated captions.



CF: a grey couch with a lot of pillows on it
C: a couch and a suitcase in a living room
BF: a grey couch with black and white pillows on top of it
B: a couch and a table in a living room

CF: a grey couch with a lot of pillows on it
C: a couch that has several pillows on top of it
BF: a white and grey couch with black and white pillows on top of it
B: a couch in a living room with a rug and a window

CF: a grey couch with a pillows on it
C: a cat sitting on top of an ottoman on the floor
BF: a grey couch with black and white pillows on it
B: a wooden bench with a blanket on it

CF: a bed with a white comforter and a wooden headboard
C: a bed in a room with two night stands and two lamps
BF: a wooden bed with a white and pink comforter on it
B: a bed with a white sheet and a wooden frame

CF: a bed with a wooden headboard and a footboard
C: a bed room with a bed and a carpet
BF: a wooden bed with a white comforter on it
B: a bedroom with a bed and a wooden bed

CF: a bed with a wooden headboard and footboard
C: a wooden bench sitting in front of a white wall
BF: a wooden bed with a white comforter and footboard
B: a bed with a wooden frame and a mattress

Figure 2. Captioning with different levels of occlusion. **C:** CoCa; **B:** BLIP2; **F:** fine-tuned (triplet loss)

in-context learning (provided examples of how to solve the task), to account for wrong or inconsistent captions.

F. Performance on occluded objects

Tab. 4 reports a comparison between off-the-shelf and fine-tuned captioners (using triplet loss) on a manually collected dataset of 50 images (25 from Gibson and 25 from HM3D) of objects, with 5 images per dataset for each category in the main paper, under varying levels of occlusion based on visual inspection. The performance is evaluated using standard captioning metrics (BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE and cosine similarity), and results are averaged across Gibson and HM3D. The results show that our fine-tuned models outperform off-the-shelf captioners for both CoCa and BLIP-2, indicating improved robustness to partial views and occlusions. Qualitative examples in Fig. 2 illustrate the increased consistency and accuracy of the generated captions in challenging visual conditions.

References

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large

Listing 2. The prompt for our proposed method.

```
You are an advanced language model tasked with generating a
concise and accurate caption for an object. You are given a list of
captions along with their frequencies. Each caption may
represent a different viewpoint and might not always be accurate.
Additionally, you are provided with the correct object class to
describe. Your goal is to generate a single, coherent caption that
accurately describes the main object, based on the provided
information. The generated caption should not exceed 20 words
and must be encapsulated within <Caption> ... </Caption>
tags.
Here is the format of the input you will receive:
[[frequency, ”caption”]]

Example Input:
[[5, ”A red apple on a table”], [3, ”A shiny red apple”], [1, ”A red
    fruit”], [2, ”A red apple”]]
Example Output:
<Caption>A shiny red apple on a table</Caption>
Example Input:
[[8, ”A small brown dog”], [3, ”A dog”], [4, ”A small dog”], [1, ”
    A brown animal”]]
Example Output:
<Caption>A small brown dog</Caption>
Example Input:
[[6, ”A blue car parked on the street”], [4, ”A car”], [2, ”A blue
    vehicle”], [1, ”A car on the street”]]
Example Output:
<Caption>A blue car parked on the street</Caption>
Example Input:
[[7, ”A cat sitting on a windowsill”], [5, ”A windowsill cat”], [2,
    ”A cat”], [1, ”A windowsill”]]
Example Output:
<Caption>A cat sitting on a windowsill</Caption>
Example Input:
[[5, ”A wooden table with a plate on it”], [2, ”A table with a plate
    and a couch in the room”],
[3, ”A wooden table”], [1, ”A plate on a wooden table”]]
Example Output:
<Caption>A wooden table with a plate on it</Caption>

Your Task:
1. Analyze the provided list of captions and their frequencies.
2. Synthesize an accurate caption that reflects the most reliable
    and frequent details.
3. Ensure the generated caption describes only the main objects
    and mentions other objects only in relation to the main
    object.
4. Ensure the generated caption is no longer than 20 words.
5. Encapsulate your generated caption within <Caption> ... </
    Caption> tags.

Input:
{str(captions_freq_list)}

Output:
```

language models. In *International Conference on Learning Representations*, 2022. 1

[2] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 1, 2

[3] OpenAI. Chatgpt conversation on citation formats. `https://chat.openai.com`, 2025. 2

[4] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. 1

[5] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv:2205.01917v2 [cs.CV]*, 2022. 1, 2