Segment and Caption Anything

Supplementary Material

A. Implementation Details

Tab. 1 presents the implementation details of our method. Note that as we scale the batch size, the learning rate is scaled linearly as well. However, during our experiments, we found that there is a maximum threshold of 4e-4. Scaling the learning rate over 4e-4 leads to underfitting of the data and degradation of performance. Since we only optimize 19.4M parameters, it costs less computation, less memory usage, and less communication bandwidth, resulting in both fast and scalable training. We attribute 200K steps of training. For only VG [11] dataset, we train the models for the full 200K steps. Otherwise, we first pre-train the models for 100K then finetune them on VG for another 100K steps. We use 64 V100 GPUs to pre-train and 32 V100 GPUs to finetune.

Table 1. The implementation details. $*$: As the batch size ramps up, the image epoch and region epochs are subjected to be changed, and the learning rate will be scaled linearly w.r.t. the batch size.

Optimization	
Optimizer	AdamW
	(0.9, 0.999)
LR *	0.0001
LR Decay Ratio	0
LR Decay	cosine
Weight Decay	0.0001
Warmup ratio	0.3333
Warmup steps	200
Gradient Clipping	1.0
Data Epoch*	
Batch Size*	8
# Reg / Img	16
Steps	200000
# Img	77398
# Reg	3684063
Img Epoch*	20.67
Reg Epoch*	6.95
GPU Type	V100-16GB
# GPUs	8
Model Details	
Input	a) 1024x1024
	Long side: 1024
	Short side: padding
	b) Large Scale Jitter
	c) Horizontal Flip
Loss	a) Cross Entropy Loss
	b) Label Smooth (0.1)
Text Decoder	a) GPT2-large
	b) Open LLAMA 3B v2
# Query Tokens	8
# Mixer Layers	12
# Task Tokens	6
Opt. Module	Text Feat. Mixer
# Opt. Params	19.4 M

B. Leveraging Other Image Features

We experiment with image features from other encoders [3–5, 7, 16, 18]. The training configuration is the same as that of the ablations, which is 8 V100 GPUs and direct decoding when inference. We use the features from the second last layer [14]. We also try to optimize the feature mixer of SAM [9] for the seek of improved performance. The results can be found in Tab. 2. The models with other image encoders perform drastically worse than those with SAM image encoders, indicating the superiority of the feature space of SAM. Please note that the image encoders are fixed, other methods like [22, 23, 26] need to fine-tune their image encoders, which increases the computation burden. While ours only fine-tune the text feature mixer. It not only achieves better performance but is cheaper for training at scale.

C. The Results of Referring VLLM

The building of referring Vision Large Language Models (VLLMs) evolves quickly [22, 26]. Here we compare our models with these referring VLLMs in Tab. 3.

D. Dataset Statistics

Tab. 4 includes the statistics of the datasets used for training.

E. Evaluation of Referring Expression Generation

Referring Expression Generation (REG) [2, 24] is closely related to regional image captioning. Regional image captioning is about depicting the regions informatively. The goal of REG is to output descriptions that discriminate the *unique* object of interest, which does not require faithfully regional descriptions. Fig. 1 illustrates the difference with two examples [2]. Despite of the textual style gaps between the two tasks, we present the zero-shot results of REG with our trained models in Tab. 5.

Table 2. Comparison of using different image encoders. "C": CIDEr-D, "M": Meteror.

Image Encoder	C	M
vit_large_patch14_clip_336.openai	67.3	10.2
vit_large_patch14_clip_224.datacompxl	59.0	9.3
eva02_large_patch14_clip_336.merged2b	53.9	8.8
vit_large_patch14_reg4_dinov2.lvd142m	76.4	11.2
vit_large_patch16_224.mae	59.6	9.4
Add optimization of sam feature mixer		
vit_large_patch14_clip_336.openai	66.7	10.1
vit_large_patch14_clip_224.datacompxl	60.3	9.5
eva02_large_patch14_clip_336.merged2b	54.2	8.8
vit_large_patch14_reg4_dinov2.lvd142m	76.1	11.1
vit_large_patch16_224.mae	59.2	9.4
SAM		
SAM-ViT-base	130.2	16.0
SAM-ViT-large	129.6	15.9
SAM-ViT-huge	130.9	16.0

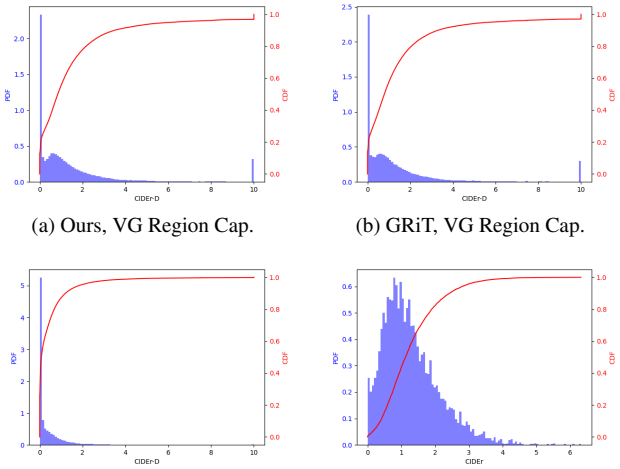
Table 3. Comparison with referring Vision Large Language Models (VLLMs). "M": Meteror, "C": CIDEr-D. †: The scores are from the papers. ‡: We reproduced the result with "GPT4RoI-7B-delta-V0" from https://github.com/jshilong/GPT4RoI. The best, the second best, the third best scores are marked as red, orange, yellow, respectively.

Method	M	C
ASM [22] (Zero-shot)†	12.6	44.2
ASM (Finetuned)†	18.0	145.1
GPT4RoI [26] (7B)†	17.4	145.2
GPT4RoI (13B)†	17.6	146.8
GPT4RoI (7B)‡	16.4	122.3
SCA (GPT2-large, VG)	17.4	148.8
SCA (LLAMA-3B, VG)	17.4	149.8
SCA (GPT2-large, Pretrain+VG)	17.5	149.8

F. Compared with Image Captioning: The Distribution of Automatic Evaluation Metrics and the Pity of the Metrics

We notice that the convention metrics based on n-gram hold a positive skewness distribution. Although some predictions perfectly match the ground truths, the overall distribution is still long-tailed. We plot the distributions of CIDER-D scores for different methods in Figs. 2a to 2c. For ours and GRiT, the distributions are similar. Whereas more scores are allocated around zero in the SAM-Captioner baseline, leading to poor a average CIDER-D score. We additionally showcase the distribution of CIDER-D on the image caption dataset COCO in Fig. 2d, which is predicted by SOTA image captioner [19] , the distribution is still skewness but is more centered compared with that of region captioning. For the majority of captions, their CIDER-D score is zero, which does not mean the predictions are wrong, it only indicates there is no n-gram matching. *e.g.*, The pair "the windshield of a bus" and "large front window on a bus"" gives zero CIDER-D. This finding intrigues us to pursue more robust and comprehensive metrics [8, 10].

Figure 1. The difference between image captioning, regional image captioning, and referring expression generation. The figures are from [2].



(a) Ours, VG Region Cap. (b) GRiT, VG Region Cap.

(c) SAM-Cap, VG Region Cap. (d) GIT, COCO Image Cap.

Figure 2. Distribution of CIDEr scores. From left to right is: ours on the VG region caption, SAM-Cap (baseline), GRiT on the VG region caption, and GIT on the COCO image caption. Note that the scales of the y-axis for each figure is *different*.

Table 4. The statistics of region-level understanding datasets used for training.

dataset	type	total samples	total regions	total sents	total tokens	total words
COCO[13]	Region recognition	117,266	860,001	860,001	1,275,513	942,822
V3Det [20]	Region recognition	183,348	1,357,351	1,357,351	3,984,388	2,126,318
Objects365 [17]	Region recognition	1,742,289	25,407,598	25,407,598	49,264,696	32,341,116
Visual Genome [11]	Region captioning	77,398	3,684,063	3,684,063	21,392,494	19,740,221
RefCOCOg [24]	Referring Expression	24,698	48,599	92,671	834,305	785,259

Table 5. The zero-shot performance on the Referring Expression Generation (REG) task. "M": Meteror, "C": CIDEr-D. $*$: "k" means the number of examples in the prompt. †: The scores are from the papers.

Method	RefCOCOg		RefCOCO+				RefCOCO			
	val		testA		testB		testA		testB	
	M	C	M	C	M	C	M	C	M	C
separate train/test										
Visdif [24]†	14.5	-	14.2	-	13.5	-	18.5	-	24.7	-
SLR [25]†	15.9	66.2	21.3	52.0	21.5	73.5	29.6	77.5	34.0	132.0
zero-shot										
Kosmos-2 [15]†	12.2	60.3	-	-	-	-	-	-	-	-
Kosmos-2 (k=2)*†	13.8	62.2	-	-	-	-	-	-	-	-
Kosmos-2 (k=4)*†	14.1	62.2	-	-	-	-	-	-	-	-
ASM [22]†	13.6	41.9	-	-	-	-	-	-	-	-
GRiT [23]	15.2	71.6	-	-	-	-	-	-	-	-
SCA (GPT2-large, Pretrain+VG)	15.4	71.9	21.7	29.2	20.4	57.2	20.4	27.0	20.2	66.4
SCA (GPT2-large, VG)	15.3	70.5	21.7	30.2	20.1	56.6	20.5	27.7	20.1	66.7
SCA (LLAMA-3B, VG)	15.6	74.0	22.0	30.0	20.2	56.1	20.7	27.3	20.3	65.3

G. Additional Visualizations

We exhibit more qualitative results in Figs. 3 and 4.

H. Failure Case Analysis and Limitations

Our model can make wrong predictions in the terms of following:

1. Wrong attribute prediction (Fig. 5). *e.g.*, the models could predict the wrong colors or textures;
2. Distinguishing similar visual concepts (Fig. 6). *e.g.*, the model may confuse "lemon" with "orange";
3. Alignment with mask predictions (Fig. 7): As we do not supervise the alignment, the model may predict mask and captions for the fore- and background separately.

We believe these drawbacks, *esp.* 1) and 2), may be addressed by weak supervision and self-training [1].

I. The Formal Definition of the Feature Mixers

Here we provide a more formalized descriptions for each variants:

1) ROI-Align operator [6]: Given image feature I, we extract the regional feature $R = \pi(I)$, where π is the ROI-Align operator. Then we project the regional feature $\hat{R} =$ Proj(R), where Proj is the project function which can be a *linear layer* or a *two-layer MLP with GELU activation* [14]. The architecture is changed in this setting, while for the rest three, The architectures are not changed.

2) Directly decoding the mask query tokens from SAM: We remove the textual query token Q. The query-based feature mixer E_R becomes:

$$E_R^j(P^{j-1}, M^{j-1}; I^{j-1}) = \{\hat{P}^j, \hat{M}^j; \hat{I}^j\}. \quad (1)$$

Then we feed the mixed mask query token $\hat{M}$ into the text decoder.

3) Learn textual query tokens without SAM's query tokens: The prompt and mask query tokens (*i.e.*, P and M) sent to the textual feature mixer are *not* encoded by the SAM feature mixer.

4) Learn textual query tokens with SAM's query tokens: The prompt and mask query tokens (*i.e.*, P and M) sent to the textual feature mixer are encoded by the SAM feature mixer.

J. Analysis of the "Verb (Fuzzy)" Metrics

For exact matching, SCA achieves the *highest* results of 4.8. While for *fuzzy* matching (the mean cosine similarity be-

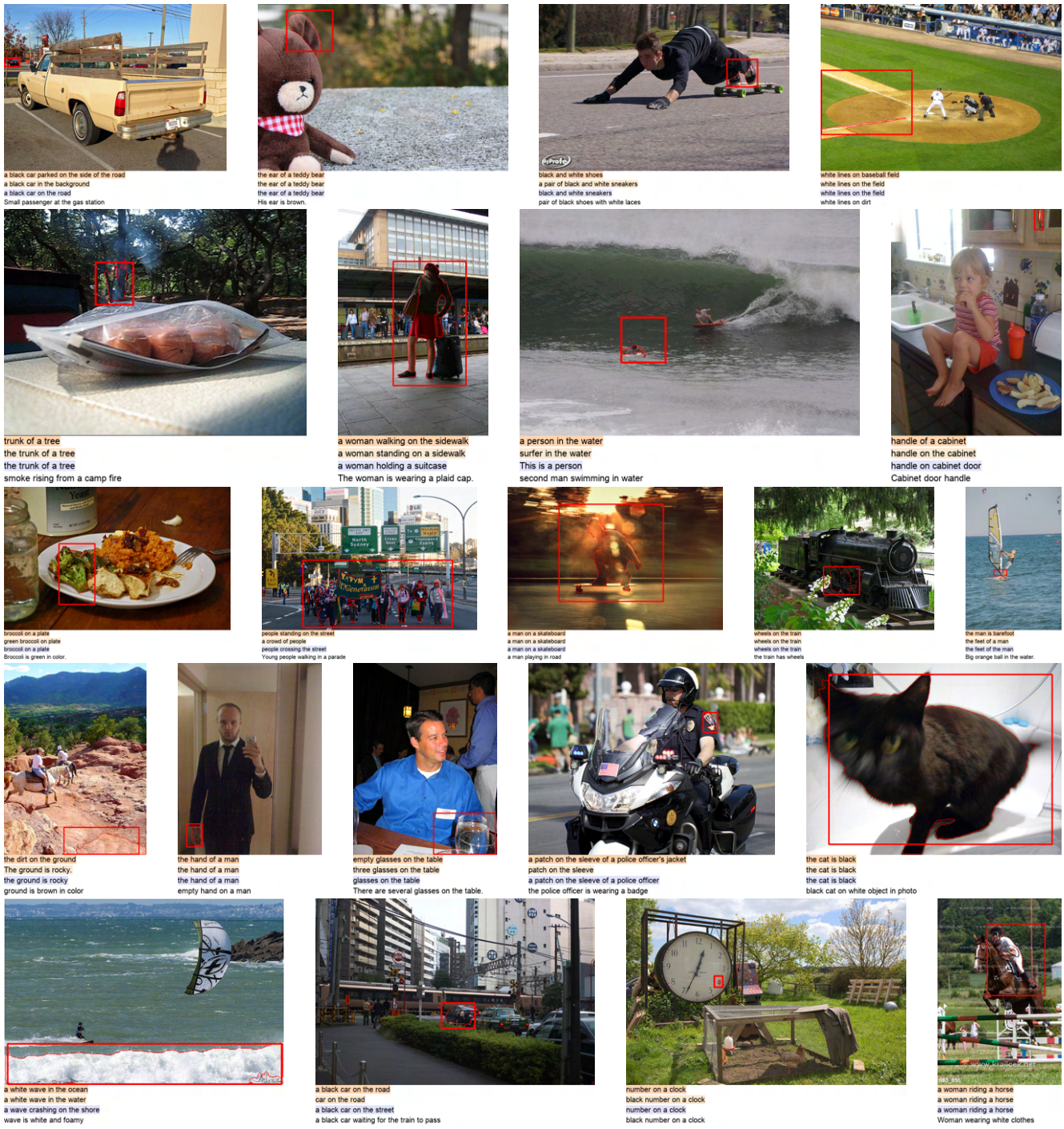


Figure 3. More visualization of the predictions. From top-to-bottom, the captions are from: *SCA* { GPT2-large+VG , LLAMA-3B+VG , GPT2-large+Pretrain+VG }, and the *ground truth*.

tween phrase embeddings), our method underperforms as it predicts noun phrases without verbs. Conversely, baseline models often predict verbs, and even incorrect verbs can have scores about 0.2-0.6 (Fig. 8).

K. The Implementation Details about the Baseline

We build the baseline models, SAM + Image Captioner as follows: 1) use input prompts to get the mask with the high-

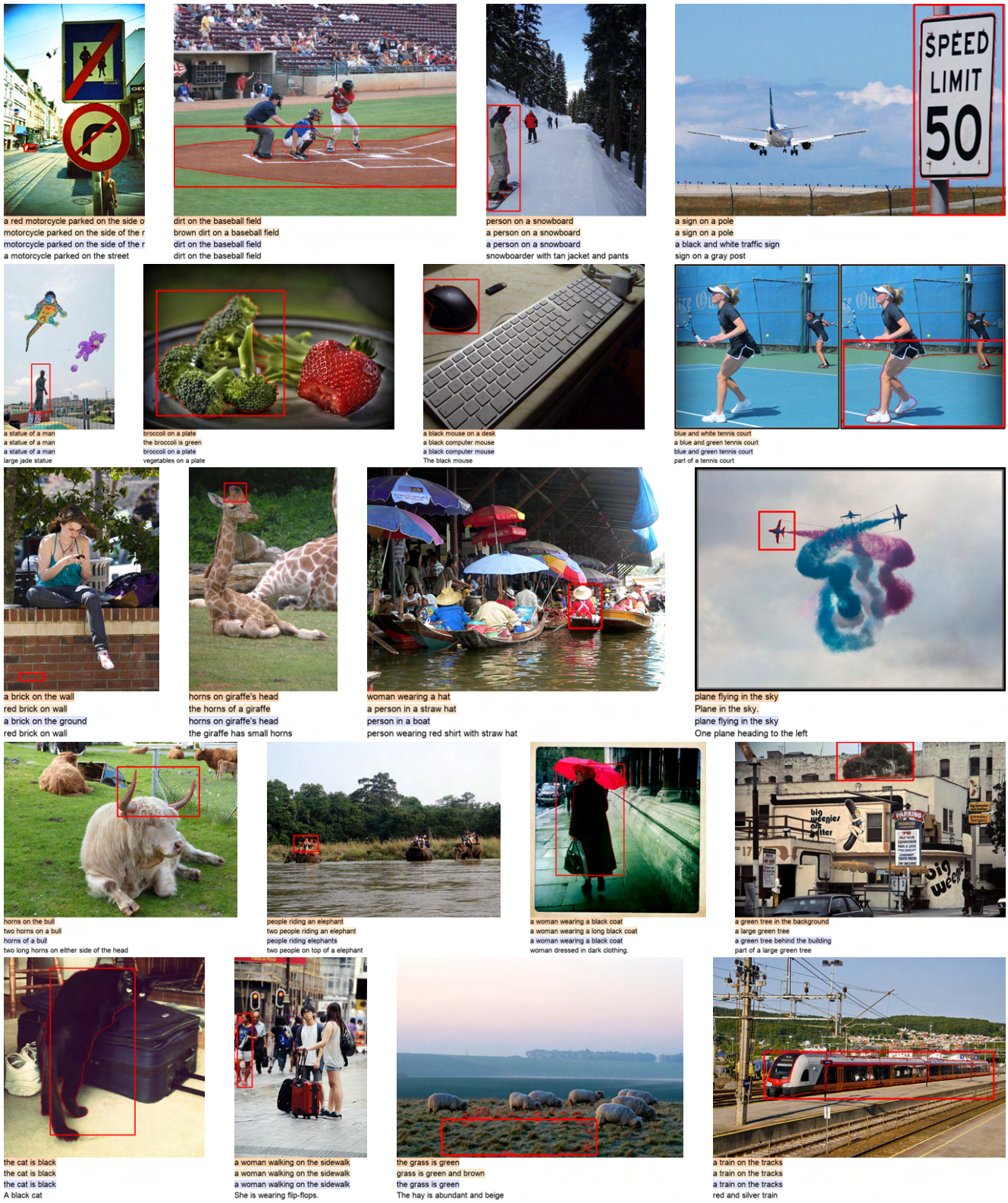


Figure 4. More visualization of the predictions. From top-to-bottom, the captions are from: *SCA* {GPT2-large+VG, LLAMA-3B+VG, GPT2-large+Pretrain+VG}, and the *ground truth*.

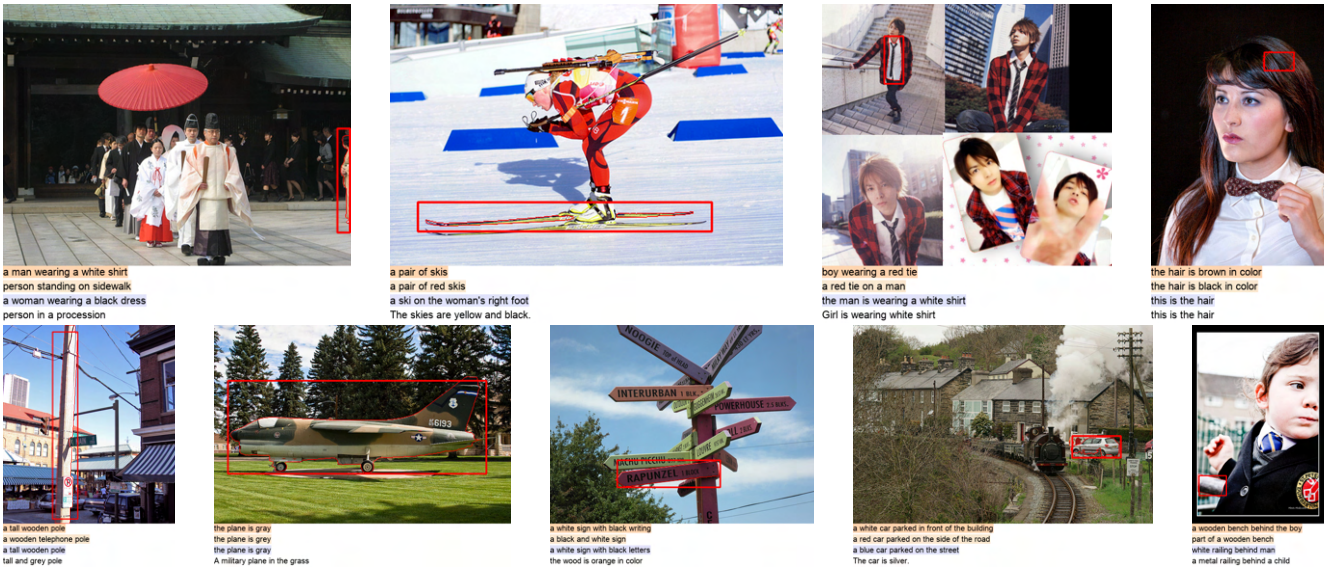


Figure 5. The predictions with wrong attributes.

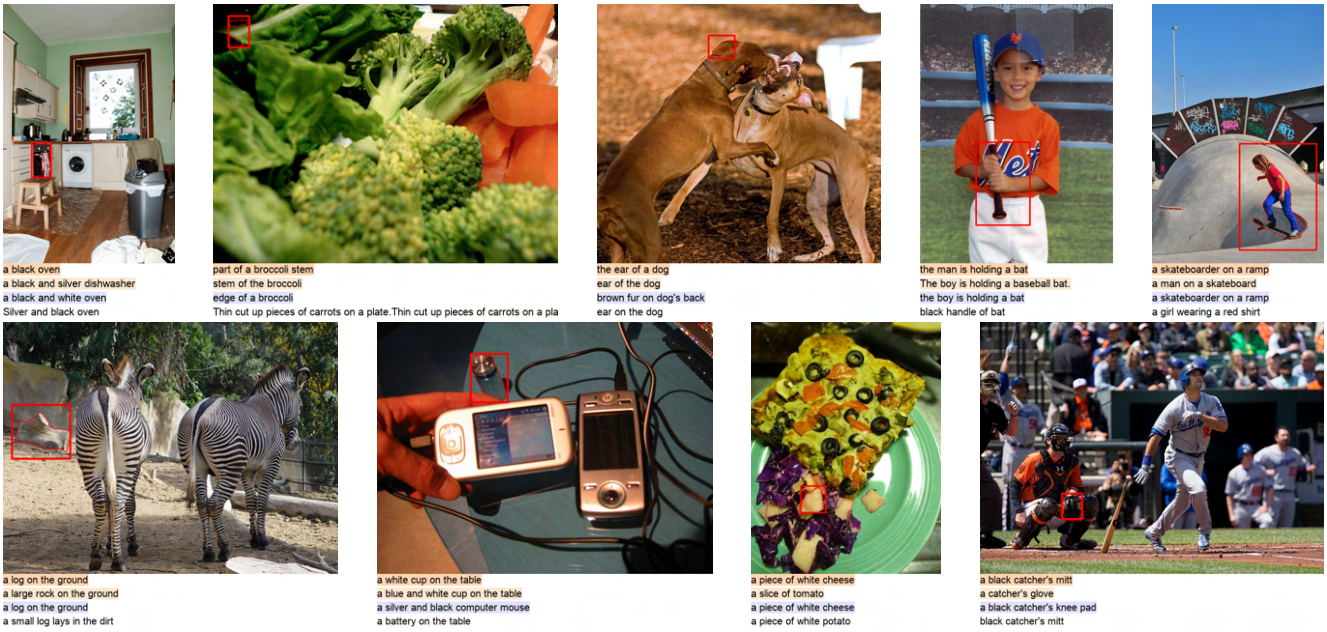


Figure 6. The predictions with wrong entities. From top-to-bottom, the captions are from: *SCA* { GPT2-large+VG , LLAMA-3B+VG , GPT2-large+Pretrain+VG }, and the *ground truth*.

est confidence score; 2) crop the image patch by the tightest box that covers the selected mask, and feed it the to the image captioner. We report the performance of the baselines with Visual Chain-of-Thought (V-CoT) [86] in Tab. 6. The V-CoT technique first uses the vision-language generative model (*e.g.*, BLIP [12]) to "recognize" the cropped subject (without the background; only the pixels inside the SAM-predicted mask). Then it re-uses the VLM to "caption" the subject with its "recognized" name and the background pixels. There are two major issues for V-CoT. 1) The prediction fails at the first "*recognition*" step due to *indistinguishable region crops* even for humans, which highlights the importance of the context (*e.g.*, the surroundings of a region, the global semantics of an image). 2) The captions based on the first-step "recoginition" are prone to hallucination. We showcase the representative qualitative results in Fig. 9.

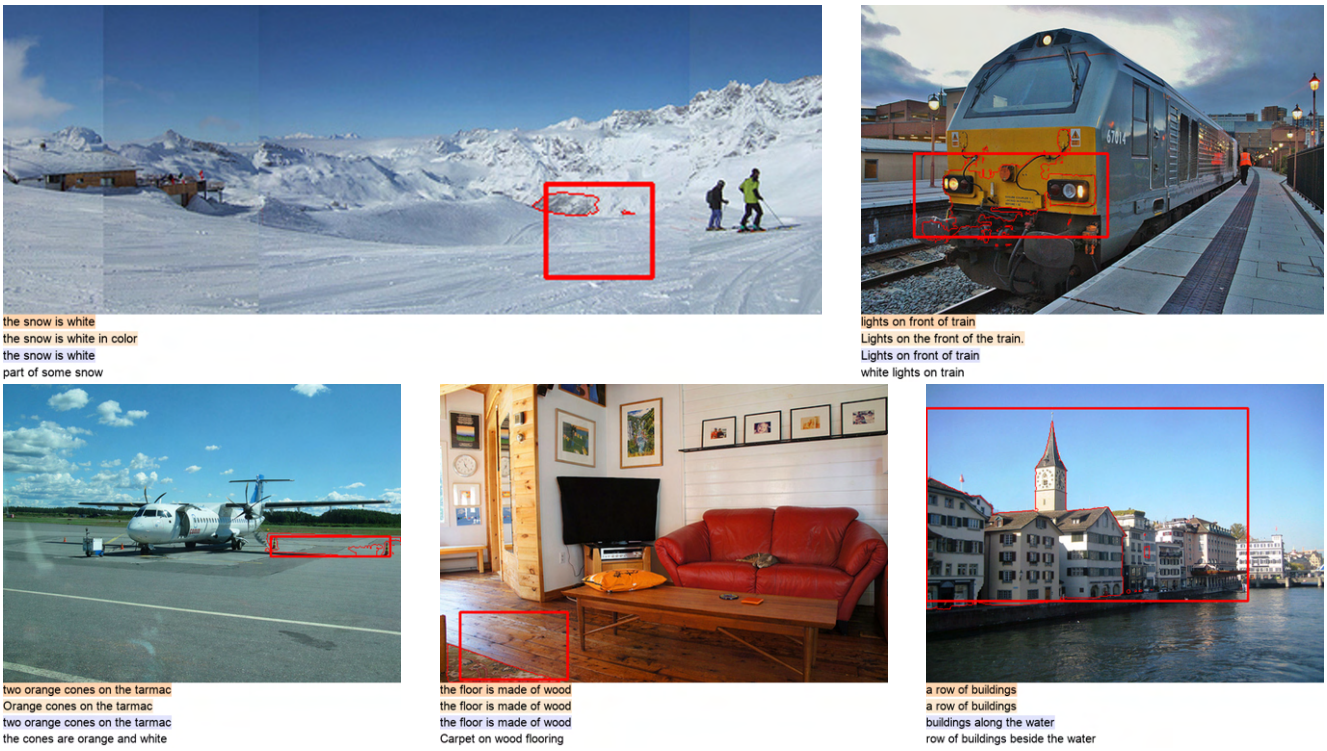
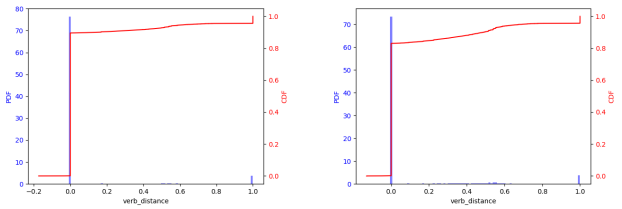
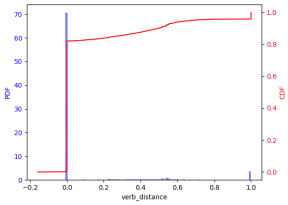


Figure 7. The predictions that are unaligned with the masks. From top-to-bottom, the captions are from: *SCA* { GPT2-large+VG , LLAMA-3B+VG , GPT2-large+Pretrain+VG }, and the *ground truth*.



(a) Ours (b) SAM-Cap + BLIP2-OPT-2.8B



(c) SAM-Cap + BLIP2-OPT-6.7B

Figure 8. Distribution of "Verb (Fuzzy)" scores on the VG region caption. From left to right is: ours, SAM-Cap (baseline) + BLIP2-OPT-2.8B, and SAM-Cap (baseline) + BLIP2-OPT-6.7B. Note the difference of the pdf ranging **from 0.2 to 0.6** in *x-axis*. Zoom-in for a clear view.

Table 6. Comparison of using V-CoT [86] or not.

Method	C	M	S	B@1	B@2	B@3	B@4	R	Noun	Verb	Noun (F)	Verb (F)
w/o V-CoT	59.7	11.7	16.7	19.6	9.8	5.3	3.0	22.7	26.6	4.5	53.7	9.7
w/ V-CoT	3.4	1.2	1.1	1.2	0.4	0.2	0.1	3.1	1.6	0.1	7.3	0.3

L. The Analysis of Different Type of Prompts

Our method supports both point and box prompts as we reuse SAM's feature mixer. However, since the dataset *exclusively* provides ground-truth boxes, we introduce *pseudo-point* prompts for *inference* derived through three strategies: 1) the box's center point (*CPB*), 2) a random point within the box (*RPB*), and 3) a random point within the highest-confidence mask predicted by SAM (*RPM*). Using points prompt performs worse due to its *absence* during training (Tab. 7).

Table 7. Comparison of different types of prompts.

Method	C	M	S	B@1	B@2	B@3	B@4	R	Noun	Verb	Noun (F)	Verb (F)
Box	149.8	17.5	31.4	38.2	24.1	16.8	12.2	35.7	41.7	4.8	65.1	7.5
CPB	86.8	12.7	22.5	29.1	15.6	9.6	6.1	26.5	29.6	3.3	56.1	5.9
RPB	68.3	10.7	18.0	25.2	12.9	7.7	4.8	23.2	23.7	2.3	50.5	4.3
RPM	90.5	13.2	23.6	29.8	16.1	9.9	6.3	27.1	30.9	3.4	57.5	5.9

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving Image Generation with Better Captions. *OpenAI blog*, 2023. 3

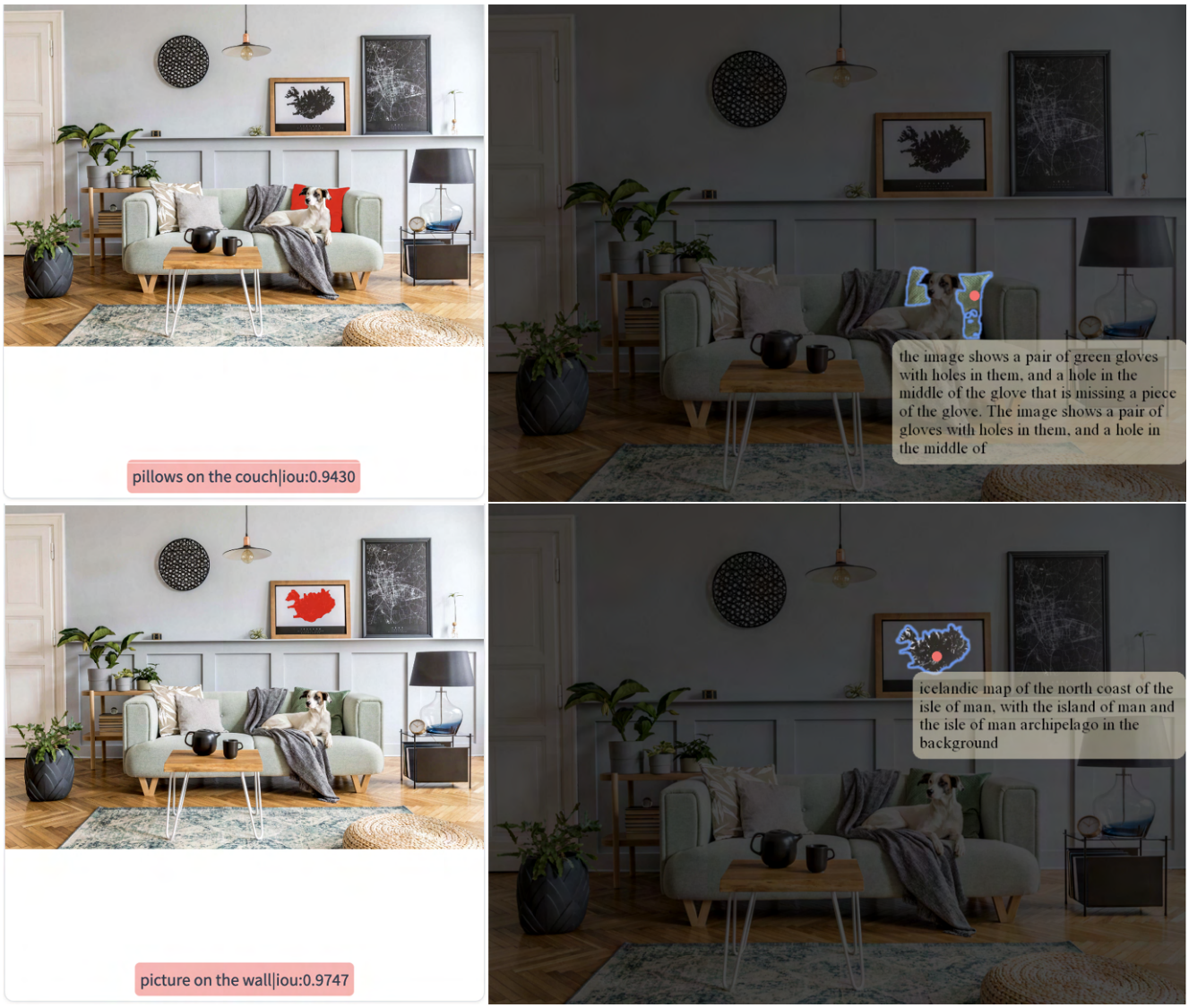


Figure 9. Qualitative comparisons between SCA (Left) and Captiong Anything [21] (Right, with V-CoT).

[2] Lior Bracha, Eitan Shaar, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Disclip: Open-vocabulary referring expression generation. *arXiv preprint arXiv:2305.19108*, 2023. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[5] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *TPMAI*, 2020. 3

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[8] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*, 2016. 2

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[10] Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. Guiding image captioning models toward more specific captions. In *ICCV*, pages 15259–15269, 2023. 2

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson,

Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1, 3

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3

[15] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[17] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3

[18] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[19] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *TMLR*, 2022. 2

[20] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3Det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*, 2023. 3

[21] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023. 8

[22] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, Yushi Chen, Tong Lu, Jifeng Dai, and Yu Qiao. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 1, 2, 3

[23] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. GRiT: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 1, 3

[24] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 1, 3

[25] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017. 3

[26] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. GPT4RoI: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1, 2