Supplementary to Pixel Aligned Language Models

Jiarui Xu[1,2*] Xingyi Zhou[1] Shen Yan[1] Xiuye Gu[1]
Anurag Arnab[1] Chen Sun[1] Xiaolong Wang[2] Cordelia Schmid[1]
[1]Google Research [2]UC San Diego

In this supplement we provide additional dataset details; quantitative experimental and qualitative visual results.

A. Dataset details

Localized Narratives [7]. We use the COCO subset of the Localized Narratives [7] for both training and evaluation. It consists of $134,272$ training images and $8,573$ validation images. $5,000$ images are annotated with 5 captions and 5 traces per image, and the rest are annotated with 1 caption and 1 trace per image. We use all the traces for evaluation of controlled trace generation when comparing with MITR [5]. We use 224×224 crop size for WebLI pre-training and 384×384 for localized narrative fine-tuning.

GoldG [2]. We use the GoldG dataset prepared in MDETR [2] for referring localization. It which consists of images from COCO [4], Visual Genome [3], and Flickr30k [6]. We filtered out all the validation and testing images of RefCOCO, RefCOCO+, and RefCOCOg from our combined training set, yielding $160,280$ training images in total.

Visual Genome [3]. We use the Visual Genome split prepared in GRiT [9] for dense object captioning, with $77,396$ training images and $5,000$ test images.

B. Experiments

Ablation on Visual Encoders. We used a trainable EVA02 backbone and a frozen SAM backbone. The main motivation for using the SAM backbone is to inherit its zero-shot segmentation ability without training. If we remove the SAM backbone, the segmentation performance drops , but the impact on the referring ability is minimal, as shown below.

backbone	RefCOCO P@0.5
EVA02	89.3
EVA02 +SAM	89.8

Table 1. **Ablation on Visual Encoders.**

*Work done during a Google internship. ✉{jiaruixu, zhouxy}@google.com

Scalability of Localized Narratives Pre-training without other localization datasets. We conduct experiments using different percentages of only LN data below. It shows the performance consistently improves when pre-trained with more data from LN, emphasizing the scalability of our approach. Note that our improvement over not pre-training with LN is a decent 2.2 points on RefCOCO.

LN data	0	10%	50%	100%
RefCOCO P@0.5	81.8	83.0	83.5	84.0

Table 2. **Ablation on Scalability of Localized Narratives Pre-training.**

Full fine-tuning T5. We add fine-tuning results and report results in Tab. 3 below. Note we couldn't fully fine-tune T5-XL due to memory limit. We observe that full fine-tuning did not significantly outperform LoRA in our attempt, likely because T5 pre-training is already strong, and LoRA fine-tuning is enough. It is also worth noting that even without LoRA, the frozen T5-XL performs on par with models that fine-tune the text encoder jointly [1, 8]. It is an evidence that the frozen large language model like T5 encompasses strong vision language ability, *e.g.* localization, which could be revealed by our PixelLLM.

Language Model	Params	Frozen	LoRA	Full fine-tuning
T5-Small	80M	67.0	76.6	76.4
T5-Base	250M	70.3	80.8	81.4
T5-Large	780M	73.6	84.8	83.1
T5-XL	3B	81.9	89.8	OOM

Table 3. **Ablation on language model size and fine-tuning approach.** We report RefCOCO official metrics under different language model sizes.

C. Qualitative results

We provide qualitative more results on pixel-aligned captioning, referring localization and segmentation, and dense object captioning in Figure 1, 2, and 3, respectively.

References

[1] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1

[2] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 1

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[5] Zihang Meng, Licheng Yu, Ning Zhang, Tamara L Berg, Babak Damavandi, Vikas Singh, and Amy Bearman. Connecting what to say with where to look by modeling human attention traces. In *CVPR*, 2021. 1

[6] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 1

[7] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 1

[8] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 1

[9] 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:2212.00280*, 2022. 1

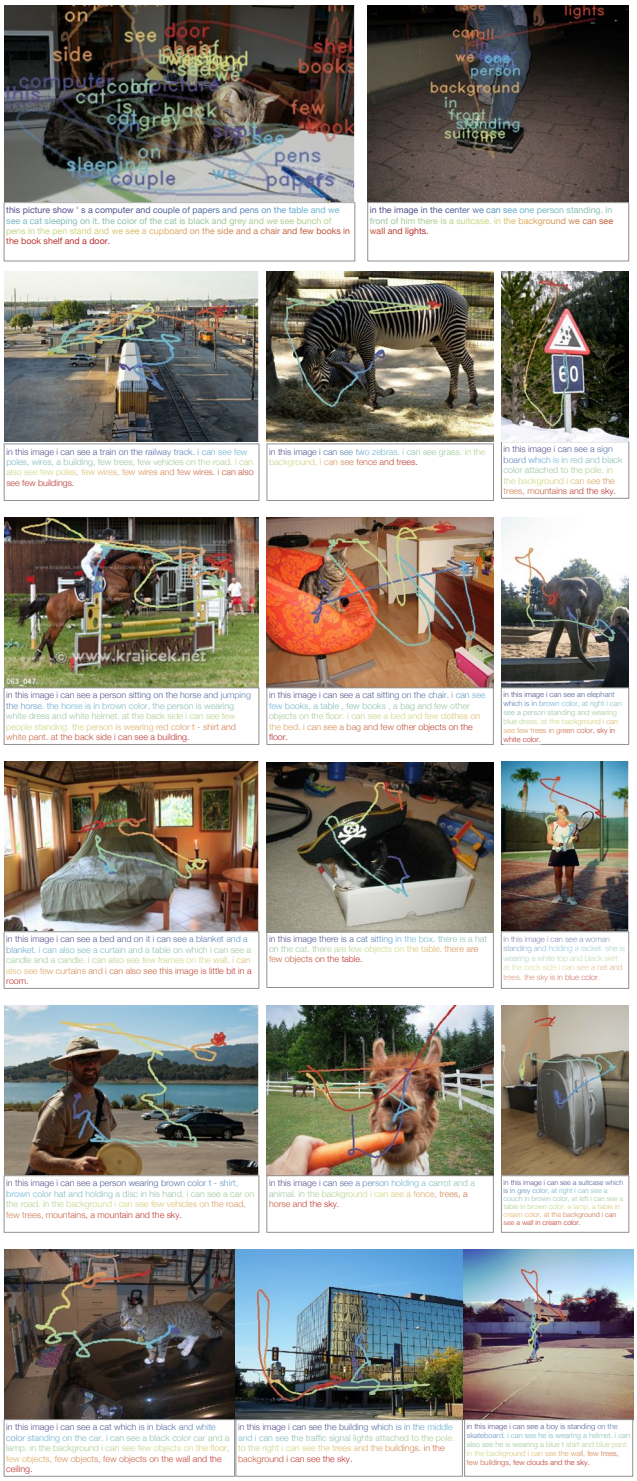

Figure 1. Qualitative results on pixel-aligned captioning. Zoom in for the best view. We show examples of our predicted coordinates of each text token in the first row. Not only noun terms (*i.e.* "papers", "pens", "cat", "suitcase", "person") are aligned to the corresponding object, but also verbs or relation terms ("sleeping", "standing") are aligned to the corresponding regions.

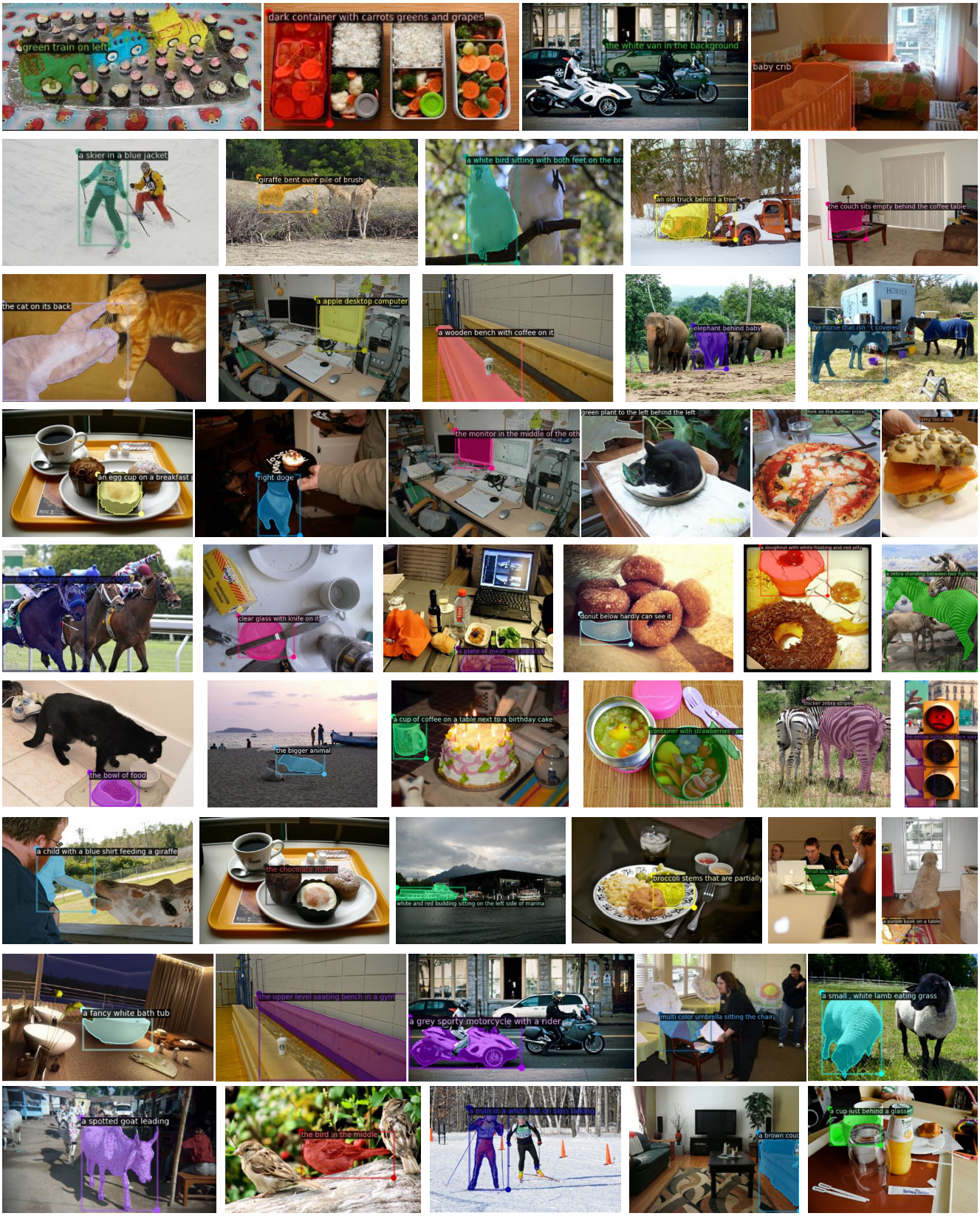

Figure 2. Qualitative results on referring segmentation. Zoom in for the best view.

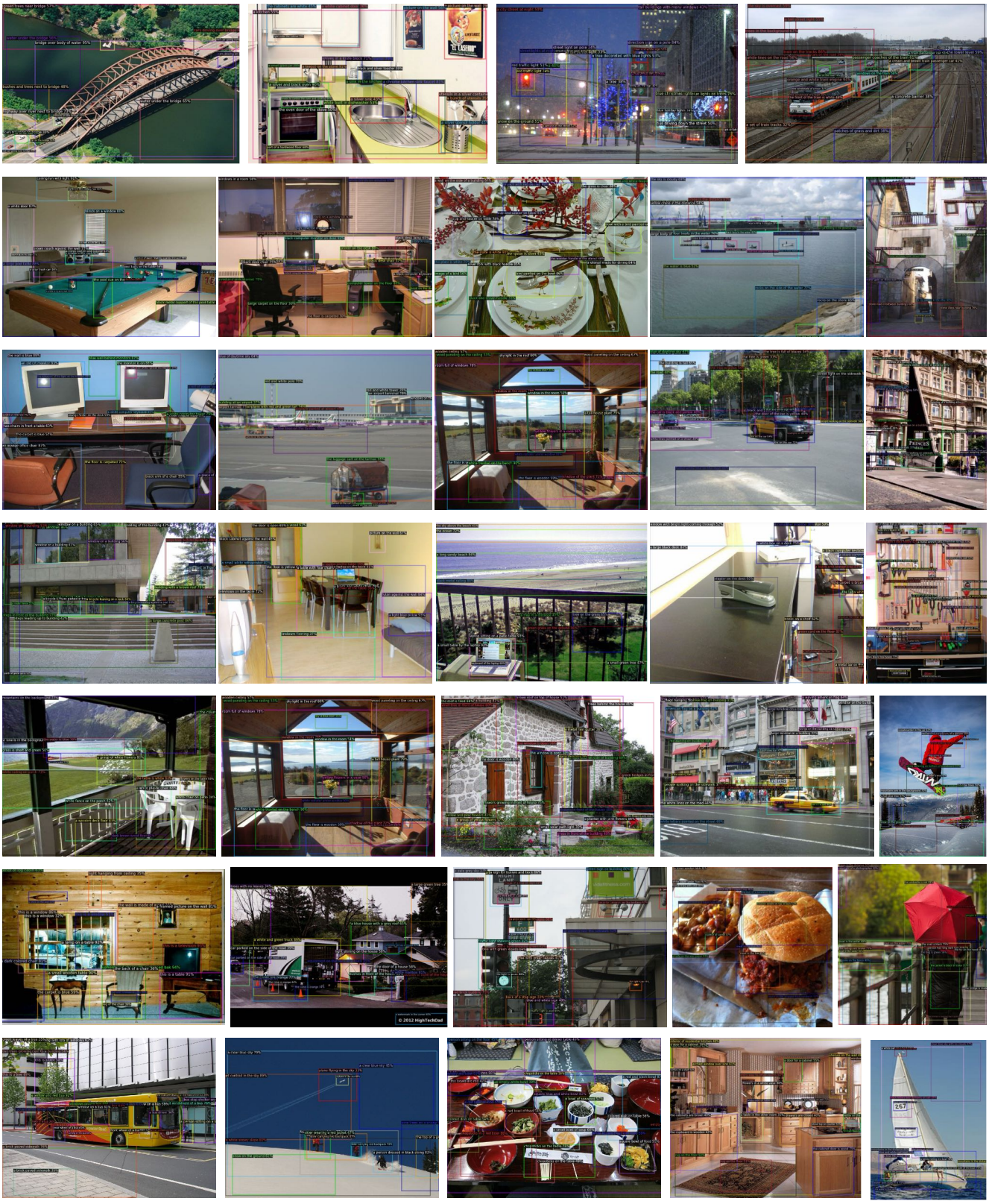

Figure 3. Qualitative results on dense object captioning. Zoom in for the best view.