1 Datasets and Tasks

In our experiments, many different datasets and tasks are involved. In training, we used a two-stage training strategy. In the first stage, we use CC3M [2] to align the vision features to the text input space. In the second stage, we train the model on four different tasks and corresponding datasets: generic segmentation(COCO Panoptic Segmentation [13]), referring segmentation(RefCOCO/+/g [18, 24]), interactive segmentation(COCO-Interactive) and the vision-language instruction task (LLaVA1.5 training data [15]). Note that COCO-Interactive is our in-house dataset, as there is no well-established public dataset that supports all four interaction types (point, scribble, box, and mask); we released this dataset, and its construction details are given in the Sec. 4.

The evaluation tasks are classified into in-domain tasks and out-of-domain tasks, according to if the evaluation task appears in the training. Specifically, we use three out-of-domain tasks in this work: generalized referring expression segmentation(gRefCOCO [14]), open-vocabulary segmentation (ADE20K [25], Cityscapes [4], Pascal Context [17], and Pascal VOC [6]) and video object segmentation (DAVIS-2017 [20]).

We list the details of the datasets used below:

COCO-Panoptic. The COCO-Panoptic dataset is an extension of the COCO dataset, specifically designed for panoptic segmentation tasks. It consists of over 200,000 images with detailed annotations that cover 80 object categories for instance segmentation and additional categories for semantic segmentation.

RefCOCO. RefCOCO is a dataset designed for the task of referring expression comprehension and segmentation. It consists of images from the COCO dataset that are annotated with referring expressions, where each expression uniquely identifies a particular object within the image. It includes three splits: RefCOCO, RefCOCO+, and RefCOCOg, each with different characteristics and annotation styles.

LVIS. The LVIS dataset is a benchmark for instance segmentation with a large vocabulary of object categories. It features high-quality instance annotations for over 1,000 object categories across a diverse set of images. LVIS is particularly known for its long-tail distribution of categories, which presents a unique challenge for segmentation models.

ADE20K. ADE20K is a widely used dataset as an open-vocabulary segmentation benchmark, it contains both things and stuffs annotations and thus can evaluate panoptic segmentation. It encompasses a diverse collection of images from various indoor and outdoor scenes. It is part of the MIT Scene Parsing Benchmark and provides dense pixel-wise annotations for 150 object categories, facilitating research in scene understanding and segmentation.

Cityscapes. Cityscape is a dataset focused on urban street scenes. The dataset contains a large number of high-quality video sequences and pixel-accurate annotations from 30 categories in 50 different urban street scenes. With its detailed instance-level annotations, Cityscapes is pivotal to advancing semantic and instance segmentation research, especially in autonomous driving and urban scene perception.

Pascal VOC. Pascal VOC contains 20 classes of semantic segmentation annotation.
Pascal Context. Pascal Context is an extension of the PASCAL VOC dataset, providing comprehensive scene understanding through detailed semantic labels for the entire scene in each image. It comes in two versions: PC-59, which focuses on the most frequent 59 categories, and PC-459 includes a broader set of 459 categories.
DAVIS-2017. DAVIS-2017 is a video segmentation benchmark that provides high-quality, full-resolution video sequences with per-pixel annotations of multiple objects. It is commonly used to evaluate the performance of video object segmentation methods, particularly in semi-supervised settings where the first-frame mask is provided.
gRefCOCO. gRefCOCO is the first large-scale Generalized Referring Expression Segmentation dataset that contains multi-target, no-target, and single-target expressions.
Ego-Exo4D. Ego-Exo4D is a diverse, large-scale multimodal multiview video dataset and benchmark challenge. Ego-Exo4D centers around simultaneously captured and time-synced egocentric and exocentric vides of skilled human activities. More than 800 participants from 13 cities worldwide performed these activities in 131 different natural scene contexts, yielding long-form captures from 1 to 42 minutes each and 1,422 hours of video combined. The Correspondence benchmark needs the model to predict the corresponding mask for the same object in each synchronized frame of the other view if it is visible.
CC3M. CC3M is a large-scale dataset of image-caption pairs designed for training and evaluating visual-language models. It contains around three million images sourced from the web, each accompanied by a descriptive caption.

2 Implementation Details

Swin-B [16] is used as a visual encoder with a Phi-1.5 1.3B [12], the architecture of the mask generator is the same as Maks2Former, the number of mask tokens is set to 100, and both Swin-B and mask generator are initialized from the pre-trained Maks2Former model weight. If not specified, the model is trained with a joint training setting and without additional task-specific fine-tuning. All experiments are run on 16×V100 GPUs.

Our model has two training stages, the first stage is the vision-language alignment stage, in which we strictly follow the default settings of LLaVA, with the only change being the adoption of Phi-1.5 as the LLM and the use of Swin as the visual encoder with the hyper-parameters as shown in Tab. 15.

In the second stage, to train the final model, we use a total of 56k training iterations. In each iteration, we randomly sample one task with equal probability from four tasks: generic segmentation (COCO Panoptic), referring segmentation (RefCOCO/+/g), interactive segmentation (COCO-Interactive), and a visual-language instruction task (LLaVA1.5 training data) and all images are resized to 1024^2 by padding the shorter side to keep the aspect ratio. During training, the visual encoder is frozen while all other model parts are trainable. In addition, we adopt the Hungarian matching to automatically assign the ground-truth of mask proposals during training. In generic segmentation tasks, we use both classification loss and mask loss as cost matrix for matching, while other tasks use only mask loss. Tab. 16 shows the hyper-parameters.

For ablation, we use the same training setting as the final model, but with the shorter 9k training iterations to reduce the cost of the experiment.

3 Prompts for Different Tasks

As discussed in Sec. 3.3, our input schema has three kinds of inputs: task instruction prompt, condition prompt, and a set of mask tokens. For different tasks, we use different instruction prompts and condition prompts, as listed in Tab. 14. Basically, tasks that use the same type of condition prompts also adopt the same task instruction prompt.

In addition, some tasks that use category condition often need to deal with background as well, such as instance segmentation, for which we specifically append a *‘background‘* at the end of the joint sentence.

Table 14: Detail prompts for all tasks.

Task	Dataset	Instruction prompt	Condition prompt
Panoptic Seg.	COCO	*You need to segment all objects.* *This is all the candidate categories:*	`[class1], [class2] ...`
OV Seg.	ADE20K, *etc.*	*You need to segment all objects.* *This is all the candidate categories:*	`[class1], [class2], ...`
Referring Seg.	RefCOCO/+/g	*Please segment according to* *the following instruction:*	`object description`
Generalized Referring Seg.	gRefCOCO	*Please segment according to* *the following instruction:*	`object description`
Interactive Seg.	COCO-Interactive	*Please segment by given regions:*	`<interaction1>, <interaction2>...`
Video Object Seg.	DAVIS	*Please segment by given regions:*	`<interaction1>, <interaction2>...`
Ego-exo Correspondence	Ego-Exo4D	*Please segment by given regions:*	`<interaction1>, <interaction2>...`

4 Details on building COCO-Interactivate Dataset

In this section, we describe in detail how to build the COCO-Interactive dataset. The COCO-Interactivate is based on image and annotations of COCO2017 instance segmentation, which provide the masks and bounding boxes for each instance, and we use the annotations to automatically generate four types of

Table 15: Hyper parameters of our model in the first stage training.

Parameters	Value
Optimizer	AdamW
Learning Rate	2×10^{-3}
Batch Size	128
Number of Iteration	4,650
Learning Rate Schedule	Cosine Decay
Weigth Decay	0.0
Warmup Steps	140
β_1	0.9
β_2	0.999
Training Data	CC3M
Image Size	1024×1024
Image Processing	Resize long edge to 1024 and padding short edge to 1024.

Table 16: Hyper parameters of our model in the second stage training.

Parameters	Value
Optimizer	AdamW
Learning Rate	4×10^{-5}
Batch Size	64
Number of Iteration	56,000
Number of Iteration (for Ablation)	9,000
Learning Rate Schedule	Cosine Decay
Weigth Decay	0.0
Warmup Steps	1680
Warmup Steps (for Ablation)	270
β_1	0.9
β_2	0.999

Training Data	COCO-Panoptic (25%); RefCOCO/+/g(25%);
	COCO-Interactive(25%); LLaVA 1.5(25%)

Image Size	1024×1024
Image Processing	Resize long edge to 1024 and padding short edge to 1024.

visual prompts: point, scribble, mask, and box. Fig. 8 shows the illustration of the visual prompts, and we will introduce each of them in the following:

Point. For each instance, we generate a point visual prompt by randomly sampling a point within a circular region centered on the bounding box with a radius of half the short side of the bounding box.

Scribble. The generation process of scribble has two steps. First, we randomly jitter the width and height of the ground-truth bounding box from a scale factor range of [0.5, 1.2], and ensure that the IoU between the jittered box with the original box is greater than 0.5. Then, given a jittered box, we randomly select

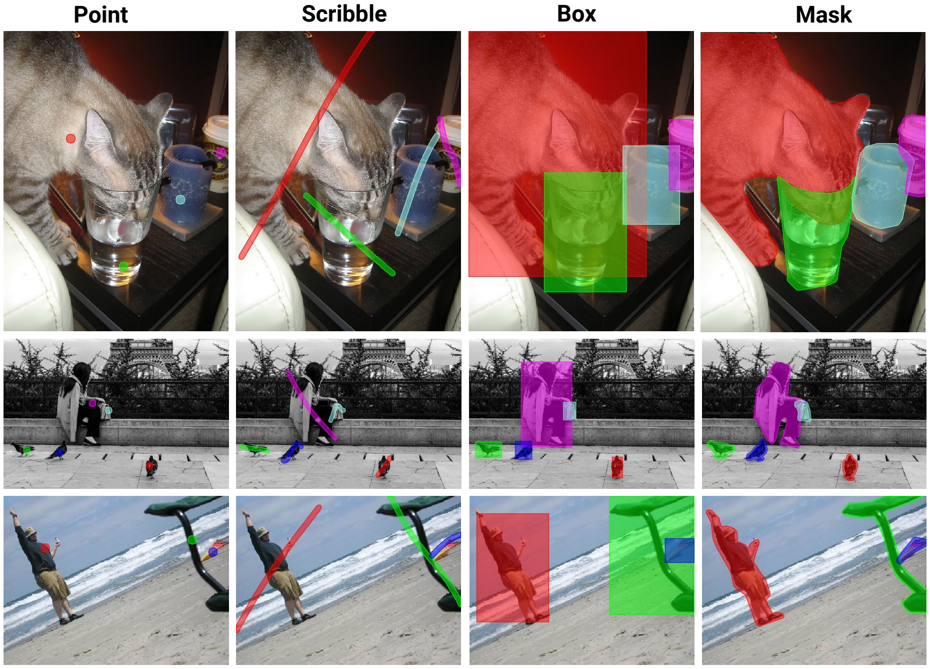


Fig. 4: Visualization of different types of visual prompts

one of its diagonals and generate a sin curve along it. The amplitude of the sin curve is randomly chosen from [10, 20], the frequency is randomly sampled from $[0.2 \times 2\pi, 2\pi]$, and the phase shift is randomly sampled from $[0, 2\pi]$.

Box. We randomly jittered ground truth boxes as box prompts, and the length and width of each jittered box were obtained by scaling from a scale sampled in the range [0.9, 1.1].

Mask. The mask visual prompts are obtained by applying a Gaussian filter on the ground truth mask at first, with the standard deviation of the Gaussian kernel set to 2, and then binarizing the blurred mask.

5 Implement Details of Decouple Ablation

In Tab. 3, we compare the decouple design and non-decouple design on COCO Semantic Segmentation. Here, we will introduce the implementation detail for the non-decouple design. Specifically, we omitted the mask token, and instead of using the category condition embeddings as the mask query, and fed into the mask generator to generate masks. In this case, the matching loss mechanism is unnecessary, the condition embedding of a category is used to predict the class and mask at the same time.

Table 17: Performance on multi-modal benchmarks.

Methods	LLM Type	VQA^{V2}	SQA	MMB	POPE
OpenFlamingo [1]	MPT-7B	51.8	-	5.7	-
Kosmos-2 [19]	-	51.1	-	-	-
BLIP-2 [10]	Vicuna-13B	41.0	61.0	-	85.3
InstructionBLIP [5]	Vicuna-7B	-	60.5	-	36.0
IDEFICS [9]	Llama-7B	50.9	44.2	48.2	-
LLaVA-1.5 [15]	Vicuna-7B	78.5	66.8	64.3	85.9
PSALM	Phi-1.5 (1.3B)	62.3	64.9	52.5	80.3

Table 18: Zero-shot performance of Correspondence benchmark on Ego-Exo4D.

Query Mask	Method	Zero-Shot	Test IoU	Val IoU
Ego	XSegTx	✓	0.6	-
Ego	XMem	✓	4.6	-
Ego	XSegTx	✗	13.9	-
Ego	XMem	✗	14.6	-
Ego	PSALM	✓	-	7.9

Query Mask	Method	Zero-Shot	Test IoU	Val IoU
Exo	XSegTx	✓	1.6	-
Exo	XMem	✓	21.8	-
Exo	XSegTx	✗	43.8	-
Exo	XMem	✗	43.4	-
Exo	PSALM	✓	-	9.6

6 Additional Experiment

Multi-modal Benchmark Evaluation. Our PSALM model is based on MLLM and thus able to deal with vision and language tasks, and therefore we evaluate our model on several commonly used multi-modal benchmarks, and results are shown in Tab. 17. PSALM achieved promising results compared with other MLLM methods, such as BLIP-2 [10] and InstructionBLIP [5]. Although our model still lags behind the official LLaVA1.5 7B model, we believe that increasing the model size can largely close the performance gap.

Given the absence of training data, our related works such as LISA [8] and PixelLM [21], despite their theoretical capability to handle such tasks, yield suboptimal results. Take LISA as an instance, it achieves a mere **0.12** on the VQA score in a zero-shot manner.

Video Object Segmentation. We evaluate PSALM on video object segmentation task on DAVIS-2017 [20], In inference, we extract the region feature based on the last-frame prediction (or first-frame segmentation reference) as the visual-prior condition and use the image of the current frame to predict the mask. Tab. 19 shows the results, without training on any video data, PSALM shows promising zero-shot performance.

Ego-Exo4D Correspondence Benchmark. Ego-Exo4D [7] is a large-scale multi-modal multiview video dataset, and its correspondence benchmark is de-

Table 19: Our method's zero-shot performance on DAVIS-2017 *val*. Note the SEEM report results on 345 randomly sampled frames, while others are evaluated on all frames.

Methods	Video Data	J&F	J	F
XMem [3]	✓	87.7	84.0	91.4
OMG-Seg [11]	✓	76.9	-	-
Painter [22]	✗	34.6	28.5	40.8
SegGPT [23]	✗	75.6	72.5	78.6
SEEM-B [26]	✗	62.8	59.5	66.2
PSALM	✗	68.8	65.9	71.7

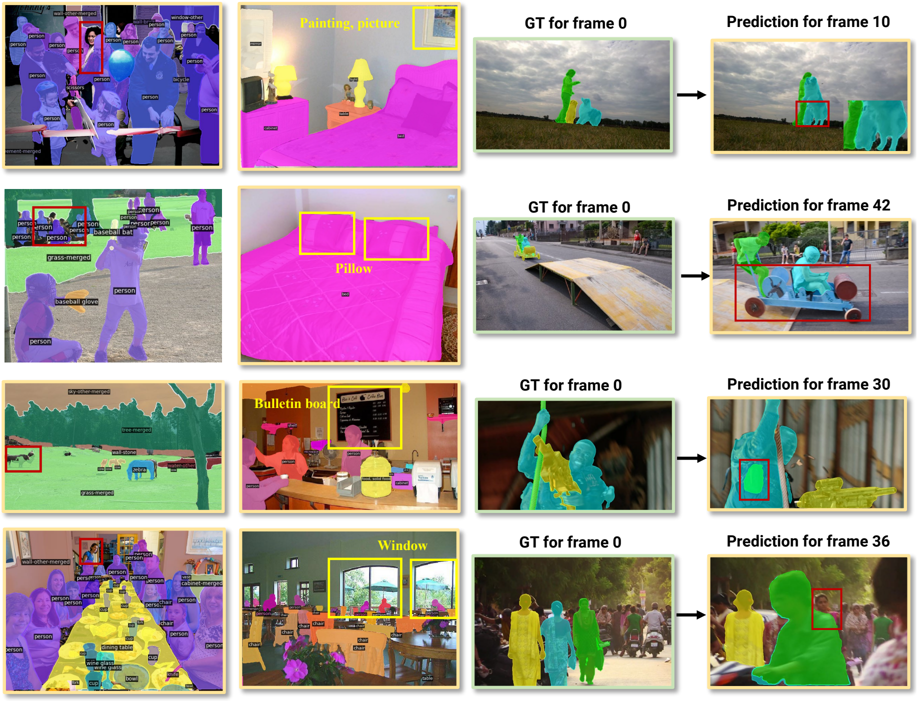
signed to predict the mask of an object in a novel view based on a given view. We evaluate this benchmark in a zero-shot manner to show the task generality of our model for such tasks. We performed the evaluation on Ego-Exo4D benchmark, since the official test set and model have not been released yet, we cannot directly compare the performance under the same setting, so we only report the quantitative results on the validation set as a reference in Tab. 18 and shows more qualitative results in Fig. 12.

7 Failure Case

As shown in Fig. 5, we observed failure cases, including: 1) Easily fails with small objects. It often presents as failing to predict small objects or connecting multiple neighboring small objects together; 2) Although PSALM performs competitively on open-vocabulary segmentation, its performance on the unseen category is far from perfect; 3) In Video Object Segmentation benchmark, model needs to understand the object relationship across different views/video frames, which is challenge for current model, and it requires better model design and involving more training tasks.

8 More Qualitative Results

Fig. 6, Fig. 7 and Fig. 8 show more qualitative examples of in-domain tasks. Fig. 9, Fig. 10 and Fig. 11 shows more examples of out-of-domain tasks.



(a) Failure cases in Panoptic seg.

(b) Failure cases in OV seg

(c) Failure cases in Video seg.

Fig. 5: Failure cases in different tasks

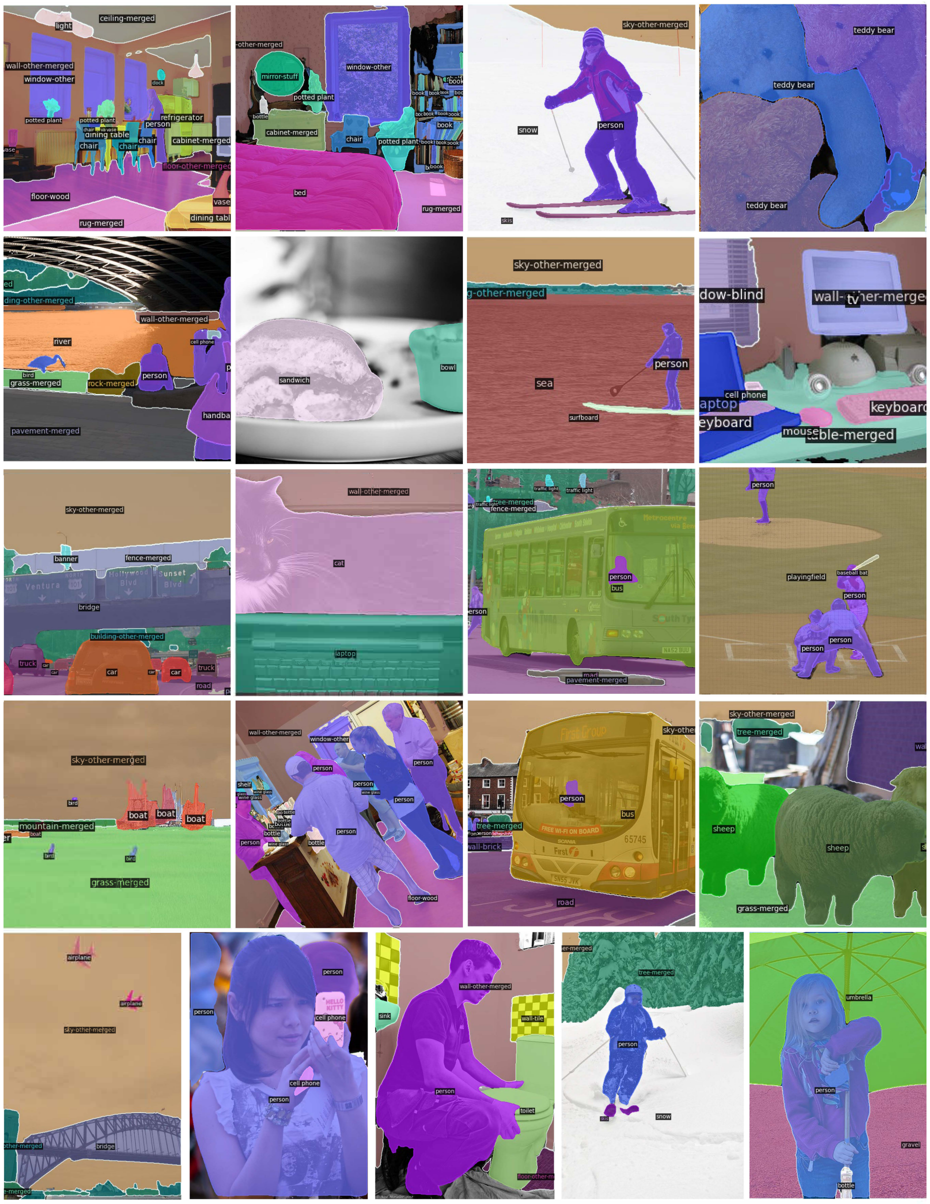


Fig. 6: More examples of panoptic segmentation in COCO [13].

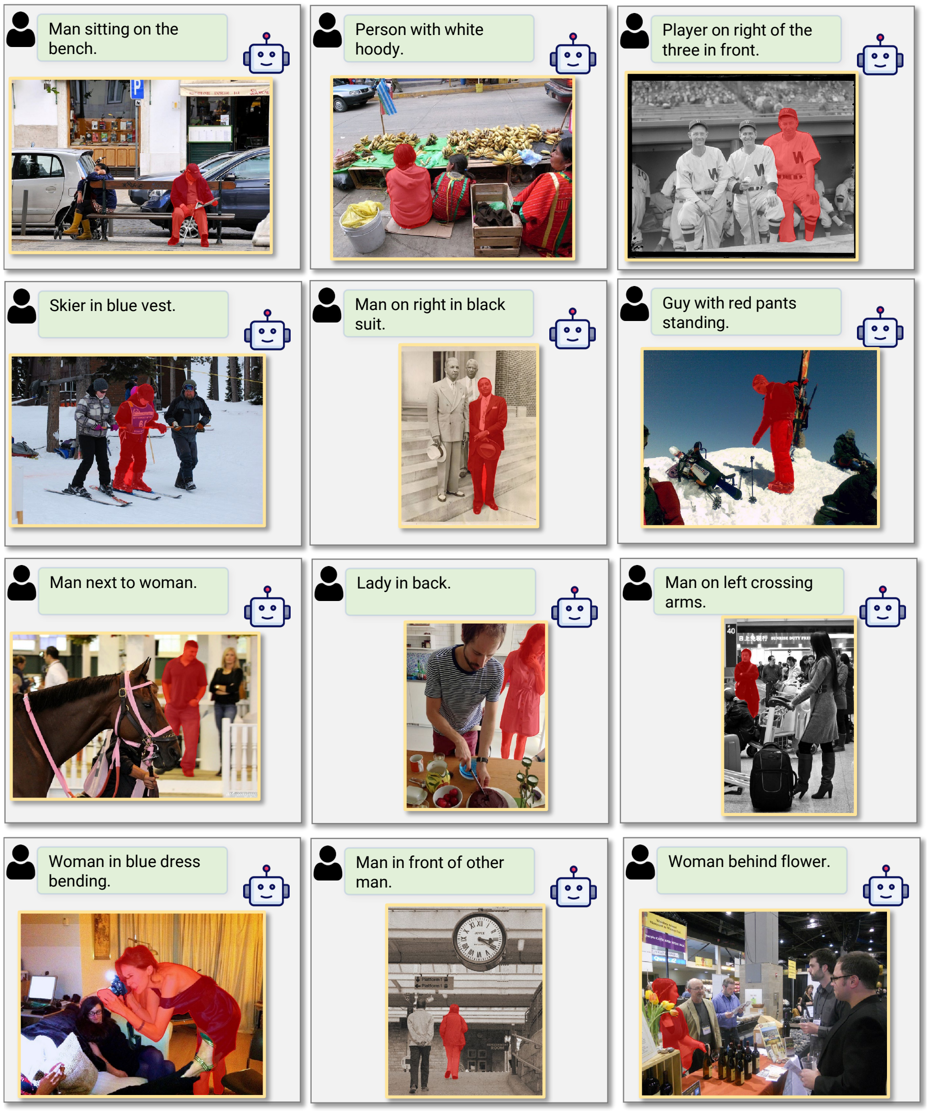

Man sitting on the bench.
Person with white hoody.
Player on right of the three in front.
Skier in blue vest.
Man on right in black suit.
Guy with red pants standing.
Man next to woman.
Lady in back.
Man on left crossing arms.
Woman in blue dress bending.
Man in front of other man.
Woman behind flower.

Fig. 7: More examples of referring segmentation in RefCOCO [24].

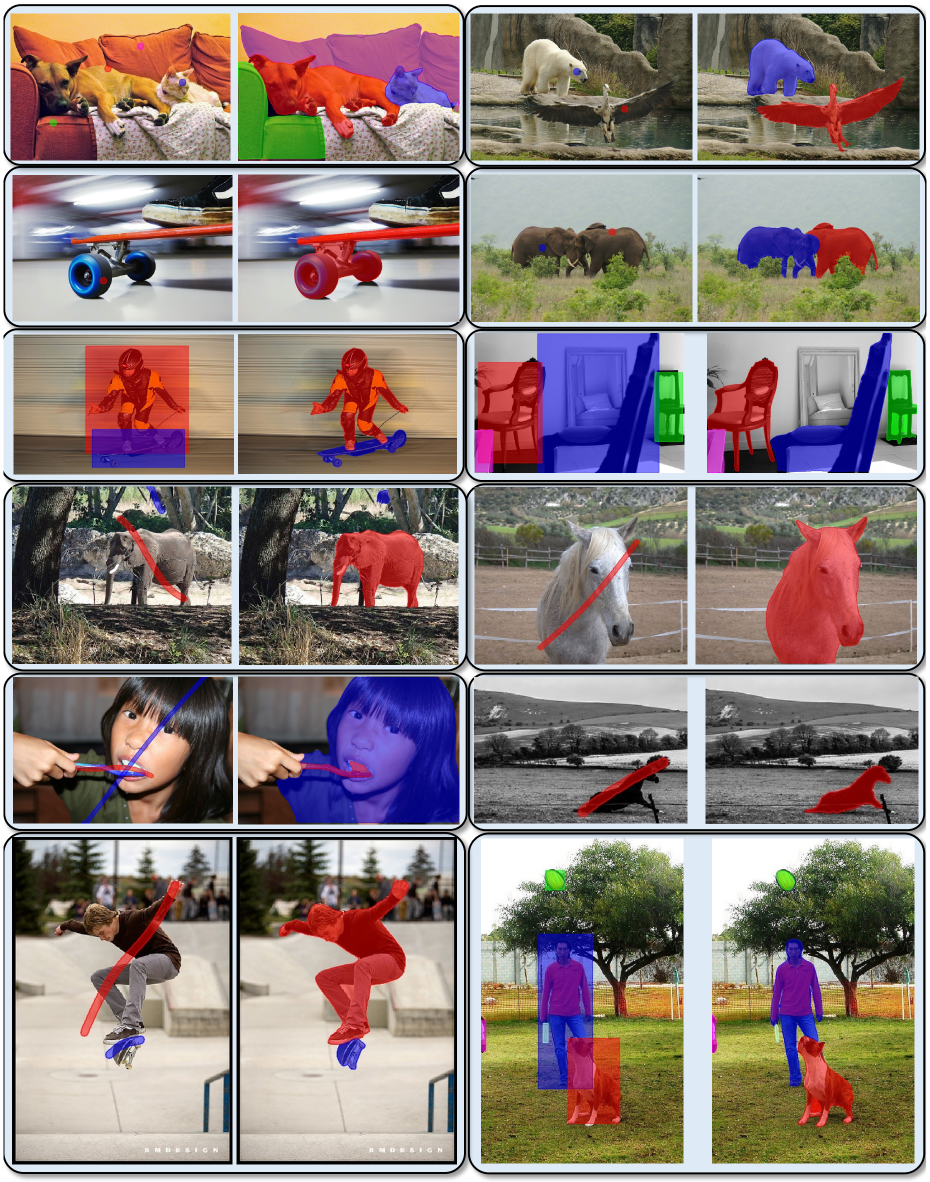


Fig. 8: More examples of interactive segmentation in COCO-Interactive.

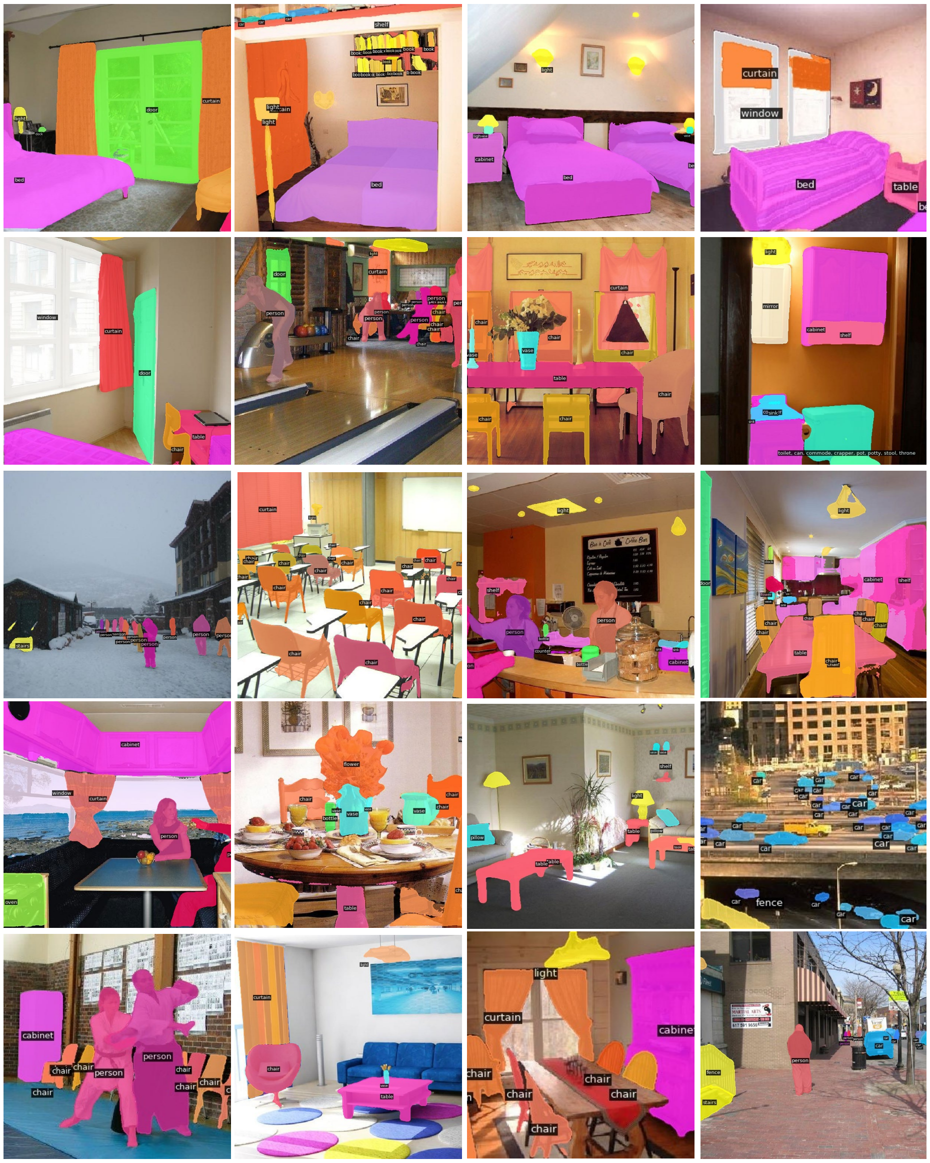


Fig. 9: More examples of open-vocabulary instance segmentation on ADE20K [25].

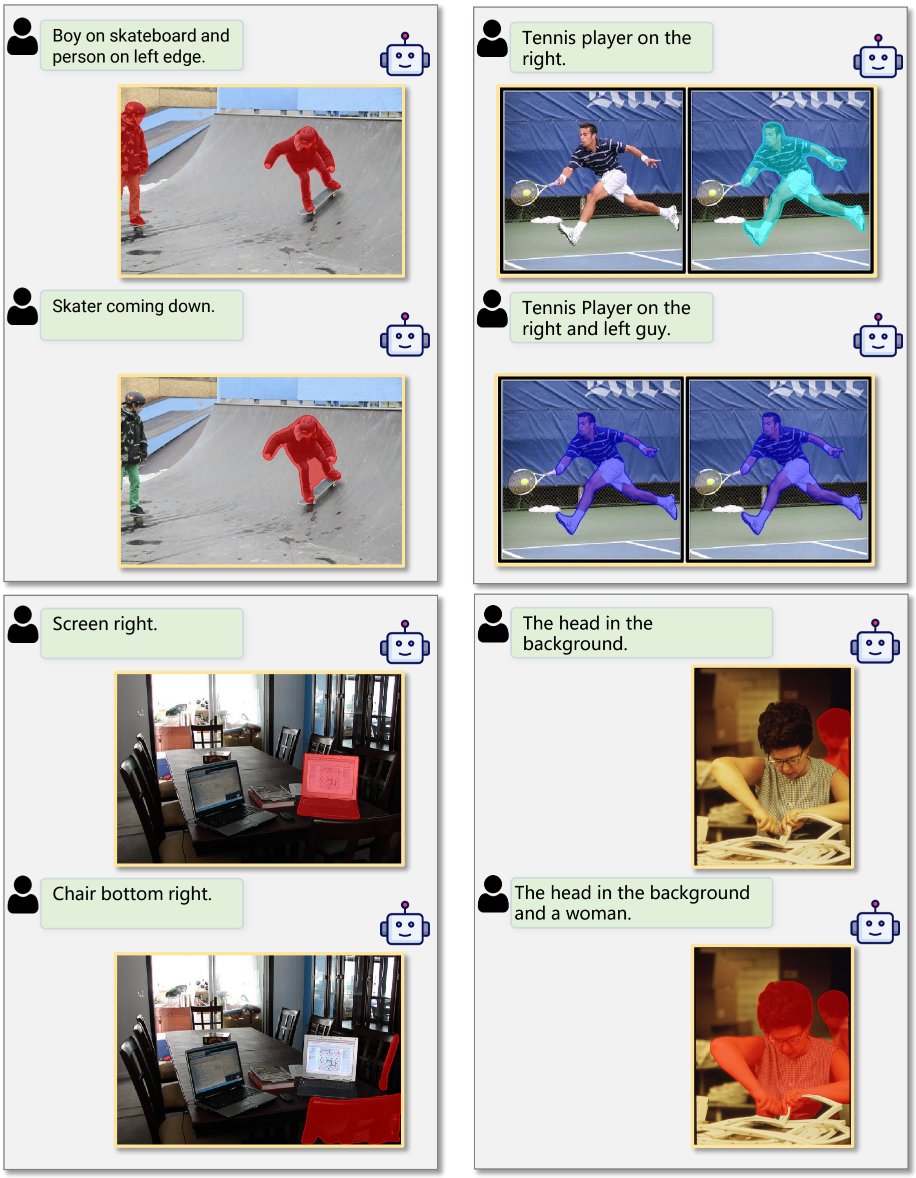
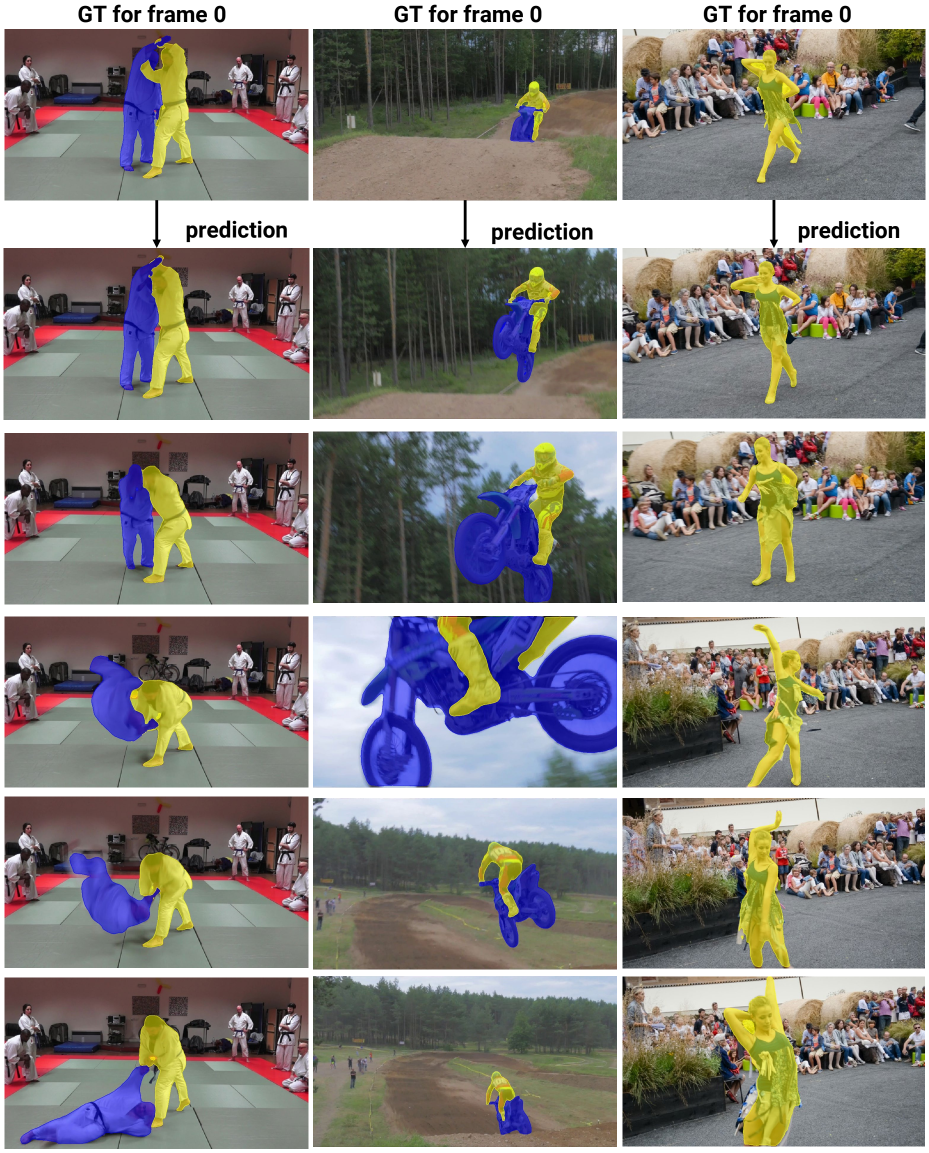

Boy on skateboard and person on left edge.
Skater coming down.
Tennis player on the right.
Tennis Player on the right and left guy.
Screen right.
Chair bottom right.
The head in the background.
The head in the background and a woman.

Fig. 10: More examples of generalized referring segmentation in gRefCOCO [14].


GT for frame 0
GT for frame 0
GT for frame 0
prediction
prediction
prediction

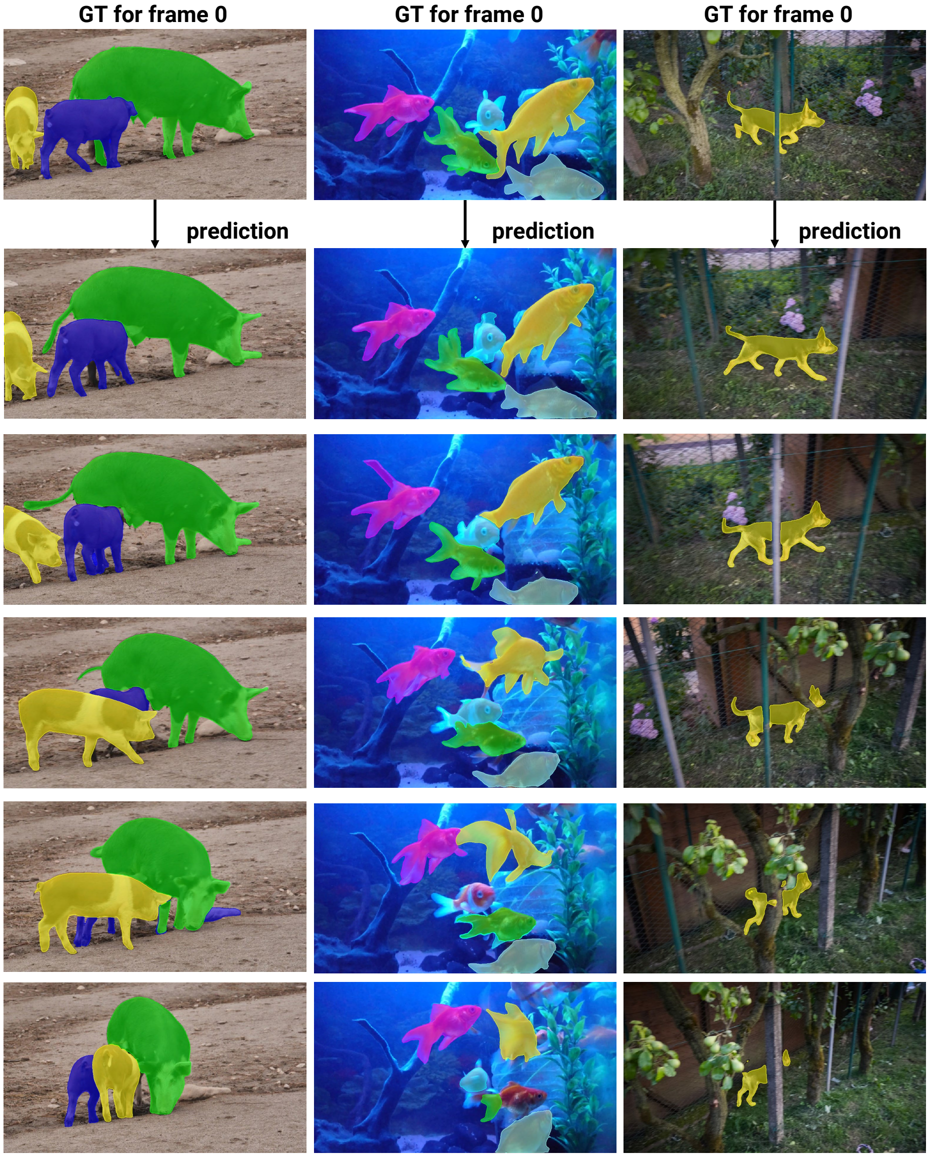


Fig. 11: More examples of video object segmentation in DAVIS *val* [20].

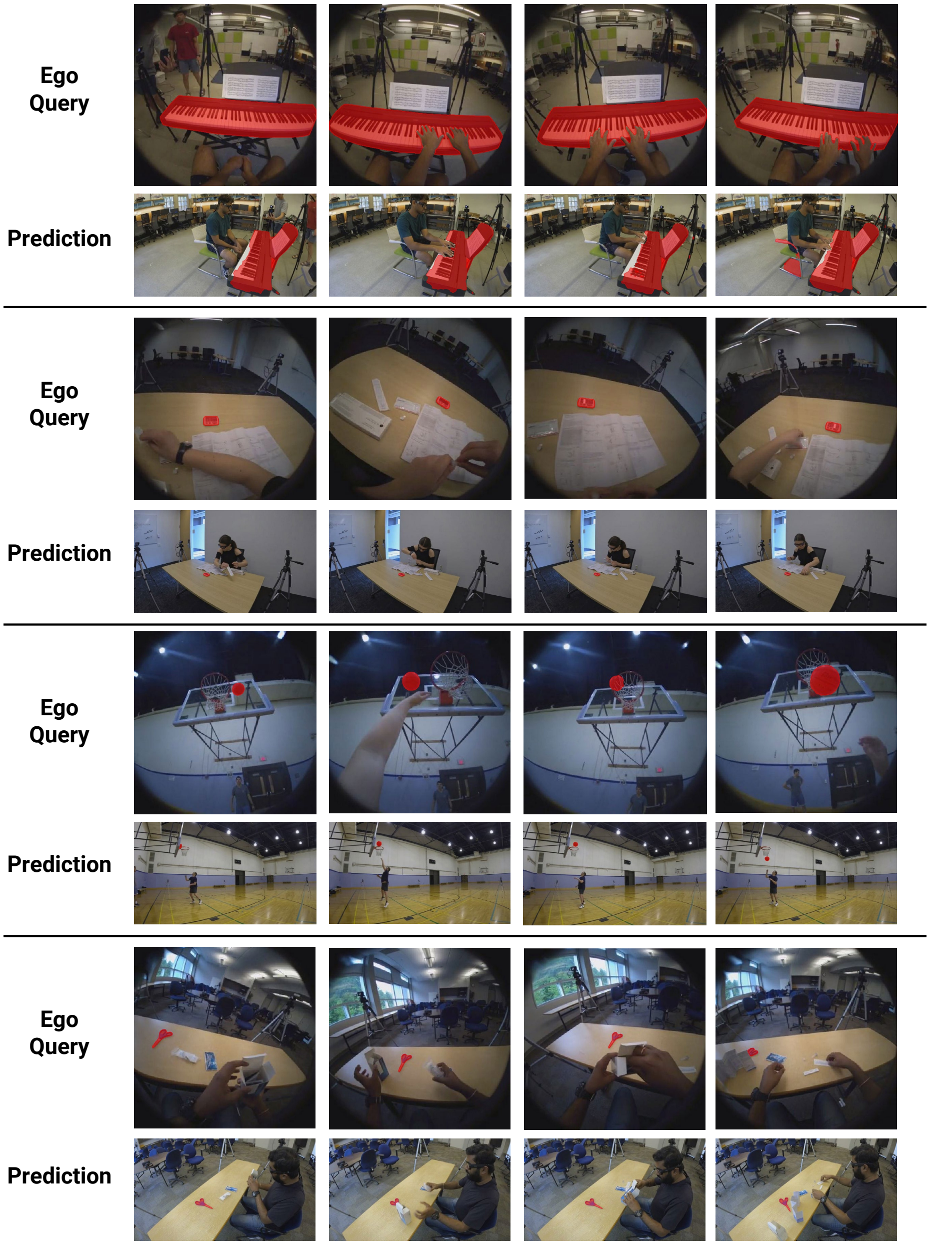


Fig. 12: More examples of Ego-exo correspondence in Ego-Exo4D [7].

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
3. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision. pp. 640–658. Springer (2022)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36** (2024)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html
7. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. arXiv preprint arXiv:2311.18259 (2023)
8. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)
9. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. Advances in Neural Information Processing Systems **36** (2024)
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
11. Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., Loy, C.C.: Omg-seg: Is one model good enough for all segmentation? arXiv preprint arXiv:2401.10229 (2024)
12. Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., Lee, Y.T.: Textbooks are all you need ii: phi-1.5 technical report. arXiv preprint arXiv:2309.05463 (2023)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
14. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23592–23601 (2023)
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)

16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
18. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 792–807. Springer (2016)
19. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
20. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
21. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. arXiv preprint arXiv:2312.02228 (2023)
22. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)
23. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
24. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 69–85. Springer (2016)
25. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
26. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)