UGLMM: Towards Unified Vision Grounding with Large Multimodal Model

Supplementary Material

Table 1. Region captioning on RefCOCOg [26].

Method	region prompt type	METEOR	CIDEr
GRiT [38]	box	15.2	71.6
Kosmos-2 [29]	box	14.1	62.3
GLaMM [32]	box	16.2	108.3
Osprey-7B [49]	mask	16.6	108.3
Groma-7B[25]	box	16.8	107.3
UGLMM-7B (Ours)	box	**17.0**	**110.0**

Table 2. Performance comparison on image captioning.

Model	Flickr30k		NoCaps	
	CIDEr	SPICE	CIDEr	SPICE
LEMON [8]	–	–	106.8	14.1
CoCa [44]	–	–	120.6	15.5
BLIP-2 [15]	–	–	**121.6**	**15.8**
LLAVA-1.5 [19]	81.3	-	103.6	-
GLaMM [32]	95.3	18.8	106.8	**15.8**
MGLMM [48]	104.6	22.7	112.6	15.2
UGLMM-7B (Ours)	**109.8**	**23.5**	110.6	14.9

Table 3. Comparison on the interactive segmentation task on COCO Interactive [51] dataset.

Method	Point		Scribble		Box		Mask	
	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU
SAM-B [11]	48.7	33.6	-	-	73.7	68.7	-	-
SAM-L [11]	51.8	37.7	-	-	76.6	71.6	-	-
SEEM-B [53]	47.8	57.8	43.0	44.0	44.9	42.1	48.4	65.0
OMG-Seg [17]	59.3	-	-	-	-	-	-	-
PSALM [51]	64.3	74.0	66.9	80.0	67.3	80.9	67.6	82.4
UGLMM-7B (Ours)	66.7	73.6	67.3	79.0	68.2	80.4	68.3	81.9

A. More Experiment Results

A.1. Results on More Tasks

In this section, we show the performance of UGLMM on more tasks, which shows that UGLMM achieves strong performance across a wide spectrum of tasks and validates its strong cross-task generalization capability.

Region Captioning. We compare the region captioning performance of different methods on RefCOCOg [26] in Tab. 1. The experiment results demonstrate that UGLMM surpasses previous counterparts on this task, demonstrating its strong regional perception ability.

Image Captioning. We present the comparative results of the image captioning task in Tab. 2. Following previous works [32, 48], we evaluate UGLMM on the commonly used Flickr30K [30] test set and NoCaps [1] validation set. The experimental results show that UGLMM has comparable image captioning capabilities compared to other counterparts. To ensure fair comparisons with counterparts like GLaMM and MGLMM, we adopt LLaVA [21] as our base LMM. It is important to note that in Tab. 2, BLIP-2 utilizes more than 400M image-text samples, including COCO Caption [4], Visual Genome [12], CC3M [35], CC12M [3], SBU [28], LAION400M [34] datasets and CoCa adopts about 3B samples from JFT-3B [50], ALIGN [9] and COCO Caption datasets. In contrast, LLaVA relies solely on the filtered CC-595K subset of CC3M and the LLaVA-Instruct-158K [21] dataset, due to which the image captioning capability of LLaVA is inferior to that of BLIP-2 and CoCa. Since our approach mainly focuses on unifying various vision grounding tasks, we do not specifically enhance the image captioning capability of the base model LLaVA during our joint training, resulting in lower performance compared to BLIP-2 and CoCa in this particular task. Moreover, as shown in Table 2 of SQ-LLaVA [36], LLaVA-1.5 [20], an improved version of LLaVA, achieves CIDEr scores of 81.3 on Flickr30K and 103.6 on NoCaps, which are lower than our model's scores of 109.8 and 110.6, respectively, which shows that our UGLMM achieves better performance than our base model LLaVA in image captioning, suggesting its good generalization ability for this task.

Interactive Segmentation. To present the interactive segmentation capability of UGLMM, we finetune and evaluate UGLMM on the COCO Interactive dataset [51] and compare it with other methods in Tab. 3. Compared to previous works, UGLMM also achieves strong performance on the interactive segmentation task.

Zero-shot Evaluation on Video Object Segmentation (VOS) and Referring Video Object Segmentation (R-VOS) Tasks. Because our work primarily focuses on image-level grounding, we have not involved any video training data or dedicated video modules. For VOS and R-VOS tasks, all other works in Tab. 4 utilize video training data, and most of them have dedicated video modules. There are only three methods in a zero-shot test setting on DAVIS-17 or Ref-DAVIS, including OMG-Seg, Video-LISA, and our UGLMM, but OMG-Seg adopts lots of video data for model training, Video-LISA utilizes both video training data and dedicated video modules, while UGLMM does not leverage any video training data or dedicated video modules. We treat the video object segmentation (VOS) task as an interactive segmentation task. During inference, we use the predicted masks of the previous frame (or GT masks of the first frame) as visual prompts to extract the

Table 4. Performance comparison on the video object segmentation (VOS) and referring video object segmentation tasks (R-VOS).

Method	zero-shot	video data	DAVIS-17			Ref-DAVIS		
			J&F	J	F	J&F	J	F
Unicorn [42]	✗	✓	69.2	65.2	73.2	-	-	-
Siam R-CNN [37]	✗	✓	70.6	66.1	75.0			
OMG-Seg [17]	✓	✓	76.9	-	-			
ReferFormer [39]	✗	✓	-	-	-	61.1	58.4	64.1
VideoLISA [2]	✓	✓	-	-	-	68.8	64.9	72.7
VideoGLaMM [27]	✗	✓	-	-	-	69.5	65.6	73.3
UGLMM-7B (Ours)	✓	✗	63.8	61.4	66.3	63.2	60.4	65.9

corresponding region embeddings of the last frame. Then, these region embeddings are leveraged to segment the objects in the current frame. For the referring video object segmentation (R-VOS) task, we treat it as the referring expression segmentation task, and obtain the segmentation results for each frame by inputting the referring expressions and each frame into the model. The results in Tab. 4 indicate that, without any video data for training or specialized video modules, our model demonstrates promising zero-shot performance on these two video segmentation tasks.

The fact that UGLMM can achieve competitive performance to specialist methods on image captioning, interactive segmentation, VOS, and R-VOS tasks is strong evidence of its cross-task generalization ability. This demonstrates our model's capability to transfer knowledge across different visual understanding tasks effectively with limited task-specific resources.

A.2. More Ablation Studies

Number of *[GND]* Tokens. For tasks that utilize the *[GND]* token, we propose to use multiple *[GND]* tokens to represent one object. This approach aims to enhance the feature representation capability of the *[GND]* tokens, thereby improving the quality of the grounding outputs. In Tab. 5, we validate the impact of different numbers of *[GND]* tokens. It can be seen that as the number of *[GND]* tokens increases, the accuracy on these tasks continues to improve. However, as the number increases, the improvement becomes smaller. By default, we set the number of *[GND]* tokens to 3.

Effect of Joint Training. Some previous works [14, 43, 51] indicate that multi-task joint training is effective for all tasks; however, some other studies [47, 52] point out that using shared model parameters for different tasks may lead to conflicts due to imbalanced parameter optimization during training. In Tab. 6, we conduct experiments to investigate the impact of joint training compared to task-specific training with our framework. Please note that, different from other ablation studies, we compare the final performance of UGLMM with task-specific trained variants in Tab. 6. The experimental results in Tab. 6 demonstrate that, compared to task-specific training, joint training improves the accuracy of all tasks, which demonstrates that our unified framework design equips UGLMM with a strong ability to perform various tasks effectively.

B. Experiment Details

B.1. Training Objective.

UGLMM is jointly trained on multiple tasks in an end-to-end manner. The training objective of our model consists of an autoregressive cross-entropy loss L_{text} for text generation and a grounding loss $L_{grounding}$ to supervise the prediction of grounding outputs. The $L_{grounding}$ contains a query classification loss L_{cls}, a mask BCE loss L_{bce} and a mask DICE loss L_{dice} for mask prediction, a box gIoU loss L_{giou} and a box L1 loss L_{l1} for box prediction. The total loss L is:

$$\begin{aligned} L &= L_{text} + L_{grounding} \\ &= L_{text} + \lambda_{cls} L_{cls} + \lambda_{bce} L_{bce} \\ &\quad + \lambda_{dice} L_{dice} + \lambda_{giou} L_{giou} + \lambda_{l1} L_{l1} \end{aligned} \tag{1}$$

Compared to existing works like LISA [13] and GLaMM [32], we only introduce two additional box-related losses (box gIoU loss and L1 loss) for box prediction, which is a common practice in traditional multimodal grounding methods, such as UNINEXT [43] and SEEM [53].

B.2. Datasets.

To equip UGLMM with various capabilities, we collect many datasets across a wide range of tasks for our joint training. These tasks and corresponding datasets are:

- image captioning and visual question answering: COCO Caption [4], Grand Caption [32], and LLaVA-150k [21].
- region captioning: RefCOCO [45], RefCOCO+ [45], RefCOCOg [26], RefCLEF [10], and GranD Referring Segment [32].
- referring expression segmentation and referring expression comprehension: RefCOCO [45], RefCOCO+ [45],

Table 5. Ablation study on the number of *[GND]* tokens. * denotes our default setting.

number	RefCOCOg val RES mask cIoU	RefCOCOg val REC box acc@0.5	gRefCOCO val GRES gIoU	Reason Seg val Reason Seg gIoU	Grandf val GCG Mask Recall
1	77.1	86.8	48.8	55.3	38.9
2	77.4	87.2	48.9	55.9	39.7
3*	77.6	87.4	49.3	56.3	40.5
5	77.6	87.5	49.5	56.6	40.1

Table 6. Ablation study on joint training.

joint training	RefCOCOg val RES mask cIoU	RefCOCOg val REC box acc@0.5	Grandf val GCG Mask Recall	COCO val Detection box AP	COCO val Pan. Seg. mask PQ	RefCOCOg Region Cap. CIDEr	NoCaps Image Cap. CIDEr
✗	78.5	89.1	44.4	52.7	55.6	105.3	101.2
✓	**79.8**	**90.0**	**47.4**	**54.6**	**56.5**	**110.0**	**110.6**

RefCOCOg [26], RefCLEF [10], and GranD Referring Segment [32].

- grounded conversation generation: MUSE [33], GranD-GCG [32], and GranDf [32].
- object detection and generic image segmentation: COCO Panoptic [18].
- interactive segmentation: COCO Interactive [51].

All these training datasets have been previously employed by existing works, which indicates that we do not introduce any unfair comparison through the use of new training data. We acknowledge that our method may use more datasets than some previous methods, but due to unifying a broader range of tasks, the inclusion of datasets from additional tasks is an inherent and necessary requirement for UGLMM. We argue that our increased training data resulting from task expansion should be viewed as an advantage of UGLMM rather than an unfair comparison, as our unified framework helps effectively unify and achieve promising performance across a wider range of tasks. Tab. 6 demonstrates the impact of joint training on the performance of individual tasks, which can effectively validate the efficiency of our approach.

B.3. Implement Details

If not specified, our model is trained with a joint training setting on all the collected datasets listed above and without additional task-specific fine-tuning. We train our model on 32 Tesla A100 GPUs (80GB) for 30k iterations with a batch size of 6 per device and gradient accumulation steps of 4. The training process takes about 5 days. During training, the image encoder and grounding encoder are kept frozen, the LLM is trained with LoRA [7], and all other parts are tunable. We utilize CLIP ViT-L/14 [31] as the image encoder, Vicuna-7B [5] as our large language model, Swin-L [23] as the grounding encoder, and MaskDINO [14] decoder as the grounding decoder. Our multimodal projector is composed of two linear layers and a GELU [6] layer. For the region encoder, we follow the design of Ospery [49].

During training, AdamW [24] optimizer is adopted with weight decay of 0, β_1 of 0.9, β_2 of 0.95, and initial learning rate of $3 \cdot 10^{-4}$. We use WarmupDecayLR as the learning rate scheduler and set warm-up steps to 200. The loss weights λ_{cls}, λ_{bce}, λ_{dice}, λ_{giou}, and λ_{l1} are set to 0.08, 0.1, 0.1, 0.04, and 0.1 respectively. We set the number of *[GND]* tokens of one object to 3. For the CLIP image encoder, we resize all input images to 336^2. For the grounding image encoder, we resize the longer side of the image to 1024 and then pad the shorter side to 1024 to keep the input resolution as 1024^2. For the grounding decoder, we set K_q to 300. For the tuning of the LLM, we utilize LoRA [7] and set the rank of LoRA to 8 and alpha to 16. To ensure the reproducibility of our experiment results, we set the same random seed of 42 during training and inference.

B.4. Data Format and Inference Process for Different Tasks

Tasks Only Requiring Textual Output. For these tasks (including image captioning and region captioning), our data format is the same as that of other LMMs. Below is an example of region captioning data samples:

> **User**: <IMAGE> Can you provide me with a detailed description of the region in the picture marked by <region>?
> **Assistant**: A black and white dog lying on a bed.

During inference, the image and text instructions are

input into the LMM, which then generates the textual responses.

Language-Referred Vision Grounding Tasks. For these tasks (including RES, REC, GRES, Reason Seg, and R-VOS), the objects that require grounding outputs are specified by the referring language expressions. Here is one data sample of this type of task:

> **User**: <IMAGE> Can you locate the man in white in this image?
> **Assistant**: Sure, it is the man in white [GND].

For these tasks, during inference, the image, text instruction, and referring expression are input into the LMM. The model then generates the text response along with grounding embeddings corresponding to *[GND]* tokens. Subsequently, the grounding embeddings are input into the decoder to obtain corresponding grounding outputs.

Visual-Prompt-Based Segmentation Tasks. For these tasks (including interactive segmentation and VOS), the objects that require grounding outputs are specified by visual prompts. One data sample is like:

> **User**: <IMAGE> Please segment by the given visual prompts. This is all the visual prompts: <prompt>, <prompt>, and <prompt>.
> **Assistant**: Sure, [GND], [GND], and [GND].

During the inference of these tasks, the image, text instruction, and visual prompts are first input into the LMM to obtain the grounding embeddings corresponding to the visual prompts, which are then input into the decoder to produce the grounding outputs corresponding to each visual prompt.

Tasks Requiring Both Textual Responses and Grounding Outputs. For the tasks (including GCG) that require both textual responses and corresponding grounding outputs, we utilize the following data format:

> **User**: <IMAGE> Could you please give me a detailed description of the image? Please respond with interleaved grounding outputs for the corresponding parts of the answer.
> **Assistant**: A person [GND] is standing on a playing field [GND], about to hit a sports ball [GND] with a baseball bat [GND]. The sports ball hovers over the playing field. Additionally, a wall [GND] is located beside the playing field, with a fence [GND] placed right next to it.

During inference, the model first receives the image and text instruction as input, then generates text responses along with interleaved *[GND]* tokens, and at last predicts the grounding outputs for each of the *[GND]* tokens with their corresponding grounding embeddings.

Category-Based Vision Grounding Tasks. For these category-based tasks, including object detection, semantic segmentation, instance segmentation, panoptic segmentation, and open-vocabulary image segmentation, we use the following data format:

> **User**: <IMAGE> Please detect and segment all the objects belonging to the category set in the image. The category set contains <categories>.
> **Assistant**: Sure, the image contains zebra-1 [GND], zebra-2 [GND], zebra-3 [GND], zebra-4 [GND], grass-merged [GND], and dirt-merged [GND].

During inference, we first input the image, text instruction, and category set into the LMM, from which we obtain the grounding embeddings corresponding to all categories in the input category set. These grounding embeddings, as well as selected grounding queries, are then fed into the grounding decoder to obtain grounding outputs corresponding to each grounding query and also the similarity scores between grounding embeddings and grounding queries. Subsequently, we utilize these similarity scores and grounding outputs to derive the correctly formatted outputs for these tasks with task-specific standardized inference processes like existing multimodal vision grounding methods [16, 22, 40, 41, 46].

C. More Qualitative Results

To better showcase the capabilities of our model, we provide more visualization results for various tasks: image captioning (Fig. 1), region captioning (Fig. 2), REC and RES (Fig. 3), Reason Seg (Fig. 4), object detection and instance segmentation (Fig. 5), panoptic segmentation (Fig. 6), open-vocabulary semantic segmentation (Fig. 7), grounded conversation generation (Fig. 8), interactive segmentation (Fig. 9), VOS (Fig. 10), and R-VOS (Fig. 11).

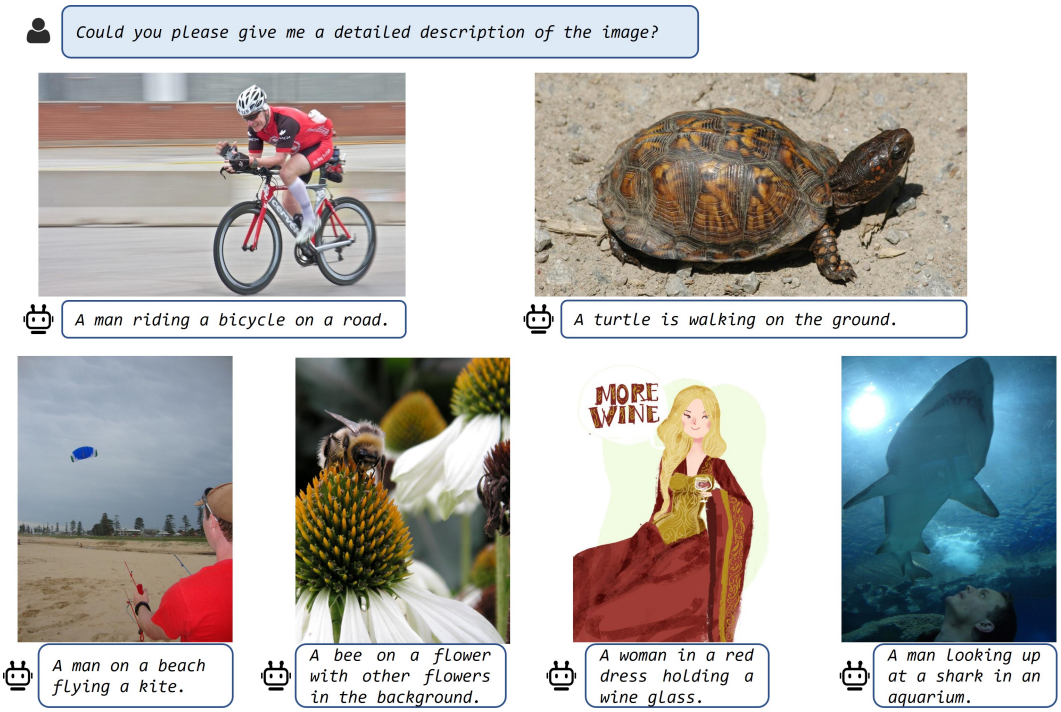

Figure 1. Visualization results on image captioning.

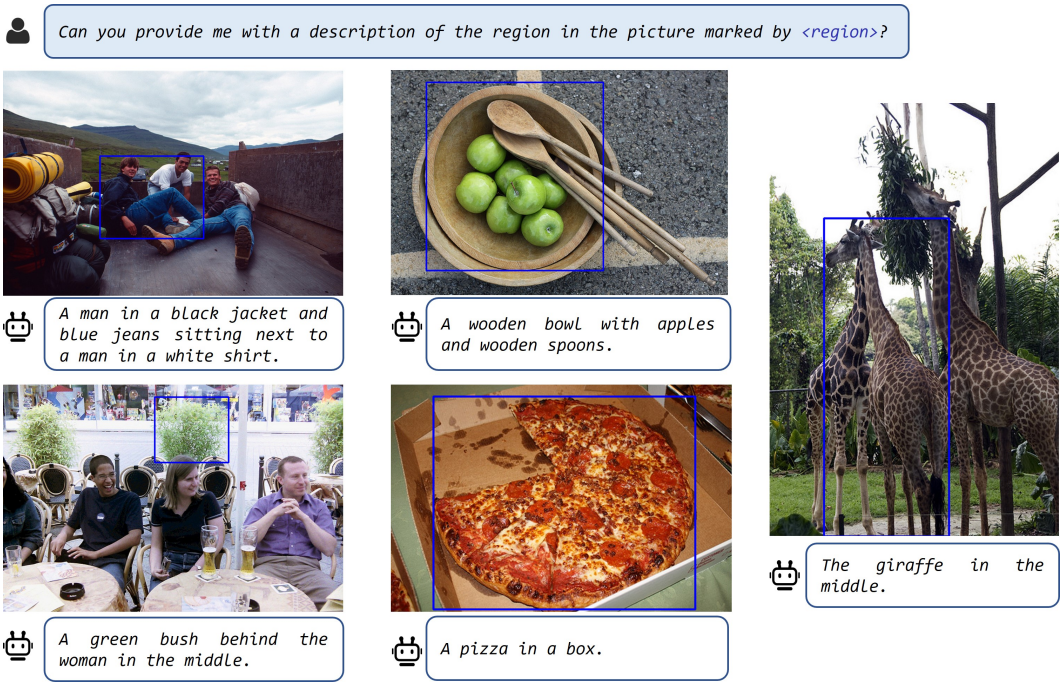

Figure 2. Visualization results on region captioning.

Figure 3. Referring expression comprehension and segmentation (REC and RES).

Figure 4. Qualitative results on reasoning segmentation (Reason Seg).

Please segment and detect all the objects belonging to the category set in the image. The category set contains [COCO categories].

Figure 5. Qualitative results on object detection and instance segmentation.

Please segment all the objects belonging to the category set in the image. The category set contains [COCO categories].

Figure 6. More visualization results on the panoptic segmentation task.

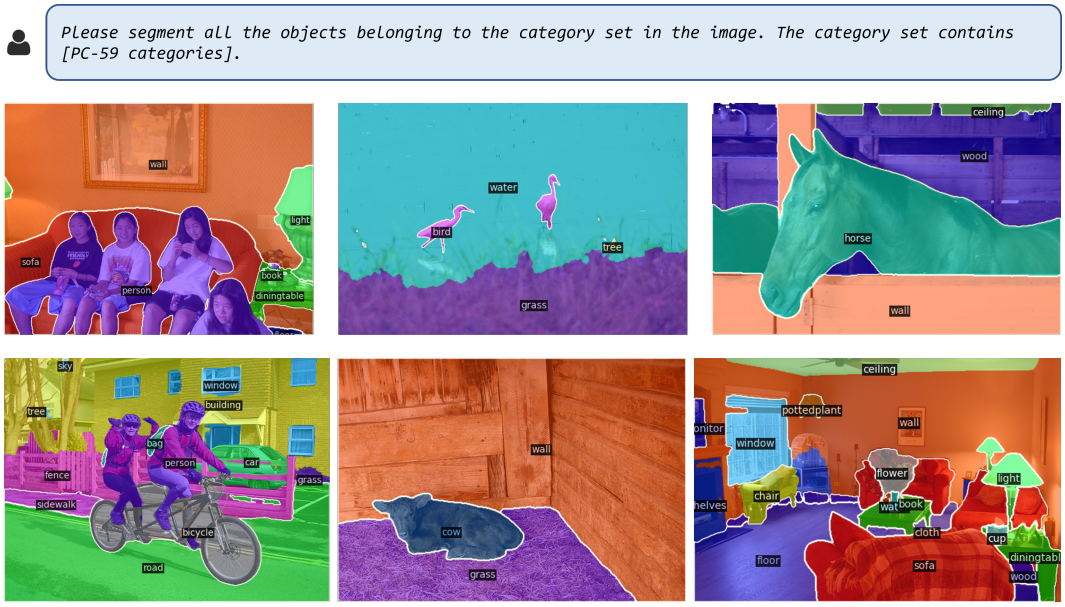

Figure 7. More examples of open-vocabulary semantic segmentation.

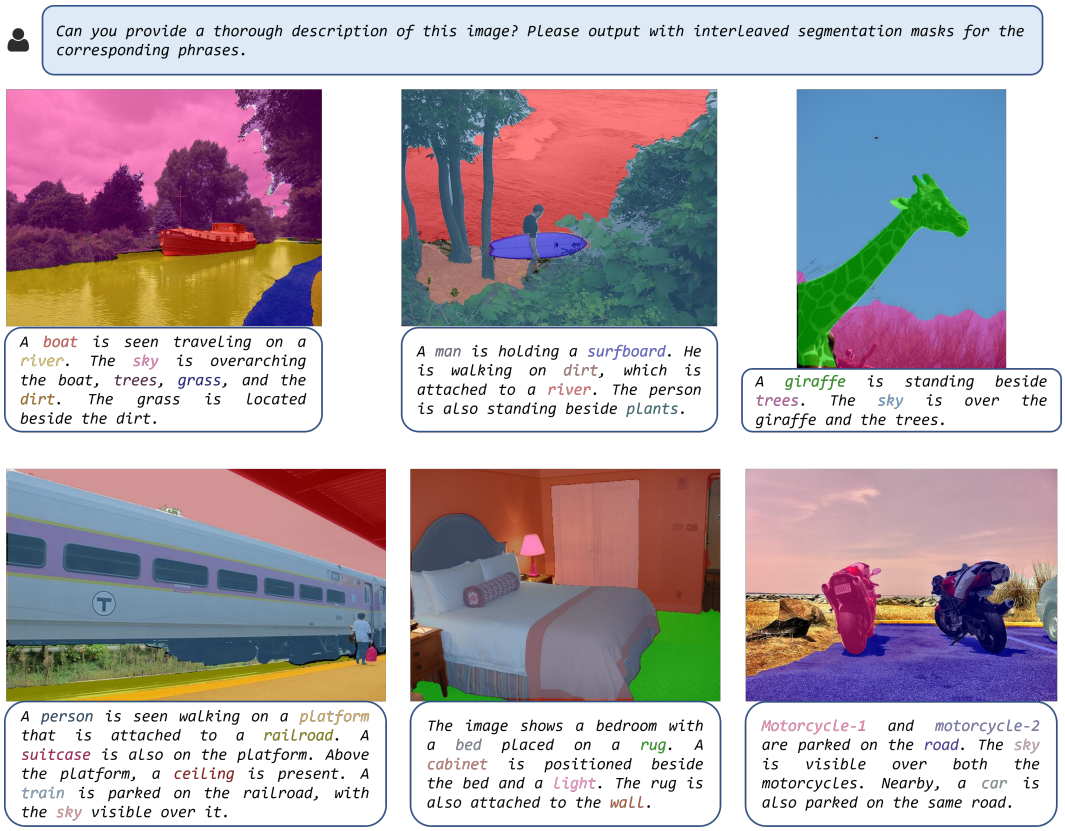

Figure 8. Qualitative results on grounded conversation generation.

Figure 9. More examples for interactive segmentation.

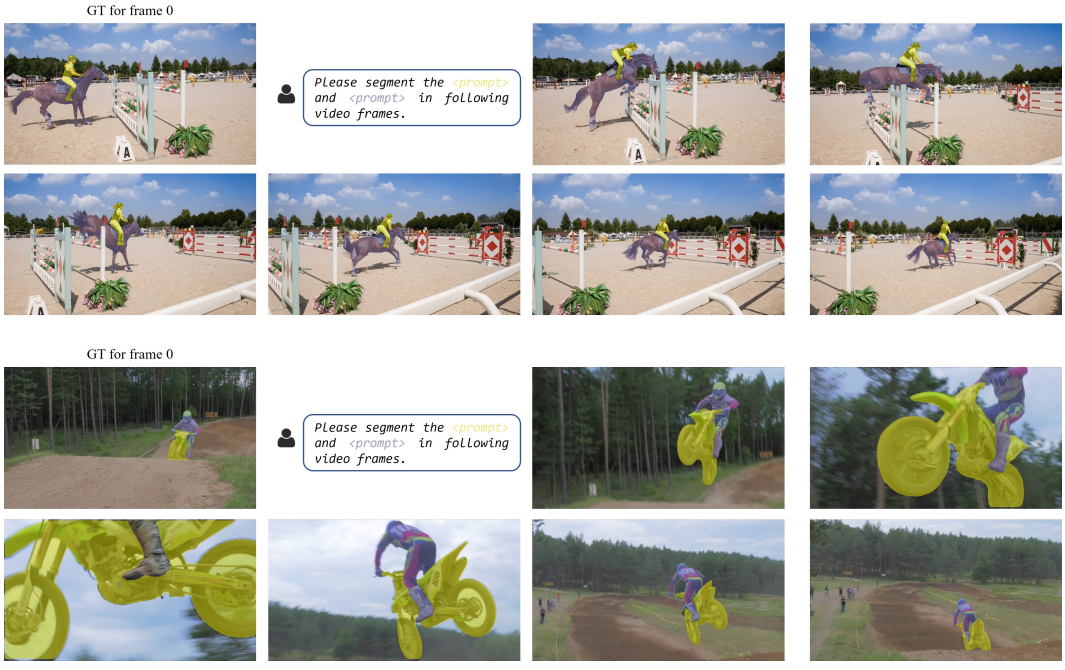

Figure 10. Visualization results of video object segmentation (VOS).

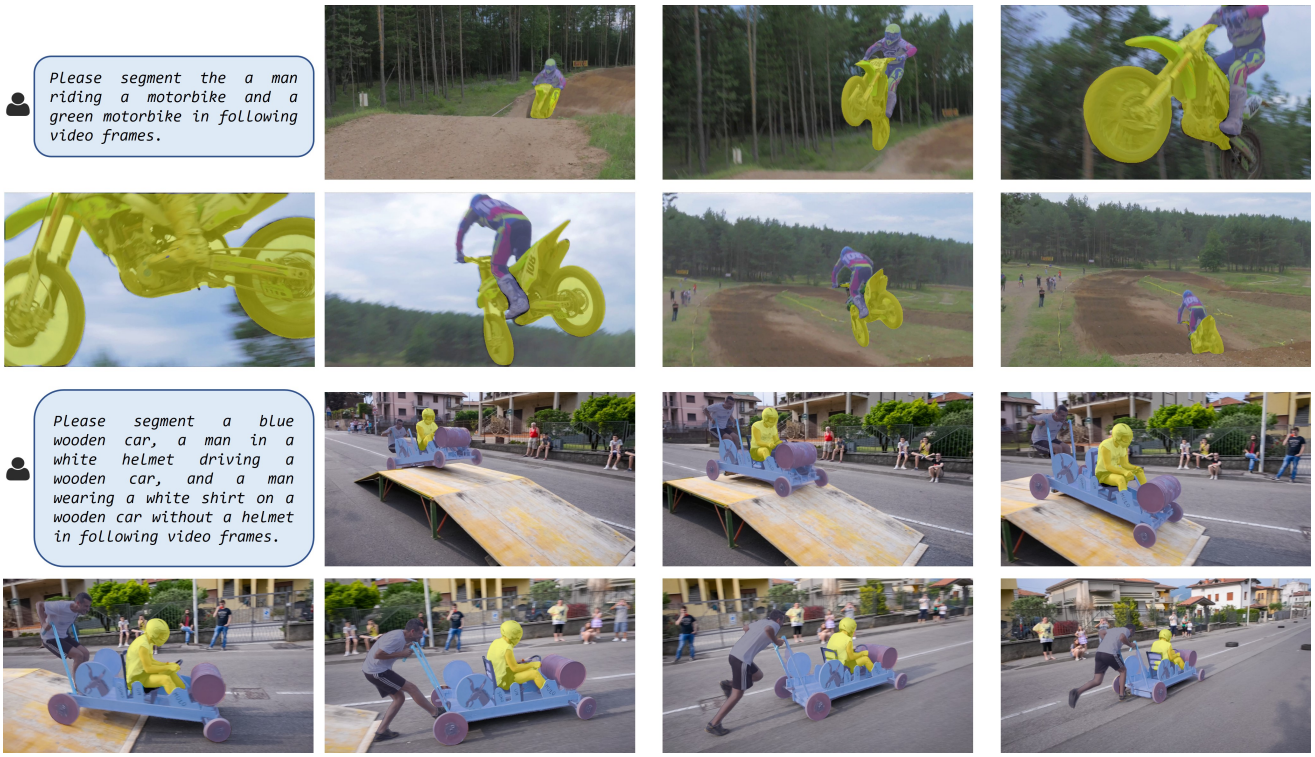

Figure 11. Qualitative results for referring video object segmentation (R-VOS).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 1

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 2

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. 1

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 3

[6] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[8] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989, 2022. 1

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 1

[10] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021. 2, 3

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[12] 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. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 1

[13] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[14] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023. 2, 3

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[16] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 4

[17] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27948–27959, 2024. 1, 2

[18] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 3

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE, 2024. 1

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE, 2024. 1

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 3

[25] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 1

[26] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2, 3

[27] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*, 2024. 2

[28] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. 1

[29] 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. 1

[30] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[32] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1, 2, 3

[33] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 3

[34] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 1

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. 1

[36] Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. Sq-llava: Self-questioning for large vision-language assistant. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, volume 15067 of *Lecture Notes in Computer Science*, pages 156–172. Springer, 2024. 1

[37] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 2

[38] 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. In *European Conference on Computer Vision*, pages 207–224. Springer, 2024. 1

[39] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2

[40] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2955–2966. IEEE, 2023. 4

[41] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2945–2954. IEEE, 2023. 4

[42] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European conference on computer vision*, pages 733–751. Springer, 2022. 2

[43] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 2

[44] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive

captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[45] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2

[46] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 4

[47] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 2

[48] Xu Yuan, Li Zhou, Zenghui Sun, Zikun Zhou, and Jinsong Lan. Instruction-guided multi-granularity segmentation and captioning with large multimodal model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9725–9733, 2025. 1

[49] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024. 1, 3

[50] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1204–1213. IEEE, 2022. 1

[51] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. *arXiv preprint arXiv:2403.14598*, 2024. 1, 2, 3

[52] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022. 2

[53] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023. 1, 2