Supplementary Materials: Scaling up Image Segmentation across Data and Tasks

Pei Wang[1], Zhaowei Cai[1], Hao Yang[1], Ashwin Swaminathan[1], R. Manmatha[2], Stefano Soatto[2]
[1]Amazon AGI [2]AWS AI Labs
{pwwng,zhaoweic,haoyng,swashwin,manmatha,soattos}@amazon.com

In this supplement, we show some other additional experimental results and details that are not present in the main paper due to the page limitation.

A. Supplementary Experimental Results

A.1. Full Results of Table 3

In Section 4.3 of the main paper, in order to investigate the generalization ability of MQ-Former for segmentation, we conduct a zero-shot evaluation of our model on the Segmentation in the Wild (SeginW) benchmark [18], which comprises 25 datasets, and report the average mAP of all the datasets. In this supplementary material, we report other additional results including median mAP and individual mAP on each dataset. The results detailed in Table A show the superiority of MQ-Former over X-Decoder [18] and OpenSeeD [17] across all datasets. This indicates that the importance of scalability across both datasets and tasks in enhancing the generalization ability of models, a capability unique to MQ-Former.

A.2. Explicit results of Figure 6

In Table B, we report the numerical results used to generate the five subfigures in Figure 6.

A.3. Ablation study

Enhancement by Synthetic Data Complementing Section 4.2, here we present more results for demonstrating the significance of synthetic data. 30% images are sampled from Objects365 [11] training set and synthetic mask is generated for each object with [4]. This subset is denoted as "Objects365-syn-m". We jointly train a model on COCO with instance annotation ("COCO ins") and Objects365-syn-m and compare with the baseline trained on "COCO ins" only. As shown in Table C, the improvement is clear, suggesting the benefit of using synthetic masks.

Similarly, synthetic object captions are generated for all COCO instances, denoted as "COCO-syn". We trained a model jointly on it with RefCOCOg. The comparison in Table D with the baseline shows that the improvement is significant (more than 4 points), indicating the benefits of synthetic captions.

The Impact of Query Numbers In this section, we ablate the impact of the number of queries. By default, we use mixture of 100 learnable and 300 conditional queries. This setting is derived from MaskDINO [7], ADE semantic setting of 100 learnable queries and COCO instance setting of 300 conditional queries. It is also the same as OpenSeeD using 100 learnable queries for stuff classes and 300 conditional queries for thing classes. Based on the Base-scale image and text encoder backbones, given different queries, we scale models with the configuration of two tasks and datasets. The training set is the combination of COCO with instance segmentation and ADE with semantic segmentation. In Table E, we observe that increasing the query number can improve the performance. However, the memory cost also increases considerably. Because such GPU memory cost is not affordable for our team when scaling up to large-scale backbones, in other experiments across the paper, we keep the "100+300" setting consistently. This also enables a fair comparison to other methods.

A.4. Qualitative Results

We present qualitative visualizations for open-set panoptic, instance and referring segmentation in Figures A, B, C and foreground segmentation in Figure D, respectively. The images are randomly selected from the web to provide a diverse and representative set for evaluation.

B. Model Size and Speed Comparison

We evaluate the model size in terms of the numbers of parameters (Params) and conduct a speed comparison by reporting frames-per-second (FPS). The speed tests are performed on A100 NVIDIA GPU with 40GB memory by taking the average computing time with batch size 1 on the entire validation set, using Detectron2 [13]. All models listed in Table F are characterized by large-scale backbone models. In general, there is no substantial difference in the forward speed across three models. The increase in parameters for both X-Decoder and our MQ-Former over OneFormer is primarily attributed to the introduction of a language encoder, given that they are open-vocabulary models.

C. Additional experimental details

Training settings For the experiments of Section 4.1, we train our model with a batch size of 32. AdamW is used as the optimizer with a base learning rate of 2e-4 for the segmentation encoder and decoder, and 2e-5 , 10 warmup iterations, and a weight decay of 0.05. We decay the learning rate at 0.9 and 0.95 fractions of the total number of training steps by a factor of 10. We train for a total of 50 epochs. On the experiments of Sections 4.2 and 4.3, we follow the same settings but the batch size is scaled up to 128. Swin-Base and CLIP-Base are used for query comparison in Table 2. Their larger-scale variants are used in other sections. The codes and models will be released upon acceptance.

Datasets In order to mitigate the data leakage issue, we implement exclusion in our training data. Specifically, for the COCO 2017 training set, examples belonging to RefCOCO, RefCOCO+, RefCOCOg validation sets are excluded. Conversely, training examples from RefCOCO, RefCOCO+, RefCOCOg that overlap with COCO 2017 validation set are also excluded. Similar exclusion procedures are applied to LVIS training set, removing examples associated with the RefCOCO, RefCOCO+, RefCOCOg validation sets. Distinct data augmentation strategies are applied based on the type of training data. For instance, semantic and panoptic data, we follow the augmentation strategy of Mask DINO [7]. For referring segmentation data, the augmentation data is the same as instance segmentation but random clip is excluded. For foreground segmentation training data, we follow the data augmentation of InSPyReNet [5]. Different upsampling ratios for each dataset are applied during joint training, which are maintained as specified in Table G. In total, the MQ-Former is trained on around 2M distinct images examples and 57M mask annotations. It is noted that the comparison in Table 3 is a system-level comparison. The training data varies across each method. For instance, X-decoder [18] additionally incorporates image-text corpora in its training process.

D. Ethical Considerations

We discuss the ethical considerations from three aspects: **Environmental Impact:** Training MQ-Former requires significant computational resources. The environmental impact of such resource-intensive processes should be taken into account, and efforts should be made to develop more energy-efficient algorithms. **Transparency and Explainability:** Like other deep learning models, MQ-Former is also considered "black boxes" because it is challenging to understand how they reach specific decisions. Ensuring transparency and explainability is essential to build trust and accountability, especially in applications with significant consequences. **Bias and Fairness:** Like other machine learning models, image segmentation models can be biased based on the data they are trained on. If the training data is not diverse and representative, the model may perform poorly on certain demographics or groups, perpetuating existing biases. However, this problem can be resolved to a certain extent by MQ-Former thanks to its versatility of joint training on multiple diverse datasets and tasks.

E. Limitations

Recently, a newly emerging reasoning segmentation task has been introduced [6]. The task is designed to output a segmentation mask given a complex and implicit query text. For example, given an image with various fruits, the query is "what is the fruit with the most Vitamin C in this image". This task demands a level of reasoning typically handled by multi-modal Large Language Models. Currently, MQ-Former does not explicitly support this task. However, addressing this limitation is part of our agenda for future research.

References

[1] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *PAMI*, pages 569–582, 2015. 4

[2] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, pages 2989–2998, 2023. 4

[3] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 4

[4] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023. 1

[5] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *ACCV*, pages 108–124, 2022. 2

[6] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

[7] 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 *CVPR*, pages 3041–3050, 2023. 1, 2

[8] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *WACV*, pages 305–314, 2021. 4

[9] Lucy AC Mansilla and Paulo AV Miranda. Oriented image foresting transform segmentation: Connectivity constraints

Table A. Open-set segmentation comparison on the SeginW benchmark. We bold the best entry in each column.

Model	Med.	Avg.	Air-Par.	Bottles	Br. Tum.	Chicken	Cows	Ele.-Sha.	Eleph.	Fruits	Gar.	Gin.-Gar.	Hand	Hand-Metal	House-Parts	HH.-Items	Nut.-Squi.	Phones	Poles	Puppies	Rail	Sal.-Fil.	Stra.	Tablets	Toolkits	Trash	W.M
X-Decoder [18]	22.3	32.3	13.1	42.1	2.2	8.6	44.9	7.5	66.0	79.2	33.0	11.6	75.9	42.1	7.0	53.0	68.4	15.6	20.1	59.0	2.3	19.0	67.1	22.5	9.9	22.3	13.8
OpenSeeD [17]	38.7	36.1	13.1	39.7	2.1	82.9	40.9	4.7	72.9	76.4	16.9	13.6	92.7	38.7	1.8	50.0	40.0	7.6	4.6	74.6	1.8	15.0	82.8	47.4	15.4	15.3	52.3
MQ-Former	**43.0**	**43.4**	**14.4**	**44.4**	**3.3**	**85.2**	**45.0**	**15.0**	**75.2**	**80.4**	**33.1**	**20.9**	**94.4**	**44.6**	**7.8**	**54.2**	**69.5**	**16.0**	**24.2**	**78.0**	**4.4**	**27.8**	**84.5**	**49.3**	**23.2**	**35.5**	**59.4**

Table B. The performance improvement with data and task scaling up.

Subfigure 1				
Data		Task		Referring segmentation
Dataset	Size (M)	Type	Number	RefCOCOg (mIoU)
RefCOCO,RefCOCO+,RefCOCOg	0.06	Referring segmentation	1	57.8
RefCOCO,RefCOCO+,RefCOCOg, COCO-syn	0.16	Referring segmentation	1	60.8
RefCOCO,RefCOCO+,RefCOCOg, COCO-syn, 30% Objects365-syn	0.67	Referring segmentation	1	62.6
Subfigure 2				
Data		Task		Open-vocabulary segmentation
Dataset	Size (M)	Type	Number	SeginW (Mask AP)
COCO	0.1	Instance segmentation	1	29.4
COCO, ADE20K	0.12	Instance segmentation	1	30.0
COCO, ADE20K, 30% Objects365-syn-m	0.63	Instance segmentation	1	35.5
Subfigure 3				
Data		Task		Open-vocabulary segmentation
Dataset	Size (M)	Type	Number	SeginW (Mask AP)
COCO	0.1	Panoptic segmentation	1	29.6
COCO	0.1	Panoptic, instance segmentation	2	32.7
COCO	0.1	Panoptic, instance, referring segmentation	3	35.2
Subfigure 4				
Data		Task		Referring segmentation
Dataset	Size (M)	Type	Number	RefCOCOg (mIoU)
COCO	0.1	Panoptic, referring segmentation	2	63.4
COCO, ADE20K, RefCOCO,RefCOCO+,RefCOCOg, LVIS, VG, COCO-syn	0.3	Panoptic, instance, referring segmentation	3	64.3
COCO, ADE20K, RefCOCO,RefCOCO+,RefCOCOg, VG, fore, LVIS, Objects365-syn	2.2	Panoptic, instance, semantic, referring, foreground segmentation, object detection	6	67.2
Subfigure 5				
Data		Task		Open-vocabulary segmentation
Dataset	Size (M)	Type	Number	SeginW (Mask AP)
COCO	0.1	Panoptic, referring segmentation	2	33.2
COCO, ADE20K, RefCOCO,RefCOCO+,RefCOCOg, VG, fore	0.4	Panoptic, instance, referring, foreground segmentation	4	37.2
COCO, ADE20K, RefCOCO,RefCOCO+,RefCOCOg, VG, fore, 30% Objects365-syn	1.0	Panoptic, instance, referring, foreground segmentation	4	39.6
COCO, Objects365-syn	1.8	Panoptic, referring	2	41.3
COCO, ADE20K, RefCOCO,RefCOCO+,RefCOCOg, VG, fore, LVIS, Objects365-syn	2.2	Panoptic, instance, semantic, referring, foreground segmentation, object detection	6	43.4

with adjustable width. In *SIBGRAPI*, pages 289–296, 2016. 4

[10] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 4

[11] 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*, pages 8430–8439, 2019. 1

[12] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 4

[13] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[14] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *CVPR*, pages 11717–11726, 2022. 4

[15] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 4

[16] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, 2019. 4

[17] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, pages 1020–1031, 2023. 1, 3

[18] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, pages 15116–15127, 2023. 1, 2, 3, 4

Table C. The impact of synthetic masks

Training data	Mask AP	Box AP
COCO ins	49.7	55.3
COCO ins + Objects365-syn-m	50.5	56.8

Table D. The impact of synthetic captions

Training data	mIoU
RefCOCOg	57.8
syn-COCO	58.8
RefCOCOg + COCO-syn	62.6

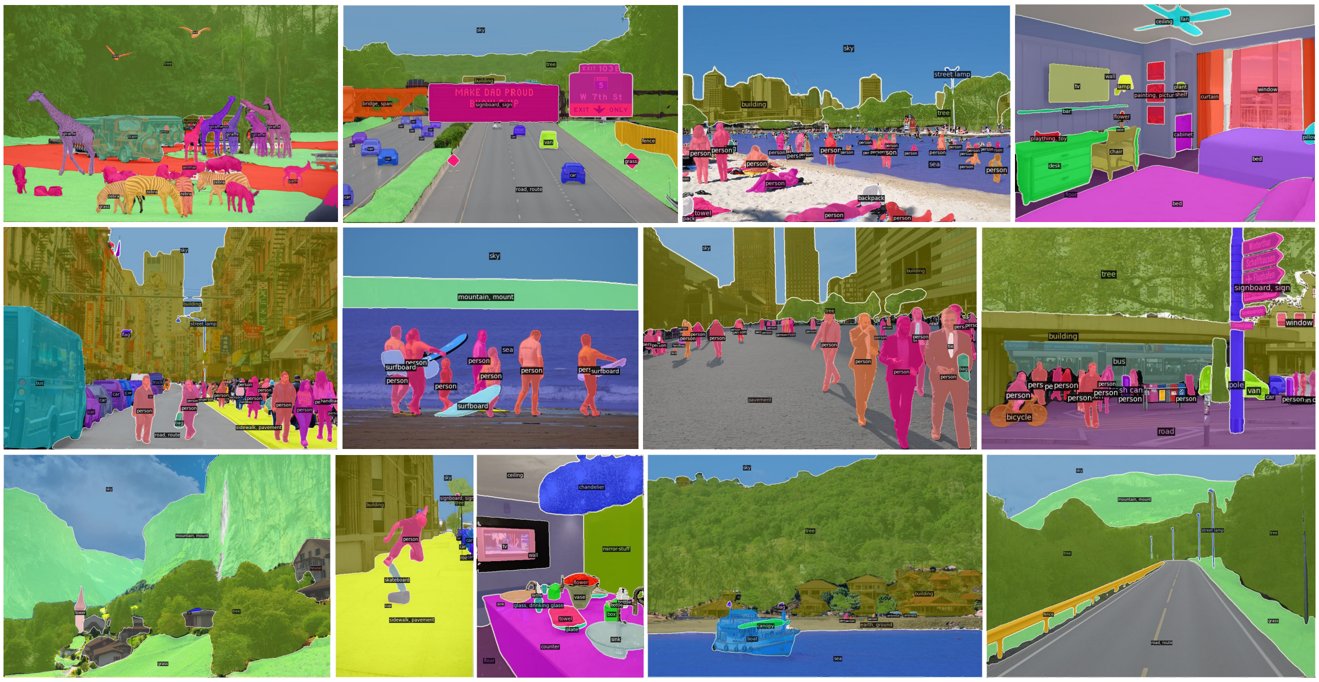

Figure A. Qualitative visualization on open-set panoptic segmentation.

Table E. The impact of query numbers.

#learnable+#conditional	ADE	COCO	
	mIoU	Mask AP	Box AP
100+300	51.7	49.6	54.9
300+900	52.0	50.7	57.4

Table F. The model size and speed comparison.

Method	Params	FPS
OneFormer [2]	219M	5.6
X-Decoder [18]	280M	6.1
MQ-Former	286M	5.1

Table G. Upsampling ratio of joint training. “referring” refers to the combination of RefCOCO, RefCOCO+, RefCOCOg [3, 15]. “foreground” refers to the combination of seven foreground datasets, HRSOD [16], DIS [10], THUS [1], COIFT [9], ThinObjects5K [8], UHRSD [14], DUTS [12].

Dataset	Ratio	#Images	#Annotations
COCO	3	100K	1.3M
ADE20K	30	20K	271K
LVIS	3	100K	1.3M
Visual Genome	9	100K	2.3M
Objects365	1	1.7M	25M
referring	6	54K	124K
syn-COCO	3	100K	1.3M
syn-Objects365	1	1.7M	25M
foreground	9	100K	100K

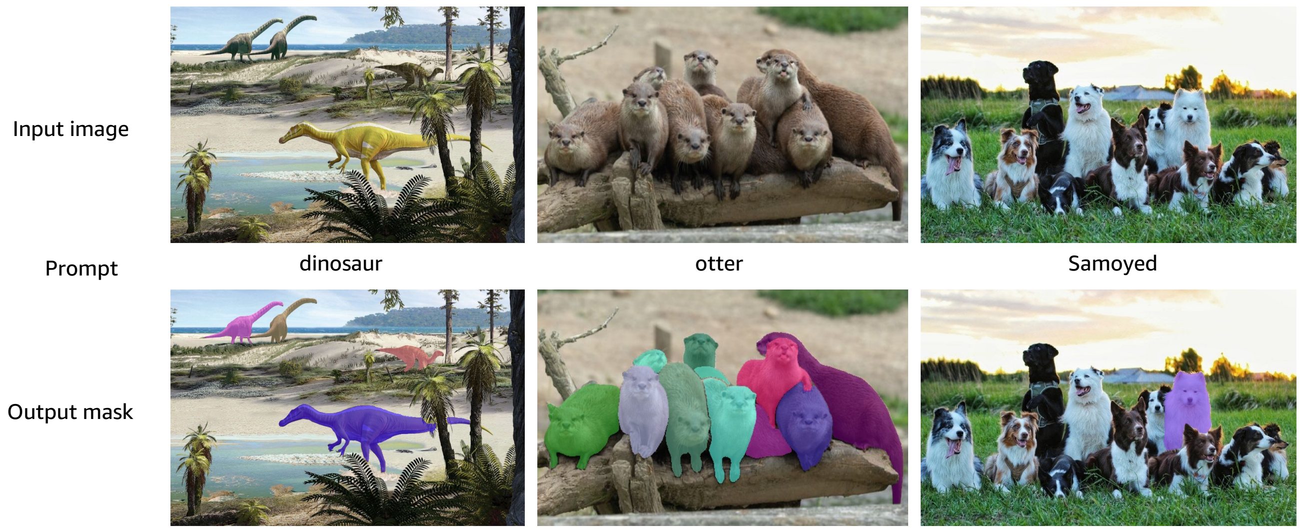

Figure B. Qualitative visualization on open-set instance segmentation.

Figure C. Qualitative visualization on open-set referring segmentation.

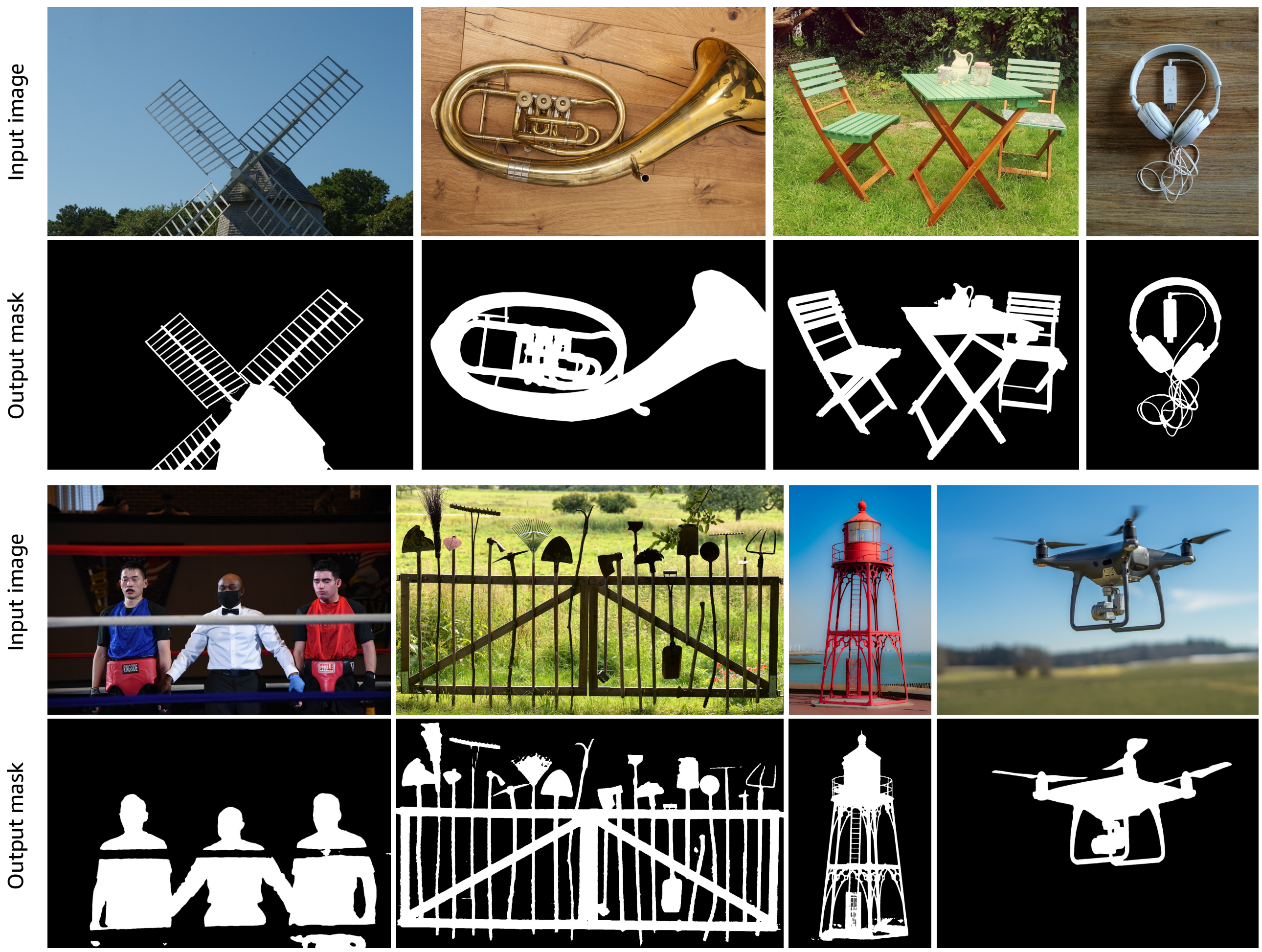

Figure D. Qualitative visualization on foreground segmentation.