Supplementary Material: Segment Anything in High Quality

Lei Ke[*1,2] **Mingqiao Ye**[*1] **Martin Danelljan**[1] **Yifan Liu**[1] **Yu-Wing Tai**[3]
Chi-Keung Tang[2] **Fisher Yu**[1]
[1]ETH Zürich [2]HKUST [3]Dartmouth College

In this supplementary material, Section 1 first presents the additional experimental analysis of our HQ-SAM, including more zero-shot transfer comparisons to SAM on both image and video benchmarks. Then, in Section 2, we describe more details of our method implementation, including the training and inference. In Section 3, we provide further details of our constructed HQSeg-44K dataset for training HQ-SAM. In Section 4, we show extensive visual results comparison between our HQ-SAM and SAM on COCO [9], DIS-test [11], HR-SOD [16], NDD20 [13], DAVIS [10], and YTVIS [15].

1 Supplementary experiments

SAM vs. HQ-SAM on Various Backbones In Table 1, we provide a comprehensive comparison between HQ-SAM and SAM using various backbones, including ViT-B, ViT-L, ViT-H and TinyViT. The comparison not only includes the numerical results on the four HQ datasets and COCO validation set, but also contains the model sizes/speed/memory. HQ-SAM consistently outperforms SAM using three different backbones, with over 10 points increase in mBIoU on the four HQ datasets. Notably, the ViT-B based HQ-SAM significantly improves the AP^B on COCO from 28.2 to 31.3 and AP from 44.4 to 46.7, with only a 1.1% increase in model parameters and negligible extra memory consumption.

Table 1: SAM vs. HQ-SAM on various ViT backbones. For the COCO dataset, we use a SOTA detector FocalNet-DINO [17] trained on the COCO dataset as our box prompt generator.

Model	Four HQ datasets		COCO					Model Params (MB)		FPS	Memory
	mIoU	mBIoU	AP_B	AP	AP_L	AP_M	AP_S	Total	Learnable		
SAM-B	70.6	62.3	28.2	44.4	57.7	48.7	32.1	358	358	10.1	5.1G
HQ-SAM-B	**86.3**	**78.1**	**31.3**	**46.7**	62.9	50.5	32.0	362.1	**4.1**	9.8	5.1G
SAM-L	79.5	71.1	33.3	48.5	63.9	53.1	34.1	1191	1191	5.0	7.6G
HQ-SAM-L	**89.1**	**81.8**	**34.4**	**49.5**	66.2	53.8	33.9	1196.1	**5.1**	4.8	7.6G
SAM-H	75.6	68.3	34.0	48.9	64.5	53.3	34.4	2446	2446	3.5	10.3G
HQ-SAM-H	**89.3**	**81.5**	**34.9**	**49.9**	66.5	54.0	34.2	2452.1	**6.1**	3.4	10.3G
MobileSAM	69.0	58.8	28.6	44.3	-	-	-	38.6	38.6	44.8	3.7G
Light HQ-SAM	**81.4**	**71.6**	**29.6**	**45.0**	-	-	-	40.3	**1.7**	41.2	3.7G

Table 2: Results on YouTubeVIS 2019 validation set and HQ-YTVIS test set using ViT-L based SAM. We adopt the SOTA detector Mask2Former [1] trained on the YouTubeVIS 2019 dataset as our video boxes prompt generator while reusing its object association prediction.

Model	YTVIS 2019						HQ-YTVIS	
	AP	AP_{50}	AP_{75}	AP_L	AP_M	AP_S	AP^B	AP^M
SAM	51.8	82.1	55.4	65.5	52.0	34.2	30.2	60.7
HQ-SAM	**53.2**	82.9	58.3	66.4	53.3	33.7	**34.0**	**63.6**

Zero-shot Video Instance Segmentation Comparison Extending from Table 8 of the paper (evaluation on the HQ-YTVIS benchmark [4]), we further perform a comparative analysis of zero-

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

shot video instance segmentation results on the popular YTVIS 2019 [15] validation set. We take the pre-trained Mask2Former [1] as our video box prompts and feed them into SAM and our HQ-SAM for mask prediction. In Table 2, HQ-SAM achieves consistent gains of 1.4 points in Tube Mask AP, increasing SAM's performance from 51.8 to 53.2. Interestingly, we find the AP_{75} improvement with a higher IoU threshold for HQ-SAM is much larger than AP_{50}, further validating the advantages of HQ-SAM in high-quality mask prediction.

Zero-shot Video Object Segmentation Comparison Besides video instance segmentation, in Table 3, we further report the comparison of video object segmentation results between HQ-SAM and SAM on DAVIS validation set in a zero-shot transfer protocol. We take the pre-trained XMem as our video box prompts and feed the same prompts into SAM and HQ-SAM. HQ-SAM improves SAM the $\mathcal{J}\&\mathcal{F}$ from 82.0 to 83.2 and the $\mathcal{F}$ score from 84.9 to 86.1, where $\mathcal{F}$ is for measuring the contour accuracy of the video objects.

Table 3: Results on DAVIS 2017 [10] validation set using ViT-L based SAM. We adopt the SOTA model XMem [2] as our video boxes prompt generator while reusing its object association prediction.

Model	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAM	82.0	79.0	84.9
HQ-SAM	**83.2**	**80.3**	**86.1**

Robustness to Input Box Prompts In Table 4, we compare HQ-SAM to SAM by adding various scales of noises to the input ground truth box prompts. In practice, we cannot expect the input box prompts provided by humans in interactive modes to be identical to the ground truth (GT) boxes or extremely accurate. We follow the data augmentation code in DN-DETR [6] to add different noise scales and identify that our HQ-SAM is much more robust compared to SAM, where the relative mBIoU advantage improves from 10.7 to 20.5 when gradually increasing the noise scales. Note that our method is not trained with noised boxes. We also visualize such noised input case in Figure 4, where SAM is more sensitive to small box location shifts that easily happened during interactive annotation.

Table 4: Comparison of segmentation accuracy on the four HQ datasets by adding various noise levels to the GT box prompts input.

Model	No Noise		Noise scale 0.2		Noise scale 0.4	
	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
SAM	79.5	71.1	65.7	57.1	46.4	39.8
HQ-SAM	89.1	**81.8**$_{\uparrow 10.7}$	82.8	**73.4**$_{\uparrow 16.3}$	69.9	**60.3**$_{\uparrow 20.5}$

2 Additional Implementation details

Training Details During training HQ-SAM on the composed HQSeg-44K, we fix the model parameters of the pre-trained SAM model while only making the proposed HQ-SAM learnable, including HQ-Output Token, its associated three-layer MLP and three convolutions for HQ-Features fusion. Two of them are transposed convolutions (size 2×2, stride 2) used to upscale encoder embedding size from 64×64 to 256×256. We treat the new HQ-Output Token as the fifth mask token compared to the original four mask tokens in SAM's mask decoder. During training, this new HQ-Output token of size 1×256 is concatenated with SAM's mask tokens (size of 4×256), iou token (size of 1×256) and prompt tokens (size of N_{prompt}×256) as the input to the SAM's mask decoder. For example, if the input image contains N box prompts (size N×2×256), the final concatenated input and output shape for the 2-layer mask decoder of SAM is N×(1+4+1+2)×256. For experiments using ViT-B, ViT-L, and ViT-H-based models on training, we adopt the same training setting, with a learning rate of 1e-3 and train our HQ-SAM for 12 epochs (learning rate drops to 1e-4 after 10 epochs). We supervise mask prediction of the new HQ-Output token with a combination of both BCE Loss and Dice Loss.

Implementation Details We follow the same inference pipeline of SAM but use the mask prediction from HQ-Output token as high-quality mask prediction. Table 1 reports the detailed inference speed

comparison using various backbones. For box-prompting-based evaluation, we feed SAM and our HQ-SAM with the same image/video bounding boxes and adopt the single mask output mode of SAM. For interactive segmentation comparison using a single point, we follow SAM and adopt the "center" point of Ground Truth (GT) masks, which is at a maximal value location in a mask's interior distance transform. For multiple-point evaluation, we randomly sample the points from the GT masks and report the averaged results with three trials.

3 More Details of HQSeg-44K

Data compostion of HQSeg-44K In Table 5, we provide more details of our composed new training dataset HQSeg-44K which contains 44,320 extremely accurate image mask annotations, where we show their annotation quality in Figure 1. HQSeg-44K is a collection of six existing image datasets including DIS [11] (train set), ThinObject-5K [8] (train set), FSS [7], ECSSD [12], MSRA-10K [3], DUT-OMRON [14] with extremely fine-grained mask labeling, where each of them contains 7.4K mask labels on average. This composed training set has no images/annotations overlapping with the zero-shot evaluation datasets adopted in our paper.

Effect of HQSeg-44K In Table 6, we show the advantage of using HQSeg-44K by comparing HQ-SAM training with 44K randomly sampled images and masks from SA-1B [5]. Using the same efficient token learning strategy, training with SA-1B (44K) decreases the averaged mBIoU on the four datasets from 71.1 to 70.1, while ours improves it from 71.1 to 81.8. This validates the effectiveness of our constructed HQSeg-44K benchmark in improving mask quality. Note that the ablation experiments in Table 2, Table 3, Table 4 and Table 9 of the paper are all based on the constructed HQSeg-44K.

Table 5: Data composition of our constructed HQ-Seg-44K.

Dataset	DIS [11]	Thin-Object 5k [8]	FSS [7]	DUTS [14]	ECSSD [12]	MSRA-10K [3]	Total
Image Num.	3000	4748	10000	15572	1000	10000	44320

Table 6: Comparison of the training dataset. For the COCO dataset using ViT-L-based SAM, we use a SOTA detector FocalNet-DINO [17] trained on the COCO dataset as our box prompt generator.

Model	Dataset	DIS		COIFT		HRSOD		ThinObject		Average	
		mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
SAM	SA-1B	62.0	52.8	92.1	86.5	90.2	83.1	73.6	61.8	79.5	71.1
HQ-SAM	+ SA-1B-44K	60.4	51.7	91.1	86.1	88.4	80.9	73.1	61.8	78.3	70.1
HQ-SAM	+ HQ-Seg-44K (Ours)	78.6	70.4	94.8	90.1	93.6	86.9	89.5	79.9	89.1	81.8

Zero-shot results on DIS and ThinObject-5K We also report zero-shot results in Table 7 on DIS and ThinObject-5K by removing the training splits of either or both datasets from the training of HQ-SAM. The improvement of HQ-SAM over SAM is still substantial on DIS or ThinObject (over 10.0 points on DIS-mIoU and 9.0 points on ThinObject-mIoU), even when the corresponding training splits are removed from training.

Table 7: Zero-shot results on DIS and ThinObject-5K by removing the training splits of either or both datasets from the training of HQ-SAM. Results not obtained in a zero-shot manner (i.e. the training split was used), are shown in parenthesis to easily compare zero-shot results.

Training Setting	DIS-mIoU	DIS-mBIoU	ThinObject-mIoU	ThinObject-mBIoU
SAM (**baseline**)	62.0	52.8	73.6	61.8
HQ-SAM (**remove both DIS and ThinObject**)	72.9	63.1	82.7	70.7
HQ-SAM (**remove DIS**)	74.7	66.2	(90.1)	(80.4)
HQ-SAM (**remove ThinObject**)	(78.4)	(70.3)	83.3	72.1
HQ-SAM (**default HQSeg-44K**)	(78.6)	(70.4)	(89.5)	(79.9)

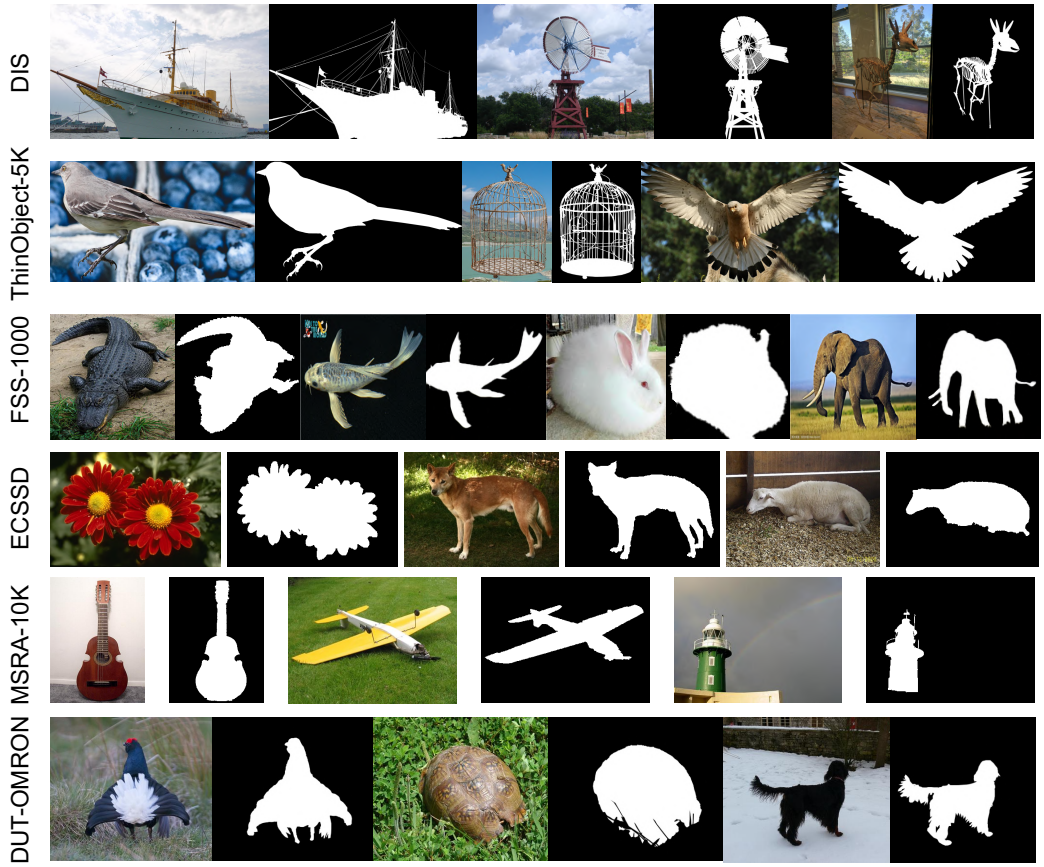

Figure 1: Visualization of annotated mask quality for randomly selected cases from the six dataset components of the HQ-Seg-44K. Zoom in for better viewing the fine-grained mask details.

4 More Visual Results Comparison

We provide more extensive visual results comparison in Figure 2 (DIS [11] test set), Figure 3 (zero-shot setting in COCO), Figure 4 (noised box input) and Figure 5 (zero-shot setting in HRSOD [16], NDD20 [13] and web images which cover objects with various structure complexities in diverse environments. In Figure 6 and Figure 7, we provide the zero-shot video segmentation results comparison on DAVIS 2017 and YTVIS 2019 benchmarks respectively. Besides, we include the dark underwater environment in NDD20 [13] and randomly selected web images in Figure 5, showing that the zero-shot segmentation power in SAM is well preserved by HQ-SAM. In Figure 5, we also include two failure cases in the rightmost two columns of the third row and bottom row, where HQ-SAM improves over SAM, but still cannot achieve fully correct mask prediction.

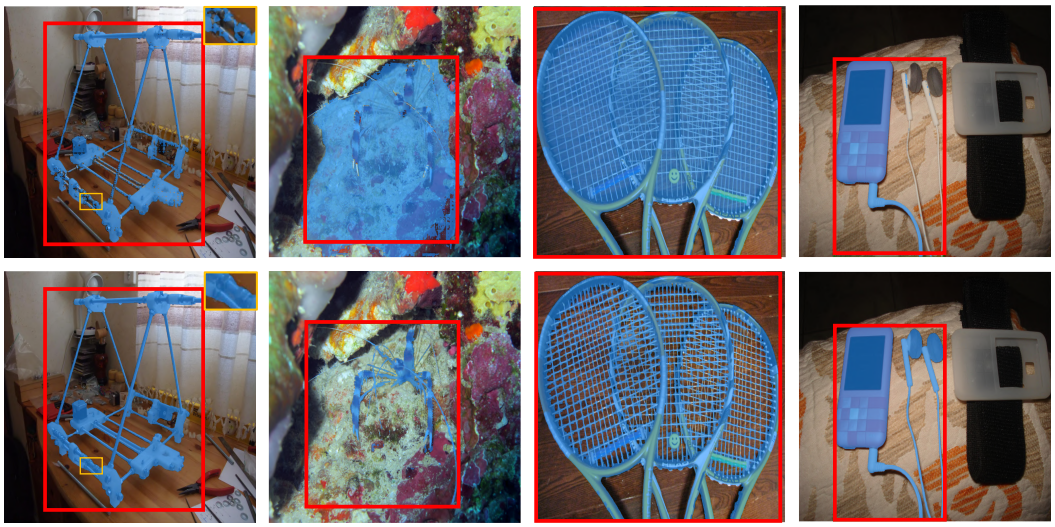

Figure 2: Visual results comparison between SAM (top row) **vs.** HQ-SAM (bottom row) on DIS test set, given the same red box prompt. HQ-SAM produces significantly more accurate boundaries.

Figure 3: Visual results comparison between SAM (top row) **vs.** HQ-SAM (bottom row) on COCO val set in *zero-shot setting*, using a SOTA detector FocalNet-DINO [17] trained on the COCO dataset as our box prompt generator. HQ-SAM predicts masks with higher quality than SAM with less mask artifacts.

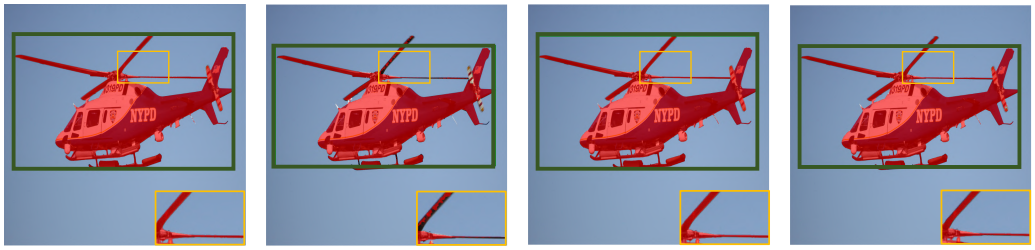

Figure 4: Visual results comparison between SAM (top row) **vs.** HQ-SAM (bottom row) with both the GT and noised green box prompt. HQ-SAM produces much more consistent and robust segmentation results regarding to the noises in the input boxes.

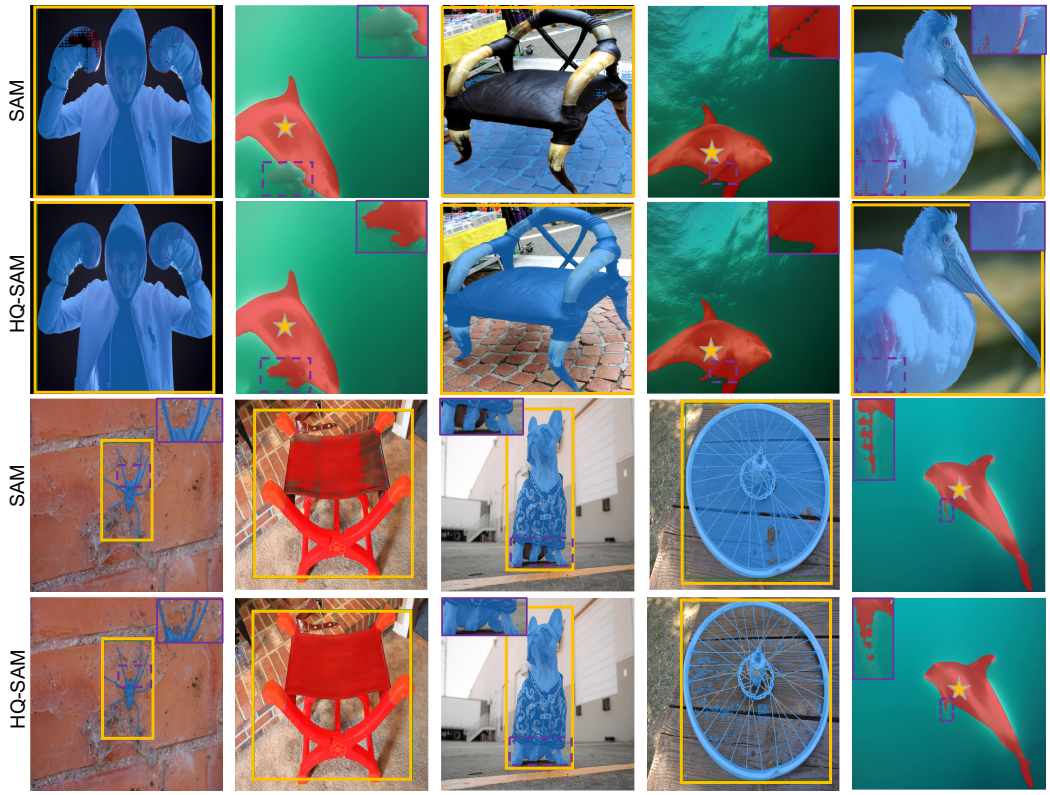

Figure 5: Visual results comparison between SAM (top row and third row) **vs.** HQ-SAM (second row and bottom row) in *zero-shot setting*, given the same yellow box or point prompt. HQ-SAM produces significantly more detailed preserving masks while fixing mask errors with broken holes. The rightmost two columns in the third row and bottom row show two *failure cases* of HQ-SAM in extremely dark environments or very tiny metal rods.

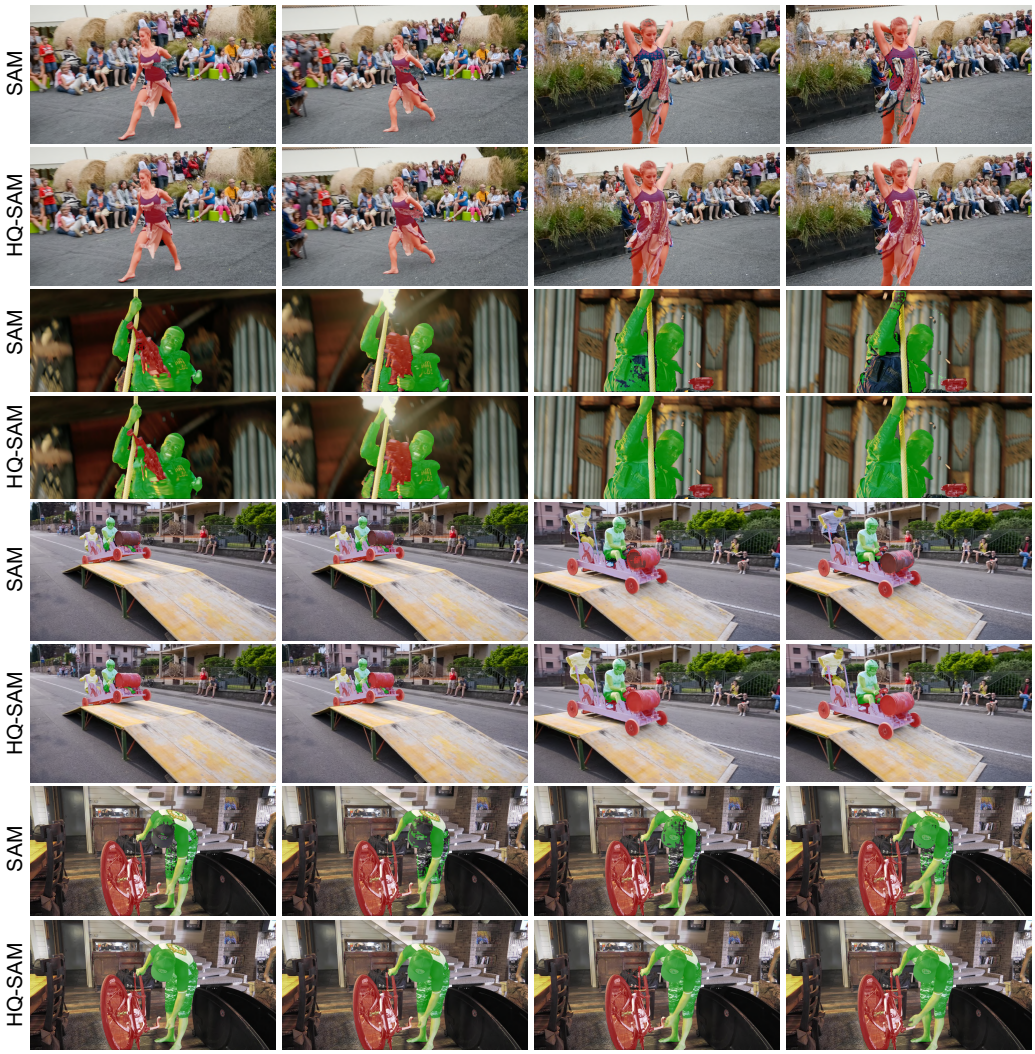

Figure 6: Visual results comparison between SAM **vs.** HQ-SAM on video object segmentation benchmark DAVIS 2017 in *zero-shot setting*, given the same video boxes prompts generated by the pre-trained XMem [2].

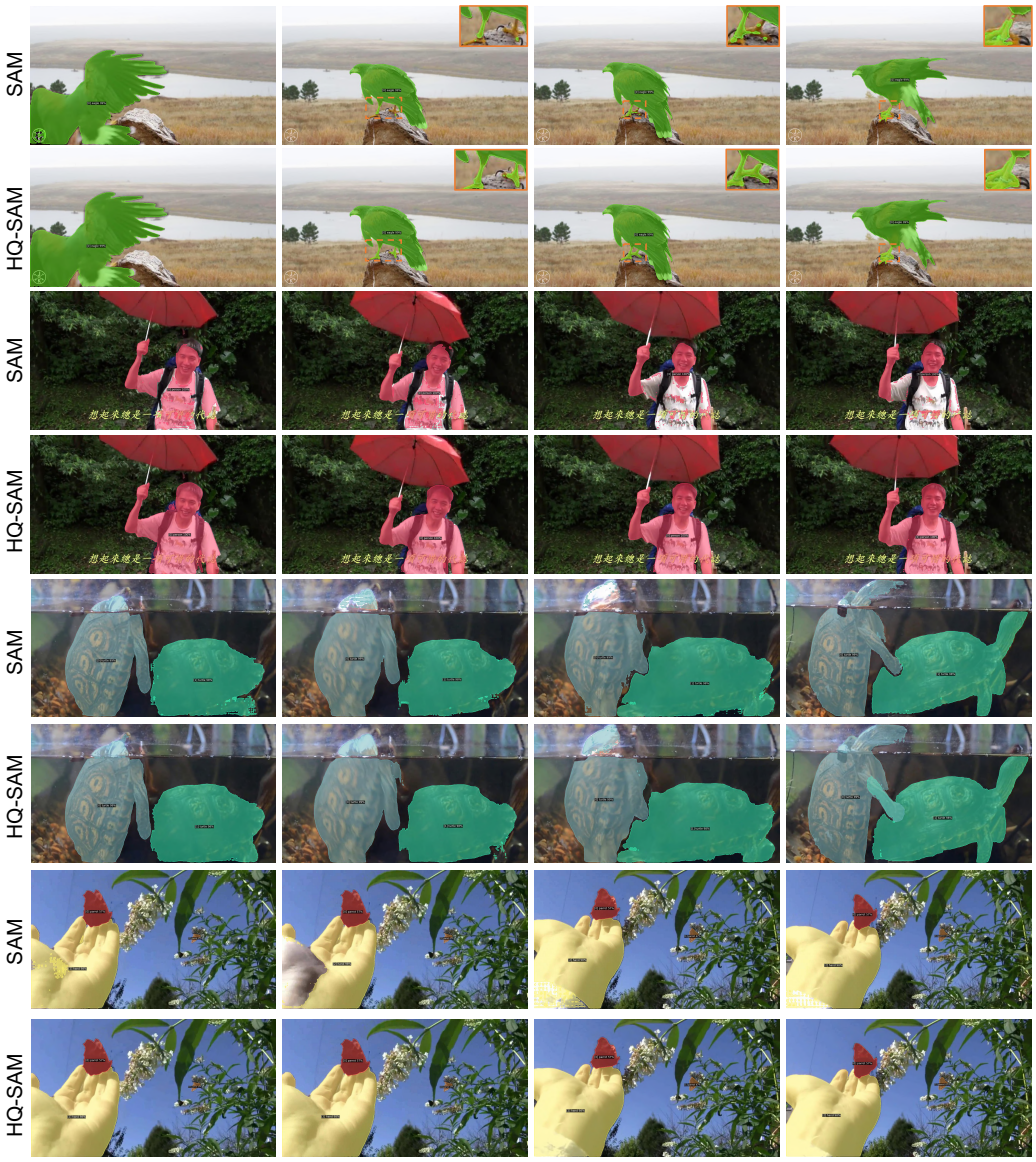

Figure 7: Visual results comparison between SAM **vs.** HQ-SAM on video instance segmentation benchmark YTVIS 2019 in *zero-shot setting*, given the same video boxes prompts generated by the pre-trained Mask2Former [1].

References

[1] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[2] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022.

[3] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2014.

[4] Lei Ke, Henghui Ding, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Video mask transfiner for high-quality video instance segmentation. In *ECCV*, 2022.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

[6] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022.

[7] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020.

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

[9] 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.

[10] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

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

[12] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 38(4):717–729, 2015.

[13] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. Ndd20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020.

[14] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[15] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019.

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

[17] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023.