A APPENDIX

A.1 CLICK SAMPLING STRATEGY.

To have fair comparison, we apply the same click sampling strategy from prior works (Sofiiuk et al., 2022; Liu et al., 2023; Huang et al., 2023; Liu et al., 2024a) for evaluation in all the experiments. This strategy generates clicks sequentially, with each new click placed at the center of the largest error region in the model's prediction.

A.2 MORE DATASETS DETAILS

We evaluate our method and comparison methods on two widely used benchmarks for interactive segmentation: HQSeg44K (Ke et al., 2024) and DAVIS (Perazzi et al., 2016). HQSeg44K is a large-scale segmentation dataset containing 44320 images with high-quality mask labels. It contains a diverse range of images spanning over 1,000 semantic classes, covering both simple and complex scenarios, and includes objects with thin shapes as well as more straightforward forms. DAVIS is a high-precision video object segmentation dataset consisting of 50 videos. It contains more complicated scenarios such as occlusion, multi-objects, motion blurs and many other challenges. To be consistent with previous works, we use a subset of 345 frames to conduct the evaluation.

A.3 ADDITIONAL RELATED WORK ON INTERACTIVE SEGMENTATION

Early interactive segmentation methods (Boykov & Jolly, 2001; Grady, 2006; Rother et al., 2004; Gulshan et al., 2010) relied on optimization techniques to solve cost functions defined by image pixels. As deep learning became more popular, approaches began incorporating user interactions directly into neural networks (Xu et al., 2016; Sofiiuk et al., 2020; Maninis et al., 2018; Jang & Kim, 2019; Lin et al., 2020). Further advancements such as RITM (Sofiiuk et al., 2022; Liu et al., 2022) leveraged large-scale data for robustness, FocalClick (Chen et al., 2022) introduced local refinement modules, and SimpleClick (Liu et al., 2023) improved performance using vision transformers. OACE (Mathur et al., 2024) is proposed to improve the foreground distinction in a contrastive learning manner. To enhance efficiency, methods like InterFormer (Huang et al., 2023) reduced interaction time by encoding the image only once during interactions, and SegNext (Liu et al., 2024a) applied similar idea with attention mechanism to achieve high quality with low latency. Following the introduction of SAM (Kirillov et al., 2023), numerous SAM-based methods have been proposed to improve interactive segmentation in aspects such as quality (Ke et al., 2024), interaction efficiency (Zhang et al., 2023a), and high-resolution image handling (Huang et al., 2024). Variants of interactive segmentation, like multi-object segmentation (Yue et al., 2023; Rana et al., 2023), medical image segmentation (Ma et al., 2024; Wong et al., 2023), and segmentation with controllable granularity (Zhao et al., 2024), have also been proposed.

A.4 COMPARISON WITH OTHER INTERACTIVE SEGMENTATION METHODS INCORPORATING 3D INFORMATION

In this section, we compare our methods to other interactive segmentation methods that integrate 3D information (MM-SAM (Xiao et al., 2024)) to show that naively incorporating depth is not as effective as our order-aware attention module discussed in Sec. 3.2 and validated in Sec. 4. Here, we take the depth maps generated from DepthAnything V2 (Yang et al., 2024a) as an input to MM-SAM and test this on the DAVIS dataset; we call this configuration MM-SAM in Table 5. Moreover, following the strategy of MM-SAM, which adds an extra encoder for other modalities such as lidar, depth, and thermal, we also train an additional Depth Encoder and add to our pipeline. Note that in this configuration, we disable the order-aware attention discussed in 3.2 to ensure fair comparison. We call this configuration DepthEncoder in Table 5. Our method outperforms this configuration and MM-SAM across all metrics. This indicates that incorporating 3D information into the model through order using our specialized order-aware attention module is a better solution than directly integrating depth maps.

Methods	NoC90 ↓	NoC95 ↓	1-mIoU ↑	5-mIoU ↑	NoF95 ↓
MM-SAM (Xiao et al., 2024)	6.22	12.26	51.41	88.19	181
DepthEncoder	3.88	10.19	85.36	92.15	140
Ours	**3.80**	**8.59**	**87.29**	**92.76**	**114**

Table 5: Comparison with other interactive segmentation methods that integrate 3D information.

A.5 Additional visualization of Attention Weights in Order-Aware Attention

In this section, we provide more visualizations of the attention weights to show the importance of proposed order-aware attention, as shown in Figure 7 and discussed in 4.5.

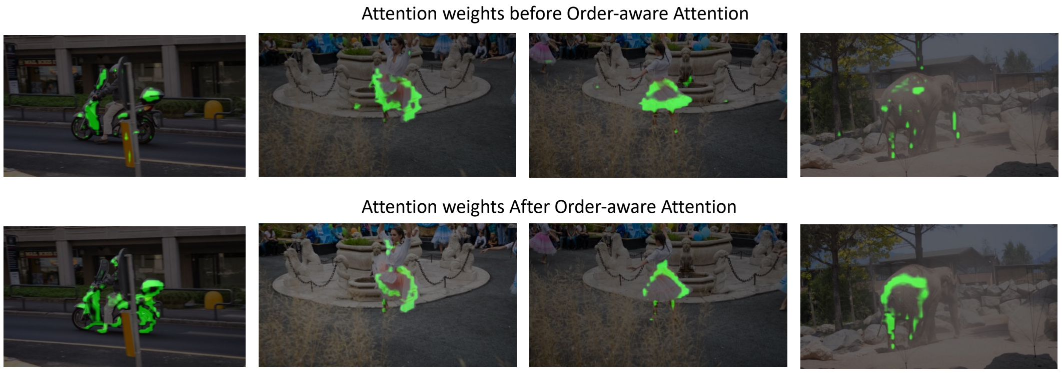
Attention weights before Order-aware Attention
Attention weights After Order-aware Attention

Figure 7: Visualization of attention weights before and after applying order-aware attention.

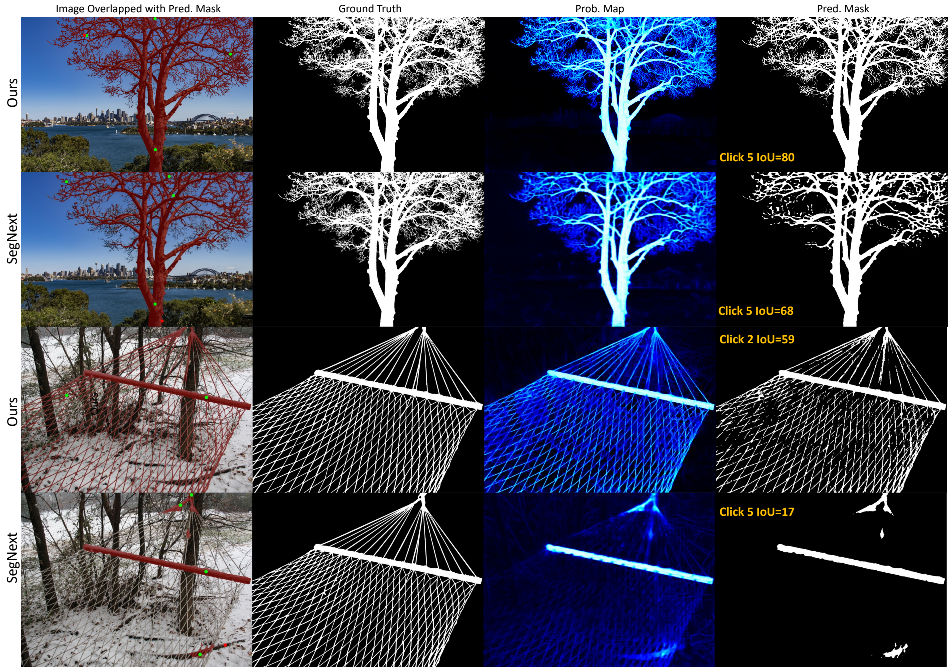

Figure 8: Qualitative results for more challenging cases with multiple clicks. Green dots mean the positive clicks, and red dots are the negative clicks.

A.6 Qualitative Results for More Challenging Cases

We provide more qualitative results of challenging cases from the HQSeg44K (Ke et al., 2024) dataset as displayed in Figure 8 and Figure 9.

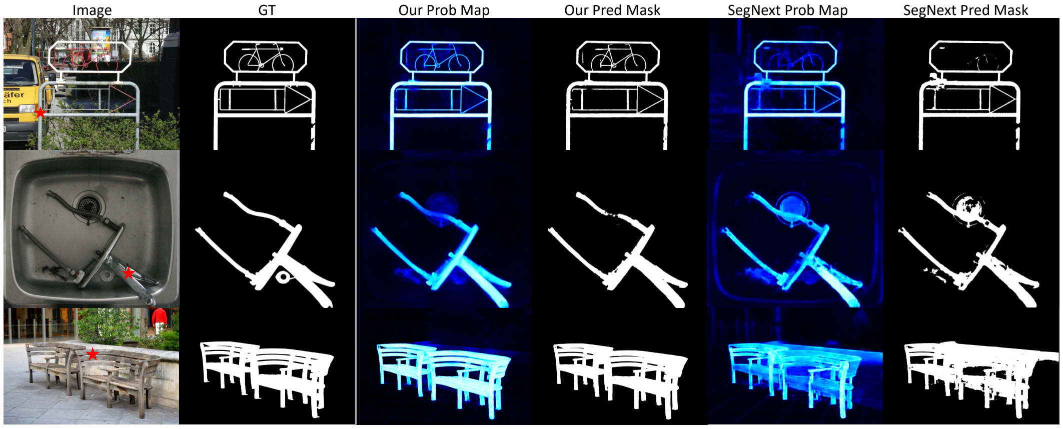

Figure 9: Qualitative results for more challenging cases with only one click. The red star represents the first click.

A.7 Analysis of multi-round interactions

Interactive segmentation is a multi-round task. The goal is to minimize the user interactions (i.e., interaction round number) and achieve a good quality segmentation mask. Hence, we evaluate the multi-round interaction performance of our method. Figure 10 illustrates that while both our method and SegNext produce some false positives in the background after the first click, our method eliminates the entire background, including regions inside the bike and adjacent to the human head, with only a single negative click. In contrast, SegNext requires 10 clicks to remove these background false positives and still can not get satisfied segmentation for the entire target. We also provide more multi-round interaction results in Figure 12. For the challenging cases, the multi-round interaction results are displayed in Figure 11 and Figure 13.

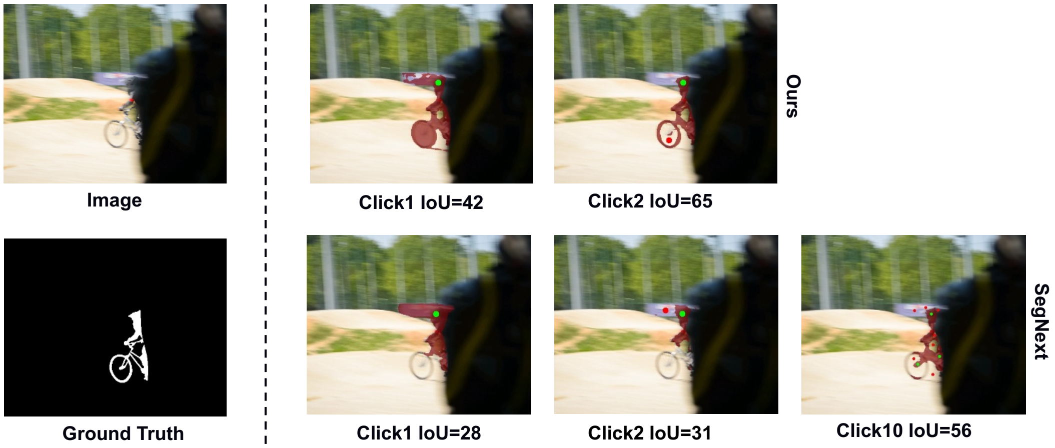

Figure 10: Multi-round interaction comparison. Green dots mean the positive clicks, and red dots are the negative clicks.

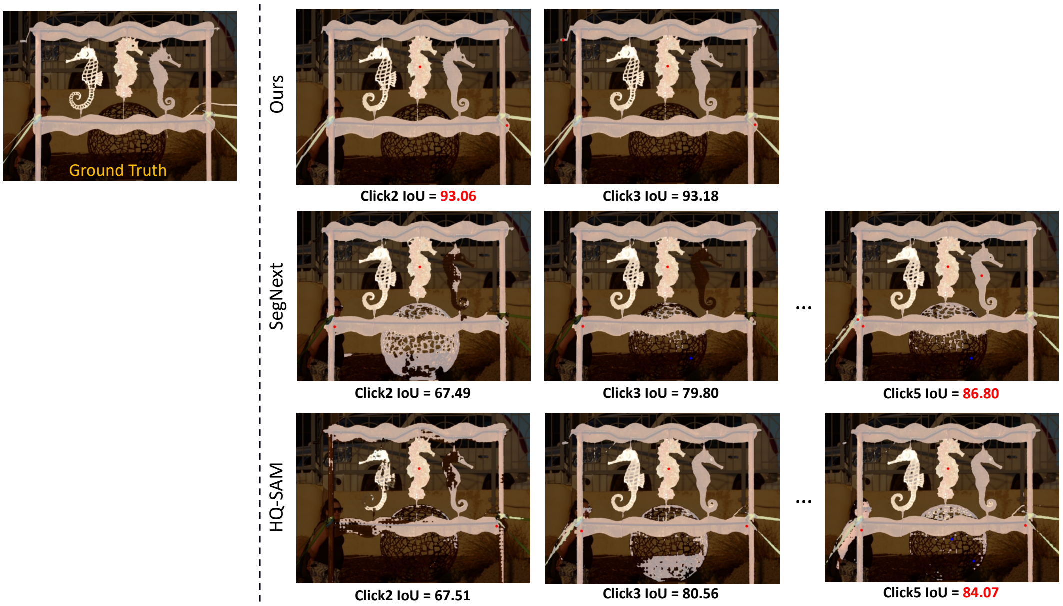

Figure 11: Multi-round interaction comparison for a challenging case.

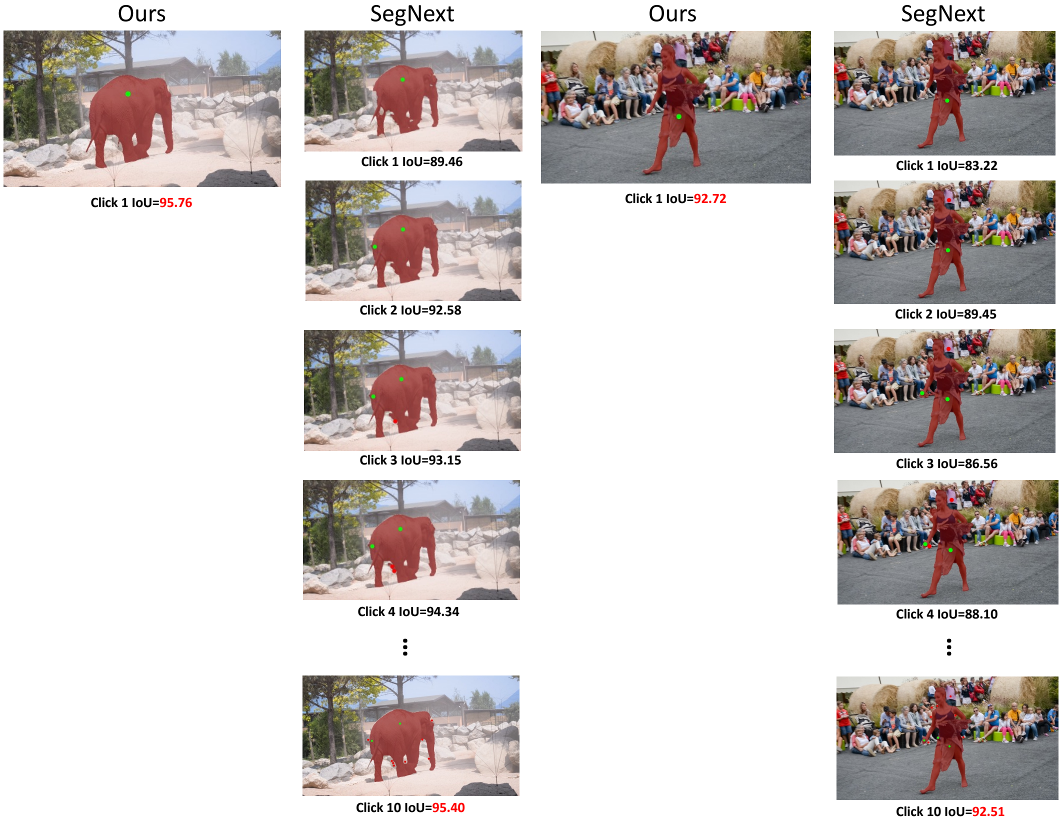

Figure 12: Multi-round interaction comparison. Green dots mean the positive clicks, and red dots are the negative clicks.

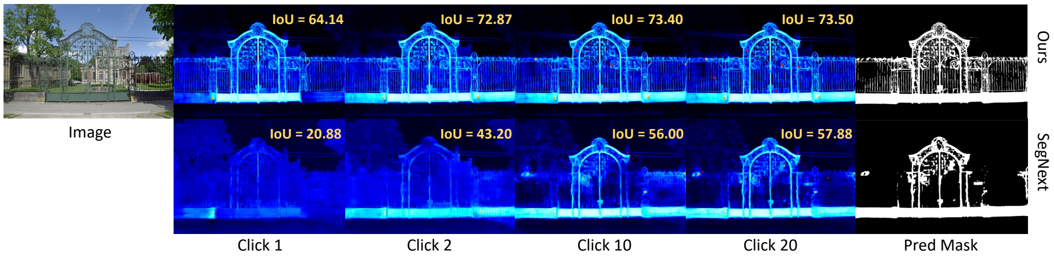

Figure 13: Multi-round interaction comparison for a challenging case.

A.8 Failure Cases

In this section, we discuss some failure cases of our method. As shown in Figure 14, our method sometimes struggles to accurately predict the segmentation masks if thin objects occlude our target. This can be seen in the first example, where thin branches and leaves occlude the target bus and in the second and third examples, where the grass occludes the target human.

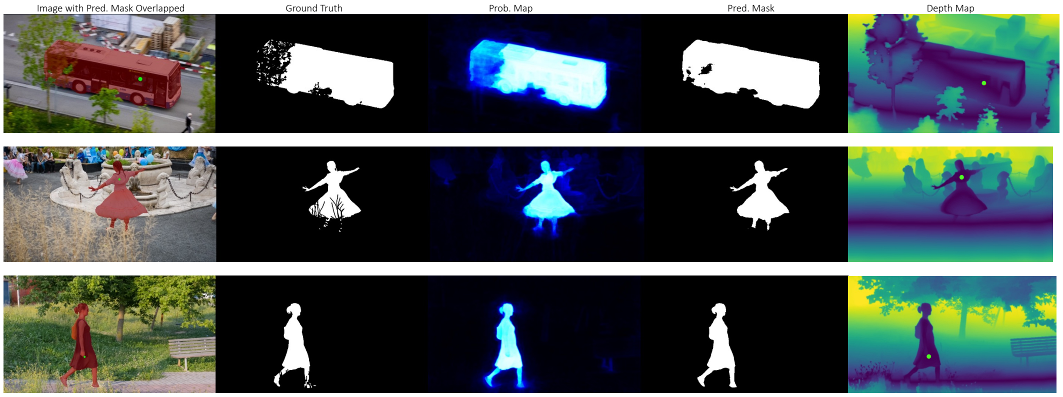

Figure 14: Failure Cases. Green dots are the first positive clicks.

A.9 Impact of Depth Map on Model Performance

To evaluate the impact of the quality of the depth map used on our model's performance, we adopt three commonly used depth prediction models: DepthAnything V2 (Yang et al., 2024b), DepthAnything V1 (Yang et al., 2024a), and ZoeDepth (Bhat et al., 2023). As illustrated in Fig. 15, DepthAnything V2 produces the most fine-grained details, such as the girl's fingers and animal's tail, while the other two methods generate lower-quality depth maps with some possible depth prediction errors. Note that these three models generate depth maps in different quality levels, allowing us to get better insights on our model's robustness to the depth quality.

Methods	**Depth Model**	**NoC90** ↓	**5-mIoU** ↑
SegNext (Liu et al., 2024a)	-	4.43	91.87
OIS	DepthAnythingV2 (Yang et al., 2024b)	3.80	**92.76**
OIS	DepthAnythingV1 (Yang et al., 2024a)	3.78	92.69
OIS	ZoeDepth (Bhat et al., 2023)	**3.75**	92.75

Table 6: Performance comparison on DAVIS using depth maps from different depth prediction models.

We compared our model's performance on the DAVIS (Perazzi et al., 2016) dataset using depth maps from these three models. The results, presented in Table 6, show minimal performance variation,

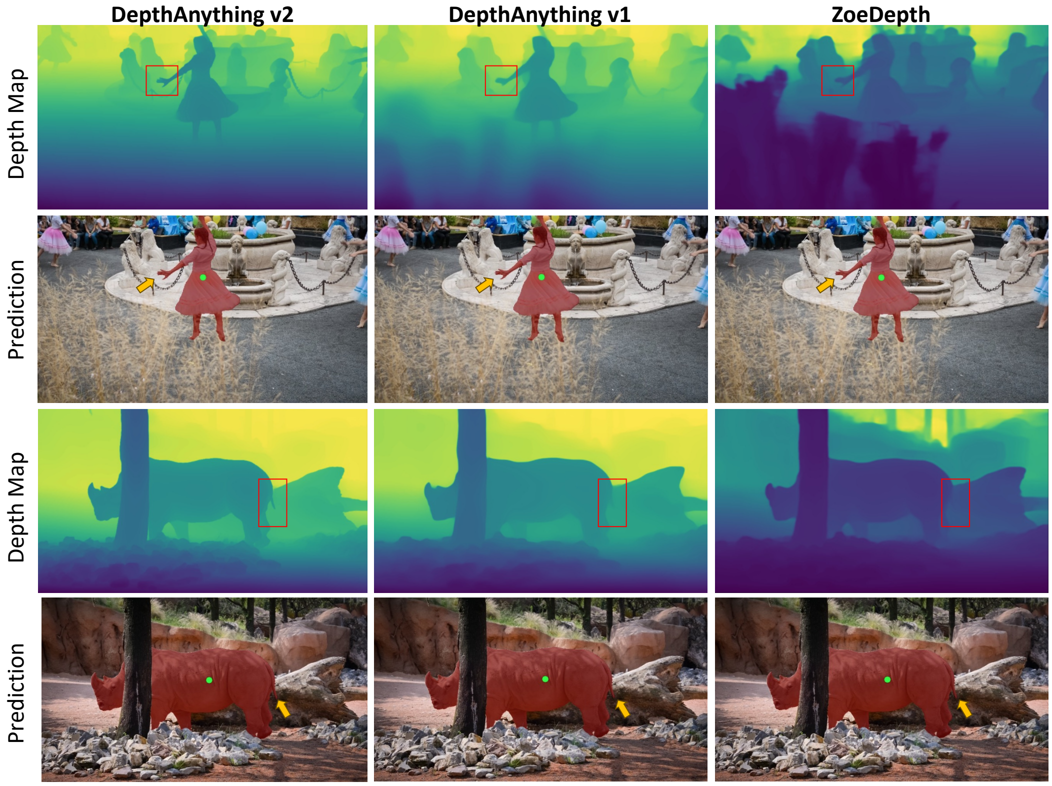

Figure 15: Qualitative results of our model with different depth map generation models. Green dots are the first positive clicks.

with our method consistently outperforming the current SOTA method, SegNext (Liu et al., 2024a). This is because the depth map is used solely to generate the order map, which guides the model's understanding of relative depth between objects. Even a lower quality depth map (DepthAnything V1) has little impact on the final segmentation performance. More importantly, our proposed object-aware attention module is able to negate the effects of erroneous order maps caused due to erroneous depth maps.

Qualitative results are provided in Fig. 15. It is observed that even though the qualities of the depth map are different, the prediction masks are nearly the same, especially the regions that depth maps have significant different (girls' hand and animal's tail), which indicates the robustness of our model with different depth map source.

Moreover, it is important for the model to remain robustness when meeting the depth prediction errors. In Fig. 16, we put two examples which contain depth prediction error. In the first case, the dancer's hat is incorrectly blending into the background audience in the predicted depth map. However, our model correctly recovers the error and accurately segments the hat. Additionally, in the second case, the duck's boundary closely mixes with the neighboring water in terms of depth, while our model successfully separates the duck from the water in the segmentation prediction. This robustness is attributed to our proposed object-aware understanding module which ensures that our model comprehends the object as a whole, enabling it to handle depth prediction errors effectively.

Besides, it is noted that our model also achieve great performance in scenarios where significant depth variation exists within the target object. As demonstrated in Fig. 17, the target objects exhibit considerable depth variation, yet our model consistently delivers accurate and high-quality segmentation. This stems from the proposed object-aware understanding module plays a key role in ensuring the model perceives the target object as a whole, enabling it to handle significant depth variation effectively.

In summary, the key role of depth in our approach is to address complex scenarios where distinguishing the target from the background is challenging, as demonstrated in Fig. 1 and Fig. 4. In these scenarios, our model significantly outperforms existing methods. For cases where depth is less critical (Fig. 17), we show that our model remains robust and unaffected by potential negative effects of depth. This demonstrates that our model can effectively solve challenging cases while maintaining strong performance in standard scenarios, highlighting its superior performance and robustness.

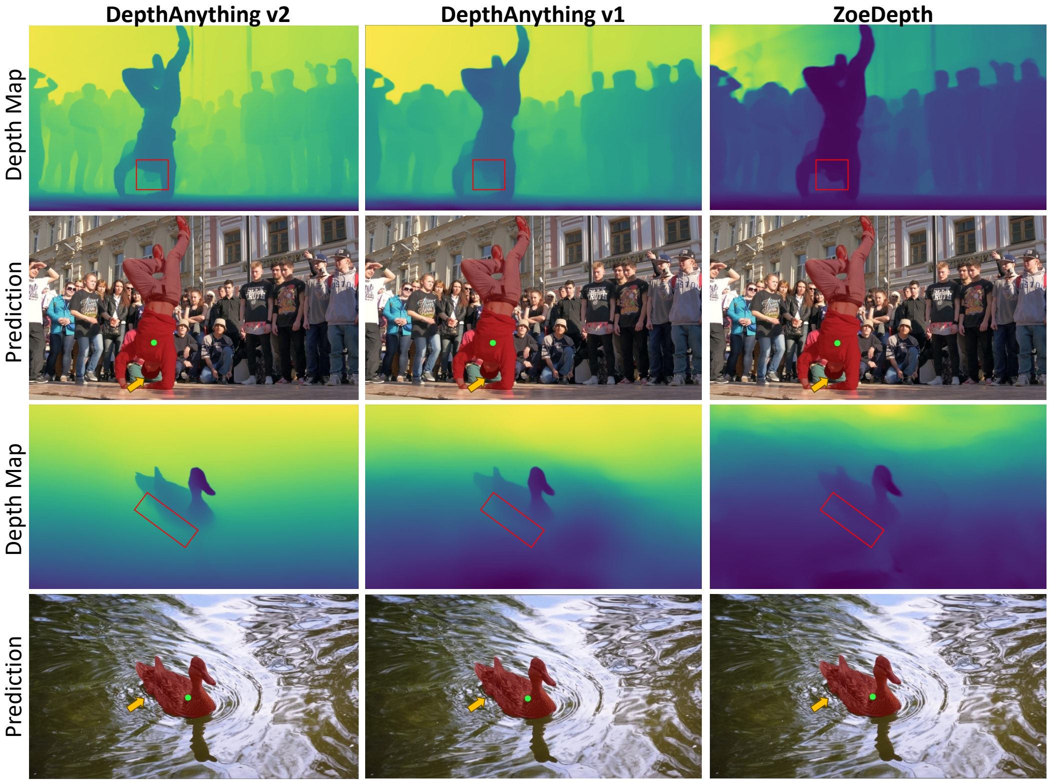

Figure 16: Qualitative results of our model in scenarios with depth prediction errors or targets with depth similar to neighboring objects. Green dots are the first positive clicks.

A.10 Ablation study about the ordering of order/object-aware understanding modules

Here, we discuss the effect of the sequence order of order-aware understanding and object-aware understanding module. We conduct the comparison experiment on DAVIS (Perazzi et al., 2016) dataset. The results in Table 7 show that having the order-aware understanding module first, followed by the object-aware understanding module slightly decreased performance than our current sequence (object-aware understanding module first, followed by the order-aware understanding module), indicating the effectiveness of our sequence choice.

Methods	NoC90 ↓	NoC95 ↓	1-mIoU ↑	5-mIoU ↑
order-aware understanding first	3.84	9.41	85.74	92.49
object-aware understanding first (current)	**3.80**	**8.59**	**87.29**	**92.76**

Table 7: Performance comparison on DAVIS with different sequence ordering of object and order-aware understanding modules.

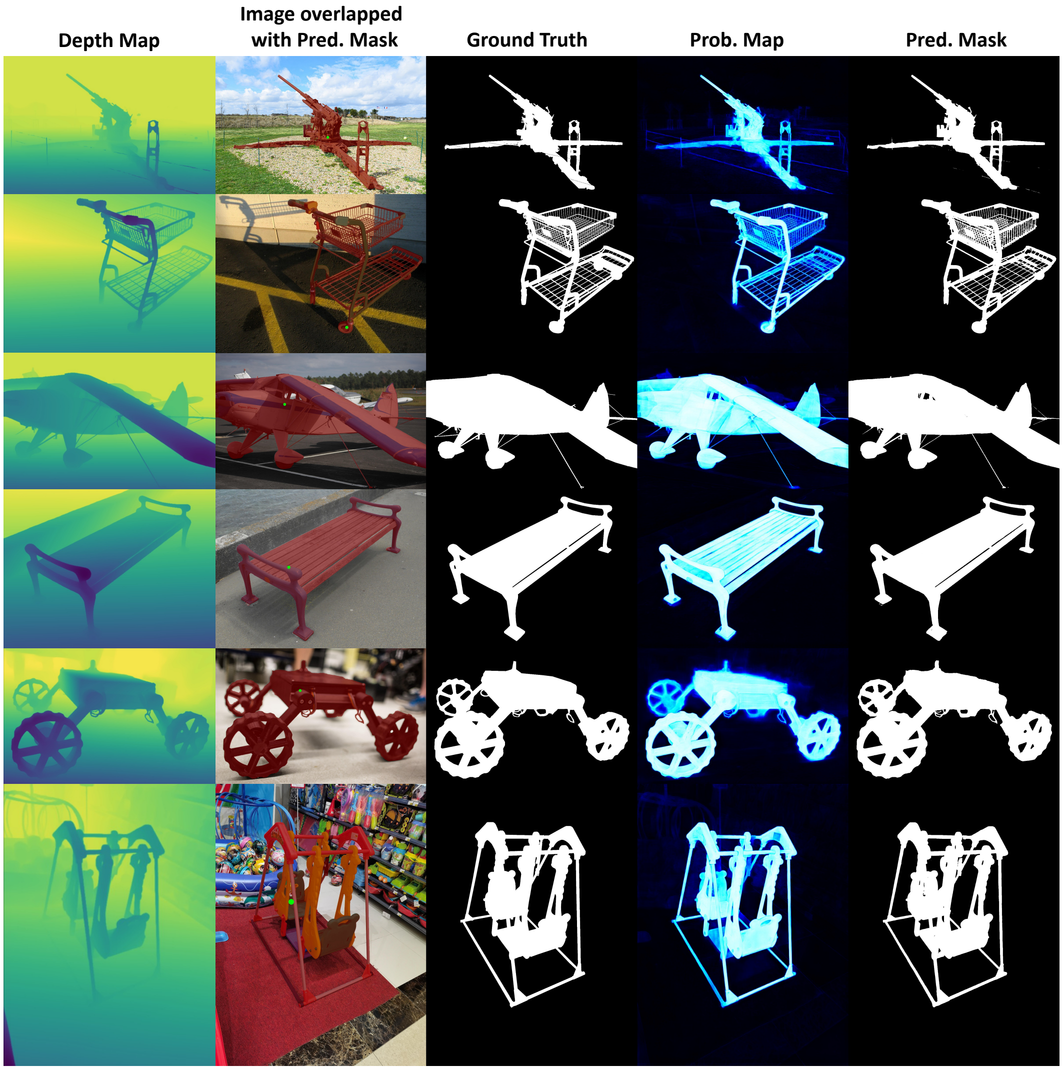

Figure 17: Qualitative results for cases with targets spanning a large depth range. Green dots are the first positive clicks.

A.11 Ablation Study on HQSeg44K dataset

To further prove the importance of each our proposed module, here we conduct the ablation experiment on HQSeg44K (Ke et al., 2024) dataset. The results, as displayed in Table 8, are consistent with the findings from the ablation study on the DAVIS dataset, as shown in Table 4, reaffirming that each proposed module plays a crucial role in enhancing overall performance.

Methods	**NoC90** ↓	**5-mIoU** ↑
Full	3.95	93.78
w/o order	4.87 (+0.98)	92.49 (-1.29)
w/o object	4.23 (+0.28)	93.28 (-0.5)
w/o sparse	5.23 (+1.28)	90.80 (-2.98)
w/o dense	4.97 (+1.02)	91.75 (-2.03)

Table 8: Ablation experiments on HQSeg44K.

A.12 Ablation study for order map with positive or negative clicks alone

We include an additional ablation study on DAVIS (Perazzi et al., 2016) dataset here to analyze the design of the order map. Table 9 shows that removing either the positive click order map or the negative click order map leads to a performance drop, confirming the effectiveness of combining both. Interestingly, we observe the following:

Impact of Positive Click Order Map When only the negative click order map is used, the 1-mIoU metric decreases more significantly. This suggests that the positive click order map is particularly beneficial during the first click, as the first click is always a positive click.

Impact of Negative Click Order Map When only the positive click order map is used, the NoC and 5-mIoU metrics see a larger decrease. This indicates that the negative click order map becomes more important as additional clicks are made. This is because subsequent clicks mainly involve adjustments and background removal, which rely heavily on negative clicks.

Methods	**NoC90** ↓	**NoC95** ↓	**1-mIoU** ↑	**5-mIoU** ↑
pos+neg	**3.80**	**8.59**	**87.29**	**92.76**
pos	4.36	9.89	85.68	92.04
neg	4.01	9.24	84.17	92.37

Table 9: Ablation experiments for order with positive or negative clicks alone.

A.13 Ablation study of image encoder backbone

To show the robustness of our model to different backbones, we conduct an ablation study on the HQSeg44K (Ke et al., 2024) dataset with different image encoder backbone. We replace the DepthAnything V2 backbone (Yang et al., 2024b) with an MAE pretrained ViT backbone (He et al., 2022) to be consistent with prior SOTA methods, SegNext (Liu et al., 2024a), InterFormer (Huang et al., 2023), and SimpleClick (Liu et al., 2023). Table 10 shows that our method still significantly outperforms the other methods with MAE ViT backbone. Furthermore, Table 11 highlights that the performance gains from our proposed order and object-aware attention far exceed those from switching backbones, representing the effectiveness and importance of our proposed methods.

A.14 Performance Comparison Trained on COCO

Since tradional interactive segmentation methods, including RITM (Sofiiuk et al., 2022), FocalClick (Chen et al., 2022), SimpleClick (Liu et al., 2023), and InterFormer (Huang et al., 2023) are purely trained on COCO+LVIS dataset (Lin et al., 2014; Gupta et al., 2019), to have a fair comparison

Methods	Backbone	**NoC90** ↓	**NoC95** ↓	**5-mIoU** ↑
SimpleClick (Liu et al., 2023)	MAE ViT-B	7.47	12.39	85.11
InterFormer (Huang et al., 2023)	MAE ViT-B	7.17	10.77	82.62
SegNext (Liu et al., 2024a)	MAE ViT-B	5.32	9.42	91.75
OIS	MAE ViT-B	**4.41**	**8.01**	**93.12**

Table 10: Comparison of performance using the same backbone with other SOTA methods.

Methods	Backbone	**NoC90** ↓	**NoC95** ↓	**5-mIoU** ↑
OIS w/o order+object	MAE ViT-B	5.54	9.57	90.58
OIS	MAE ViT-B	4.41	8.01	93.12
OIS w/o order+object	DepthAnythingV2 ViT-B	5.23	8.91	90.80
OIS	DepthAnythingV2 ViT-B	3.95	7.50	93.78

Table 11: Comparison of performance improvement of order and object-aware attention with the same backbone.

with them, we also train our model using only COCO+LVIS dataset. In Table 12, we provide the comparison results on HQSeg44K (Ke et al., 2024) and DAVIS (Perazzi et al., 2016) dataset. The results demonstrate that our method still outperforms other methods by a large margin, which indicates the effectiveness of our model.

	HQSeg44K			**DAVIS**		
	NoC90 ↓	**NoC95** ↓	**1-mIoU** ↑	**NoC90** ↓	**NoC95** ↓	**1-mIoU** ↑
RITM (Sofiiuk et al., 2022)	10.01	14.58	36.03	5.34	11.45	72.53
FocalClick (Chen et al., 2022)	7.03	10.74	61.92	5.17	11.42	76.28
SimpleClick (Liu et al., 2023)	7.47	12.39	65.54	5.06	10.37	72.90
InterFormer (Huang et al., 2023)	7.17	10.77	64.40	5.45	11.88	64.40
OIS	**5.16**	**9.18**	**85.36**	**4.41**	**9.87**	**87.21**

Table 12: Performance comparison with methods trained on COCO+LVIS.