Inter2Former: Dynamic Hybrid Attention for Efficient High-Precision Interactive Segmentation - Supplementary Materials

You Huang, Lichao Chen, Jiayi Ji, Liujuan Cao, Shengchuan Zhang*, Rongrong Ji
Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
youhuang0607@gmail.com, fpg2012@foxmail.com, jjyxmu@gmail.com
caoliujuan@xmu.edu.cn, zsc_2016@xmu.edu.cn, rrji@xmu.edu.cn

A. Ablation Study

To validate the effectiveness of each key component in Inter2Former, we conduct extensive ablation studies. We train different architectural variants from scratch for 80k iterations on HQSeg44K [2], using HRSAM++ encoder [1] distilled from SAM as the unified image encoder.

As shown in Table 1, replacing DPE with full prompt embedding (Non-DPE) shows marginal impact on model performance. In terms of attention mechanisms, our proposed hybrid attention design (DHA) achieves comparable results with Full Attention (All FA), while significantly outperforming pure BSQ Attention (All BSQA). For the upsampling module, although DLU shows slightly lower performance than full upsampling at 1024^2 resolution on DAVIS, this gap becomes negligible at 2048^2 resolution. These results demonstrate the effectiveness of our proposed lightweight modules.

Additionally, we compare our BSQA with VQA [3] by replacing BSQA in our model, as shown in the last row of each input resolution in Table 1. The results indicate that VQA leads to performance degradation due to convergence issues during training, which validates the effectiveness of our BSQA design.

Configuration	Input Image Size	HQSeg44K Max H/W>4000			DAVIS Max H/W<1000		
		5-mIoU ↑	NoC90 ↓	NoC95 ↓	5-mIoU ↑	NoC90 ↓	NoC95 ↓
Inter2Former-Base	1024 × 1024	91.32	5.46	9.34	90.36	4.96	12.72
DPE → Non-DPE	1024 × 1024	91.33	5.44	9.38	90.56	4.94	12.48
DHA → All FA	1024 × 1024	91.37	5.38	9.31	90.35	5.01	12.24
DHA → All BSQA	1024 × 1024	89.71	6.19	10.12	88.20	5.83	13.53
DLU → Non-DLU	1024 × 1024	91.52	5.35	9.16	90.90	4.89	11.81
BSQA → VQA	1024 × 1024	90.91	5.59	9.58	89.59	5.31	12.75
Inter2Former-Base	2048 × 2048	92.68	4.24	7.39	92.00	4.29	7.82
DPE → Non-DPE	2048 × 2048	92.86	4.19	7.33	92.17	4.28	7.94
DHA → All FA	2048 × 2048	92.61	4.24	7.39	92.26	4.20	7.78
DHA → All BSQA	2048 × 2048	90.12	5.64	8.80	89.31	5.37	9.75
DLU → Non-DLU	2048 × 2048	92.76	4.22	7.32	92.13	4.30	7.90
BSQA → VQA	2048 × 2048	91.07	4.82	8.01	90.31	4.73	8.86

Table 1. Ablation study. Inter2Former-Base represents our complete model configuration. The arrow (→) indicates replacement of our proposed module with alternatives: Non-DPE replaces DPE with full prompt embedding, All FA and All BSQA replace DHA with Full Attention and BSQ Attention respectively, Non-DLU substitutes DLU with full upsampling module, and VQA replaces BSQA in DHA.

*Corresponding author

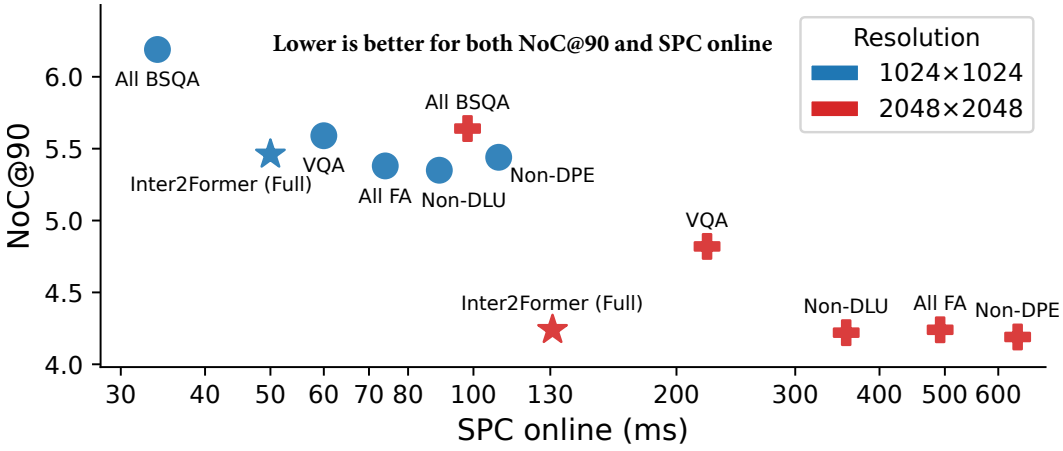

Figure 1. Performance-efficiency trade-off in the ablation study. We plot segmentation quality (NoC@90) against online inference latency (SPC online) for the architectural variants. The comparison, shown for both 1024×1024 and 2048×2048 resolutions, visually demonstrates that the full Inter2Former configuration achieves a superior balance of performance and efficiency compared to the ablated versions.

References

[1] You Huang, Wenbin Lai, Jiayi Ji, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Hrsam: Efficiently segment anything in high-resolution images. *arXiv preprint arXiv:2407.02109*, 2024. 1

[2] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[3] Lucas D Lingle. Transformer-vq: Linear-time transformers via vector quantization. *arXiv preprint arXiv:2309.16354*, 2023. 1