Rethinking Interactive Image Segmentation with Low Latency, High Quality, and Diverse Prompts

Supplementary Material

1. Evaluation with the Text Prompt

This section supplements the "Experiments" section in the main paper. Table 1 quantitatively evaluates the text prompt. The models were trained on RefCOCO [2] and evaluated on its testA subset across three settings: text-only, click-only, and a combination of text and click (text+click). Following PhraseClick [1], we used three clicks for the click-only setting and two clicks for the text+click setting. While our text prompt had much room to improve, it yielded promising results combined with visual prompts.

Method	Interaction	Text	Click	mIoU (%)
PhraseClick [1]	Text-only	✓	✗	50.98
Ours (SA×2)	Text-only	✓	✗	58.32
Ours (SA×2)	Click-only	✗	✓	82.79
Ours (SA×2)	Text+Click	✓	✓	85.95

Table 1. Evaluation with text prompt on the testA of RefCOCO. Our model attained enhanced performance by integrating text and click prompts, surpassing the results achieved with clicks alone.

References

[1] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 417–435. Springer, 2020. 1

[2] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1