Supplementary

Data augmentation

In our data augmentation process, we utilized Albumentation version 1.3.1 and employed three functions with their respective parameters as follows:

1. RandomResizedCrop, with parameters scale=[0.5, 1.0], ratio=[0.75, 1.33], and interpolation=1.
2. ShiftScaleRotate, with parameters shift_limit=[-0.05, 0.05], scale_limit=(-0.05, 0.05), rotate_limit=10, border_mode=0, and value=0.
3. ColorJitter, with parameters brightness=0.3, contrast=0.3, saturation=0, and hue=0.

Parameter of FFM

We employ the BERT_12_768_12 model for text information conversion, aligning the BERT output feature length with the ViT-B's transformer block feature length at 768. Within the FFM structure, the cross-attention Q, K, V feature mapping length is set to 96, without utilizing a multi-head mechanism. For the MLP part, two fully connected layers are used with input and output feature lengths of (768, 192) and (192, 768), respectively, and a ReLU activation function between them.

Ablation study about the effects of LGA and textual features

In this ablation study, we use fine-tuning only the mask decoder as the baseline and compare the effects of adding only textual information, only incorporating the LGA module, and integrating both textual information and the LGA module. When only adding textual information, we process the BERT-extracted features through a fully connected layer to reduce the feature length to 256, then concatenate them with the output query of the mask decoder. This is followed by subsequent self-attention and cross-attention with image information, mirroring the original SAM's treatment of prompt encoder's prompt features. When only incorporating the LGA module, the input to the first FFM in LGA is replaced with a learnable query instead of the original BERT output text features, maintaining the computational approach without textual features. The method that uses both LGA and textual features sets the input to the first FFM in LGA as the BERT-output textual information, as previously described.

Ablation study about the number of FFM in SPA

In this ablation experiment, the incorporation of Feature Fusion Modules (FFMs) within the image encoder is systematically varied to examine their impact on model performance:

Adding One FFM: The FFM is placed after the last transformer block in the image encoder.

Adding Two FFMs: FFMs are placed after the 6th and the last transformer blocks.

Adding Three FFMs: FFMs are placed before the first transformer block and after the 6th and last transformer blocks.

Adding Four FFMs: FFMs are placed before the 1st, 5th, and 9th transformer blocks, and one more is added after the last transformer block.

Adding Five FFMs: FFMs are placed before the 1st, 4th, 7th, and 10th transformer blocks, and another one is added after the last transformer block.