Segment Anything in High Quality

2024.04.15 김채원

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

Published date: 2023.06.02

Citation : 80

Github : https://github.com/SysCV/SAM-HQ

1.Abstract

Segment Anything Model (SAM) show powerful zero-shot capabilities and flexible prompting.

Despite being trained with 1.1 billon masks, When dealing with objects with intricate structures, it exhibits low prediction quality in many cases.

So, We propose HQ-SAM, equipping SAM with the ability to accurately segment any object, while maintaining SAM’s original promptable design, efficiency, and zero-shot generalizability

1.Abstract

Our careful design reuses and preserves the pre-trained model weights of SAM, while only introducing minimal additional parameters and computation.

We design a learnable High-Quality Output Token, which is injected into SAM’s mask decoder and is responsible for predicting the high-quality mask.

Instead of only applying it on mask-decoder features, we first fuse them with early and final ViT features for improved mask details.

To train our introduced learnable parameters, we compose a dataset of 44K fine-grained masks from several sources.

HQ-SAM is only trained on the introduced dataset of 44k masks, which takes only 4 hours on 8 GPUs (RTX3090)

2.Introduction

Segment Anything Model (SAM) was recently released as a foundational vision model for general image segmentation. But its segmentation results are still unsatisfactory in many cases.

SAM model’s Key problems

1. Coarse mask boundaries, often even neglecting the segmentation of thin object structures

2. Incorrect predictions, broken masks, or large errors in challenging cases.

2.Introduction

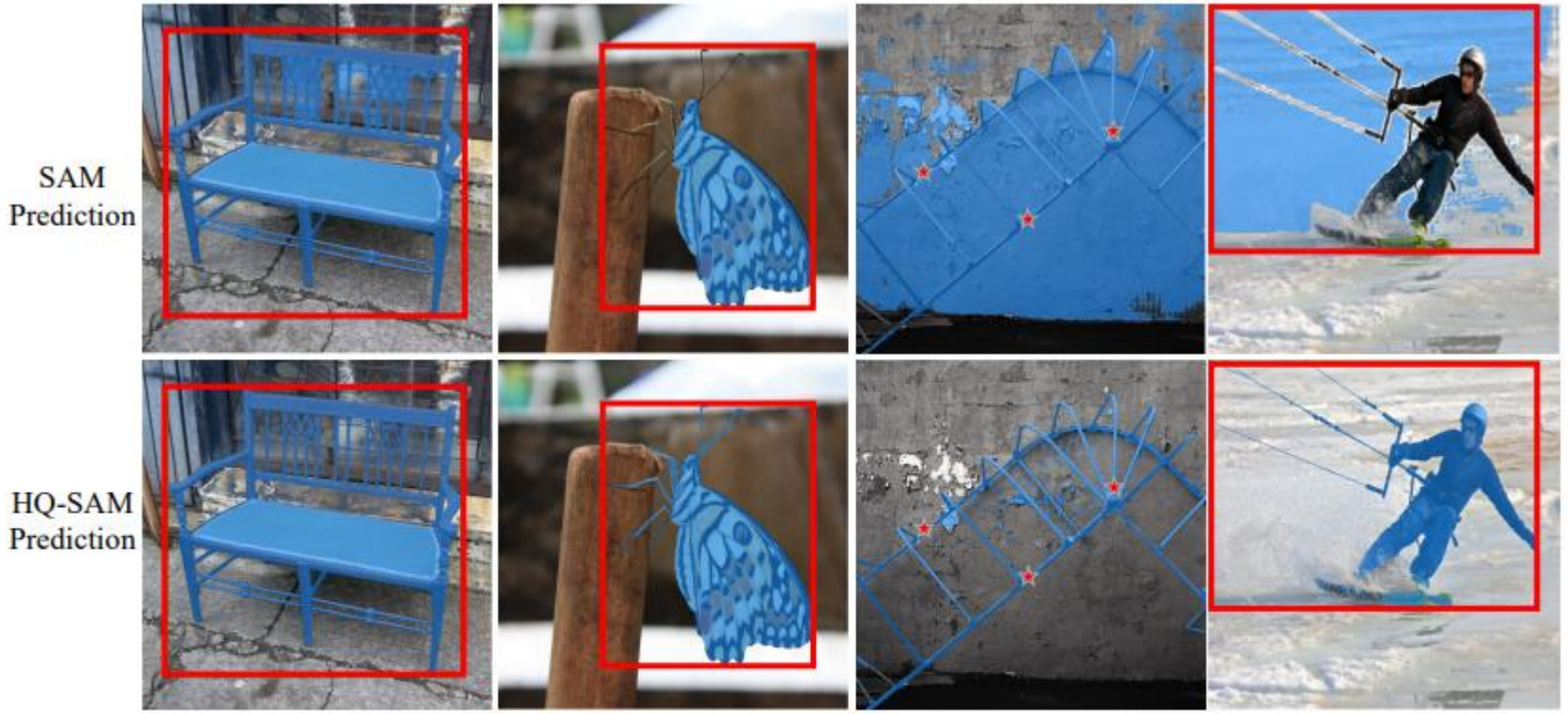
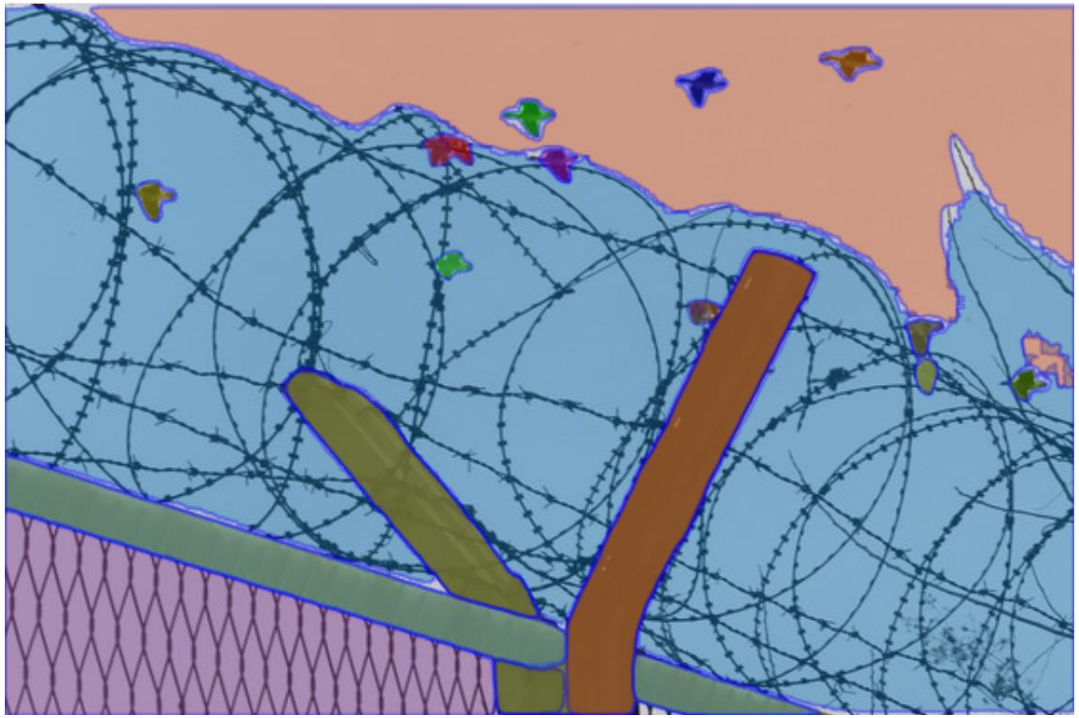


Figure 1: The predicted masks of SAM **vs.** our HQ-SAM, given the same red box or several points on the object as input prompts. HQ-SAM produces significantly more detailed results with very accurate boundaries. In the rightmost column, SAM misinterprets the thin structure of the kite lines, and produces a large portion of errors with broken holes for the input box prompt.

2.Introduction



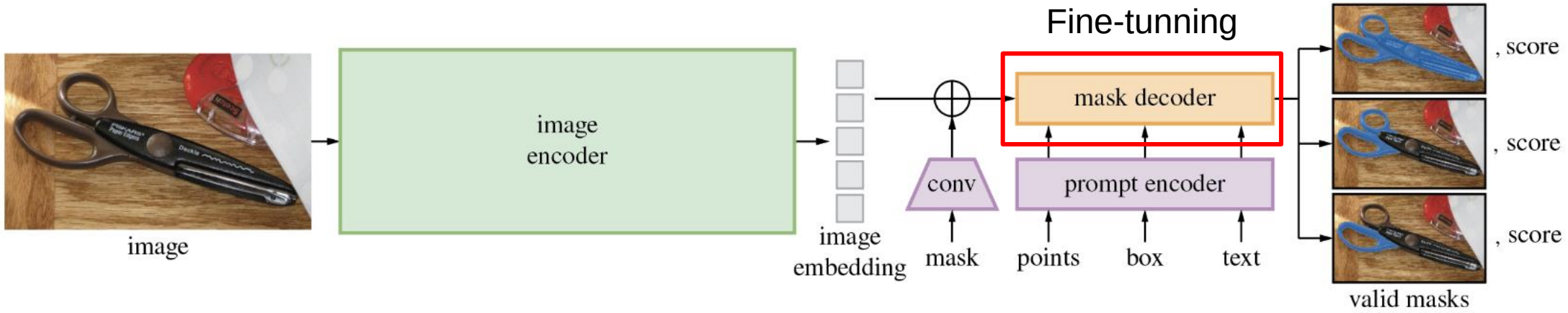
2.Introduction

HQ-SAM

It can predict highly accurate segmentation masks, even in very challenging cases without compromising the strong zero-shot capabilities and flexibility of the original SAM.

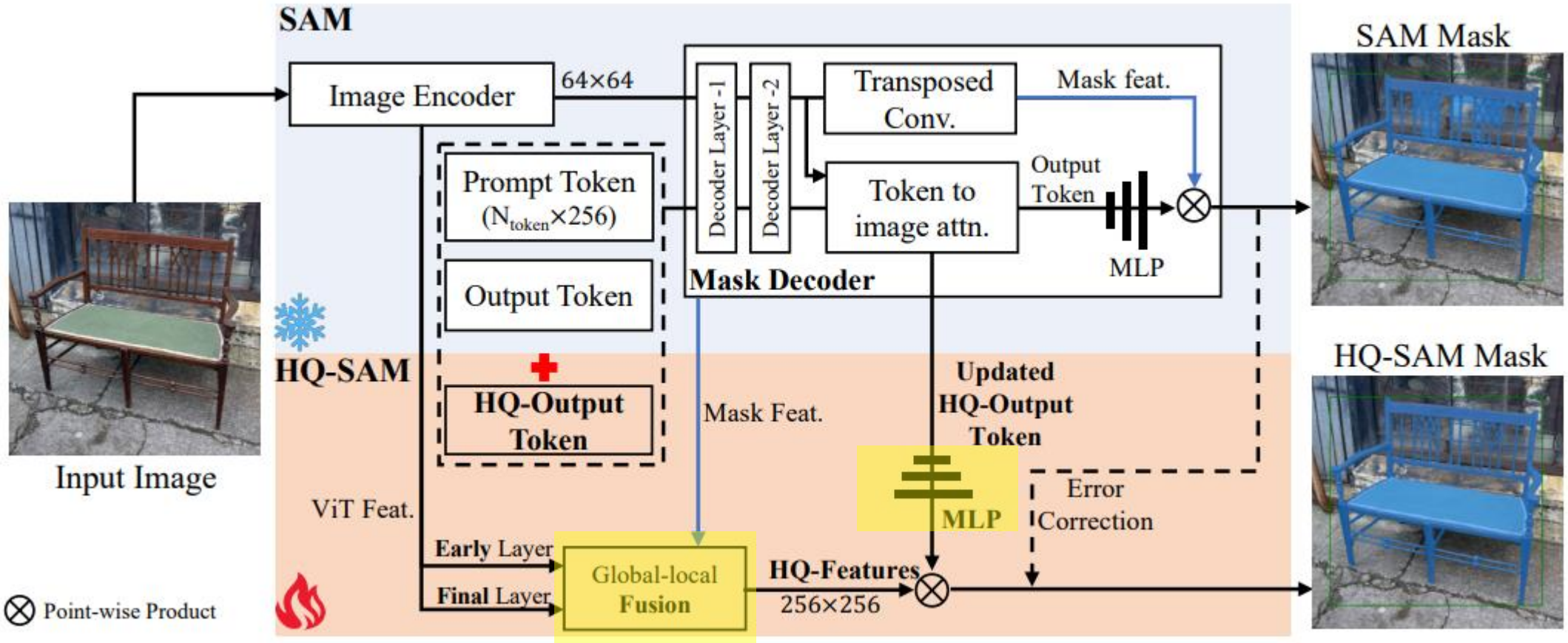
To preserve the efficiency and zero-shot performance, we propose a minimal adaptation of SAM, adding less than 0.5% parameters, to extend its capability to high-quality segmentation.

2.Introduction



Reduce the general zero-shot segmentation performance.

2.Introduction



Update : HQ-Output token, related 3-layer MLP, feature fusion block

2.Introduction

Learning accurate segmentation requires a dataset with accurate mask annotations of diverse objects with complex and detailed geometries.

SAM is trained on the SA-1B dataset, which contains 11M images with 1.1 billion masks automatically generated by a SAM-like model.

However, using this extensive dataset presents significant cost implications and falls short of achieving the desired high-quality mask generations pursued in our work.

2.Introduction

HQSeg - 44K

- 44K extremely fine-grained image mask annotation.
- Constructed by merging six existing image dataset with highly accurate mask labels, covering over 1000 diverse semantic classes.

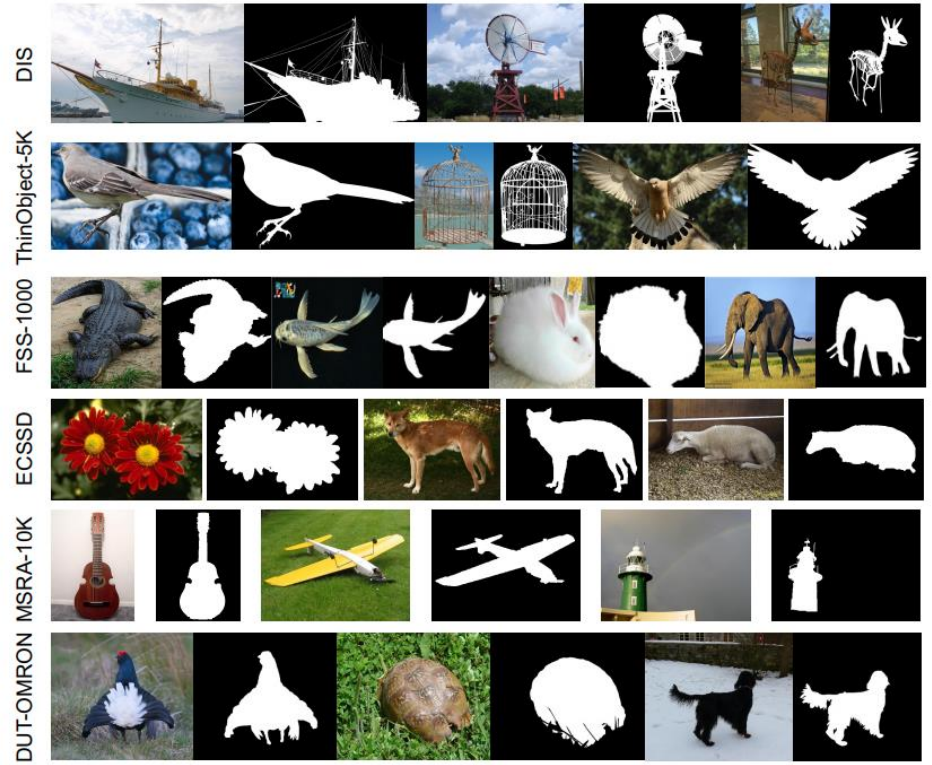


Figure 8: Visualization of annotated mask quality for randomly selected cases from the six dataset components of the HQ-Seg-44K. Zoom in for better viewing the fine-grained mask details.

2.Introduction

To validate the effectiveness of HQ-SAM, we perform extensive quantitative and qualitative experimental analysis.

We compare HQ-SAM with SAM on a suite of 10 diverse segmentation datasets across different downstream tasks,
where 8 out of them are under a zero-shot transfer protocol

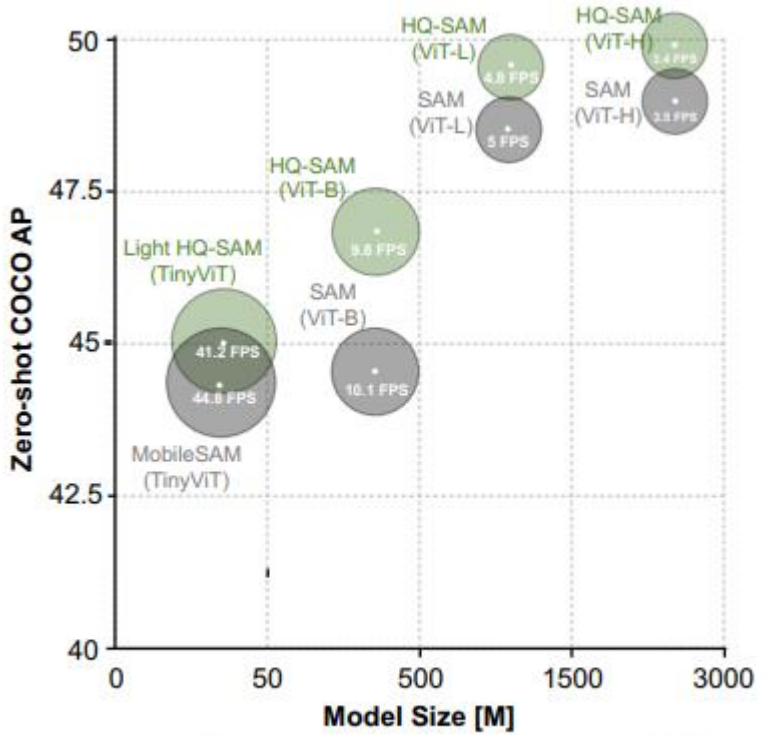
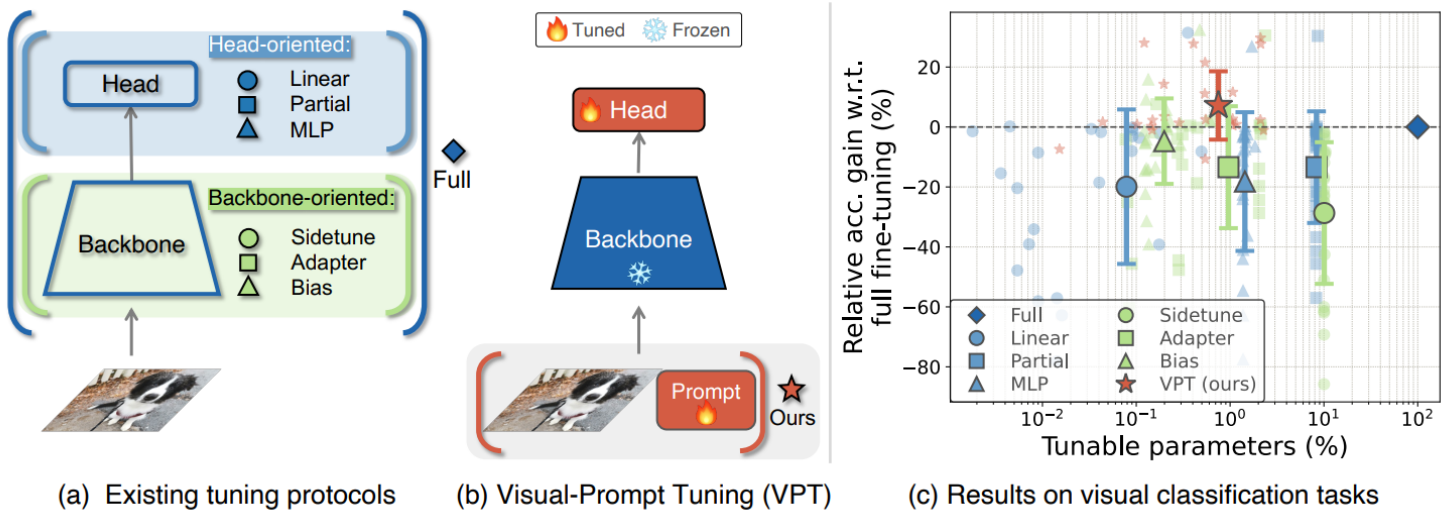


Figure 2: Performance vs. speed vs. model size for an array of SAM variants [21, 52].

2.Related work

- Prompt Tuning

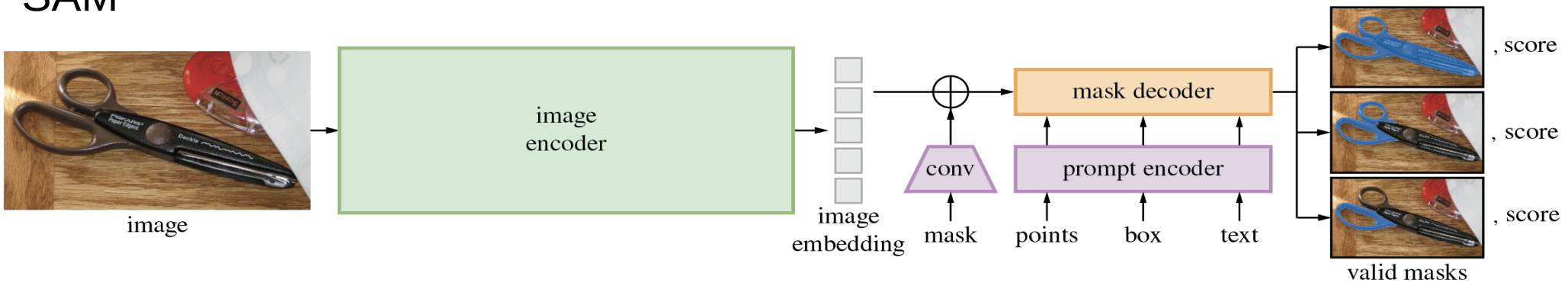


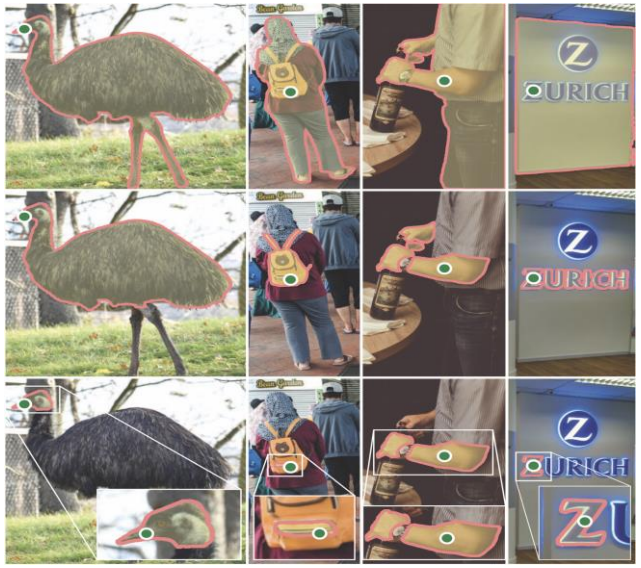
(a) Existing tuning protocols (b) Visual-Prompt Tuning (VPT) (c) Results on visual classification tasks

Fig. 1. Visual-Prompt Tuning (VPT) *vs*. other transfer learning methods. (a) Current transfer learning protocols are grouped based on the tuning scope: Full fine-tuning, Head-oriented, and Backbone-oriented approaches. (b) VPT instead adds extra parameters in the input space. (c) Performance of different methods on a wide range of downstream classification tasks adapting a pre-trained ViT-B backbone, with mean and standard deviation annotated. VPT outperforms Full fine-tuning 20 out of 24 cases while using less than 1% of all model parameters

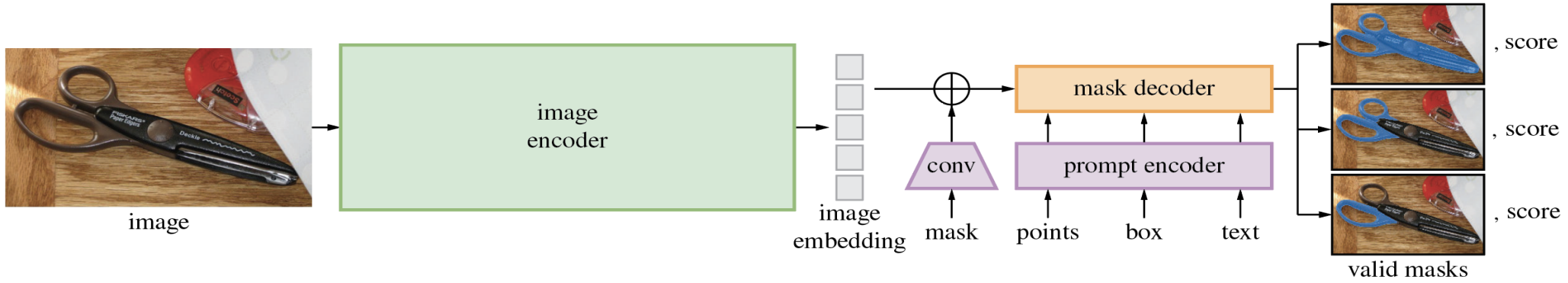
2.Related work

- SAM





3.Method

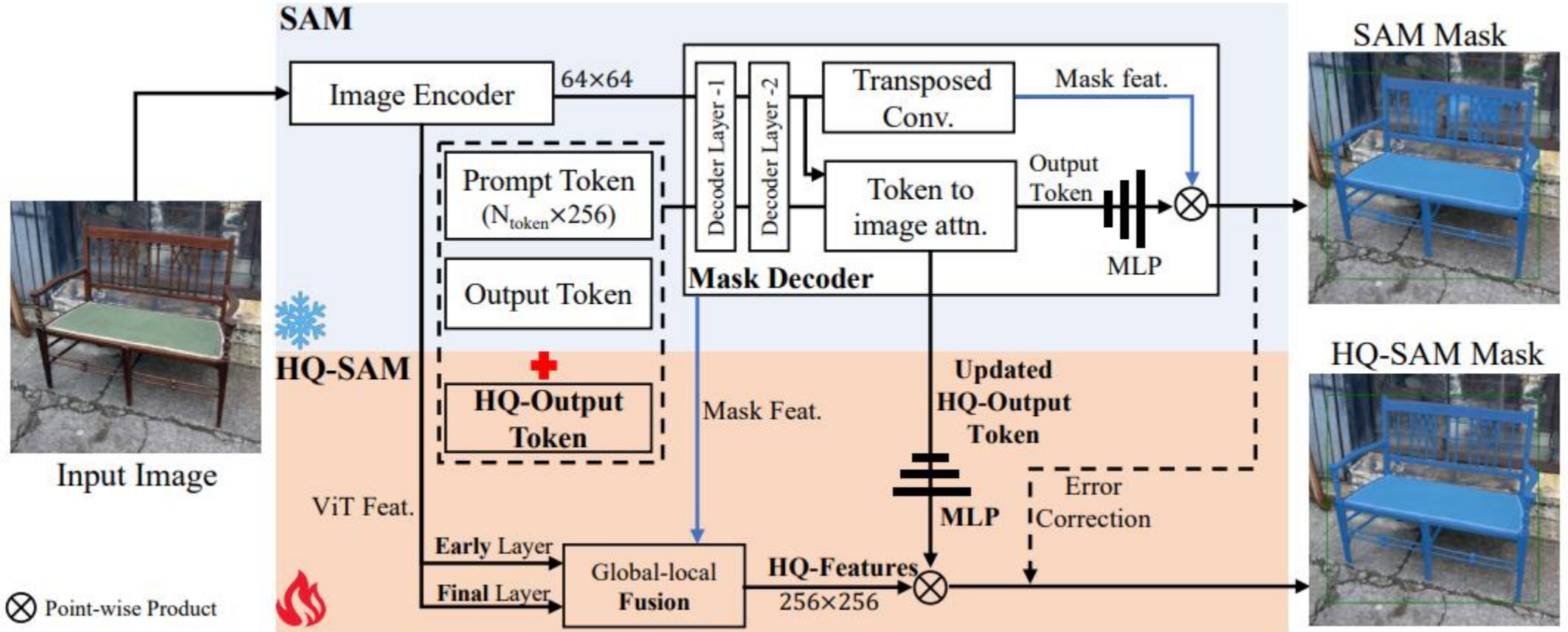


Preliminaries : SAM model

1. Image Encoder : a heavy ViT-based backbone for image feature extraction, resulting in image embedding in spatial size 64×64.
2. Prompt encoder: encoding the interactive positional information from the input points, boxes, masks to provide for the mask decoder.
3. Mask decoder: a two-layer transformer-based decoder takes both the extracted image embedding with the concatenated output and prompt tokens for final mask prediction.

3.Method

2. HQ-SAM



3.Method

2. HQ-SAM

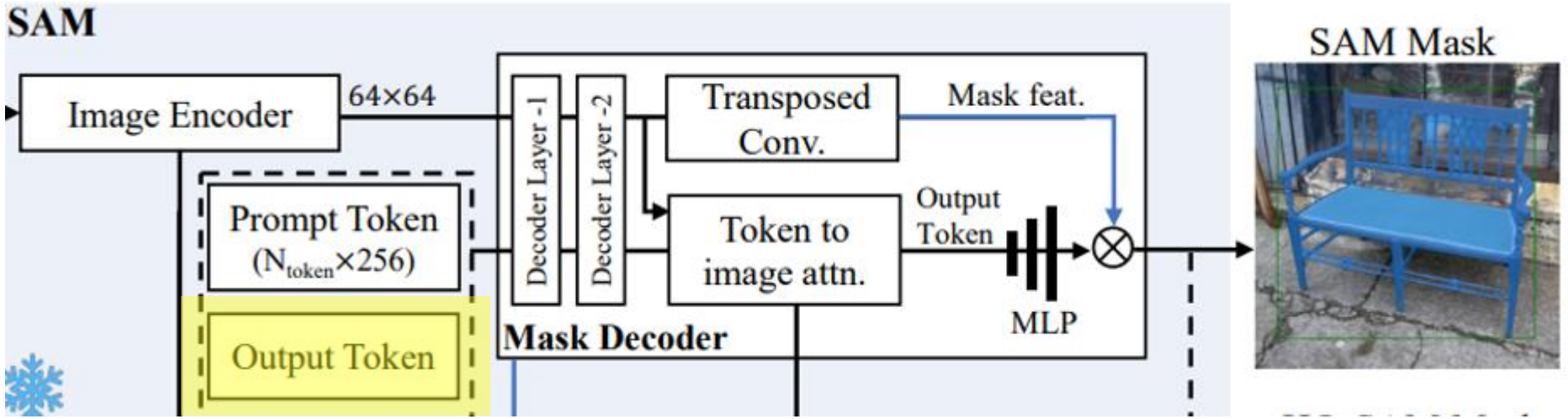
To preserve the zero-shot transfer capability of SAM, while preventing model overfitting or catastrophic forgetting, instead of directly fine-tuning SAM or adding a new heavy decoder network, we take a minimal adaptation approach as much as possible.

To this end, HQ-SAM reuses the pre-trained model weights of SAM as much as possible with only two new key components, namely, High-Quality Output Token and Global-local Feature Fusion.

3.Method

2-1. High-Quality Output Token

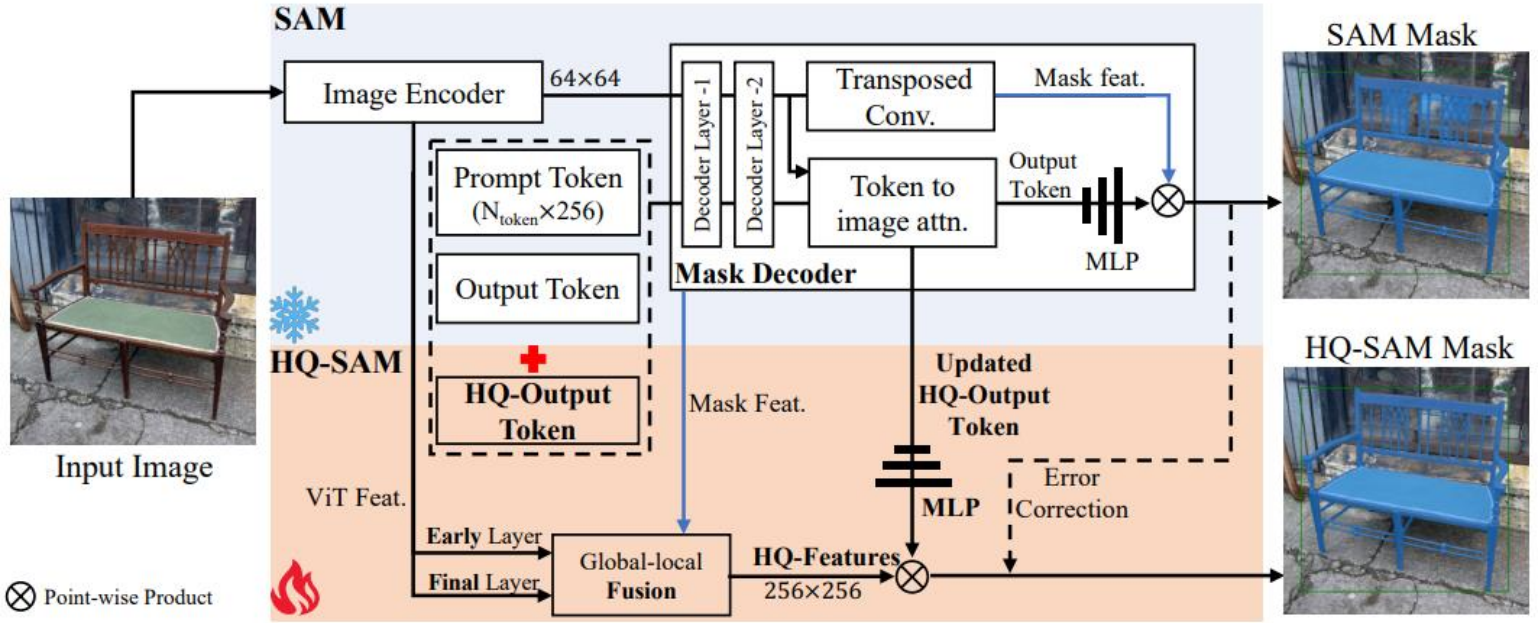
In SAM’s original mask decoder design, the output token is adopted for mask prediction, which predicts dynamic MLP weights and then performs point-wise product with the mask features.



3.Method

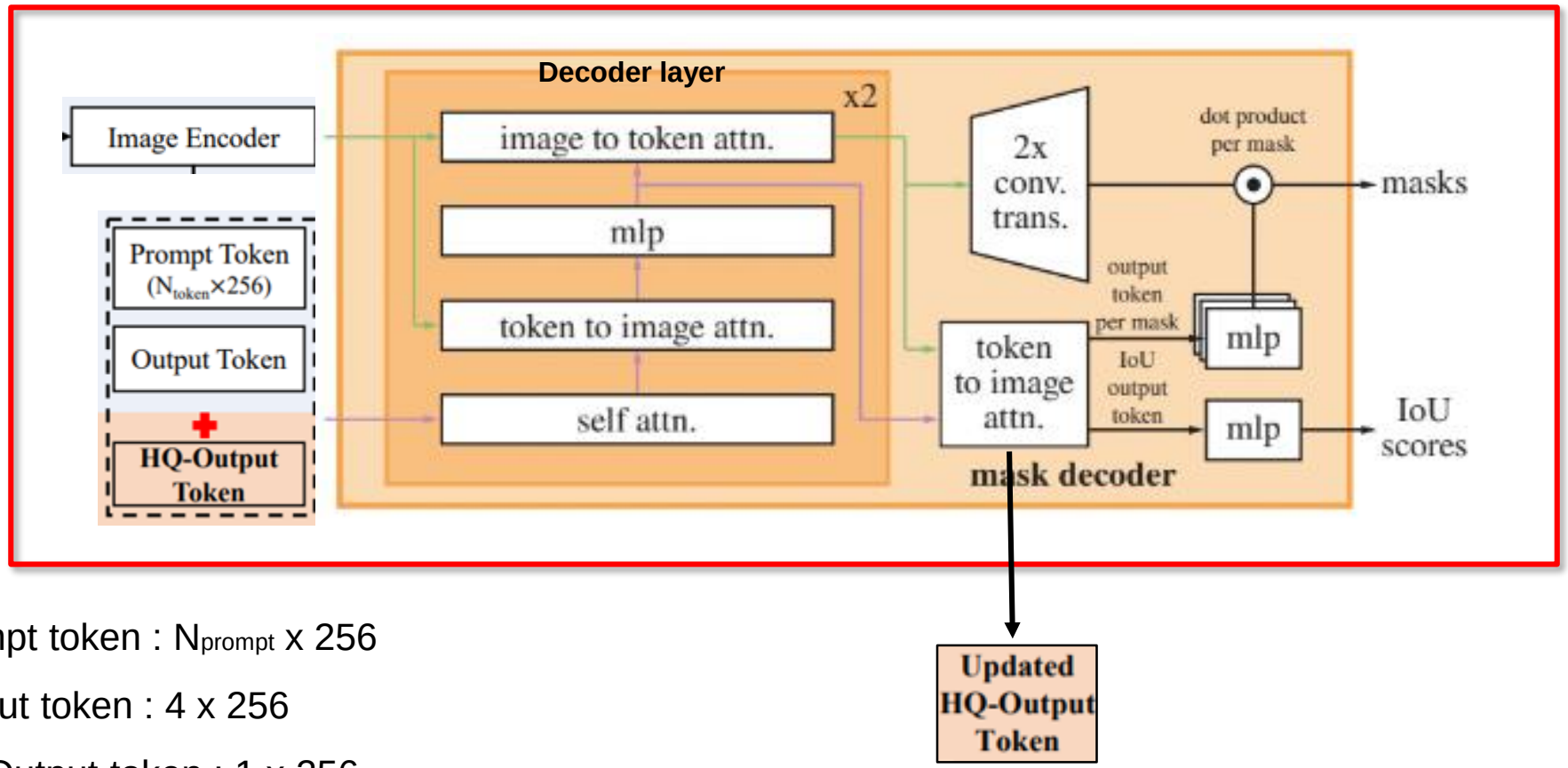
2-1. High-Quality Output Token

To promote SAM's mask quality in HQ-SAM, instead of directly taking SAM's coarse masks as input, we introduce the HQ-Output token and a new mask prediction layer for high-quality mask prediction.



3.Method

2-1. High-Quality Output Token



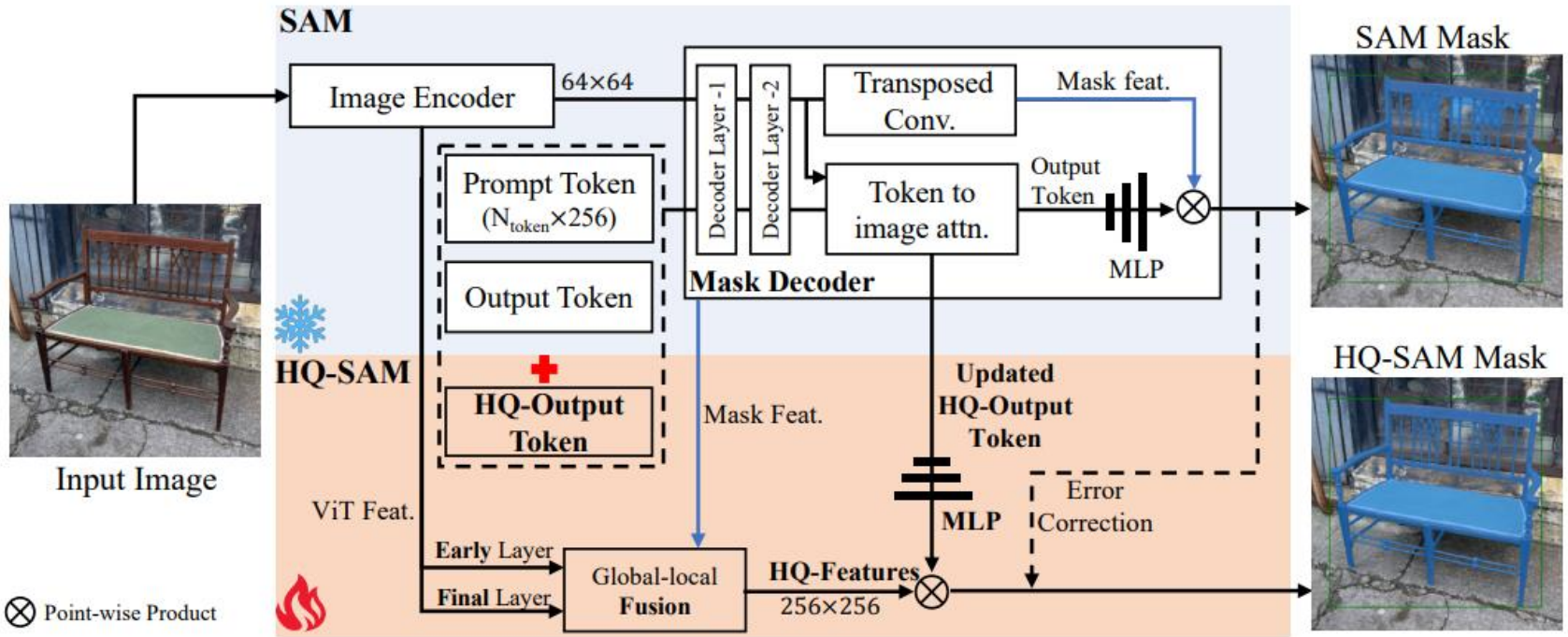
Prompt token : N_{prompt} x 256

Output token : 4 x 256

HQ-Output token : 1 x 256

3.Method

2-1. High-Quality Output Token



3.Method

2-1. High-Quality Output Token

Advantages

1. This strategy significantly improves SAM's mask quality while only 4 introducing negligible parameters compared to original SAM, making HQ-SAM training extremely time and data-efficient.

2. The learned token and MLP layers do not overfit to mask the annotation bias of a specific dataset, thus keeping SAM's strong zero-shot segmentation capability on new images without catastrophic knowledge forgetting.

3.Method

2-2. Global-local Fusion for High-quality Features

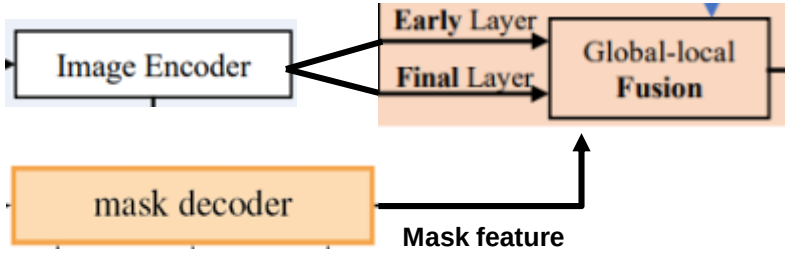
Very accurate segmentation also requires input image feature with both rich global semantic context and local boundary details.

To further promote mask quality, we enrich both the high-level object context and low-level boundary/edge information in the mask decoder features of SAM.

Instead of directly using SAM’s mask decoder feature, we compose the new high-quality features (HQFeatures) by extracting and fusing features from different stages of the SAM model.

3.Method

2-2. Global-local Fusion for High-quality Features



1. The early layer local feature of SAM's ViT encoder with spatial shape 64×64

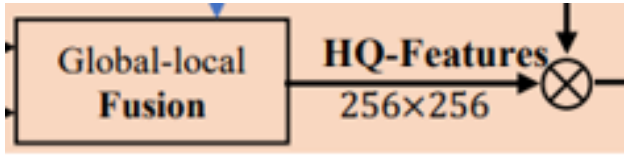
   Concretely, we extract the feature after the first global attention block of the ViT encoder, and for ViT-Large based SAM, this is the 6th block output for the 24 blocks in total

2. The final layer global feature of SAM's ViT encoder with shape 64×64, which has more global image context information

3. The mask feature in SAM's mask decoder with size 256×256, which is also shared by the output tokens, contains strong mask shape information

3.Method

2-2. Global-local Fusion for High-quality Features



To obtain the input HQ-Features

1. upsample the early-layer and final-layer encoder features to the spatial size 256×256 by transposed convolution
2. sum up these three types of features in an element-wise manner after simple convolutional processing.

Global-local feature fusion is simple while effective, yielding detail-preserving segmentation results with a small memory footprint and computation burden

3.Method

3-1. Training Data Construction

- Training Dataset : HQSeg-44k
- SA-1B dataset only contains automatically generated mask labels, missing very accurate manual annotation on objects with complex structures.
- HQSeg-44k leverages a collection of six existing image datasets including DIS, ThinObject-5K, FSS-1000, ECSSD, MSRA10K, DUT-OMRON with extremely fine-grained mask labeling
- HQSeg-44k contains diverse semantic classes of more than 1,000

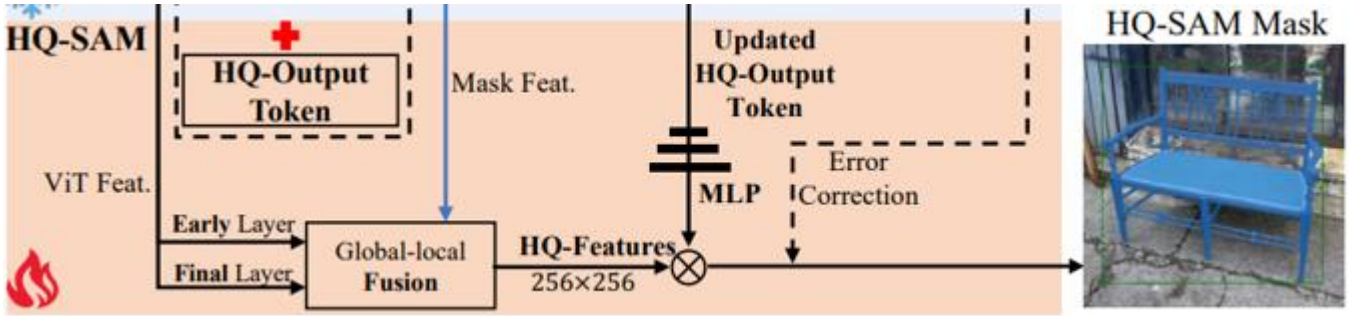
3.Method

3-2. HQ-SAM Training

- Learnable parameters: HQ-Output token, associated 3-layer MLP, 3 simple convolutions for HQ-Feature fusion
- Sampling a mixed type of prompts including bounding box, randomly sampled points, and rough mask inputs
- Generating degraded masks by adding random Gaussian noise to the boundary regions of GT masks

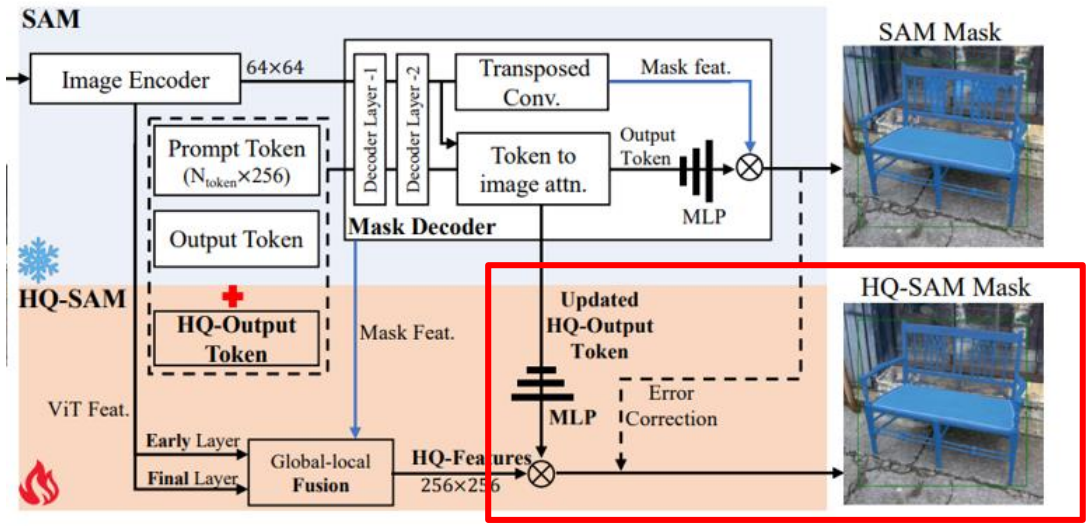
▶ Parameter

- Learning rate : 0.001 (10epochs 마다 감소)
- Epochs : 12
- Batch size : 32



3.Method

3-3. Inference



Predicted logits of the SAM mask
(by Output Token)

+

Our predicted mask(256x256)
(by HQ-Output Token)

Resize (1024x1024)

HQ-SAM Mask

4.Results

- SAM vs. HQ-SAM on Training and Inference

Table 1: Training and inference comparison between ViT-L [11] based SAM and HQ-SAM. HQ-SAM brings negligible extra computation burden to SAM, with *less than 0.5% increase* in model parameters and reaching 96% of its original speed. SAM-L is trained on 128 A100 GPUs for 180k iterations. Based on SAM-L, we only need to train our HQ-SAM on 8 RTX3090 GPUs for 4 hours.

Method	Learnable Params (M)	Training # GPU	Batch Size	Time (h)	Inference FPS	Inference Mem.
SAM [21]	1191	128	128	N/A	5.0	7.6G
HQ-SAM	**5.1**	**8**	**32**	**4**	**4.8**	**7.6G**

4.Results

- Effect of the High-Quality Output Token

Table 2: Ablation study of the HQ-Output Token on four extremely fine-grained segmentation datasets. We adopt the boxes converted from their GT masks as the box prompt input. By default, we train the predicted mask of HQ Output-Token by computing full GT mask loss.

Model	DIS [35]		COIFT [29]		HRSOD [51]		ThinObject [29]		Average	
	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
SAM (**baseline**)	62.0	52.8	92.1	86.5	90.2	83.1	73.6	61.8	79.5	71.1
Using SAM's mask decoder feature:										
SAM + Context Token [56]	71.5	62.2	93.0	87.7	91.8	85.0	84.5	73.1	85.2	77.0
SAM + HQ-Output Token (× Output Token)	75.1	65.8	93.9	88.9	93.0	86.1	86.1	74.6	87.0	78.9
SAM + HQ-Output Token (Boundary Loss)	75.2	66.4	94.0	88.9	92.1	85.7	87.3	76.0	87.2	79.3
SAM + HQ-Output Token	75.3	66.0	94.2	89.2	93.0	86.1	86.8	75.4	87.3	79.2
Using Our HQ-Feature:										
SAM + HQ-Output Token (+ Context Token)	78.5	70.4	94.6	89.6	93.6	**87.0**	88.9	79.3	88.9	81.6
SAM + HQ-Output Token	**78.6**	**70.4**	**94.8**	**90.1**	**93.6**	86.9	**89.5**	**79.9**	**89.1**	**81.8**

Context token : learnable vectors into the SAM's mask decoder for better context learning

4.Results

- Ablation on the Global-local Fusion for HQ-Features

Table 3: Ablation study on the HQ-Features sources. Early-layer denotes the feature after the first global attention block of the ViT encoder, while final-layer denotes the output of the last ViT block. Four HQ datasets denote DIS (val) [35], ThinObject-5K (test) [29], COIFT [29] and HR-SOD [51].

Model	Fusion conv	Decoder Mask feature	ViT Encoder Final-layer	ViT Encoder Early-layer	Four HQ datasets mIoU	Four HQ datasets mBIoU
SAM [21]		✓			79.5	71.1
HQ-SAM (Ours)		✓			87.3	79.2
	✓	✓			87.8	80.1
	✓		✓		15.1	9.0
	✓	✓	✓		88.6	81.3
		✓	✓	✓	88.6	81.1
	✓	✓	✓	✓	**89.1**	**81.8**

4.Results

- Zero-shot Comparison with SAM

Zero-Shot Open-world Segmentation

Model	AP_B^{strict}	AP_{B75}^{strict}	AP_{B50}^{strict}	AP_B	AP_{B75}	AP_{B50}	AP
SAM	8.6	3.7	25.6	17.3	14.4	37.7	29.7
HQ-SAM	**9.9**	**5.0**	**28.2**	**18.5**	**16.3**	**38.6**	**30.1**

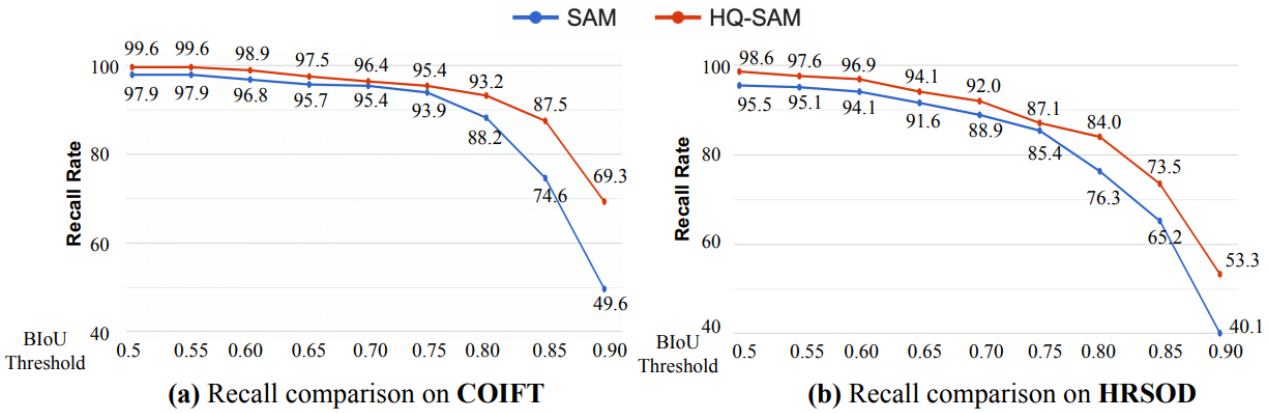
Zero-Shot Segmentation on High-resolution BIG Dataset

Model	GT Box Prompt		Mask Prompt	
	mIoU	mBIoU	mIoU	mBIoU
SAM	81.1	70.4	66.6	41.8
HQ-SAM	**86.0**	**75.3**	**86.9**	**75.1**

Zero-shot Instance Segmentation on COCO and LVIS

Model	COCO		LVIS				
	AP_B	AP	AP_B^{strict}	AP_{B75}^{strict}	AP_B	AP_{B75}	AP
SAM	33.3	48.5	32.1	32.8	38.5	40.9	43.6
HQ-SAM	**34.4**	**49.5**	**32.5**	**33.5**	**38.8**	**41.2**	**43.9**

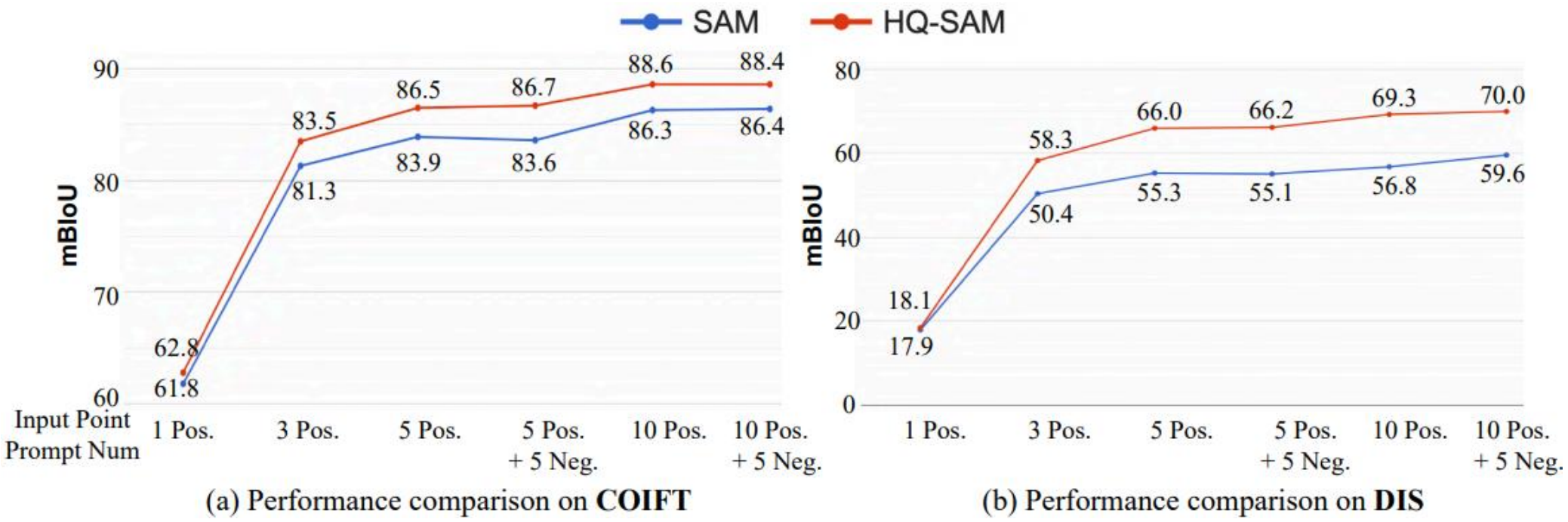
Recall rate comparison between COIFT and HRSOD under the zero-shot protoc



(a) Recall comparison on **COIFT** **(b)** Recall comparison on **HRSOD**

4.Results

- Point-based Interactive Segmentation Comparison (with different Prompt Num)



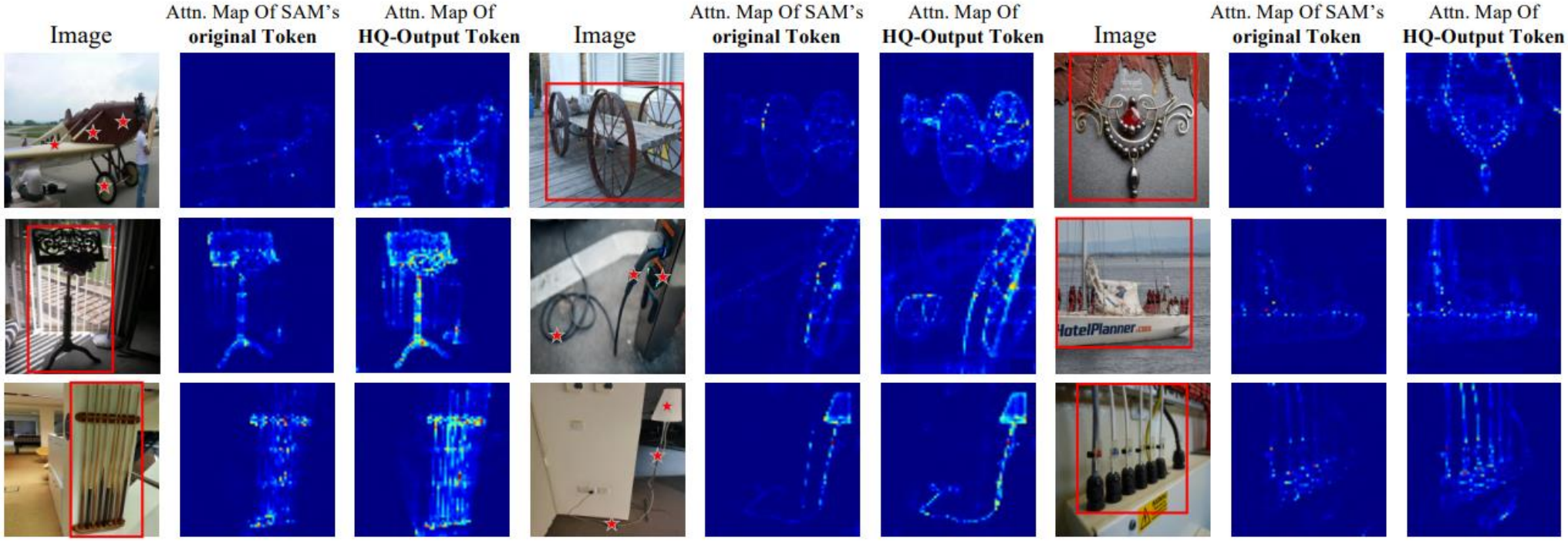
(a) Performance comparison on **COIFT**

(b) Performance comparison on **DIS**

4.Results

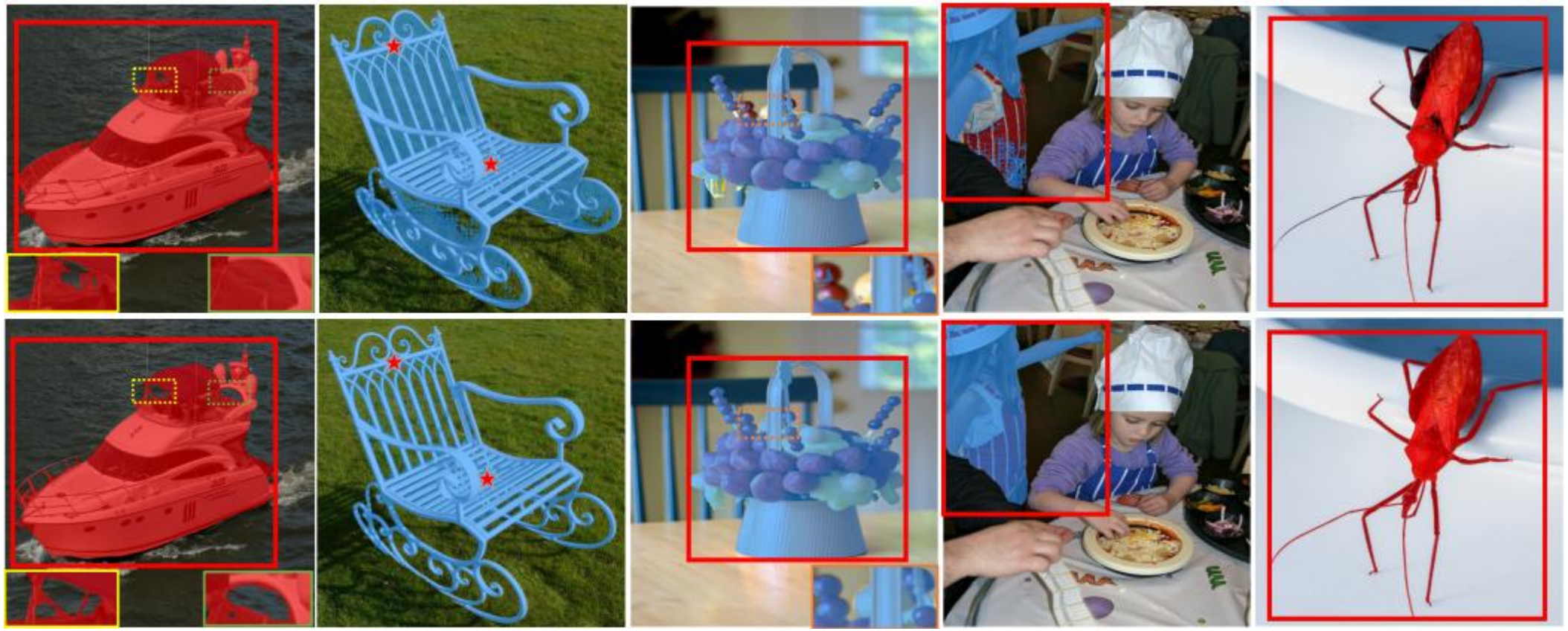
- Cross-attention of SAM's original token vs. HQ-Output Token in the last decoder layer.

HQ-Token attends to the boundary and thin structure regions that are missed by the original token



4.Results

- Visual results comparison between SAM (top row) vs. HQ-SAM (bottom row) in a zero-shot transfer setting



4.Results

- Visual results comparison between SAM (top row) vs. HQ-SAM (bottom row) in a zero-shot transfer setting