Segment and Matte Anything in a Unified Model

Zezhong Fan*, **Xiaohan Li***, **Topojoy Biswas**, **Kaushiki Nag**, **Kannan Achan**

Personalization Team, Walmart Global Tech
Sunnyvale, CA, USA
{zezhong.fan, xiaohan.li, topojoy.biswas, kaushiki.nag, kannan.achan}@walmart.com

Abstract

Segment Anything (SAM) has recently pushed the boundaries of segmentation by demonstrating zero-shot generalization and flexible prompting after training on over one billion masks. Despite this, its mask prediction accuracy often falls short of the precision required in real-world applications. While several refinement modules have been proposed to boost SAM's segmentation quality, achieving highly accurate object delineation within a single, unified framework remains an open challenge. Furthermore, interactive image matting—which aims to generate fine-grained alpha mattes guided by diverse user hints—has not yet been explored in the context of SAM. Insights from recent studies highlight strong correlations between segmentation and matting, suggesting the feasibility of a unified model capable of both tasks.

In this paper, we introduce Segment And Matte Anything (SAMA), a lightweight extension of SAM that delivers high-quality interactive image segmentation and matting with minimal extra parameters. Our Multi-View Localization Encoder (MVLE) captures detailed features from local views, while the Localization Adapter (Local-Adapter) refines mask outputs by recovering subtle boundary details. We also incorporate two prediction heads for each task into the architecture to generate segmentation and matting masks, simultaneously. Trained on a diverse dataset aggregated from publicly available sources, SAMA achieves state-of-the-art performance across multiple segmentation and matting benchmarks, showcasing its adaptability and effectiveness in a wide range of downstream tasks.

Introduction

Precise object segmentation lies at the heart of computer-vision applications, from photo editing and augmented reality to autonomous driving and medical analysis. Two complementary problems dominate this landscape. Semantic/instance segmentation assigns a class label to every pixel, while natural image matting predicts a continuous alpha matte that captures fine, semi-transparent boundaries such as hair or glass. Segment Anything (SAM) (Kirillov et al. 2023) represents a milestone in segmentation research: trained on over one billion masks, it exhibits remarkable zero-shot generalization and supports diverse prompting modalities (points, boxes, text). Nevertheless, SAM's raw masks often lack the tight boundaries, sub-pixel accuracy and detail preservation as discussed in (Ke et al. 2023; Liu, Fu, and Zhao 2023; Liu et al. 2024a; Fan et al. 2024).

Recently, researchers improve SAM with dedicated refinement modules. For example, HQ-SAM (Ke et al. 2023) extends the original SAM by introducing a learnable High-Quality (HQ) output token into the mask decoder to enhance the quality of mask prediction. Several other approaches, such as DIS-SAM (Liu, Fu, and Zhao 2023), SAMRefiner (Lin et al. 2025) and Pi-SAM (Liu et al. 2024a), have attempted to address this limitation. However, these methods typically require additional post-processing models or rely on extra human interactions to refine the input prompts, thereby increasing model complexity and reducing practicality. We identify two primary challenges that hinder these SAM-based models from achieving accurate segmentation. First, interactive segmentation models like SAM struggle to capture detailed structures of target objects due to limited fine-grained perception. Second, it remains difficult to integrate high-resolution detail into the decoding process without compromising SAM's strong zero-shot generalization ability. Addressing these challenges is critical for advancing fine-grained, high-quality segmentation in complex scenes.

Interactive matting (Li, Jain, and Shi 2024; Yao et al. 2024b), in contrast, focuses on estimating accurate alpha mattes under sparse user guidance (e.g., trimaps, scribbles, or clicks). While classical matting networks achieve remarkable boundary detail, they struggle with object-level reasoning and cannot generalize across categories without extensive retraining. Importantly, recent studies reveal strong structural correlations between segmentation and matting (Wang and Cohen 2005; Zheng et al. 2024): segmentation offers global object cues, whereas matting supplies local boundary precision. Leveraging these synergies within a unified model promises both practical simplicity and performance gains, yet remains largely unexplored.

To address these challenges, we present Segment And Matte Anything (SAMA), a lightweight extension of SAM that unifies high-accuracy segmentation and interactive matting in a single framework. It includes three key components. First, the Multi-View Localization Encoder (MVLE) enhances spatial precision by aggregating localized details from the multiple local views, capturing fine structures that the coarse global encoder may overlook. Second, the Local-

*These authors contributed equally.

ization Adapter (Local-Adapter) refines mask predictions by injecting local features into the decoding process to integrate fine-grained local features into the decoding process. Furthermore, we extend our framework on both image segmentation and matting tasks with two prediction heads for each task. Specifically, we introduce a lightweight up-sampling module, enabling SAMA to produce both high-quality segmentation and matting masks simultaneously. This unified design allows seamless task transfer without architectural modification of the encoder and decoder. Importantly, all SAM parameters are kept frozen during training, and only the proposed modules are fine-tuned. This strategy ensures that our approach remains both data-efficient and computationally lightweight. Collectively, these modules add only 1.8% to SAM's parameter count and impose only marginal latency.

We utilize publicly open-sourced datasets including segmentation masks with high-quality alpha mattes and train SAMA end-to-end on both tasks. Comprehensive experiments on standard segmentation suites and matting benchmarks demonstrate that SAMA outperforms prior interactive segmentation and matting networks, while retaining SAM's advantage of prompting flexibility.

Our contributions can be summarized as follows:

- ☐ Unified framework: We propose the first SAM-based model that jointly performs interactive segmentation and matting with minimal overhead.
- ☐ Architectural advances: We design a Multi-View Localization Encoder, Localization Adapter, and Prediction Heads to bridge object-level context and boundary-level detail.
- ☐ State-of-the-art results: SAMA achieves new performance records across diverse segmentation and matting benchmarks without sacrificing inference speed or prompting versatility.

Related Works

Interactive Segmentation and Matting

Interactive segmentation allows users to steer the extraction of target regions through prompts such as bounding boxes (Kirillov et al. 2023; Ke et al. 2023; Liu et al. 2024a), points (Kirillov et al. 2023; Ke et al. 2023; Yao et al. 2025), or natural-language descriptions (Zou et al. 2023; Nguyen et al. 2024; Fan et al. 2025). Recent work embeds these prompts directly into the network to condition its predictions, with the Segment Anything Model (SAM) (Kirillov et al. 2023)—pretrained on over one billion masks—emerging as the de-facto benchmark. Some methods refine SAM to boost either accuracy, such as HQ-SAM (Ke et al. 2023), SAMRefiner (Lin et al. 2025) and DIS-SAM (Liu, Fu, and Zhao 2023). Meanwhile, there are also some papers to extend the functionality of SAM in semantic segmentation (Zou et al. 2023; Li et al. 2024), iterative click-based refinement (Liu et al. 2024a), cross-modal inputs (Xiao et al. 2024) and segmentation with large vision–language models (Nguyen et al. 2024).

Variants of SAM have also been adapted for image matting. MAM (Li, Jain, and Shi 2024) transforms SAM features into alpha mattes using a lightweight mask-to-matte (M2M) head, while MatAny (Yao et al. 2024b) generates a trimap with SAM and feeds it to VitMatte (Yao et al. 2024a) for high-quality results. Despite their effectiveness, these approaches still depend on additional heavy models (Yao et al. 2024a) or cascaded modules.

Recognizing the strong synergy between segmentation and matting (Wang and Cohen 2005; Zheng et al. 2024), we present a unified framework that augments SAM with a lightweight matting head, delivering precise segmentation masks and high-fidelity alpha mattes with minimal computational overhead.

High-Quality Segmentation and Matting

Segmentation: Accurate delineation of fine-grained, complex objects underpins numerous sub-tasks, including dichotomous image segmentation (DIS) (Qin et al. 2022; Yu et al. 2024; Zheng et al. 2024), semantic segmentation (Long, Shelhamer, and Darrell 2015; Zhao et al. 2017; Cheng, Schwing, and Kirillov 2021), instance segmentation (He et al. 2017; Dai, He, and Sun 2016), and panoptic segmentation (Kirillov et al. 2019; Cheng et al. 2020). Classic CNN-based approaches (He et al. 2017; Qin et al. 2022; Long, Shelhamer, and Darrell 2015; Zhao et al. 2017; Dai, He, and Sun 2016; Yu et al. 2024) design sophisticated multi-scale modules to fuse low-level texture with high-level semantics via diverse receptive fields. Transformer-based models push this further with self-attention windows to capture local details while retaining global context (Kirillov et al. 2023; Zheng et al. 2024; Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022).

Matting: Image matting methods fall into two streams. (1) *Trimap-based matting* supplies a foreground/background/unknown trimap to networks, enabling deep models to produce precise alpha mattes (Xu et al. 2017; Lutz, Amplianitis, and Smolic 2018; Tang et al. 2019; Lu et al. 2019; Hou and Liu 2019; Li and Lu 2020; Yao et al. 2024a; Park et al. 2022). (2) *Trimap-free matting* predicts the matte directly. Although more convenient, these methods still need auxiliary cues such as segmentation masks (Yu et al. 2021; Huynh et al. 2024), motion information (Sengupta et al. 2020), or prompt signals (Li, Jain, and Shi 2024; Yao et al. 2024b).

As segmentation and matting are inherently complementary (Wang and Cohen 2005; Zheng et al. 2024), we propose a single unified network that includes lightweight prediction heads with a high-quality interactive segmentation backbone based on SAM, enabling both tasks to share features and mutually enhance each other, while incurring only minimal computational overhead.

Methodology

We propose a unified model Segment And Matte Anything (SAMA) to leverage SAM achieving both highly accurate segmentation and interactive image matting.

Preliminary

SAM (Kirillov et al. 2023) consists of three components: An image encoder which is a VIT (Dosovitskiy et al. 2020) backbone produces a 64×64 spatial feature map for the input image. A prompt encoder embeds user interactions (points, boxes, or masks) as positional tokens. Then combining image features with prompt tokens, a mask decoder, which is a two-layer transformer, is used to predict the final segmentation mask. To acquire the zero-shot transfer capability, SAM is trained on SA-1B, a dataset containing 11 million images and over 1 billion automatically generated masks.

Image matting is to estimate alpha matte α given only image I as the input. Formally, given an image I, which can be viewed as a combination of foreground image F and background image B with coefficient α,

$$I = \alpha F + (1 - \alpha) B \tag{1}$$

SAMA

SAMA is a unified model that enables both highly accurate image segmentation and matting while preserving the zero-shot generalization capabilities of SAM. Unlike conventional single-image inputs, SAMA introduces a multi-view input strategy by treating the original image as a global view and incorporating additional local views to capture fine-grained object details. To effectively extract and fuse high-resolution features, we design three key components: Multi-view Localization Encoder (MVLE), Localization Adapter (Local-Adapter) and Matting Module, as illustrated in Figure 1.

Multi-view Localization Encoder (MVLE) In the SAM framework, an input image $I \in \mathbb{R}^{3\times H\times W}$ is passed through an pre-trained image encoder $\mathcal{E}$ to produce a global feature map $F^I \in \mathbb{R}^{C\times \frac{H}{16}\times\frac{W}{16}}$. However, relying solely on this single global representation limits the model's ability to capture fine-grained visual details. To address this limitation, we introduce the Multi-View Localization Encoder (MVLE), which enhances object localization through the use of high-resolution local views.

Inspired by human vision, we divide high-resolution inputs into distant-view global contexts and close-view local details, following MVANet (Yu et al. 2024), to promote comprehensive scene understanding. Specifically, we evenly crop the input image I into four non-overlapping local patches $\{L_m\}_{m=1}^4 \in \mathbb{R}^{3\times h\times w}$, such that $(H, W) = (2h, 2w)$ as (Yu et al. 2024). Each cropped patch is then up-sampled back to the original resolution and passed through the same encoder $\mathcal{E}$, yielding m high-resolution local feature maps $F^{L_m} \in \mathbb{R}^{B\times C\times \frac{H}{16}\times\frac{W}{16}}$. The resulting local feature maps are stacked to form a m-layer (m = 4 here) local feature map $F^L \in \mathbb{R}^{B\times 4\times C\times \frac{H}{16}\times\frac{W}{16}}$.

To effectively align local features with their global context, we apply a cross-attention mechanism between local features F^{L_m} and the global feature F^I. First, we apply average pooling with multiple receptive fields (e.g., 4, 8, 16) to F^I to obtain a multi-scale context representation F^I_{pool}. We then partition F^I_{pool} into four spatial regions $\{I_m\}_{m=1}^4$, each corresponding to the position of a local patch. Within each region, we perform multi-head cross-attention, treating the local features as queries and the pooled global features as keys and values:

$$\mathbf{Q}_m = F^{L_m}\mathbf{W}^{Q_m}, \mathbf{K}_m = F^{I_m}_{pool}\mathbf{W}^{K_m}, \mathbf{V}_m = F^{I_m}_{pool}\mathbf{W}^{V_m} \tag{2}$$

$$F'^{P_m} = \text{Cross-Attn}(\mathbf{Q}_m, \mathbf{K}_m, \mathbf{V}_m), \tag{3}$$

where $\mathbf{W}^Q_m, \mathbf{W}^K_m, \mathbf{W}^V_m$ are learnable projection matrices for the m-th patch. The output of the cross-attention layer in MVLE is the updated local features F'^{P_m}. These refined features are then passed to decoders to support precise mask prediction.

Localization Adapter (Local-Adapter) Local-Adapter is a dedicated module designed to inject fine-grained visual information from high-resolution local features into the SAM decoder via a specialized local attention mechanism. Specifically, after processing the local features through one layer of the SAM decoder, we divert the output to Local-Adapter instead of directly passing it to the next decoder layer.

SAM original decoder is a two-way Transformer. The input of the decoder layer is composed of two parts. The first part is the global feature map $F^{I_{\text{token}}}$ to serve as image tokens to the decoder layer. The other one is the concatenation of the prompt token, SAM token provided by SAM and our proposed SAMA token initialized in the model. Specifically, To enable the model to perform segmentation and matting simultaneously, we replace SAM's original output token with learnable SAMA tokens—two tokens dedicated to segmentation and matting, respectively. These SAMA tokens are concatenated with the original SAM tokens and fed jointly into the decoder layers. The prompt token and SAM token are frozen while only SAMA token is trainable. Please note that the SAMA token is distinct in segmentation and matting tasks, i.e. two SAMA tokens are used separately when training and inference on both tasks.

After each decoder layer, our proposed Local-Adapter is followed to enhance the local fine-grained features. There are three steps in Local-Adapter:

1) First cross-attention layer. We use the output of the cross-attention layer in MVLE as the input in this layer. To further enhance boundary sensitivity, we extract early-layer features from the image encoder as mentioned in (Ke et al. 2023), denoted as F_{early}, and then they are fused with local features via a residual connection. The fused tokens are served as the keys and values in this layer, while the decoder outputs F_{out} provide the queries, allowing local and global information to be integrated seamlessly.

$$\mathbf{Q}_\mathbf{A} = F_{\text{out}}\mathbf{W}^A, \tag{4}$$

$$\mathbf{K}_{\mathbf{A}m} = (F'^{P_m} + F_{\text{early}})\mathbf{W}^{K_A}, \tag{5}$$

$$\mathbf{V}_{\mathbf{A}m} = (F'^{P_m} + F_{\text{early}})\mathbf{W}^{V_A}, \tag{6}$$

$$F''^{P_m} = \text{Cross-Attn}(\mathbf{Q}_\mathbf{A}, \mathbf{K}_{\mathbf{A}m}, \mathbf{V}_{\mathbf{A}m}), \tag{7}$$

2) Second cross-attention layer. Inspired by GLIP(Li et al. 2022) and GroundingDINO (Liu et al. 2024b), we introduce a feature fusion with global-local and local-global cross-attention modules by swapping keys and values in the previous layer, enabling bidirectional interaction between global context and local details. This allows the Local-Adapter to become both globally and locally aware within the decoders. Similar to the first layer, we swap roles: the keys and values produced in the previous layer become the queries for

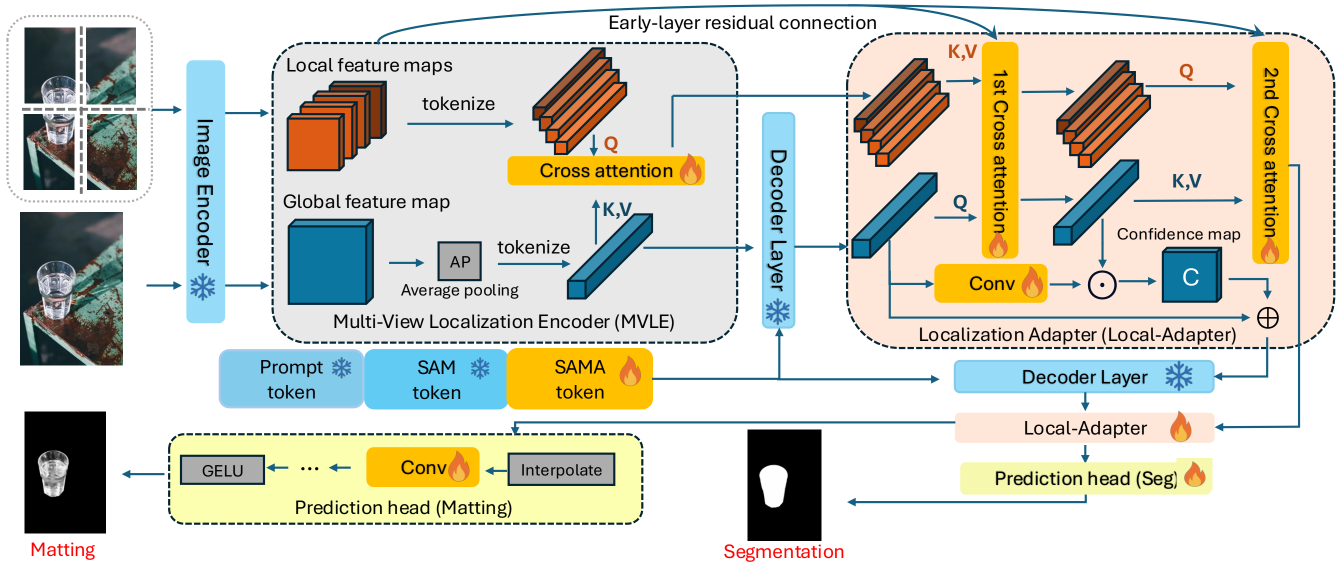

Figure 1: SAMA Overall Framework.

this layer, while the previous queries now serve as keys and values.

3) Generation of output features. There are two output features from Local-Adapter. The first is the tokens from the output of the second cross-attention layer, which will be used as the input of the second Local-Adapter in the SAMA model. The second feature is the updated global feature map F'_{out}, which is used as the input of the second decoder layer. It is the combination of a confidence map C and the global feature from the output of the decoder layer F_{out}. The confidence map is to maintain SAM's strong zero-shot generalization and mitigate the risks of overfitting or catastrophic forgetting. A confidence map C is generated using a 1×1 convolution followed on the F_{out} and activated by a sigmoid function, and multiple the output of the first cross attention layer F''^{P_m}. The second output feature is calculated from

$$C = \sigma\left(\text{Conv}(F_{\text{out}})\right) \odot F''^{P_m}, \tag{8}$$

$$F'_{\text{out}} = F_{\text{out}} + C, \tag{9}$$

where $\odot$ denotes element-wise product. This formulation allows the model to adaptively blend detailed information from the local features with the original decoder output, thereby achieving a better balance between precision and generalization. After the Local-Adapter, there is another set of decoder layer and Local-Adapter to increase the depth of the model and improve the performance.

Prediction Heads

Here we introduce how SAMA predicts final segmentation and alpha matting masks. First, following (Ke et al. 2023), we introduce two trainable output tokens-a segmentation token and a matting token (all shown as SAMA token in Figure 1)—designed to generate high-quality outputs for their respective tasks. These tokens are processed by the SAMA decoder, which provides semantic features as global priors for final prediction heads. To enhance fine-grained predictions, we employ two lightweight task-specific prediction heads for segmentation and matting. Each head integrates an interpolation operation for upsampling with convolutional layers that collaboratively reconstruct and enhance details. The convolutional layers, along with batch normalization and GeLU activation function, are to generate fine-grained feature maps from the output of Local-Adapter. This design enables SAMA to produce high-resolution segmentation and matting masks simultaneously, achieving both semantic coherence and boundary-level precision.

Training of SAMA

Training Data Construction To enable efficient and effective training of SAMA, we opt for high-quality segmentation datasets with exceptionally accurate mask annotations- DIS-5K (Qin et al. 2022) and ThinObject-5K (Liew et al. 2021), instead of relying on the large-scale but noisier SAM-1B dataset. For training the matting task, we utilize a combination of Adobe Image Matting (AIM) (Xu et al. 2017) and AIM-500 (Li, Zhang, and Tao 2021), which together provide diverse and representative foreground objects across a range of natural scenes.

SAMA Training During training, we freeze all parameters of the pre-trained SAM backbone and update only the our proposed modules. When optimizing for the segmentation task, the matting prediction head remains frozen, and vice versa during matting training. Additional implementation details are provided in the supplementary material.

Training Loss We train SAMA model end-to-end using a multi-task loss to learn segmentation and matting tasks concurrently:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{matting}}$$

For segmentation training, we employ a composite loss function that integrates pixel, region, and boundary-aware supervision:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{SSIM}}$$

where binary cross-entropy (BCE) loss provides pixel-level supervision to guide the generation of binary masks, intersection over union (IoU) loss introduces region-level constraints to enhance the overall segmentation quality, and structural similarity index measure (SSIM) loss encourages structural similarity, particularly improving mask accuracy near object boundaries.

For matting, we adopt a more fine-grained objective that accounts for subtle variations in transparency and edge quality:

$$\mathcal{L}_{\text{matting}} = \mathcal{L}_{l_1} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{Grad}} + \mathcal{L}_{\text{Laplacian}} \qquad (10)$$

where ℓ_1 loss ensures global consistency, SSIM preserves structural similarity, gradient loss (Dai et al. 2022) improves edge sharpness, and Laplacian loss (Hou and Liu 2019) captures high-frequency details.

Experiments

In this section, we present experiment settings and comparisons for both segmentation and matting across multiple benchmarks. We further evaluate SAMA's performance under point-based interactive segmentation and zero-shot semantic image matting settings. To assess the contribution of the different modules, we conduct ablation studies.

Comparison Study on Segmentation task

We conduct experiments on an extremely fine-grained segmentation datasets: DIS-5K (Qin et al. 2022) including a validation set with 470 images (DIS-VD), four subsets DIS-TE1 $\sim$ TE4 (500 images in each set) with increasing shape complexities, and DIS-TE (All 2000 testing images). We compare SAMA with SAM (Kirillov et al. 2023), HQ-SAM (Ke et al. 2023), Pi-SAM (Liu et al. 2024a), DIS-SAM (Liu, Fu, and Zhao 2023) IS-Net (Qin et al. 2022), UDUN (Pei et al. 2023) and BiRefNet (Zheng et al. 2024) tailored for the task of dichotomous image segmentation. For SAM, HQ-SAM, Pi-SAM, DIS-SAM and our SAMA, bounding boxes are provided as prompts for images, while ISNet, UDUN and BiRefNet only take images as the input. For a thorough performance evaluation of segmentation, we report maximum F-measure (F_β^{max} (Achanta et al. 2009)), weighted F-measure (F_β^w), mean absolute error (MAE), S-measure (S_α (Fan et al. 2017)), and average enhanced alignment measure (E_ϕ^m (Fan et al. 2018)) as evaluation metrics.

As shown in Table 1, our proposed SAMA consistently outperforms other models that are built based on the original SAM architecture, highlighting the effectiveness of our newly introduced modules and fine-tuning strategy. Furthermore, when compared with models specifically designed and extensively trained on the DIS dataset, our SAMA achieves competitive performance across multiple evaluation metrics. Even though the dataset is more complex on DIS TE3 and TE4, our SAMA still achieves close to SOTA performance. It should be noted that these baseline models are often trained with significantly more epochs, allowing them to better fit the inherent distributional characteristics of the dataset. However, due to their reliance on fully automatic segmentation pipelines, such models lack the flexibility to support interactive segmentation, which remains a key strength of our approach.

Comparison Study on Matting

To assess the general matting performance of our proposed models, we conduct evaluations on two widely adopted benchmarks: Composition-1K (Xu et al. 2017) and Distinctions-646 (Qiao et al. 2020). We compare our SAMA with two categories of matting approaches: (1) trimap-free methods such as LFM (Zhang et al. 2019), MODNet (Ke et al. 2022), and MFC-Net (Zhao 2024), and (2) trimap-based methods that leverage additional trimap guidance, including Information-Flow (I-F) (Aksoy, Ozan Aydin, and Pollefeys 2017), DIM (Xu et al. 2017), DCNN (Cho, Tai, and Kweon 2016), MGMatting (Yu et al. 2021), and VITMatte (Yao et al. 2024a). To quantify performance on alpha matting task, we use commonly adopted evaluation metrics including Sum of Absolute Difference (SAD) and Mean Squared Error (MSE).

As reported in Table 2, our model achieves state-of-the-art performance among trimap-free methods on both benchmarks. Notably, SAMA, without relying on any trimap input, demonstrates substantial improvements across diverse visual scenes. When compared to leading trimap-based approaches, such as VITMatte, our SAMA achieves comparable results, highlighting its strong potential to generalize well in real-world matting scenarios while maintaining the advantages of a trimap-free framework.

Visual Results Comparison

From Figure 3, we show the visual results comparison between SAM, HQ-SAM, and our SAMA, given the same red box prompt. Our SAMA produces significantly more detailed results. For example, SAMA's result for the chair on the first row exhibits clear mesh in the mask. On the fourth row, the thin lines on the gate are also segmented out from the input image. These examples demonstrate SAMA's ability in segmenting detailed features from images.

Additionally, Figure 4 shows the visual results comparison between MatAny, MAM and our SAMA. SAMA also presents visible improvements on the matting task. For instance, the matting mask from SAMA on the second row clearly shows details of the woman's hair. Besides, on the fifth row, the transparency of the glasses is also obvious on SAMA's output mask. These results indicate that SAMA can handle the transparency and hair/fur in the matting mask.

Point Prompts Effect Comparison

Followed (Ke et al. 2023), to analyze how interactive point prompts affect the performance of segmentation, we evaluate SAMA with varying numbers of input points on COIFT (Liew et al. 2021), which is a zero-shot interactive segmentation dataset about thin objects. As illustrated in Figure 2, SAMA consistently achieves higher mean Intersection over Union (mIoU) scores than both HQ-SAM and

Method	DIS-VD					DIS-TE1					DIS-TE2				
	F_β^{max}	F_β^w	M	S_α	E_ϕ^m	F_β^{max}	F_β^w	M	S_α	E_ϕ^m	F_β^{max}	F_β^w	M	S_α	E_ϕ^m
IS-Net	0.791	0.717	0.074	0.813	0.856	0.740	0.662	0.074	0.787	0.820	0.799	0.728	0.070	0.823	0.858
UDUN	0.823	0.763	0.059	0.838	0.892	0.784	0.720	0.059	0.817	0.860	0.829	0.768	0.058	0.843	0.886
BiRefNet	0.891	0.854	0.038	0.898	0.931	0.860	0.819	0.037	0.885	0.911	0.894	0.857	0.036	0.900	0.930
SAM	0.835	0.782	0.069	0.808	0.889	0.838	0.807	0.047	0.843	0.805	0.803	0.758	0.081	0.792	0.863
HQ-SAM	0.851	0.829	0.045	0.848	0.919	0.903	0.888	0.019	0.907	0.959	0.895	0.874	0.029	0.883	0.950
Pi-SAM	0.883	0.866	0.035	0.889	0.945	0.890	0.869	0.027	0.894	0.947	0.903	0.887	0.027	0.907	0.953
DIS-SAM	0.920	0.877	0.031	0.909	0.948	0.929	0.897	0.019	0.929	0.960	0.924	0.889	0.025	0.921	0.955
SAMA	**0.942**	**0.885**	**0.021**	**0.930**	**0.962**	**0.940**	**0.911**	**0.012**	**0.947**	**0.977**	**0.932**	**0.904**	**0.019**	**0.934**	**0.962**

Method	DIS-TE3					DIS-TE4					DIS-TE (ALL)				
	F_β^{max}	F_β^w	M	S_α	E_ϕ^m	F_β^{max}	F_β^w	M	S_α	E_ϕ^m	F_β^{max}	F_β^w	M	S_α	E_ϕ^m
IS-Net	0.830	0.758	0.064	0.836	0.883	0.827	0.753	0.072	0.830	0.870	0.799	0.725	0.070	0.819	0.858
UDUN	0.865	0.809	0.050	0.865	0.917	0.846	0.792	0.059	0.849	0.901	0.831	0.772	0.057	0.844	0.891
BiRefNet	**0.925**	**0.893**	**0.028**	0.919	**0.955**	0.904	**0.864**	**0.039**	**0.869**	**0.939**	0.896	0.858	0.035	0.901	0.934
SAM	0.773	0.724	0.094	0.761	0.848	0.677	0.634	0.162	0.697	0.762	0.773	0.731	0.096	0.773	0.845
HQ-SAM	0.860	0.853	0.045	0.851	0.926	0.786	0.748	0.088	0.799	0.863	0.859	0.835	0.045	0.860	0.924
Pi-SAM	0.899	0.882	0.030	0.901	0.953	0.893	0.870	0.039	0.893	0.948	0.893	0.873	0.033	0.893	0.948
DIS-SAM	0.918	0.877	0.030	0.908	0.948	0.899	0.849	0.043	0.888	0.932	0.917	0.872	0.029	0.911	0.949
SAMA	0.920	0.889	0.032	**0.924**	0.949	**0.917**	0.857	0.041	0.897	0.937	**0.926**	**0.897**	**0.026**	**0.925**	**0.956**

Table 1: Comparison on DIS datasets. Higher ↑ is better except for M, where lower ↓ is better.

Dataset	Metric	Trimap-based					Trimap-free			
		I-F	DIM	DCNN	MGMatting	VITMatte	LFM	MODNet	MFC-Net	SAMA
Composition-1K	SAD↓	70.3	50.4	115.8	32.1	21.5	58.4	47.1	35.6	22.8
	MSE↓	13	14	23	7.0	3.3	11.8	12.3	8.7	2.9
Distinction-646	SAD↓	78.9	47.6	103.8	36.6	21.22	44.6	41.7	34.5	22.4
	MSE↓	16	9	20	7.2	2.1	12.8	9.0	7.8	2.2

Table 2: Quantitative results on Composition-1K and Distinction-646 test sets. Models before VITMatte are trimap-based, while those after are trimap-free. Lower values indicate better performance.

the original SAM across all prompt configurations. Notably, in comparison to HQ-SAM, which is also trained on the DIS-5K dataset (Qin et al. 2022)—our SAMA exhibits more significant improvements on COIFT under zero-shot conditions, particularly when fewer point prompts (1, 3, or 5) are provided. These results highlight the superior generalization ability of SAMA in interactive segmentation scenarios with limited user input.

MVLE	L-A	$\mathbf{F}_\beta^{\max}\uparrow$	$\mathbf{F}_\beta^{\mathbf{w}}\uparrow$	$\mathbf{M}\downarrow$	$\mathbf{S}_\alpha\uparrow$	$\mathbf{E}_\phi^{\mathbf{m}}\uparrow$
–	–	0.872	0.849	0.038	0.868	0.932
–	✓	0.893	0.876	0.029	0.912	0.944
✓	–	0.882	0.869	0.027	0.903	0.942
✓	✓	**0.942**	**0.885**	**0.021**	**0.930**	**0.962**

Table 3: Ablation study on MVLE and Local-Adapter (L-A) on segmentation dataset DIS-VD. Higher↑ is better except for M, where lower↓ is better.

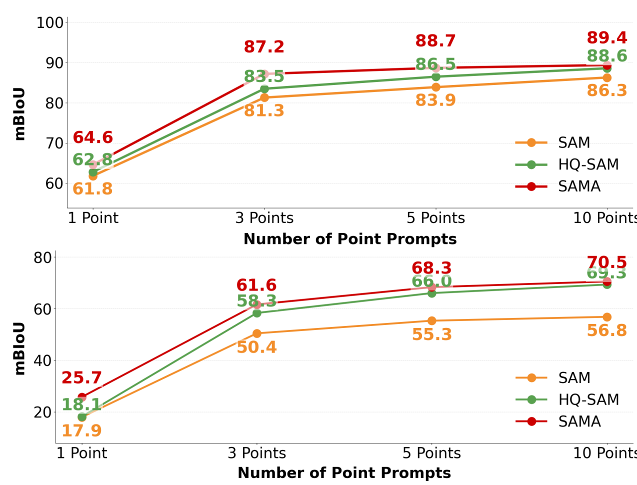

Figure 2: Results of Interactive Segmentation with varying point prompts

Ablation Study

Ablation on MVLE and Local-Adapter Module We conduct ablation experiments on our SAMA to evaluate the ef-

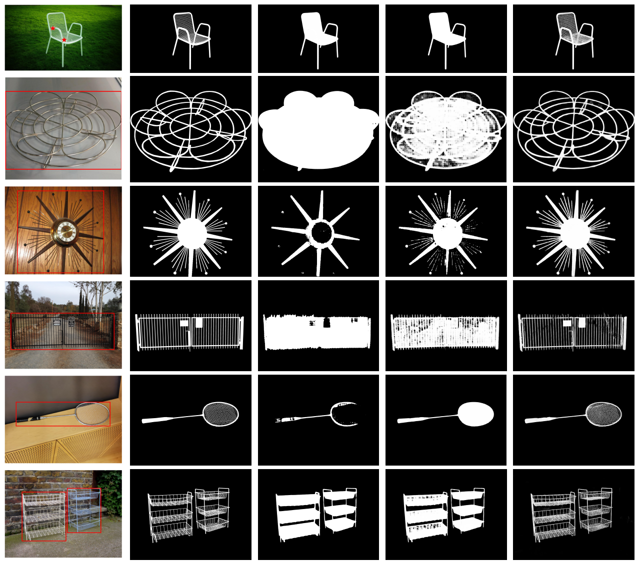

Figure 3: Comparison of segmentation results.

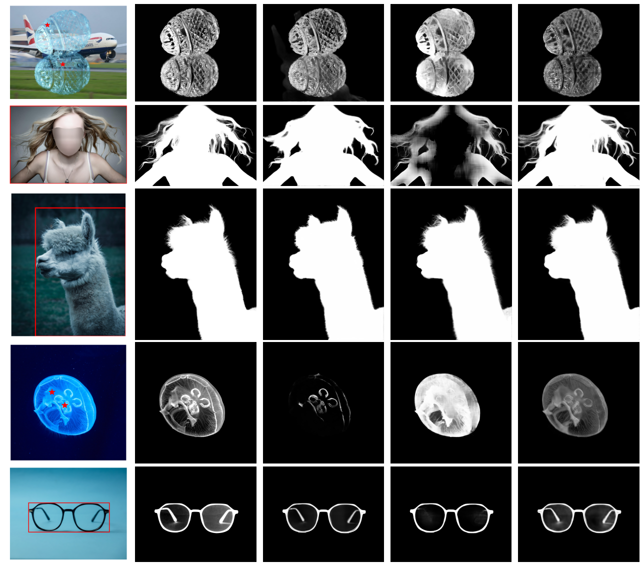

Figure 4: Comparison of matting results.

MVLE	L-A	AM2K		P3M-500	
		SAD↓	MSE↓	SAD↓	MSE↓
–	–	19.79	0.028	16.06	0.0346
–	✓	12.74	0.007	12.15	0.0057
✓	–	13.12	0.011	13.09	0.0063
✓	✓	**8.04**	**0.003**	**9.08**	**0.0028**

Table 4: Ablation study on MVLE and Local-Adapter (L-A) on matting datasets AM2K and P3M-500. Lower values indicate better performance for SAD and MSE.

fectiveness of our proposed modules in SAMA. Specifically, for the segmentation task, we use two segmentation datasets: DIS-VD for highly accurate segmentation and COIFT for interactive segmentation to conduct our experiments. We show our results in Table 3.

From Table 3, we find that the MVLE module boosts performance across all metrics on the baseline model, showing its role in offering complementary information from precise local features for fine-grained segmentation. Besides, Local-Adapter improves significantly on DIS-VD including more complex structures in images, which means it is useful to provide information about local details of objects. We also analyze the contribution of MVLE and Local-Adapter to the matting task. We adopt the same ablation protocol as used in the segmentation task to evaluate the contributions of MVLE and LA to matting performance. As reported in Table 4, the exclusion of either MVLE or LA leads to a substantial decline in matting accuracy, highlighting the critical role of both components. These findings demonstrate that the integration of MVLE and LA significantly enhances the model's ability to capture fine-grained details essential for high-quality alpha matte prediction.

Seg	Matting	DIS-VD					RefMatte-RW100	
		$F_\beta^{\max}\uparrow$	$F_\beta^{w}\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^{m}\uparrow$	SAD↓	MSE↓
✓	–	0.917	0.876	0.027	0.916	0.946	62.70	0.054
–	✓	0.682	0.709	0.071	0.807	0.855	34.25	0.021
✓	✓	**0.942**	**0.885**	**0.021**	**0.930**	**0.962**	**25.69**	**0.0100**

Table 5: Multi-task learning study on datasets DIS-VD and RefMatte-RW100. Higher↑ is better except for M, where lower↓ is better.

Ablation on Multi-task Learning We evaluate the effectiveness of our SAMA framework in jointly training both segmentation and matting tasks. To assess the benefits of multi-task learning, we compare the proposed joint training approach with models trained exclusively on either the segmentation or the matting task. As shown in Table 5, models trained jointly outperform those trained on either task alone. Incorporating matting benefits segmentation by providing fine-grained boundary details, while joint training with segmentation significantly boosts matting accuracy even without trimaps. This demonstrates that large-scale interactive segmentation data effectively supports interactive matting, especially when matting annotations are limited.

Conclusion

We propose SAMA, a lightweight extension of SAM that jointly performs image segmentation and matting. Using a Multi-View Localization Encoder, Local-Adapter, and task-specific prediction heads, SAMA improves performance on both segmentation and matting with minimal overhead. Future work includes runtime optimization and extending SAMA to video segmentation.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, 1597–1604. IEEE.

Aksoy, Y.; Ozan Aydin, T.; and Pollefeys, M. 2017. Designing effective inter-pixel information flow for natural image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 29–37.

Cheng, B.; Collins, M. D.; Zhu, Y.; Liu, T.; Huang, T. S.; Adam, H.; and Chen, L.-C. 2020. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12475–12485.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875.

Cho, D.; Tai, Y.-W.; and Kweon, I. 2016. Natural image matting using deep convolutional neural networks. In *European conference on computer vision*, 626–643. Springer.

Dai, J.; He, K.; and Sun, J. 2016. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3150–3158.

Dai, Y.; Price, B.; Zhang, H.; and Shen, C. 2022. Boosting robustness of image matting with context assembling and strong data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11707–11716.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, Z.; Li, X.; Ma, L.; Zhao, K.; Peng, L.; Biswas, T.; Korpeoglu, E.; Nag, K.; and Achan, K. 2025. LayoutAgent: A Vision-Language Agent Guided Compositional Diffusion for Spatial Layout Planning. *arXiv preprint arXiv:2509.22720*.

Fan, Z.; Li, X.; Nag, K.; Fang, C.; Biswas, T.; Xu, J.; and Achan, K. 2024. Prompt optimizer of text-to-image diffusion models for abstract concept understanding. In *Companion Proceedings of the ACM Web Conference 2024*, 1530–1537.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hou, Q.; and Liu, F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4130–4139.

Huynh, C.; Oh, S. W.; Shrivastava, A.; and Lee, J.-Y. 2024. Maggie: Masked guided gradual human instance matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3870–3879.

Ke, L.; Ye, M.; Danelljan, M.; Tai, Y.-W.; Tang, C.-K.; Yu, F.; et al. 2023. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36: 29914–29934.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1140–1147.

Kirillov, A.; He, K.; Girshick, R.; Rother, C.; and Dollár, P. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9404–9413.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Li, C.; Yang, J.; Zhang, L.; and Gao, J. 2024. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, 467–484. Springer.

Li, J.; Jain, J.; and Shi, H. 2024. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1775–1785.

Li, J.; Zhang, J.; and Tao, D. 2021. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Li, Y.; and Lu, H. 2020. Natural image matting via guided contextual attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11450–11457.

Liew, J. H.; Cohen, S.; Price, B.; Mai, L.; and Feng, J. 2021. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 305–314.

Lin, Y.; Li, H.; Shao, W.; Yang, Z.; Zhao, J.; He, X.; Luo, P.; and Zhang, K. 2025. SAMRefiner: Taming Segment Anything Model for Universal Mask Refinement. *arXiv preprint arXiv:2502.06756*.

Liu, M.; Wang, M.; Ding, H.; Xu, Y.; Zhao, Y.; and Wei, Y. 2024a. Segment anything with precise interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3790–3799.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Liu, X.; Fu, K.; and Zhao, Q. 2023. Promoting segment anything model towards highly accurate dichotomous image segmentation. *arXiv preprint arXiv:2401.00248*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3266–3275.

Lutz, S.; Amplianitis, K.; and Smolic, A. 2018. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*.

Nguyen, K. A.; Juvekar, A.; Yu, T.; Wahed, M.; and Lourentzou, I. 2024. CALICO: Part-Focused Semantic Co-Segmentation with Large Vision-Language Models. *arXiv preprint arXiv:2412.19331*.

Park, G.; Son, S.; Yoo, J.; Kim, S.; and Kwak, N. 2022. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11696–11706.

Pei, J.; Zhou, Z.; Jin, Y.; Tang, H.; and Heng, P.-A. 2023. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2139–2147.

Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; and Wei, X. 2020. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13676–13685.

Qin, X.; Dai, H.; Hu, X.; Fan, D.-P.; Shao, L.; and Van Gool, L. 2022. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, 38–56. Springer.

Sengupta, S.; Jayaram, V.; Curless, B.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2020. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2291–2300.

Tang, J.; Aksoy, Y.; Oztireli, C.; Gross, M.; and Aydin, T. O. 2019. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3055–3063.

Wang, J.; and Cohen, M. F. 2005. An iterative optimization approach for unified image segmentation and matting. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, 936–943. IEEE.

Xiao, A.; Xuan, W.; Qi, H.; Xing, Y.; Yokoya, N.; and Lu, S. 2024. Segment anything with multiple modalities. *arXiv preprint arXiv:2408.09085*.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2970–2979.

Yao, J.; Wang, X.; Yang, S.; and Wang, B. 2024a. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103: 102091.

Yao, J.; Wang, X.; Ye, L.; and Liu, W. 2024b. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147: 105067.

Yao, Y.; Yang, Q.; Cui, M.; and Bo, L. 2025. Towards Fine-grained Interactive Segmentation in Images and Videos. *arXiv preprint arXiv:2502.09660*.

Yu, Q.; Zhang, J.; Zhang, H.; Wang, Y.; Lin, Z.; Xu, N.; Bai, Y.; and Yuille, A. 2021. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1154–1163.

Yu, Q.; Zhao, X.; Pang, Y.; Zhang, L.; and Lu, H. 2024. Multi-view aggregation network for dichotomous image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3921–3930.

Zhang, Y.; Gong, L.; Fan, L.; Ren, P.; Huang, Q.; Bao, H.; and Xu, W. 2019. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7469–7478.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhao, L. S. W. Z. G. 2024. Boosting General Trimap-free Matting in the Real-World Image. *arXiv preprint arXiv:2405.17916*.

Zheng, P.; Gao, D.; Fan, D.-P.; Liu, L.; Laaksonen, J.; Ouyang, W.; and Sebe, N. 2024. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.