

ZIM: Zero-Shot Image Matting for Anything

Beomyoung Kim
Se-Yun Lee

Chanyong Shin
Sewhan Chun

Joonhyun Jeong
Dong-Hyun Hwang

Hyungsik Jung
Joonsang Yu

NAVER Cloud, ImageVision

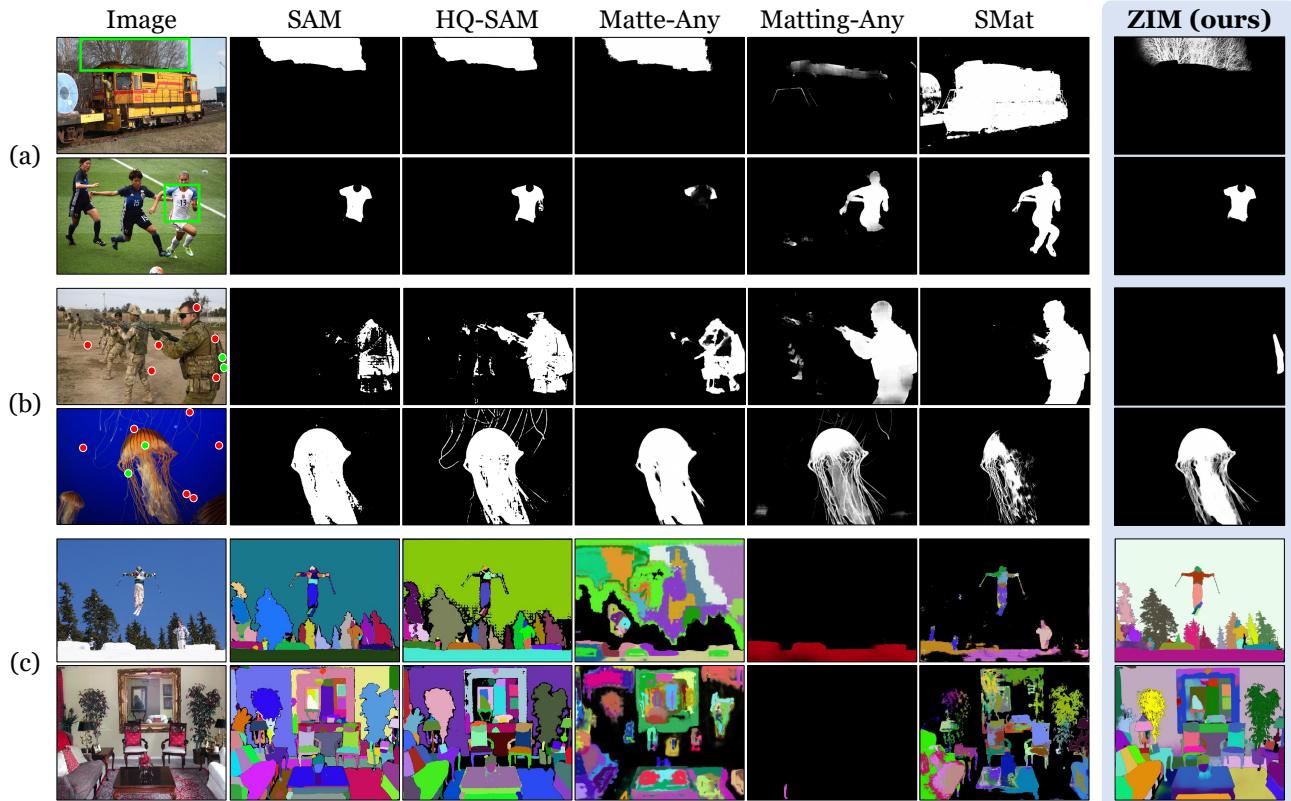


Figure 1. **Qualitative comparison** of ours with five existing zero-shot models (SAM [27], HQ-SAM [23], Matte-Any [59], Matting-Any [33], and SMat [60]). It showcases (a) box prompting results, (b) point prompting results, and (c) automatic mask generation results.

Abstract

The recent segmentation foundation model, Segment Anything Model (SAM), exhibits strong zero-shot segmentation capabilities, but it falls short in generating fine-grained precise masks. To address this limitation, we propose a novel zero-shot image matting model, called ZIM, with two key contributions: First, we develop a label converter that transforms segmentation labels into detailed matte labels, constructing the new SA1B-Matte dataset without costly manual annotations. Training SAM with this dataset enables it to generate precise matte masks while maintaining its zero-shot capability. Second, we design the zero-shot matting model equipped with a hierarchical pixel decoder

to enhance mask representation, along with a prompt-aware masked attention mechanism to improve performance by enabling the model to focus on regions specified by visual prompts. We evaluate ZIM using the newly introduced MicroMat-3K test set, which contains high-quality micro-level matte labels. Experimental results show that ZIM outperforms existing methods in fine-grained mask generation and zero-shot generalization. Furthermore, we demonstrate the versatility of ZIM in various downstream tasks requiring precise masks, such as image inpainting and 3D NeRF. Our contributions provide a robust foundation for advancing zero-shot matting and its downstream applications across a wide range of computer vision tasks. The code is available at <https://github.com/naver-ai/ZIM>.

1. Introduction

Image segmentation, which divides an image into distinct regions to facilitate subsequent analysis, is a fundamental task in computer vision. Recent breakthroughs in segmentation models have made significant strides in this area, particularly with the emergence of the segmentation foundation model, Segment Anything Model (SAM) [27]. SAM is trained on the SA1B dataset [27] containing 1 billion micro-level segmentation labels, where its extensiveness enables SAM to generalize effectively across a broad range of tasks. Its strong zero-shot capabilities, powered by visual prompts, have redefined the state of the art in zero-shot interactive segmentation and opened new avenues for tackling more complex tasks within the zero-shot paradigm.

Despite these achievements, SAM often struggles to generate masks with fine-grained precision (see Figure 1). To address this limitation, recent studies [33, 59, 60] have extended SAM to the image matting task, which focuses on capturing highly detailed boundaries and intricate details such as individual hair strands. These approaches achieve enhanced mask precision by fine-tuning SAM on publicly available matting datasets [30, 44, 56]. However, this fine-tuning process can undermine the zero-shot potential of SAM, since most public matting datasets contain only macro-level labels (*e.g.*, entire human portrait) rather than the more detailed micro-level labels (*e.g.*, individual body parts), as illustrated in Figure 2. Fine-tuning with macro-level labels can cause SAM to overfit to this macro-level granularity, resulting in catastrophic forgetting of its ability to generalize at the micro-level granularity, as shown in Figure 1. Moreover, the scarcity of large-scale matting datasets with micro-level matte labels poses a significant obstacle in developing effective zero-shot matting solutions.

In this paper, we introduce a pioneering **Zero-shot Image Matting** model, dubbed **ZIM**, that retains strong zero-shot capabilities while generating high-quality micro-level matting masks. A key challenge in this domain is the need for a matting dataset with extensive micro-level matte labels, which are costly and labor-intensive to annotate. To address this challenge, we propose a novel label conversion method that transforms any segmentation label into a detailed (pseudo) matte label. For more reliable label transformation, we design two effective strategies, *i.e.*, Spatial Generalization Augmentation and Selective Transformation Learning, to reduce noise and yield high-fidelity matte labels (Section 3.1). Subsequently, we construct a new dataset, called SA1B-Matte, which contains an extensive set of micro-level matte labels generated by transforming segmentation labels from the SA1B dataset via the proposed converter (see Figure 2). By training SAM on the SA1B-Matte dataset, we introduce an effective foundational matting model with micro-level granularity while preserving the zero-shot ability of SAM (see Figure 1).

To further ensure effective interactive image matting, we enhance the major bottleneck in the network architecture of SAM that impedes capturing robust and detailed feature maps. Specifically, SAM employs a pixel decoder with simple two transposed convolutional layers to generate mask feature maps with a stride of 4, which is susceptible to checkerboard artifacts and often falls short in capturing fine details. To mitigate this, we implement a more elaborated pixel decoder with a hierarchical feature pyramid design, inspired by [58], enabling more robust and richer mask feature representations (Section 3.2). Furthermore, inspired by the Mask2Former [8] framework, we introduce a prompt-aware masked attention mechanism that leads to the improvement of interactive matting performance by allowing the model to focus on regions specified by visual prompts.

To validate our zero-shot matting model, we present a new test set, called **MicroMat-3K**, consisting of 3,000 high-quality micro-level matte labels. Our experiments on this dataset demonstrate that while SAM exhibits strong zero-shot capabilities, it struggles to deliver precise mask outputs. In contrast, existing matting models show limited zero-shot performance. ZIM, however, not only maintains robust zero-shot functionality but also provides superior precision in mask generation. Additionally, we highlight the foundational applicability of ZIM in several downstream tasks requiring precise masks, such as image inpainting [64] and 3D NeRF [6]. We hope this work provides valuable insights to the research community, encouraging further development and utilization of zero-shot matting models.

2. Related Work

Image Segmentation. Image segmentation is a fundamental task in computer vision, enabling the division of an image into distinct regions. Recent advancements in segmentation models [8, 22, 28] have significantly improved the accuracy of segmentation tasks, including semantic, instance, and panoptic segmentation. The emergence of Segment Anything Model (SAM) [27] introduced a new paradigm in segmentation by leveraging visual prompts (*e.g.*, points or boxes). SAM is designed as a foundational segmentation model capable of handling diverse tasks due to its robust zero-shot capabilities, showing remarkable versatility across a wide range of tasks and domains. However, despite its strengths, SAM struggles to produce high-precision masks. In this paper, we address this limitation by developing a novel zero-shot model that enhances mask precision while maintaining SAM’s generalization capabilities.

Image Matting. Image matting is a more complex task than image segmentation, as it focuses on estimating the soft transparency of object boundaries to capture fine details, which is critical in tasks like image compositing and background removal. Unlike segmentation, which assigns hard

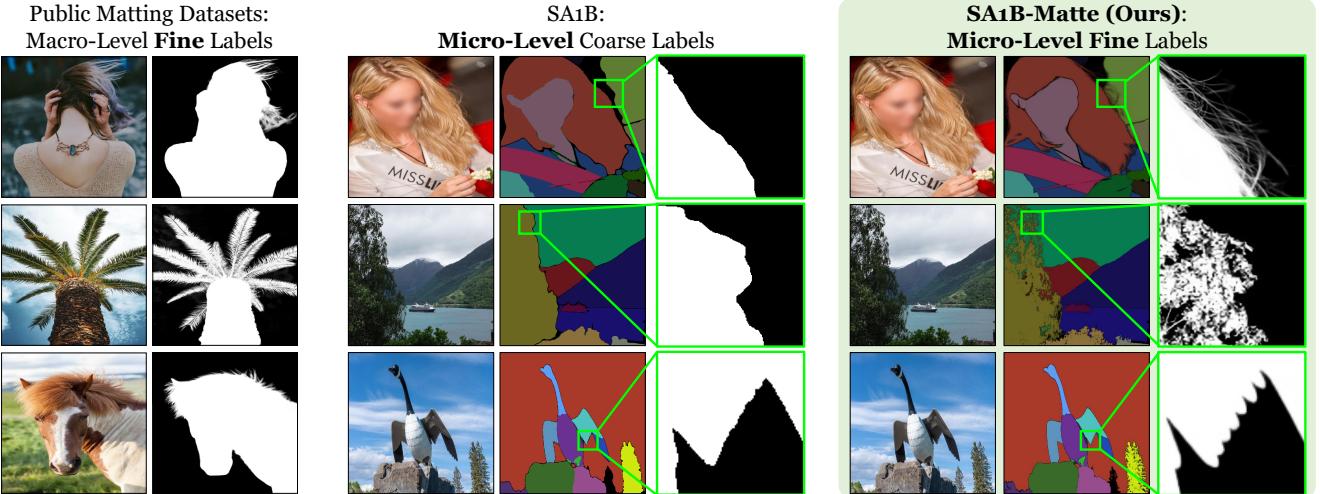


Figure 2. **Qualitative samples from each dataset:** Public matting datasets [30, 31, 44, 56] with macro-level fine labels, the SA1B dataset [27] with micro-level coarse labels, and our proposed SA1B-Matte dataset incorporating the micro-level labels with fine details.

labels to each pixel, matting requires precise edge detection and soft labeling for smooth blending between objects and their background. Recent developments in zero-shot matting have aimed to build upon the foundational segmentation capabilities of SAM. Most approaches [33, 59, 60] fine-tune SAM on public matting datasets [30, 31, 44, 56]. However, these datasets predominantly contain macro-level labels, degrading SAM’s ability to generalize on micro-level structures, such as individual body parts of a human. The reliance on these datasets can deteriorate the zero-shot generalization of the model. Furthermore, the lack of large-scale matting datasets with micro-level labels restricts progress in developing matting models with truly effective zero-shot ability. In this paper, we correspondingly construct a large-scale micro-level labeled matting dataset via our proposed label converter without laborious annotation procedures, enabling effective zero-shot matting modeling.

3. Methodology

Our contributions can be divided into two components: matting dataset construction (Section 3.1) and network architecture enhancements (Section 3.2).

3.1. Constructing the Zero-Shot Matting Dataset

Motivation. For effective zero-shot matting, a dataset with micro-level matte labels is essential. However, manually annotating matte labels at the micro-level requires extensive human labor and cost. To this end, we present an innovative **Label Converter** that transforms any segment label into a matte label, motivated by mask-guided matting works [41, 63]. We first collect public image matting datasets [29–32, 53, 63] to train the converter. Since these

datasets provide only matte labels, we derive coarse segmentation labels from matte labels by applying image processing techniques such as thresholding, resolution downscaling, Gaussian blurring, dilation, erosion, and convex hull transformations. The converter takes an image and segmentation label as input source and is trained to produce a corresponding matte label, as illustrated in Figure 3a.

Challenges. This approach poses two key challenges: (1) Generalization to unseen patterns. Public matting datasets predominantly contain macro-level labels (*e.g.*, entire portraits or large object masks), as shown in Figure 2. Consequently, the converter trained on these datasets often struggles to generalize to unseen micro-level objects, such as individual body parts or detailed components. This limitation leads to the generation of noisy matte labels when applied to micro-level segmentation (see the 4th column in Figure 5a). (2) Unnecessary fine-grained representation. Some objects, such as cars or boxes, commonly do not require fine-grained representation. However, since the converter is trained to always transform segmentation labels into fine-grained matte labels, it often generates unnecessary noise into the output matte, particularly for objects that do not benefit from fine-grained representation (see the 4th column in Figure 5b). To address these challenges, we propose two simple yet effective strategies: Spatial Generalization Augmentation and Selective Transformation Learning.

Spatial Generalization Augmentation. To improve the converter’s ability to generalize to diverse segmentation labels, we design Spatial Generalization Augmentation. This approach introduces variability into the training data by applying a random cut-out technique, as shown in Figure 3a. During training, both the segmentation label and the corresponding matte label are randomly cropped in the same

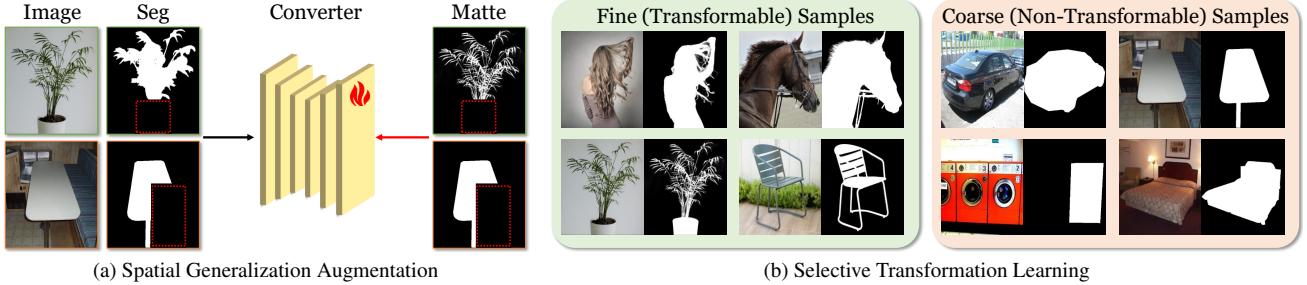


Figure 3. **Illustration of the key components of the Label Converter.** (a) Overview of the training procedure of the converter using Spatial Generalization Augmentation (indicated by red dotted boxes). (b) Examples of transformable (fine) and non-transformable (coarse) samples used in Selective Transformation Learning for the converter.

regions. By exposing the converter to irregular and incomplete input patterns, this augmentation forces the converter to adapt to diverse spatial structures and unseen patterns, thus enhancing its generalization capability. This method ensures that the converter can better handle a variety of input segmentation labels, even those that deviate from training patterns (see the 3rd column in Figure 5a).

Selective Transformation Learning. To prevent the unnecessary transformation of objects that do not require fine-grained details (*e.g.*, cars or desks), we introduce Selective Transformation Learning. This technique enables the converter to selectively focus on objects requiring detailed matte conversion (*e.g.*, hair, trees) while skipping finer transformations for coarse-grained objects. We incorporate these non-transformable samples into the training process by collecting coarse-grained object masks from public segmentation datasets [66] (see Figure 3b). During training, the ground-truth matte label for the non-transformable samples is set to identical to the original segmentation label, allowing the converter to learn that no transformation is required. This selective approach reduces noise in the output and ensures that fine-grained transformations are applied only when needed (see the 3rd column in Figure 5b).

Training. We employ standard loss functions commonly used in matting tasks, namely using a linear combination of L1 and Gradient losses [24, 30, 31] to minimize pixel-wise differences between the ground-truth and predicted matte:

$$L = L_{l1} + \lambda L_{grad} \quad (1)$$

$$L_{l1} = |M - M'| \quad (2)$$

$$L_{grad} = |\nabla_x(M) - \nabla_x(M')| + |\nabla_y(M) - \nabla_y(M')| \quad (3)$$

where M and M' represent the ground-truth and predicted matte label, respectively, and λ is a loss weighting factor hyperparameter. In addition, ∇_x and ∇_y represent the gradients along the horizontal and vertical axes, respectively. Moreover, we set a probability parameter p to control the random application of Spatial Generalization Augmentation during training.

SA1B-Matte Dataset. After training the label converter, we transform segmentation labels in the SA1B dataset [27] to matte labels using the converter, constructing a new SA1B-Matte dataset. As shown in Figure 2, the coarse labels in the SA1B dataset are successfully transformed into high-quality precise matte labels. Compared to existing public matting datasets consisting of macro-level fine labels, the SA1B-Matte dataset is a large-scale image matting dataset with micro-level fine labels, providing an ideal foundation for developing zero-shot matting models.

3.2. ZIM: Zero-Shot Image Matting Model

Overview of ZIM. Our proposed model, ZIM, builds upon the network architecture of SAM [27]. As illustrated in Figure 4, ZIM consists of four components: (1) Image Encoder: extracts image features from the input image, producing an image embedding with a stride of 16. (2) Prompt Encoder: encodes point or box inputs into prompt embeddings. The prompt embeddings are concatenated with learnable token embeddings, serving a role similar to the `[cls]` token in ViT [15]. (3) Transformer Decoder: takes the image and token embeddings to generate output token embeddings. It performs four operations: self-attention on the tokens, token-to-image cross-attention, an MLP layer, and image-to-token cross-attention that updates the image embedding. (4) Pixel Decoder: upsamples the output image embedding with a stride of 2. Lastly, the model produces matte masks by computing a dot product between the upsampled image embedding and output token embeddings.

Motivation. While SAM has shown success in segmentation tasks, its pixel decoder, which comprises two straightforward transposed convolutional layers, is prone to generating checkerboard artifacts, especially when handling challenging visual prompts, such as multiple positive and negative points placed near object boundaries or box prompts with imprecise object region delineation, as shown in Figure 1. Furthermore, their upsampled embeddings with a stride of 4 are often insufficient for image matting, which benefits from finer mask feature representations.

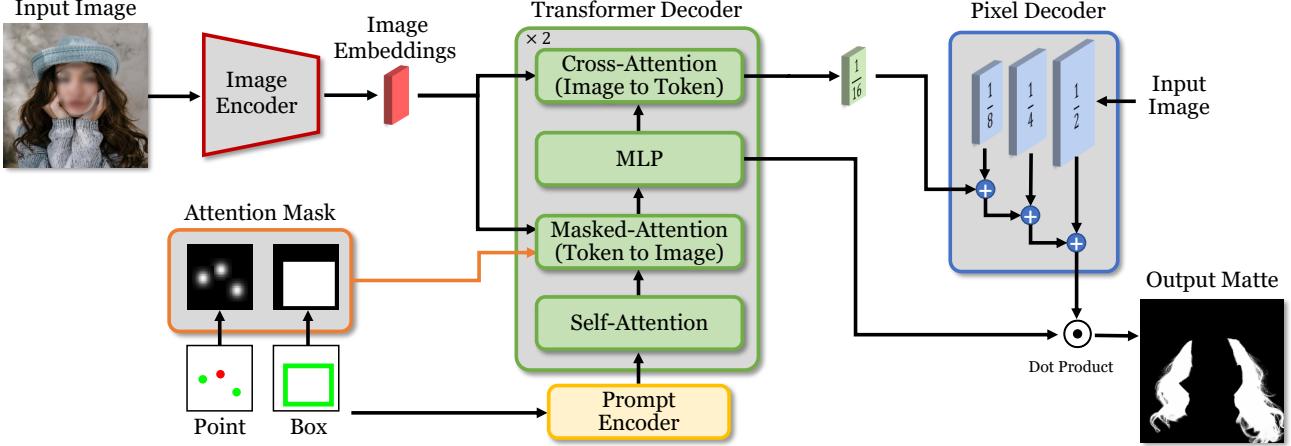


Figure 4. **Overview of the ZIM architecture.** Based on the SAM network architecture [27], we introduce two key improvements: (1) Hierarchical Pixel Decoder for more robust and higher-resolution mask feature map generation, and (2) Prompt-Aware Masked Attention mechanism to enhance interactive matting performance.

Hierarchical Pixel Decoder. To address these limitations, we introduce a hierarchical pixel decoder with a multi-level feature pyramid design, motivated by [58], as illustrated in Figure 4. The pixel decoder takes an input image and generates multi-resolution feature maps at strides 2, 4, and 8 using a series of simple convolutional layers. The image embedding is sequentially upsampled and concatenated with the corresponding feature maps at each resolution. The decoder is designed to be highly lightweight, namely adding only 10 ms of computational overhead compared to the original pixel decoder of SAM on a V100 GPU.

Our hierarchical design serves two key purposes. First, by integrating multi-level skip connections between earlier and deeper network layers, the hierarchical decoder preserves high-level semantic information while progressively refining spatial details. This combination of semantic preservation and spatial refinement results in more accurate matte outputs and reduces the potential risk of checkerboard artifacts, making the architecture more robust to challenging input prompts. Second, it enables the decoder to solidly generate high-resolution feature maps with a stride of 2, rather than the coarser stride of 4 in SAM. This finer resolution is crucial for achieving detailed matte outputs in image matting, where capturing the intricate structures of objects requires high spatial precision.

Prompt-Aware Masked Attention. To further boost the interactive matting performance, we propose a Prompt-Aware Masked Attention mechanism, inspired by Mask2Former [8] (See Figure 4). This mechanism allows the model to dynamically focus on the relevant regions within the image based on visual prompts (*e.g.*, points or boxes), enabling more attention to the areas of interest.

For box prompts, we generate a binary attention mask \mathcal{M}^b that indicates the specific bounding box region. The

binary attention mask $\mathcal{M}^b \in \{0, -\infty\}$ is defined as:

$$\mathcal{M}^b(x, y) = \begin{cases} 0 & \text{if } (x, y) \in \text{box region} \\ -\infty & \text{otherwise} \end{cases} \quad (4)$$

where (x, y) represents the pixel coordinates. This forces the model to prioritize the region within the box prompt.

For point prompts, we generate a soft attention mask using a 2D Gaussian map distribution with standard deviation σ . The soft attention mask, $\mathcal{M}^p \in [0, 1]$, smoothly weighs the region around the point of interest, ensuring a graded focus that transitions smoothly to the surrounding regions.

The attention mask is incorporated into the cross-attention blocks of the transformer decoder. Specifically, the attention mask modulates the attention map as follows:

$$X_l = \begin{cases} \text{softmax}(\mathcal{M}^b + Q_l K_l^\top) V_l + X_{l-1} & \text{(box prompt)} \\ \text{softmax}(\mathcal{M}^p \odot Q_l K_l^\top) V_l + X_{l-1} & \text{(point prompt)} \end{cases} \quad (5)$$

where \odot denotes element-wise multiplication, X_l represents the query feature maps at the l^{th} layer of the decoder, and Q_l , K_l , and V_l implies the query, key, and value matrices, respectively, at the l^{th} layer. This mechanism dynamically adjusts the model's attention according to the visual prompt, leading to performance improvement in prompt-driven interactive scenarios (see Table 4a).

Training. We train our ZIM model using the SA1B-Matte dataset. For each ground-truth matte label, we extract the corresponding box prompt from the given min-max coordinates, where the size of the box prompt is randomly perturbed up to 10% of the original size during training. Additionally, we randomly sample positive and negative point prompts following [51]. The model is optimized using the same matte loss functions defined in Eq. (1).

Method	Prompt	Fine-grained						Coarse-grained					
		SAD↓	MSE↓	MAE↓	Grad↓	Conn↓	SAD↓	MSE↓	MAE↓	Grad↓	Conn↓		
SAM [27]	point box	68.076 36.086	21.651 11.057	23.307 12.714	16.496 14.867	67.730 35.834	17.093 3.516	5.569 1.044	5.756 1.231	4.800 2.551	17.035 3.450		
SAM2 [45]	point box	77.552 38.386	25.296 12.024	26.952 13.680	20.691 14.536	77.182 38.123	44.538 2.301	14.794 0.613	14.982 0.800	8.191 2.440	44.474 2.236		
HQ-SAM [23]	point box	110.681 124.262	36.674 42.457	38.331 44.144	16.855 13.673	110.421 124.113	18.842 8.458	6.457 2.733	6.645 2.920	4.599 2.472	18.792 8.400		
Matte-Any [59]	point box	68.797 34.661	20.844 9.746	23.564 12.182	8.118 7.021	68.939 34.856	19.717 6.950	6.053 1.983	6.675 2.445	2.633 2.142	19.506 6.905		
Matting-Any [33]	point box	275.398 246.214	77.335 68.372	97.141 87.617	20.019 19.185	270.722 241.597	164.145 109.639	36.187 23.780	55.943 38.662	23.244 15.841	155.780 102.439		
SMat [60]	point box	363.821 390.360	123.664 133.515	128.651 138.330	28.669 28.418	362.767 389.478	177.392 183.200	59.113 61.157	62.622 64.678	20.420 20.614	176.125 181.927		
ZIM (ours)	point box	31.286 9.961	8.213 1.893	10.740 3.426	5.324 4.813	31.009 9.655	6.645 1.860	1.788 0.448	2.320 0.659	1.469 1.281	6.472 1.807		

Table 1. **Quantitative comparison** of our ZIM model and six existing methods on the MicroMat-3K test set. By dividing the dataset into fine-grained and coarse-grained categories, all models are evaluated on each category. Results are reported for each type of prompt (point and box) across five evaluation metrics.

4. MicroMat-3K: Zero-Shot Matting Test Set

We introduce a new test set, named MicroMat-3K, to evaluate zero-shot interactive matting models. It consists of 3,000 high-resolution images paired with micro-level matte labels, providing a comprehensive benchmark for testing various matting models under different levels of detail (see Figure 12). It includes two types of matte labels: (1) Fine-grained labels (*e.g.*, hair, tree branches) to primarily evaluate zero-shot matting performance, where capturing intricate details is critical. (2) coarse-grained labels (*e.g.*, cars, desks) to allow comparison with zero-shot segmentation models, which is still essential in zero-shot matting tasks. Moreover, It provides pre-defined point prompt sets for positive and negative points and box prompt sets for evaluating interactive scenarios. More detailed information about the MicroMat-3K is described in the supplementary material.

5. Experiments

5.1. Experimental Setting.

Training Dataset for Label Converter. To train the label converter, we collect six publicly available matting datasets (*i.e.*, AIM-500 [30], AM-2K [31], P3M-10K [29], RWP-636 [63], HIM-2K [53], and RefMatte [32]), consisting of 20,591 natural images and 118,749 synthetic images in total. For non-transformable samples, we extract coarse object categories (*e.g.*, car and desk) from the ADE20K dataset [66], sampling 187,063 masks from 17,768 images.

Training Dataset for ZIM. To train the ZIM model, we

convert the segmentation labels from the SA1B dataset [27] into matte labels using the label converter, constructing a new dataset called SA1B-Matte dataset.

Test Dataset. Models are evaluated on the MicroMat-3K, which is described in Section 4. We report results separately for fine-grained and coarse-grained object masks.

Evaluation Metrics. We use widely adopted evaluation metrics for the image matting task, including Sum of Absolute Difference (SAD), Mean Squared Error (MSE), Gradient Error (Grad), and Connectivity Error (Conn).

Implementation Details for Label Converter. The label converter model is based on MGMatting [63] with Hierarchical-plus [49] backbone network. For training the converter, we set the input size to 1024×1024 , a batch size of 16, and a learning rate of 0.001 with cosine decay scheduling using the AdamW optimizer [36]. The training process runs for 500K iterations with the probability parameter p for Selective Transformation Learning of 0.5 and the loss weight λ of 10.

Implementation Details for ZIM. For the ZIM model, we use the same image encoder (*i.e.*, ViT-B [15]) and prompt encoder as SAM. Leveraging the pre-trained weights from SAM, we fine-tune the ZIM model on 1% of the SA1B-Matte dataset, which amounts to approximately 2.2M matte labels. We set the input size to 1024×1024 , batch size to 16, a learning rate to 0.00001 with cosine decay scheduling using the AdamW optimizer [36], and training iterations to 500K. The loss weight λ is set to 10 and the σ for the point-based attention mask is set to 21 by default.

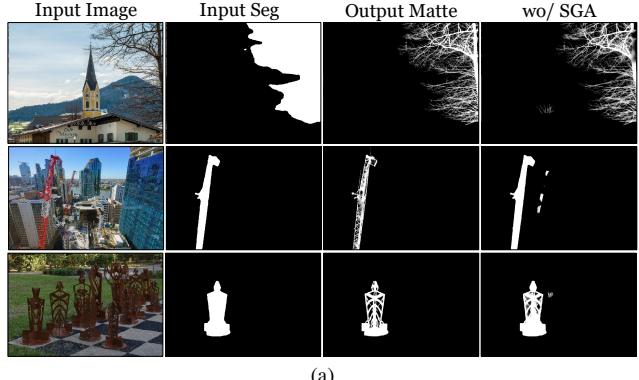
Method	Prompt	AM-2K [31]		AIM-500 [30]	
		MSE↓	Grad↓	MSE↓	Grad↓
SAM [27]	point	82.495	48.716	88.648	56.331
	box	29.061	45.136	51.486	56.885
SAM2 [45]	point	55.614	49.568	84.867	62.052
	box	25.132	45.339	39.539	55.519
HQ-SAM [23]	point	47.911	41.366	63.222	49.596
	box	11.212	33.890	18.687	46.190
Matte-Any [59]	point	82.079	21.677	83.310	33.330
	box	27.293	18.790	38.497	30.333
Matting-Any [33]	point	29.144	19.199	44.443	31.193
	box	6.969	16.711	12.786	28.214
SMat [60]	point	104.492	37.493	83.180	42.936
	box	3.788	17.288	7.869	27.238
ZIM (ours)	point	22.600	15.682	42.422	27.819
	box	3.601	14.378	24.337	25.492

Table 2. **Quantitative comparison** of ZIM and existing methods on two public matting datasets, AM-2K [31] and AIM-500 [30].

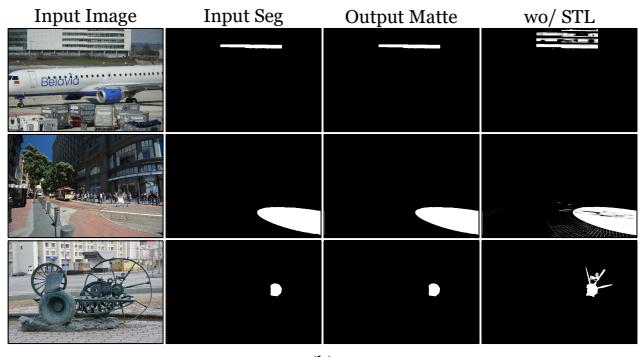
5.2. Experimental Results

We evaluate ZIM against six related methods: SAM [27], SAM2 [45], HQ-SAM [23], Matte-Any [59], Matting-Any [33], and SMat [60]. All methods use the ViT-B [15] backbone network, except SAM2, which uses Hierarchical-base-plus [49]. Table 1 presents the evaluation scores across five metrics for both point and box prompts on fine-grained and coarse-grained masks. For coarse-grained results, SAM achieves reasonable zero-shot performance, while other methods (*e.g.*, Matting-Any and SMat) struggle to generalize to unseen objects. This is likely due to their fine-tuning on macro-level labeled datasets, which degrades their zero-shot capabilities. The qualitative results in Figure 1 show that these methods often produce macro-level outputs, even when provided with micro-level prompts. Moreover, SAM tends to suffer from checkerboard artifacts when challenging prompts are introduced. In contrast, ZIM generates more robust and higher-quality masks, due to our hierarchical feature pyramid decoder, as shown in Figure 1. The fine-grained results in Table 1 highlight the superiority of our method in producing high-quality matting outputs while maintaining strong zero-shot capabilities.

In addition, Table 2 provides quantitative results on two public matting datasets (*i.e.*, AM-2K [31] and AIM [30]), which contain macro-level matte labels only. Here, existing matting methods (*e.g.*, Matte-Any, Matting-Any, and SMat) outperform SAM, as they are specifically fine-tuned on macro-level labeled training datasets that are closely aligned with the domain of the test set. Despite being trained on the micro-level SA1B-Matte dataset, ZIM also achieves highly competitive performance on these matting datasets, showing superior generalization capability.



(a)



(b)

Figure 5. **Qualitative analysis of key components of the label converter:** qualitative samples (a) without Spatial Generalization Augmentation (SGA) and (b) without Selective Transformation Learning (STL).

SGA	STL	Fine-grained			Coarse-grained		
		SAD↓	MSE↓	Grad↓	SAD↓	MSE↓	Grad↓
✓	✓	3.324	0.276	2.664	0.716	0.117	0.684
		2.440	0.122	2.139	0.697	0.092	0.634
		3.153	0.239	2.457	0.635	0.089	0.653
✓	✓	1.999	0.080	1.771	0.281	0.021	0.399

Table 3. **Quantitative analysis of key components of the label converter:** Spatial Generalization Augmentation (SGA) and Selective Transformation Learning (STL).

Lastly, Table 7 provides an in-depth analysis, including throughput, backbone network effects, and detailed scores based on object size. This analysis demonstrates that ZIM’s components (hierarchical mask decoder and prompt-aware masked attention) are lightweight, adding only 10 ms more throughput compared to SAM, while making ZIM more robust to checkerboard artifacts, particularly in larger objects.

6. Ablation Study

In this section, we analyze the impact of the key components of our method using the MicroMat-3K test set.

Attn	Dec	Fine-grained			Coarse-grained		
		SAD \downarrow	MSE \downarrow	Grad \downarrow	SAD \downarrow	MSE \downarrow	Grad \downarrow
\checkmark		13.623	2.718	6.516	2.071	0.474	1.526
	\checkmark	13.198	2.504	6.445	2.049	0.471	1.486
	\checkmark	11.074	2.094	5.401	2.069	0.487	1.355
\checkmark	\checkmark	9.961	1.893	4.813	1.860	0.448	1.281

(a)

Attn Mask	Fine-grained			Coarse-grained				
	T2I	I2T	SAD \downarrow	MSE \downarrow	Grad \downarrow	SAD \downarrow	MSE \downarrow	Grad \downarrow
\checkmark			11.074	2.094	5.401	2.069	0.487	1.355
		\checkmark	9.961	1.893	4.813	1.860	0.448	1.281
	\checkmark	\checkmark	12.526	2.658	6.032	2.353	0.554	1.481
\checkmark	\checkmark		10.437	1.997	5.066	1.999	0.470	1.306

(b)

Table 4. **Analysis of ZIM** using box prompt evaluations: (a) Effect of Attn (prompt-aware masked attention) and Dec (hierarchical pixel decoder). (b) Effect of the masked attention in T2I (token to image) and I2T (image to token) cross-attention layers.

Analysis of Label Converter. To analyze the effect of Spatial Generalization Augmentation (SGA) and Selective Transformation Learning (STL) strategies, we conduct an ablation study by removing each component individually. The SGA is designed to enhance the generalization ability of the converter by simulating diverse input patterns, particularly beneficial given that the converter is trained on macro-level labeled datasets. Without the SGA, the converter struggles to produce clear matte labels for unseen objects, as shown in Figure 5a. In addition, the STL is designed to help the converter avoid unnecessary label conversion for coarse objects. Without the STL, the converter attempts to transform every segmentation label into a matte label, resulting in noisy outputs for unseen coarse objects, as shown in Figure 5b. The quantitative results in Table 3 confirm that using both strategies yields the best label conversion performance on the MicroMat-3K test set.

Analysis of ZIM Model. We conduct experiments to analyze the effect of the prompt-aware masked attention and hierarchical mask decoder. The prompt-aware masked attention is designed to direct the model’s focus on the regions of interest to improve the promptable matting performance. Table 4a shows that leveraging the masked attention yields a substantial improvement to our ZIM model. In addition, the hierarchical mask decoder is designed to produce more robust and higher-resolution mask feature maps to alleviate checkerboard artifacts and capture finer representation, simultaneously. Its effectiveness is particularly evident in reducing the gradient error for fine-grained objects in Table 4a, since the enhanced pixel decoder generates more solid and detailed mask outputs. Notably, the decoder remains lightweight, adding only 10 ms of additional inference time.

Mask	CLIP Dist \uparrow	CLIP Acc \uparrow		
		Top-1	Top-3	Top-5
COCO GT [35]	67.07	0.7767	0.6236	0.5415
SAM [27]	68.17	0.7940	0.6521	0.5753
ZIM (ours)	73.11	0.8616	0.7543	0.6855

Table 5. **Quantitative results of image inpainting** using the Inpainting Anything framework [64]. The inpainting model takes three types of input masks (COCO ground-truth [35], SAM [27], and ZIM) and we evaluate the corresponding inpainting results using CLIP distance and accuracy metrics [16, 61].

Moreover, we delve into the effect of prompt-aware masked attention in our transformer decoder, which comprises two kinds of cross-attention layers (see Figure 4): token-to-image (t2i) updating token embeddings (as queries) and image-to-token (i2t) updating the image embedding (as queries). As a result in Table 4b, applying masked attention to only the t2i layer leads to a meaningful improvement. This suggests that focusing attention on tokens based on visual prompts in the t2i layer enhances their ability to capture relevant features. In contrast, applying attention to specific regions within the image embedding in the i2t layer may disturb the capture of global features.

7. Downstream Task

In this section, we demonstrate the versatility of our zero-shot image matting model, ZIM, by applying it to a variety of downstream tasks. We show that ZIM surpasses SAM in these tasks, especially in scenarios requiring higher precision in mask generation.

7.1. Zero-Shot Image Matting

To showcase the generalizability of ZIM, we evaluate it across 23 diverse datasets, including ADE20K [66], BBB038v1 [5], Cityscapes [12], DOORS [42], Ego-HOS [65], DRAM [11], GTEA [17, 34], Hypersim [48], IBD [7], iShape [57], COCO [35], NDD20 [54], NDSPark [9, 10], OVIS [43], PIDRay [55], Plittersdorf [20], PPDLS [39], STREETS [50], TimberSeg [18], Trash-Can [21], VISOR [13, 14], WoodScape [62], and Zero Waste-f [4]. Using the Automatic Mask Generation strategy introduced by SAM, we apply a regular grid of point prompts to each image and perform post-processing with thresholding and non-maximum suppression (NMS) to generate the final matting masks. Figure 10 and Figure 11 show that ZIM produces high-quality *matte anything* results for all datasets with considerably detailed matte quality and powerful generalization capability. Although SAM shows powerful generalization capability, the output mask is the coarse quality. In addition, existing interactive matting methods (*i.e.*, Matte-Any, Matting-Any, and SMat) often fail to generalize the unseen data.

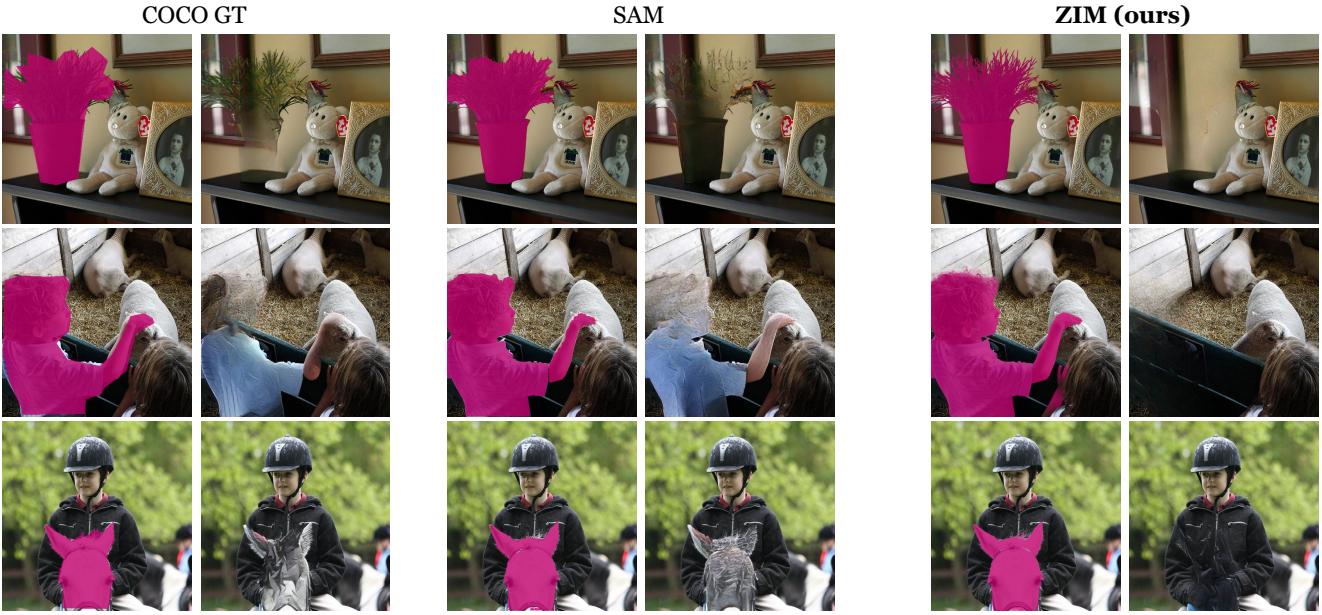


Figure 6. **Qualitative results** of three kinds of input masks (*i.e.*, COCO ground-truth [35], SAM [27], and ZIM) along with their corresponding image inpainting results using the Inpainting Anything framework [64].

7.2. Image Inpainting

Image inpainting is an important application in generative AI, where precise mask generation plays a critical role in removing or reconstructing parts of an image. We follow the Inpaint Anything framework [64], allowing users to interactively select objects to be erased by clicking on them. In this experiment, we compare the performance of SAM and ZIM on the COCO 2017 validation set [35]. To generate object masks, we randomly sample 10 points from the COCO ground-truth segmentation masks and use them as point prompts for both models. The resulting masks are then passed to an inpainting model. As shown in Figure 6, ZIM produces more accurate object masks than SAM, leading to significantly better inpainting results. For complex objects like flowerpots and hair, SAM’s coarse masks fail to remove the object cleanly, leaving noticeable artifacts. In contrast, our precise matting masks enable the inpainting model to smoothly remove objects without artifacts.

Furthermore, to quantify this performance, we use CLIP Distance and CLIP Accuracy [16, 61] as evaluation metrics. CLIP Distance measures the similarity between the source and inpainted regions, where a larger distance indicates better removal. CLIP Accuracy evaluates the change in class predictions after removal, considering the task successful if the original class is absent from Top-1/3/5 predictions. Following [61], we use the text prompt a photo of a {category name}. Table 5 shows that ZIM outperforms SAM on both metrics by better preserving surrounding context after inpainting through enhanced mask quality.

7.3. 3D Object Segmentation with NeRF

The quality of the segmentation mask is crucial when converting a 2D mask into a 3D representation. In this work, we adopt the SA3D framework [6], which utilizes SAM to segment 3D objects from 2D masks by manually prompting the target object in a single view. By replacing SAM with ZIM in the 2D mask segmentation process, we significantly improve the quality of the resulting 3D objects. Since the SA3D framework relies on binary masks for projecting 2D masks into 3D space, we binarized the ZIM results using a threshold of 0.3. For qualitative evaluation, we use the LLFF [38] and 360° [3] datasets, which contain fine-grained object details. For quantitative evaluation, we employ the NVOS [47] dataset, which includes finely annotated 2D masks. We fixed the number of self-prompting points at 10 and followed the experimental setup of SA3D [6] for pre-trained NeRFs and manual prompts.

Figure 7 presents the qualitative results of segmented 3D objects guided by the SAM and ZIM models on the LLFF-trex, LLFF-horns [38], and 360°-kitchen (Lego) [3] datasets. Compared to SAM, which often misses finer details due to its coarse-level mask generation, ZIM captures more intricate object features. Furthermore, as shown in Table 6, ZIM outperforms SAM in quantitative 3D segmentation results on the target view, reporting higher mask IoU scores. These findings demonstrate that the precise matting capabilities of ZIM extend beyond 2D tasks, significantly enhancing the quality of 3D object segmentation.

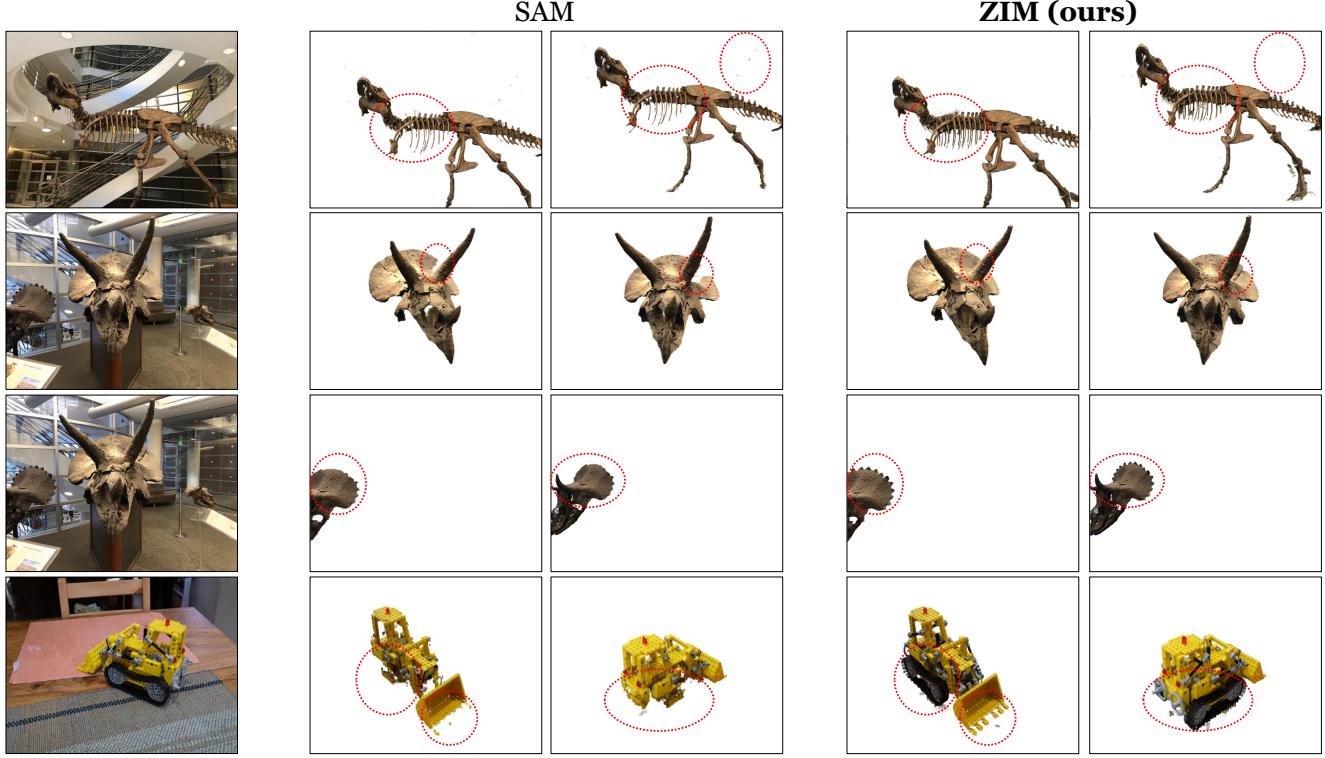


Figure 7. **Qualitative samples** of 3D object segmentation results guided by SAM [27] and ZIM models within the SA3D framework [6] for the LLFF-trex, LLFF-horns [38], and 360°-kitchen (Lego) [3] datasets.

Scenes	Mask IoU (%) ↑	
	SAM	ZIM
Fern	84.9	86.6
Flower	94.0	95.7
Fortress	96.5	98.1
Horns-center	94.0	97.4
Horns-left	92.9	94.7
Leaves	92.2	92.9
Orchids	89.9	91.8
Trex	83.7	85.4

Table 6. **Quantitative results** of mask IoU scores on the target view for the NVOS dataset [47].

7.4. Medical Image Segmentation

Segmentation models are essential in medical image analysis, where they assist in identifying key anatomical structures and abnormalities. Building on the recent evaluation of SAM’s performance in medical imaging by [37], we explore the applicability of ZIM for zero-shot medical image segmentation. Given that neither SAM nor ZIM has been trained on medical image datasets, this experiment focuses on evaluating their zero-shot segmentation capabili-

ties. Both SAM and ZIM, using the ViT-B backbone, are evaluated across five medical imaging datasets: the hippocampus and spleen datasets from the Medical Image Decathlon [2], an ultrasonic kidney dataset [52], an ultrasonic nerve dataset [40], and an X-ray hip dataset [19]. Since these datasets comprise binary ground-truth masks, we apply a threshold of 0.3 to the matte output of ZIM. Following the evaluation protocol from [37], we employ five prompt modes: (1) a single point at the center of the largest contiguous region, (2) multiple points centered on up to three regions, (3) a box surrounding the largest region, (4) multiple boxes around up to three regions, and (5) a box encompassing the entire object.

Figure 8 illustrates the distribution of IoU scores across the five datasets for each prompt mode. These results show that ZIM consistently outperforms SAM, particularly in point-based prompts (modes 1 and 2). In addition, as shown in Figure 9, SAM frequently exhibits checkerboard artifacts when dealing with the indistinct visual details of medical images. In contrast, ZIM produces more robust and precise segmentation masks due to our advanced pixel decoder. This demonstrates ZIM’s strong generalization on unseen and complex medical image data, highlighting its superior zero-shot capabilities compared to SAM.

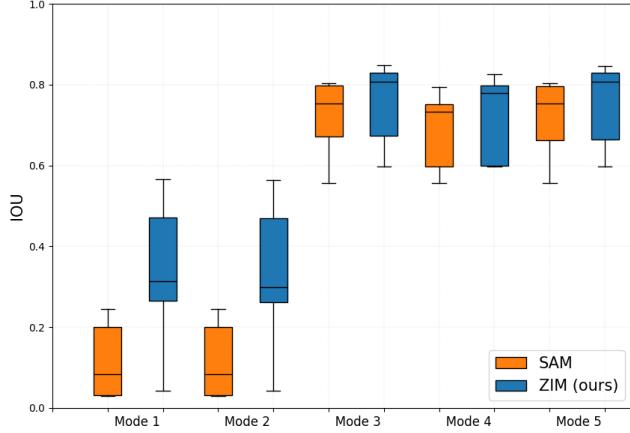


Figure 8. **Mask IoU distribution** across the five medical image analysis datasets [2, 19, 40, 52] for the five prompt modes.

8. Conclusion

In this paper, we introduced ZIM, a pioneering zero-shot image matting model designed to address the limitations of SAM and existing zero-shot matting approaches in generating high-quality matte masks. While SAM’s extensive training on the SA1B dataset enables broad generalization across segmentation tasks, it lacks the mask precision needed for tasks like image matting. Existing matting approaches that fine-tune SAM on public matting datasets achieve higher precision but often sacrifice zero-shot versatility due to reliance on macro-level labels. To overcome these challenges, we developed ZIM by constructing the SA1B-Matte dataset through a novel label conversion method, allowing for micro-level matting without extensive manual annotation. Supported by strategies like Spatial Generalization Augmentation and Selective Transformation Learning, ZIM achieves high-fidelity matte labels while preserving SAM’s zero-shot potential. Additionally, we enhanced ZIM with a hierarchical pixel decoder and prompt-aware masked attention mechanism, allowing it to produce robust, detailed masks that mitigate checkerboard artifacts and accurately respond to visual prompts. Experimental results on the MicroMat-3K dataset confirm ZIM’s effectiveness in producing high-quality matting results in a zero-shot setting, consistently outperforming SAM and other matting models on micro-level matting tasks. Furthermore, evaluations across downstream tasks, including image inpainting, 3D NeRF, and medical image analysis, demonstrate ZIM’s versatility and applicability in fields requiring precise mask generation. In each task, ZIM not only preserves context and fine details but also improves over SAM and existing zero-shot matting approaches by delivering high-resolution masks that better capture intricate structures. We believe ZIM’s strong performance in zero-shot matting marks an important step forward in interactive im-

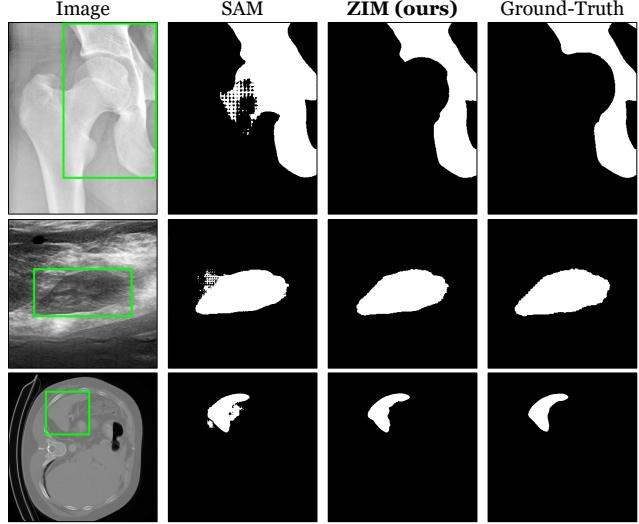


Figure 9. **Qualitative samples** of SAM and ZIM output masks on the medical image datasets [2, 19, 40, 52] using the box prompt.

age matting and inspires further exploration in developing more refined zero-shot solutions. Future work could explore ZIM’s adaptability in additional fields such as video or 3D domains, potentially extending its impact across both established and emerging areas in visual understanding.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. [17](#)
- [2] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), 2022. [10, 11](#)
- [3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. [9, 10](#)
- [4] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James

- Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157, 2022. [8](#), [16](#)
- [5] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghghi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. [8](#), [15](#)
- [6] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. [2](#), [9](#), [10](#)
- [7] Jiazhou Chen, Yanghui Xu, Shufang Lu, Ronghua Liang, and Liangliang Nan. 3-d instance segmentation of mvs buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022. [8](#), [15](#)
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. [2](#), [5](#)
- [9] Luca Ciampi, Carlos Santiago, Joao Paulo Costeira, Claudio Gennaro, and Giuseppe Amato. Domain adaptation for traffic density estimation. In *VISIGRAPP (5: VISAPP)*, pages 185–195, 2021. [8](#), [16](#)
- [10] Luca Ciampi, Carlos Santiago, Joao Costeira, Claudio Gennaro, and Giuseppe Amato. Night and Day Instance Segmented Park (NDISPark) Dataset: a Collection of Images taken by Day and by Night for Vehicle Detection, Segmentation and Counting in Parking Areas, 2022. [8](#), [16](#)
- [11] Nadav Cohen, Yael Newman, and Ariel Shamir. Semantic segmentation in art paintings. In *Computer graphics forum*, pages 261–275. Wiley Online Library, 2022. [8](#), [15](#)
- [12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. [8](#), [15](#)
- [13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. [8](#), [16](#)
- [14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. [8](#), [16](#)
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [4](#), [6](#), [7](#), [17](#), [18](#)
- [16] Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models, 2024. [8](#), [9](#)
- [17] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. [8](#), [15](#)
- [18] Jean-Michel Fortin, Olivier Gamache, Vincent Grondin, François Pomerleau, and Philippe Giguère. Instance segmentation for autonomous log grasping in forestry operations. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6064–6071. IEEE, 2022. [8](#), [16](#)
- [19] Daniel Gut. X-ray images of the hip joints, 2021., Mendeley Data, V1, doi: 10.17632/zm6bxzhmfz.1. [10](#), [11](#)
- [20] Timm Haucke, Hjalmar S Kühl, and Volker Steinlage. Socrates: Introducing depth in visual wildlife monitoring using stereo vision. *Sensors*, 22(23):9082, 2022. [8](#), [16](#)
- [21] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. [8](#), [16](#)
- [22] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. [2](#)
- [23] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. [1](#), [6](#), [7](#), [17](#), [18](#), [21](#)
- [24] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. [4](#)
- [25] Beomyoung Kim, Joonhyun Jeong, Dongyoon Han, and Sung Ju Hwang. The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11360–11370, 2023.
- [26] Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3346–3356, 2024.
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#), [8](#), [9](#), [10](#), [17](#), [18](#), [20](#), [21](#)
- [28] Feng Li, Hao Zhang, Huaijhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards

- a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 2
- [29] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3501–3509, 2021. 3, 6
- [30] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235*, 2021. 2, 3, 4, 6, 7
- [31] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, 130(2):246–266, 2022. 3, 4, 6, 7
- [32] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22448–22457, 2023. 3, 6, 19
- [33] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1785, 2024. 1, 2, 3, 6, 7, 18, 21
- [34] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015. 8, 15
- [35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 8, 9, 15
- [36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6
- [37] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 10
- [38] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 9, 10
- [39] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016. 8, 16
- [40] Anna Montoya, Hasnín, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation. <https://kaggle.com/competitions/ultrasound-nerve-segmentation>, 2016. Kaggle. 10, 11
- [41] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Mask-guided matting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1992–2001, 2023. 3
- [42] Mattia Pugliatti and Francesco Toppo. Doors: Dataset for boulders segmentation. *Zenodo*, 9(20):6, 2022. 8, 15
- [43] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 8, 16
- [44] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 2, 3
- [45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädele, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6, 7, 17, 18
- [46] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 18
- [47] Zhongzheng Ren, Aseem Agarwala[†], Bryan Russell[†], Alexander G. Schwing[†], and Oliver Wang[†]. Neural volumetric object selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. ([†] alphabetic ordering). 9, 10
- [48] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 8, 15
- [49] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 6, 7, 17, 18
- [50] Corey Snyder and Minh Do. Streets: A novel camera network dataset for traffic flow. *Advances in Neural Information Processing Systems*, 32, 2019. 8, 16
- [51] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 5
- [52] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 10, 11
- [53] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2656, 2022. 3, 6
- [54] Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A Stephen McGough, Nick Wright, Ben

- Burville, and Per Berggren. Ndd20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020. 8, 16
- [55] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5412–5421, 2021. 8, 16
- [56] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 2, 3
- [57] Lei Yang, Yan Zi Wei, Yisheng He, Wei Sun, Zhenhang Huang, Haibin Huang, and Haoqiang Fan. ishape: A first step towards irregular shape instance segmentation. *arXiv preprint arXiv:2109.15068*, 2021. 8, 15
- [58] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 2, 5
- [59] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147: 105067, 2024. 1, 2, 3, 6, 7, 17, 18, 21
- [60] Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, and Zhiguo Cao. Unifying automatic and interactive matting with pretrained vits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25585–25594, 2024. 1, 2, 3, 6, 7, 18, 21
- [61] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models, 2023. 8, 9
- [62] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O’Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318, 2019. 8, 16
- [63] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1154–1163, 2021. 3, 6
- [64] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2, 8, 9
- [65] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 8, 15
- [66] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4, 6, 8, 15

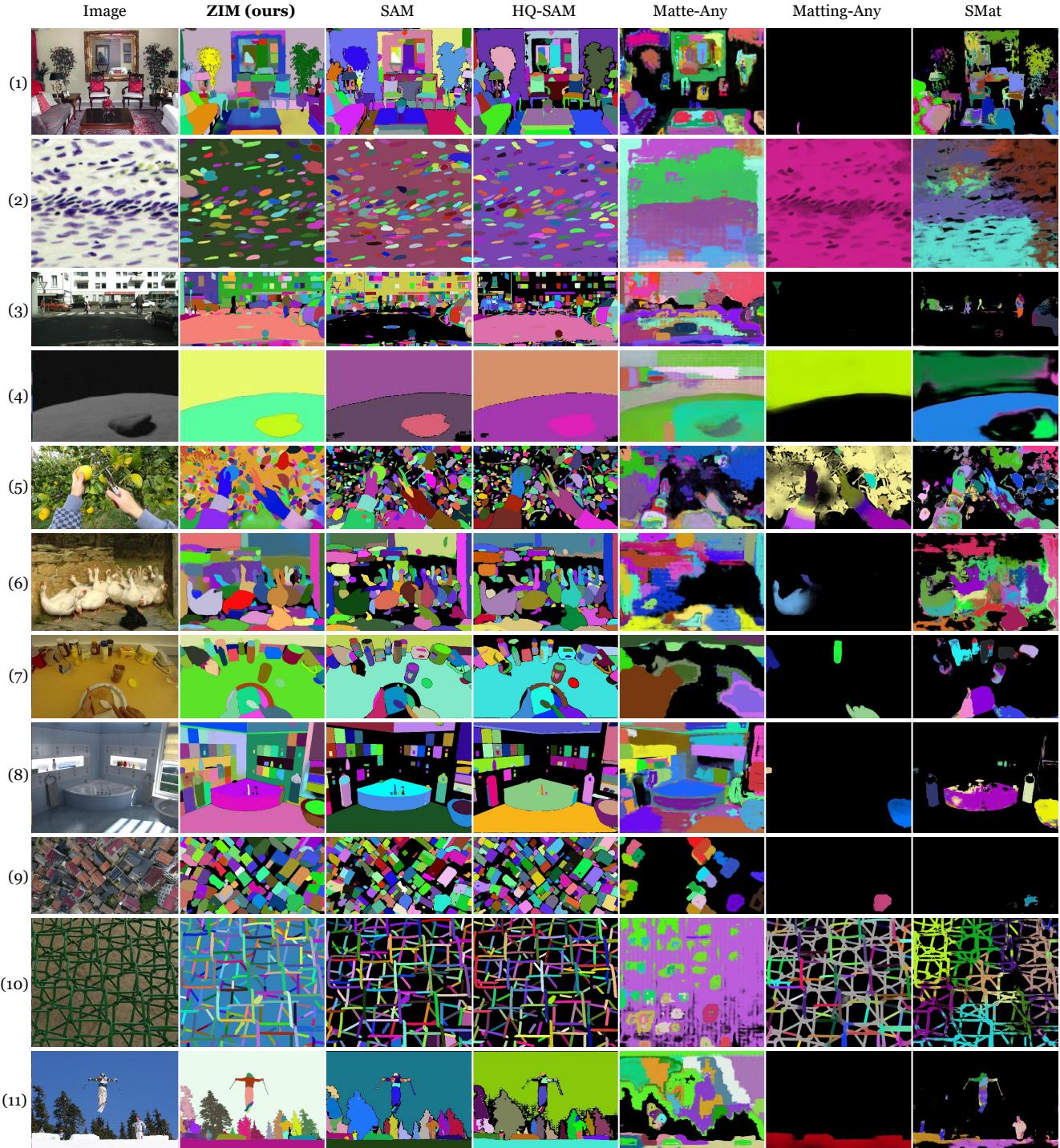


Figure 10. Qualitative samples of automatic mask generation results on (1) ADE20K [66], (2) BBBC038v1 [5], (3) Cityscapes [12], (4) DOORS [42], (5) EgoHOS [65], (6) DRAM [11], (7) GTEA [17, 34], (8) Hypersim [48], (9) IBD [7], (10) iShape [57], and (11) COCO [35] datasets.

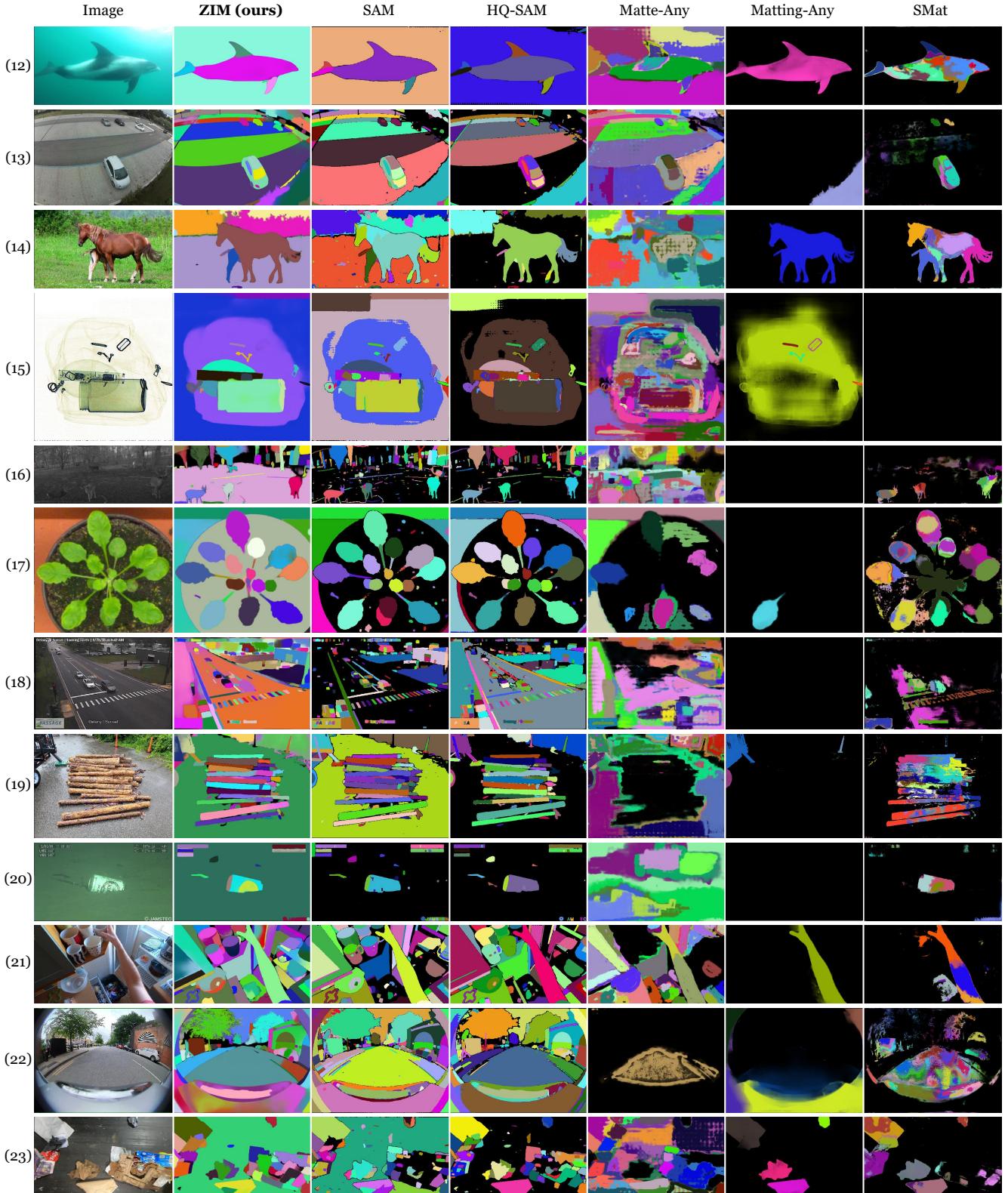


Figure 11. (continue) **Qualitative samples of automatic mask generation results** on (12) NDD20 [54], (13) NDISPark [9, 10], (14) OVIS [43], (15) PIDRay [55], (16) Plittersdorf [20], (17) PPDLS [39], (18) STREETS [50], (19) TimberSeg [18], (20) TrashCan [21], (21) VISOR [13, 14], (22) WoodScape [62], and (23) ZeroWaste-f [4] datasets.



Figure 12. **Visualization of samples** from the MicroMat-3K test set, providing a high-quality benchmark for zero-shot matting models.

Appendix

A. MicroMat-3K Details

Inspired by the data engine procedure in the SAM [27], we constructed the MicroMat-3K test set. The dataset construction involved four key steps: (1) We collected high-resolution images from the DIV2K dataset [1], which is originally intended for super-resolution tasks. (2) We generated pseudo-segmentation labels using automatic mask generation of SAM, powered by a large backbone model to ensure high-quality segmentation. (3) We transformed these segmentation labels into matte labels using our label converter, which is also powered by a large backbone network. This step provided an initial set of pseudo-matte labels for each image. (4) Our annotators inspected the pseudo-matte labels. High-quality labels were directly retained as ground-truth annotations, while any low-quality matte labels were manually revised to ensure high-fidelity ground-truth labels. Additionally, we categorized the final matte labels into two classes: fine-grained masks (750 samples) and coarse-grained masks (2250 samples). Figure 12 showcases visualization examples from MicroMat-3K, which provides diverse and high-quality micro-level matte labels.

B. Detailed Experimental Result

Table 7 provides an in-depth evaluation of zero-shot matting models, with additional experiments utilizing various backbone networks: ViT-L and ViT-H [15] for SAM [27], HQ-SAM [23], and Matte-Any [59], and Hiera-L [49] for SAM2 [45]. Furthermore, we investigate model behavior based on the size of the target object by defining three object size groups, according to the ratio of the foreground region in the image: small (ratio $< 1\%$), medium ($1\% \leq \text{ratio} < 10\%$), and large ($\text{ratio} \geq 10\%$). The MSE error is reported for each object size group, offering a detailed understanding of model performance across varying object sizes. Additionally, we measure the model throughput using an NVIDIA V100 GPU to assess computational efficiency.

The results in Table 7 provide some meaningful insights: (1) ZIM consistently outperforms SAM, especially for larger objects, as indicated by the MSE_L metric. This improvement is likely due to the reduction of checkerboard artifacts, a known issue in SAM’s pixel decoder, which our advanced decoder addresses effectively, as evidenced in Figures 1, 10, and 11. (2) ZIM demonstrates highly competitive results even with the smaller ViT-B backbone, outperforming models like SAM and HQ-SAM with larger

Method	Backbone	Latency (ms)↓	Prompt	MSE	Fine-grained↓			MSE	Coarse-grained↓		
					MSE _S	MSE _M	MSE _L		MSE _S	MSE _M	MSE _L
SAM [27]	ViT-B [15]	172.5	point box	21.651 11.057	2.717 0.329	7.145 2.983	70.502 38.501	5.569 1.044	4.092 0.181	5.405 2.456	77.761 24.519
	ViT-L [15]	361.6	point box	15.663 7.989	3.312 0.320	5.165 2.423	49.202 27.276	4.293 0.534	3.640 0.145	3.389 1.606	46.278 5.575
	ViT-H [15]	605.2	point box	14.534 6.188	3.619 0.281	5.753 1.687	43.371 21.389	2.100 0.468	0.653 0.152	3.278 1.334	56.086 4.610
SAM2 [45]	Hiera-B+ [49]	147.4	point box	25.296 12.024	15.657 0.322	15.970 2.155	53.346 43.670	14.794 0.613	14.786 0.152	12.128 1.600	49.058 10.194
	Hiera-L [49]	195.0	point box	16.937 10.616	0.871 0.263	5.613 1.888	56.702 38.636	2.572 0.704	0.954 0.149	4.308 1.424	57.804 18.019
HQ-SAM [23]	ViT-B [15]	177.2	point box	36.674 42.457	5.094 0.392	9.356 4.369	123.199 160.456	6.457 2.733	3.602 0.230	9.596 2.521	102.834 124.372
	ViT-L [15]	368.1	point box	20.481 19.881	1.928 0.326	6.496 2.683	67.979 73.913	3.046 0.762	1.200 0.163	4.987 1.408	66.373 21.129
	ViT-H [15]	608.1	point box	22.547 23.743	5.308 0.263	6.601 2.538	71.446 89.518	3.599 0.789	1.659 0.164	5.034 1.266	77.643 24.469
Matte-Any [59]	ViT-B [15]	668.5	point box	20.844 9.746	3.381 0.918	7.953 3.856	65.108 31.109	6.053 1.983	4.739 1.235	5.897 3.039	70.443 24.426
	ViT-L [15]	814.1	point box	15.230 7.323	3.985 0.975	6.243 3.692	44.842 21.711	5.116 1.597	4.570 1.222	3.996 2.418	44.606 9.119
	ViT-H [15]	1036.9	point box	14.119 6.048	4.270 0.917	7.085 2.992	38.704 17.872	3.022 1.571	1.669 1.228	3.913 2.294	56.044 8.836
Matting-Any [33]	ViT-B [15]	200.3	point box	77.335 68.372	27.256 18.286	41.349 33.111	202.692 192.566	36.187 23.780	31.549 17.344	40.993 36.194	197.340 175.418
SMat [60]	ViT-B [15]	263.4	point box	123.664 133.515	31.549 17.344	40.993 36.194	197.340 175.418	59.113 61.157	52.305 53.280	68.709 74.099	250.587 261.544
ZIM (ours)	ViT-B [15]	187.8	point box	8.213 1.893	0.870 0.205	3.962 2.228	24.934 3.617	1.788 0.448	1.444 0.200	1.731 1.382	18.983 0.632
	ViT-L [15]	373.4	point box	5.825 1.589	0.563 0.191	2.724 1.888	17.898 2.982	1.719 0.446	1.041 0.175	1.574 1.458	35.687 0.686

Table 7. **Detailed Quantitative comparison** of our ZIM model and six existing methods on the MicroMat-3K dataset. Results are presented for different backbone networks, model throughput, and MSE scores across object sizes (small, medium, and large). The latency is measured on the NVIDIA V100 GPU.

backbones such as ViT-H. Additionally, the performance of ZIM with the ViT-L backbone suggests that further improvements could be achieved with more powerful architectures. (3) Despite our advanced decoder, ZIM introduces only a marginal increase in latency (just 10ms more than SAM) making it a lightweight and efficient option for zero-shot matting tasks. (4) Compared to existing matting models (*e.g.*, Matte-Any, Matting-Any, and SMat), ZIM delivers superior performance while maintaining efficiency.

C. Expanding Prompt Sources

Interactive models, such as SAM, commonly support only point and box prompts. Here, we demonstrate the potential ZIM offers a more flexible approach by expanding the variety of prompt types, including text and scribble prompts.

Text Prompt. To enable text prompts, we integrate ZIM with the Grounded-SAM framework [46]. Grounded-SAM uses a grounding object detection model that processes an image-text pair and returns bounding boxes for objects mentioned in the text. ZIM then uses these bounding boxes as prompts to produce detailed matte outputs. We refer to this combined model as Grounded-ZIM. As shown in Figure 13, Grounded-ZIM provides high-quality outputs with a simple text prompting pipeline, offering more precise and robust mask generation than Grounded-SAM.

Scribble Prompt. In addition, ZIM can support scribble prompts, which provide users with an intuitive way to mark regions of interest. We implement this functionality by sampling points along the scribble path. To ensure comprehen-

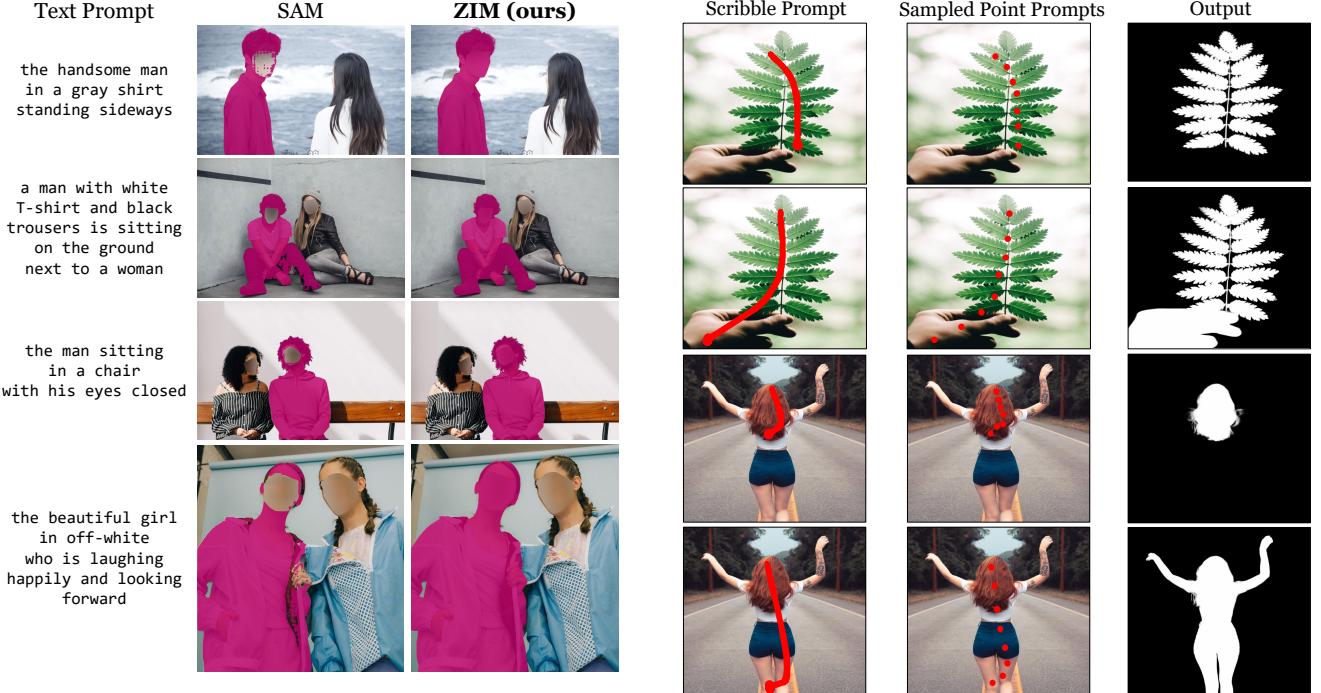


Figure 13. Qualitative samples of text prompting results on the RefMatte-RW100 [32] dataset.

sive coverage of the scribble region, we employ uniform sampling with setting the maximum number of sampled points to 24. It allows ZIM to effectively handle the scribble input and generate high-quality matte outputs. Figure 14 shows an example of how the scribble prompt leads to accurate and stable results. These expansions highlight the versatility of ZIM in accommodating diverse input prompts.

D. Additional Ablation Study

In this section, we explore the impact of hyperparameters on the performance of the ZIM model, focusing on σ and λ .

Effect of Hyperparameter σ . The hyperparameter σ controls the standard deviation of the 2D Gaussian map used to create the soft attention mask for point prompts. A larger σ results in a wider spread of the Gaussian, covering a broader region around the point. Table 8a presents the performance of the ZIM model using point prompts on the MicroMat-3K dataset for varying values of σ . Through experimentation, we found that setting σ to 21 strikes a balance by generating an appropriately sized soft attention mask that effectively captures relevant features while minimizing unnecessary coverage.

Effect of Hyperparameter λ . The hyperparameter λ , as defined in Eq (1), controls the weight assigned to the Gradient loss, influencing the emphasis on edge detail during

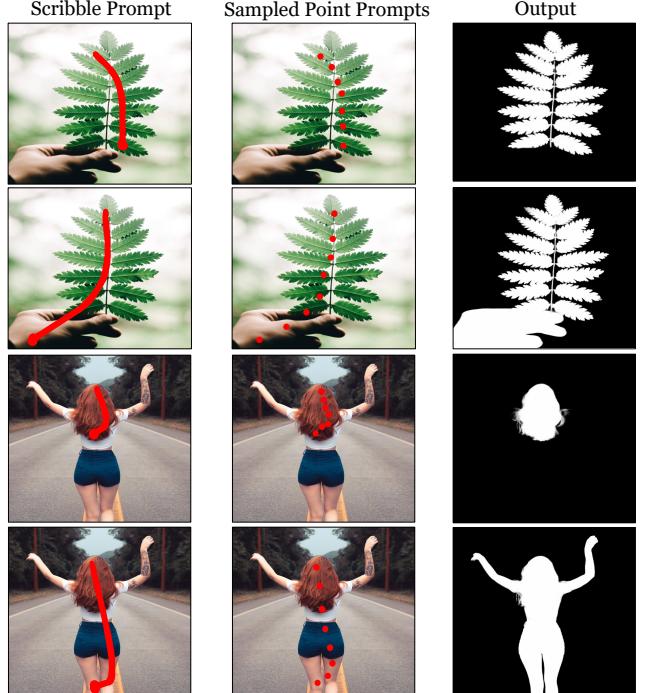


Figure 14. Qualitative examples of scribble prompting results.

σ	Fine-grained			Coarse-grained		
	SAD \downarrow	MSE \downarrow	Grad \downarrow	SAD \downarrow	MSE \downarrow	Grad \downarrow
11	31.476	8.381	5.505	6.741	1.816	1.538
21	31.286	8.213	5.324	6.645	1.788	1.469
41	31.341	8.298	5.476	6.686	1.807	1.493

(a)

λ	Fine-grained			Coarse-grained		
	SAD \downarrow	MSE \downarrow	Grad \downarrow	SAD \downarrow	MSE \downarrow	Grad \downarrow
5	33.086	9.056	7.392	6.966	1.970	1.817
10	31.286	8.213	5.324	6.645	1.788	1.469
20	31.402	8.295	5.356	6.712	1.838	1.475

(b)

Table 8. Analysis of ZIM using point prompt evaluations: (a) Effect of the hyperparameter σ . (b) Effect of the hyperparameter λ .

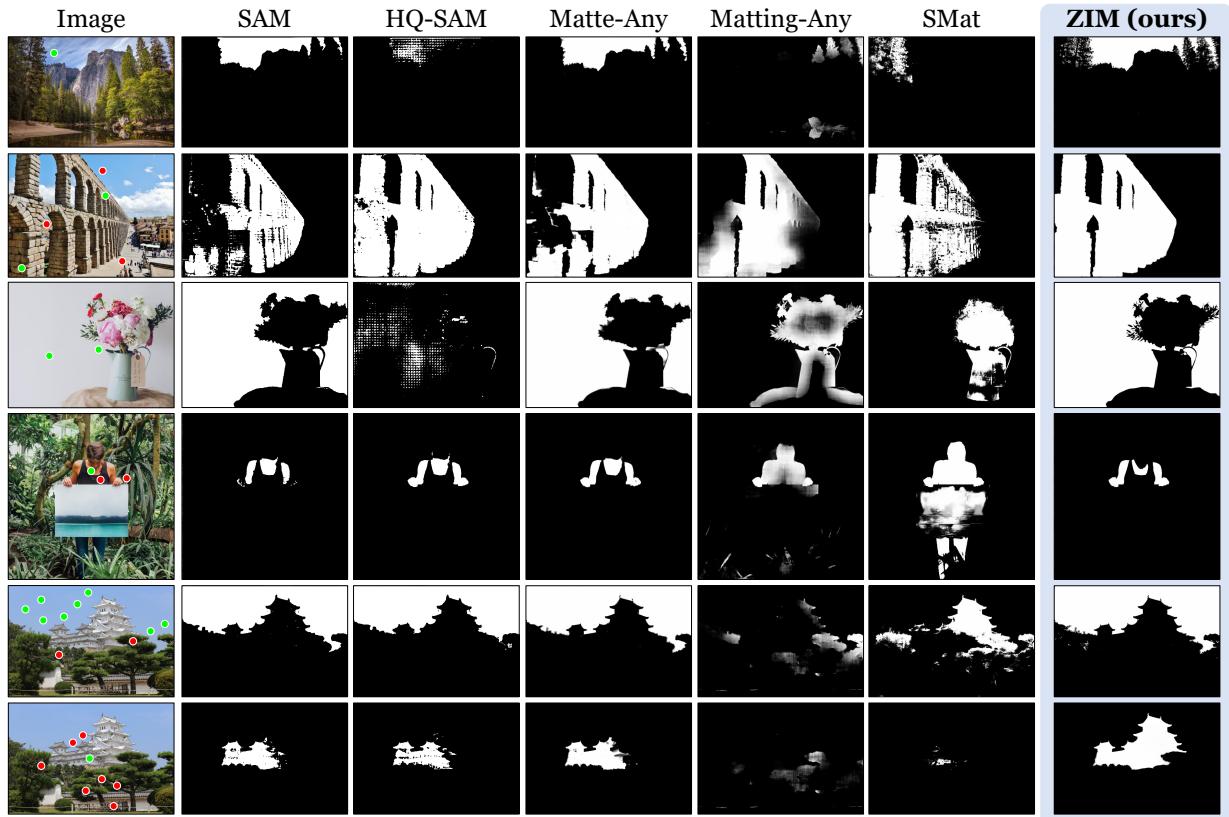
training. We evaluate how different values of λ affect the performance of the ZIM on the MicroMat-3K dataset. The results in Table 8b indicate that a λ value of 10 achieves optimal performance, providing a balanced trade-off between smoothness and edge accuracy.

E. Additional Qualitative Samples

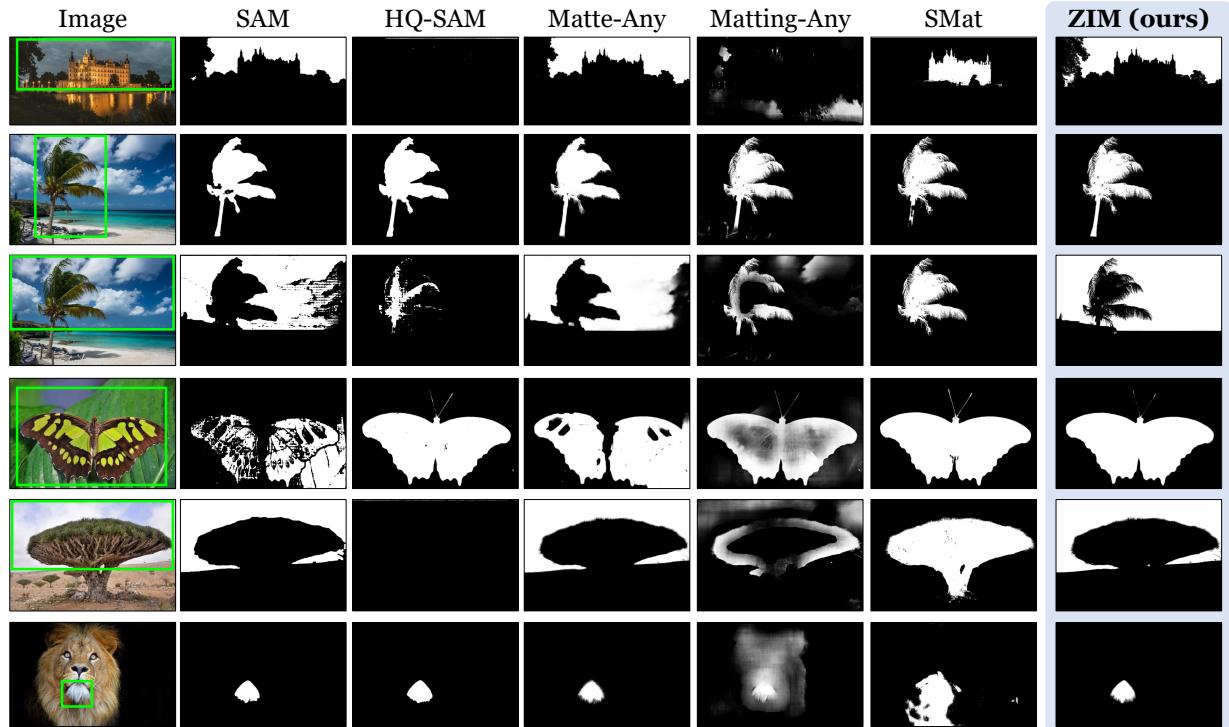
We provide more qualitative samples for the SA1B-Matte dataset (Figure 15) and ZIM output mattes (Figure 16).



Figure 15. **Additional qualitative samples** of the SA1B dataset [27] with micro-level coarse labels and our SA1B-Matte dataset with micro-level fine labels.



(a)



(b)

Figure 16. **Additional qualitative samples** of ZIM with five existing zero-shot models (SAM [27], HQ-SAM [23], Matte-Any [59], Matting-Any [33], and SMat [60]) based on (a) point prompts and (b) box prompts.