

# DFIMat: Decoupled Flexible Interactive Matting in Multi-Person Scenarios

Siyi Jiao<sup>\*</sup>, Wenzheng Zeng<sup>\*</sup>, Changxin Gao, and Nong Sang<sup>†</sup>

School of AIA, Huazhong University of Science and Technology  
{m202173030,m202173066}@alumni.hust.edu.cn, {cgao,nsang}@hust.edu.cn

**Abstract.** Interactive portrait matting refers to extracting the soft portrait from a given image that best meets the user’s intent through their inputs. Existing methods often underperform in complex scenarios, mainly due to three factors. (1) Most works apply a tightly coupled network that directly predicts matting results, lacking interpretability and resulting in inadequate modeling. (2) Existing works are limited to a single type of user input, which is ineffective for intention understanding and also inefficient for user operation. (3) The multi-round characteristics have been under-explored, which is crucial for user interaction. To alleviate these limitations, we propose DFIMat, a decoupled framework that enables flexible interactive matting. Specifically, we first decouple the task into 2 sub-ones: localizing target instances by understanding scene semantics and the flexible user inputs, and conducting refinement for instance-level matting. We observe a clear performance gain from decoupling, as it makes sub-tasks easier to learn, and the flexible multi-type input further enhances both effectiveness and efficiency. DFIMat also considers the multi-round interaction property, where a contrastive reasoning module is designed to enhance cross-round refinement. Another limitation for multi-person matting task is the lack of training data. We address this by introducing a new synthetic data generation pipeline that can generate much more realistic samples than previous arts. A new large-scale dataset SMPMat is subsequently established. Experiments verify the significant superiority of DFIMat. With it, we also investigate the roles of different input types, providing valuable principles for users. Our code and dataset will be public.

**Keywords:** Interactive matting · Multi-modal learning · SMPMat dataset

## 1 Introduction

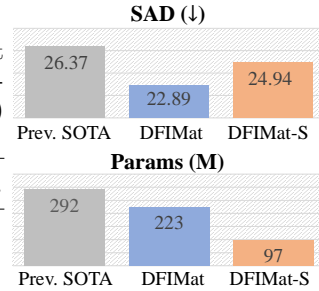
Interactive portrait matting (IPM) is a crucial computer vision task that aims to extract the fine-grained alpha matte of the specific foreground instance that can best match the users’ intention given from their interactive inputs (e.g., clicks, scribbles, texts). The significance of IPM lies in its wide downstream application values, such as image editing, advertisement production, and video conferencing.

<sup>\*</sup> Equal contribution.

<sup>†</sup> Corresponding author.

**Table 1:** Comparison of different methods. DFIMat supports (1) multi-type user inputs, (2) any combination of different types of input at each time, and (3) multi-round iteration.

Design	Approach	Supported input type				Mixed type of input?	Multi-round?
		click	scribble	box	text		
Coupled	[6, 7, 32]	✓					
	[35, 36]		✓				✓
	FGI [5]	✓	✓				
	RIM [17]				✓		
Decoupled	MatAny [37]	✓	✓	✓	✓		
	DFIMat (ours)	✓	✓	✓	✓	✓	✓



**Fig. 1:** Performance and complexity comparison.

Previous studies [5–7, 17, 32, 35, 36], have demonstrated successful performance validation in relatively idealized scenarios (e.g., clear background and single instance without occlusions). However, in real-world scenarios, images often present highly complex backgrounds, with multiple instances and even severe instance occlusions [30]. As a result, the existing approaches have exhibited poor performance on images that are more representative of real-world scenarios [30].

We argue that there are three crucial reasons that potentially cause the failure of the aforementioned methods in challenging real-world scenarios. Firstly, they generally use a coupled network to directly predict matting results. From a top-down perspective, the task actually consists of multiple steps: we need to localize the targeted instance conditioning on the understanding of both user intention and scene semantics, and then conduct fine-grained matting on the corresponding instance. The coupled design lacks interpretability and makes the network difficult to learn each sub-task well especially in complex scenarios.

The second issue is that existing works only consider a single type of user input for the model to understand the user intention, which is inefficient for instance matting tasks that need both global scene understanding for instance localization and local awareness for fine-grained matting at boundary. Specifically, the bounding box input is a suitable way to quickly localize target instance, but lack of ability to focus on fine-grained boundary details. While click and scribble inputs are good at distinguishing local fine-grained details but inefficient for instance localization. One optimal way is first using a bounding box to localize instance and then applying click and scribble to refine local boundary. Such a multiple-type input will also make the user operation more flexible. Therefore, enabling a universal matting interface that is capable of accommodating various types of human prompts is more preferred.

Thirdly, for practical usage, multiple rounds of interaction are typically necessary to refine the matting result until it is satisfactory. However, most works only support feeding user inputs in one step, ignoring the cross-round information that can be useful clues for refinement.

Based on the aforementioned analysis, we propose DFIMat, a decoupled framework that enables flexible interactive matting. Particularly, we propose

to decouple the IPM task into 2 sub-ones: localizing target instances by understanding scene semantics and the flexible user inputs, and conducting refinement for instance-level matting. Following this rule, we subsequently design an interactive semantic capture network (ISCN) and a matting refinement network (MRN) to address these two tasks respectively.

Within ISCN, we propose to enable multi-modal user inputs such as clicks, scribbles, boxes, texts, or any combination of them, resulting in a more concise, flexible, and efficient interaction. This is achieved by encoding various inputs into a unified visual-semantic space, and building strong interactions in the decoder to understand the user intents and predict the target instance mask for instance localization. To meet the practical needs, DFIMat also consider the multi-round interaction property, where we design a contrastive reasoning module to evaluate the consistency between the model prediction and user intention while also explicitly identifying and reasoning the conflict areas during each round’s interaction, providing valuable auxiliary guidance for cross-around refinement.

For MRN, we build a dual-branch network to effectively capture fine-grained local details with global instance-level consideration. As summarized in Tab. 1, our DFIMat distinguishes existing works in: (1) supporting multi-types of user inputs; (2) allowing any combination of different input types (including a single input) at each time; (3) with multi-round iteration ability. Those properties make it more user-friendly and with better effectiveness as verified by experiments.

Data is another important point for method training and evaluation. The volume of real-image datasets for multi-person matting remains relatively small due to the cost of data collection and annotation. In order to obtain a large amount of matting data that contains multi-instance scenes, previous methods [17, 30] adopt a simple synthesis strategy to iteratively add portrait foregrounds to non-portrait backgrounds. Due to the randomness of the adding positions and the lack of instance-scene prior consideration, there is often a large gap between the synthetic images and natural images, it is more preferred to utilize more realistic and complex images for training and evaluation. To fill this gap, we further design a new synthetic data generation pipeline that can generate much more diverse and realistic samples, and build a new large-scale dataset SPMMat, which consists of 40,000 realistic multi-instance images with high-quality matte GT.

Our extensive experiments verify the superiority of DFIMat over representative methods. Notably, DFIMat outperforms previous SOTA by 3.48 SAD on the challenging SMPMat dataset with higher efficiency. We also provide a more lightweight version, DFIMat-S, with only 33% of the parameters of SOTA methods, while still achieving higher matting accuracy, as shown in Fig. 1. By utilizing DFIMat, we also investigate the roles of different input types, providing valuable principles for users on more effective interaction. Our main contributions are:

- We propose a decoupled network for IPM task, which decomposes the task based on a top-down perspective, resulting in a clear performance gain.
- We propose to enable flexible and multi-type user input for interactive matting, making it more effective, efficient, and user-friendly. This is achieved by encoding different inputs into a unified visual-semantic space.

- Concerning the multi-round feature of interaction, we design a contrastive reasoning module to enhance cross-round refinement.
- We propose a new synthetic data generation pipeline that can generate diverse and high-quality image-matte-text pairs in multi-person scenarios. A large-scale dataset is further introduced to facilitate relevant research.
- We investigate the roles of different input types and provide valuable principles for users on more effective interaction.

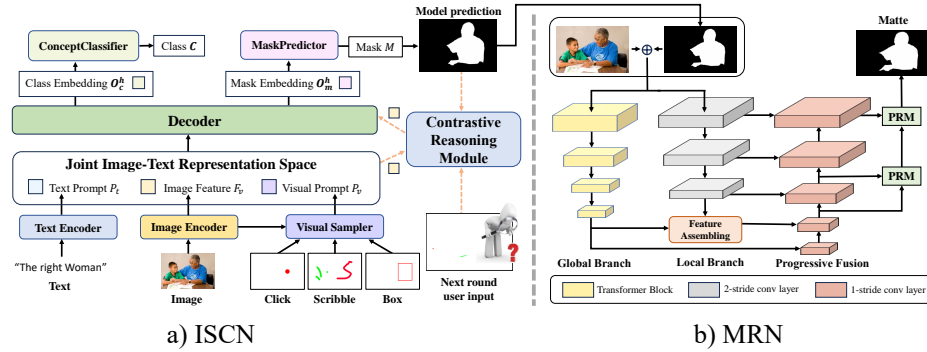
## 2 Related Work

*Interactive Image matting.* Existing methods [5–7, 17, 32, 35, 36] adopt user inputs to identify the foreground and background region, which can usually obtain much better matting results than the automatic ones [2, 3, 8, 10, 14–16, 18–26, 31, 34, 39]. Most of the existing interactive matting methods [5–7, 17, 32, 35, 36] adopt an encoder-decoder-like architecture that takes image and user input as input, and directly predicts matting results. Such a coupled network design makes the model difficult to adapt well in complex real-world scenarios, such as multi-person scenes with severe occlusions. The reason is that the coupled design lacks interpretability and thus increasing the learning difficulty.

Another limitation is that existing works generally only consider a single type of user input (e.g., click, scribble, box, or text), which is ineffective and not user-friendly, as different types of inputs can play different roles and can complement each other. Although a very recent work (i.e., MatAny [37]) expands the input type by using a segmentation foundation model (SAM [13]) to receive different types of input, it still can not support mixed types of input at one time, which failed to exploit complement information from different types of input during user interactions. Moreover, most of the existing works do not consider the multi-round interaction property that is necessary for practical usage and thus ignore the cross-round information that can be useful clues for refinement.

Here we proposed a decoupled network DFIMat that is of better interpretability and performance. It also enables truly flexible inputs by encoding different types of inputs into a unified visual-semantic space, resulting in a more effective and user-friendly matting experience. We also consider the multi-round interaction characteristic and design a contrastive reasoning module to enhance cross-round refinement. A summary of different methods can be seen in Tab. 1.

*Matting datasets.* Numerous matting datasets [14, 26, 34] have been introduced to propel advancements in the field of image matting. Typical matting datasets contain high-resolution images belonging to some specific object categories that have lots of details like hair, accessories, fur, and net, as well as transparent objects. Besides, some other matting datasets focus on a specific category of object, e.g., humans [14, 21] and animals [29]. To generate images containing multiple foreground objects, a typical solution in previous matting methods [17, 30] is to iteratively composite the foreground onto the background image sequentially. Although augmentation strategies have been proposed to reduce the domain gap



**Fig. 2:** The overall framework of DFIMat, which consists of two components: (a) Interactive semantic capture network (ISCN), and (b) Matting refinement network (MRN).

between the real-world images and the composite ones, there is still an urgent need to generate synthetic images that are closer to the real world.

In our work, we build a multi-person matting dataset and ensure its diversity and high quality by designing a new synthetic data generation pipeline.

### 3 Method

#### 3.1 Overview

By reflecting on and summarizing the shortcomings of the existing works, we propose DFIMat, a novel decoupled framework for flexible interactive matting. It consists of two independent components: the interactive semantic capture network (ISCN) and the matting refinement network (MRN), as illustrated in Fig. 2. The ISCN is responsible for understanding the user intention from their various inputs, and localizing the interested instance in the image (Sec. 3.2). The MRN takes the prediction result of ISCN (i.e., a coarse mask of the interested instance) as well as the original image, and performs refinement to produce the final alpha matte for the corresponding instance-level matting (Sec. 3.3).

#### 3.2 Interactive Semantic Capture Network (ISCN)

Here we introduce the proposed ISCN, a model that understands scene semantics and multiple/flexible user inputs to localize the target instance, in an interactive manner. The ISCN design takes inspiration from the recent success of multi-modal learning methods [9, 11, 41] and encodes the different types of user inputs and image into a unified visual-semantic space. Then, strong interactions among them are introduced to better understand the user intentions and then predict the target instance mask for instance localization. Leveraging the characteristic of multi-round interaction, we further design a simple but effective contrastive reasoning module to evaluate the consistency between the model prediction and

user intention while also explicitly identifying and reasoning the conflict areas during each round’s interaction, providing valuable auxiliary guidance for the model to further refine its outputs. The overall architecture is shown in Fig. 2, which will be introduced in detail.

**Unified visual-semantic space.** Given an input image  $I \in \mathbb{R}^{H \times W \times 3}$ , we first extract image feature  $F_v$  by an image encoder. Visual inputs  $s (s \in \{\textit{clicks}, \textit{scribbles}, \textit{boxes}\})$  are converted to visual prompts  $P_v$  through a visual sampler:

$$P_v = \mathbf{VisualSampler}(s, F_v). \quad (1)$$

The visual sampler performs feature point sampling on the corresponding locations in image features based on user inputs. Textual inputs are fed into a text encoder for text prompts  $P_t$ . Then, a decoder is used to build strong interactions between the image feature  $F_v$  and user prompts  $P_v, P_t$  to understand both scene semantics and user intentions, and finally predict the target instance mask for instance localization. This can be formulated as follows:

$$\langle M, C \rangle = \mathbf{Decoder}(\langle P_t, P_v \rangle | F_v), \quad (2)$$

where  $M$  is the predicted instance mask and  $C$  is its class (is human or not). More specifically, we first initialize three types of learnable query: object queries  $Q_o$ , text queries  $Q_t$  and visual queries  $Q_v$ . Each decoder stage contains cross-attention operations on images and learnable queries, and a prompt self-attention block to perform the interaction between queries and prompts:

$$\begin{aligned} Q_x &= \mathbf{CrossAttention}(Q = F_v, K = V = Q_x), \quad x \in \{o, v, t\}, \\ Q_o, Q_t, Q_v &= \mathbf{SelfAttention}(Q_o, Q_t, Q_v, P_v, P_t). \end{aligned} \quad (3)$$

The output of last decoder stage and image features are passed to FFN to obtain the mask embeddings  $O_h^m$  and class embeddings  $O_h^c$ :

$$O_h^m, O_h^c = \mathbf{FFN}(F_v, Q_o, Q_t, Q_v). \quad (4)$$

Finally, ISCN predicts the masks  $M$  and the classes  $C$  based on  $O_h^m$  and  $O_h^c$ :

$$\begin{aligned} \mathbf{M} &= \mathbf{MaskPredictor}(O_h^m), \\ \mathbf{C} &= \mathbf{ConceptClassifier}(O_h^c). \end{aligned} \quad (5)$$

Here MaskPredictor (and ConceptClassifier) are task-specific heads and we follow the design of X-decoder [40] for its simplicity.

**Contrastive reasoning module.** We propose a contrastive reasoning module (CRM) to evaluate the consistency between the model prediction and user intention while also explicitly identifying and reasoning the conflict areas during each round’s interaction, which can serve as a useful clue for cross-round refinement. Specifically, we maintain a mask  $M_{ref} \in \mathbb{R}^{H \times W}$ , where each pixel has one of the following three values based on the difference between new user input and the previous model prediction: (1)  $D_r = 0$  indicates no conflict; (2)  $D_{fg} = 1$  means

previous prediction classified the area as background but new user input suggest it is foreground, and (3)  $D_{gf} = 2$  means conversely with  $D_{fg}$ .  $M_{ref}$  is initialized with all pixels set to  $D_r$  and recalculated upon receiving new user input. A convolution layer is utilized to transform  $M_{ref}$  into an embedding  $E_c \in \mathbb{R}^{H \times W \times C_1}$ , with each pixel value corresponding to a learnable  $C_1$ -dimensional vector. Then, we resize and combine  $E_c$  with the image feature  $F_v$  to get the conflict-involved feature  $F_v = \text{resize}(E_c) + F_v$  for cross-round refinement. In addition, the previous prediction result is also sent to the decoder as a mask to participate in the calculation of masked multi-head attention in prompt self-attention block.

**Loss.** We train ISCN with standard segmentation loss:

$$\mathcal{L}_{\text{ISCN}} = \mathcal{L}_{c\_ce}(C, \hat{C}) + \mathcal{L}_{m\_bce}(M, \hat{M}) + \mathcal{L}_{m\_dice}(M, \hat{M}), \quad (6)$$

where  $\hat{C}, \hat{M}$  is GT category and mask respectively.  $\mathcal{L}_{c\_ce}$ ,  $\mathcal{L}_{m\_bce}$ , and  $\mathcal{L}_{m\_dice}$  denote cross-entropy, binary cross-entropy, and dice loss, with weights 0.1:1:1.

### 3.3 Matting Refinement Network (MRN)

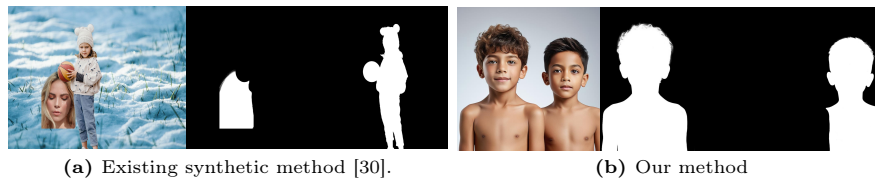
MRN aims to refine the mask prediction from ISCN to obtain accurate alpha matte predictions. As ISCN has already given a relatively good instance mask as a good beginning, the task difficulty for MRN is largely reduced. Here, we design a simple dual-branch network as our MRN, to capture fine-grained local details while simultaneously considering global instance-level semantics, as shown in Fig. 2. Our insight is that images for multi-instance matting tasks often contain complex human interaction and background, so a global encoder is needed to better capture the overall structure and background information, while a local encoder can focus more on details. Then, a subsequent progressive feature fusion should be built to fuse the features and decode them to final matte. Taking those things in mind, we design a simple network containing a global encoder, a local encoder, and a progressive feature fusion module, as shown in Fig. 2. Specifically, we choose a CNN as the local encoder, as it can effectively exploit local features, and we choose a transformer-based encoder as our global encoder, as the self-attention operation can build strong non-local interactions in the images to form a better global instance-level understanding. For progressive fusion, we start from the latent feature from the global branch, as it contains rich global instance-level representation. We then fuse it with the feature from local branch in a progressive manner (from low to high resolution), we also utilize PRM [38] to further refine the result.

**Loss.** Since most pixels in the coarse mask are already predicted correctly, only a few “hard” pixels need significant refinement. To make the network pay more attention to those “hard” pixels, we adopt a simplified hard-sample mining objective function  $\mathcal{L}_{\text{MRN}}$  as follows:

$$\mathcal{L}_{\text{MRN}} = \frac{1}{|C|} \sum_{i \in C} |\alpha_p^i - \alpha_g^i| + \lambda \frac{1}{|H|} \sum_{j \in H} |\alpha_p^j - \alpha_g^j|, \quad (7)$$

**Table 2:** Comparison with previous multi-instance matting dataset and ours.

Datasets	Image Number	Instance Number	Annotation	
			Matte	Text description
HIM2K(nature)	320	830	✓	
HIM2K(synthetic)	1,680	5,884	✓	
SMPMat	40,000	142,357	✓	✓

**Fig. 3:** Visual comparison of synthetic datasets.

where  $C$  represents the whole pixel-set and  $H$  denotes “hard” pixel-set whose error to corresponding ground truth ranks in the top 30% of the matte.  $\lambda$  denotes the weight that emphasizes the hard samples and is set as 1 by default.

## 4 The SMPMat Dataset

We propose a synthetic multi-person matting dataset called SMPMat to facilitate the research of instance matting task. To our knowledge, the existing multi-instance dataset from natural images [30] suffers from low data scale as well as diversity. Specifically, HIM2K [30] is the mainly used dataset that focuses on instance-level matting under multi-person scenarios. As can be observed in Tab. 2, the HIM2K dataset only contains a limited scale of data that is collected from natural scenes (i.e., only contains 320 natural images with 930 instances in total), making it only serve as a validation set.

In order to obtain a large amount of matting data that contains multi-instance scenes, previous methods [17, 30] adopt a simple synthesis strategy to iteratively add portrait foregrounds to no-portrait background. Due to the randomness of the adding positions and the lack of instance-scene prior consideration, there is often a large gap between the synthetic images and natural images, as shown in Fig. 3. To fill this gap, we design a new synthetic data generation pipeline that can generate much more diverse and realistic samples, and build a new large-scale dataset SMPMat, which consists of 40,000 realistic multi-instance images with high-quality matte GT.

**The synthetic data generation pipeline.** Inspired by work [27, 28, 33] using diffusion models for image synthesis, we build a new synthetic data generation pipeline that can synthesize an infinite amount of realistic and diverse images with high-quality matte ground truth. Our insight is that the latent feature within the well-trained diffusion process (e.g., Stable Diffusion [27]) contains rich



semantic contexts of the corresponding generated image, so ideally, the matte ground truth of the generated image can be well interpreted from such latent feature. Thus, we build an interpreter to derive the matte ground truth from the latent feature of the diffusion process. Based on the good performance of the recent text-to-image diffusion models, all we need to do is just train the interpreter for this interpretation task, using a small amount of annotated real data. After that, we can synthesize an infinite amount of images with high diversity based on the well-trained text-to-image diffusion models (e.g., Stable Diffusion [27]), and simultaneously obtain their matte ground truth by our trained interpreter.

Specifically, our method is illustrated in Fig. 4. During training, given a real image  $I$  and the paired text description  $H$ , we feed them into a pre-trained text-to-image diffusion model (i.e., Stable Diffusion [27] in our implementation) and acquire the multi-scale latent feature map  $\mathcal{F}$  as well as the text-visual cross-attention map  $\mathcal{M}$  in it (i.e., the denoising U-Net). We concat those intermediate representations  $\hat{\mathcal{F}} = \text{Concat}([\mathcal{F}, \mathcal{M}])$  and send  $\hat{\mathcal{F}}$  to our interpreter and interpret them into GT matte. For the detailed architecture of the interpreter, any decoder for dense prediction task can be used, and here we adopt Mask2Former [4], which contains a transformer decoder and a pixel decoder. Given  $\hat{\mathcal{F}}$ , and  $N$  learnable queues  $\{Q_0, Q_1 \dots Q_T\}$  as input, it outputs  $N$  foreground alpha matte  $A$  and their corresponding categories  $L$  (is a human instance or not). We train the interpreter with binary cross-entropy loss and alpha loss:

$$\mathcal{L}_{\text{P-decoder}} = \lambda \mathcal{L}_{c\_bce}(\hat{L}, L) + \mathcal{L}_{a\_alpha}(\hat{A}, A), \quad (8)$$

where  $\lambda = 0.1$ ,  $\hat{L}, \hat{A}$  means the ground truth category and alpha matte respectively. To enable the interpreter training, we collected 400 real images and labeled them with matte ground truth as well as a text description. Once the training finishes, we use the Stable Diffusion to generate realistic images, and use the trained interpreter to obtain the matte ground truth at the same time. To let the Stable Diffusion generate diverse images, we design a prompt to instruct GPT-4 [1] to generate an infinite amount of diverse and semantic-rich text descriptions (see supplementary for details), and send them to the Stable Diffusion for text-to-image generation. We give a visual comparison in Fig. 3. It can be observed that the image synthesized by existing data synthetic algorithm [30] often lacks of realistic instance lay-out with scene-instance prior consideration, while our new data synthetic pipeline enables a much more realistic data generation. Please also refer to Sec. 5.4 for quantitative evaluation.

**The SMPMat dataset.** We followed the above generation pipeline to generate a large-scale multi-person matting dataset SMPMat, in which we carefully select 40,000 high-quality multi-person scene images from our generated samples to form the dataset. We generated diverse text descriptions for the people in each image through GPT-4 [1], as the additional text annotations to broaden the usage of the proposed dataset. Compared with existing multi-person matting datasets, the SMPMat dataset shows superiority in both data diversity (see Tab. 2) and image quality (see Fig. 3).

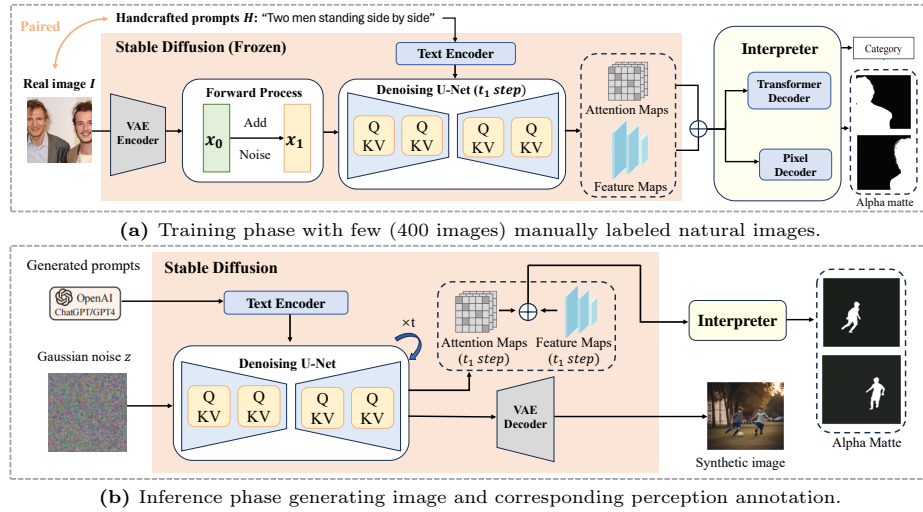


Fig. 4: The synthetic data generation pipeline.

## 5 Experiments

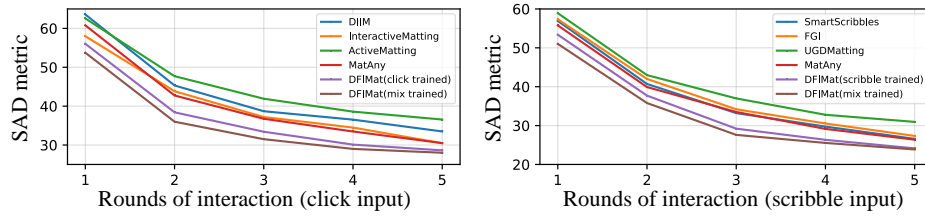
### 5.1 Implementation Details

We first train the semantic capture network alone without the matting refinement network. After the semantic network converges, we freeze it and then train the matting refinement network. For all the network training, Adam optimizer [12] is used and the base learning rate is set to  $5 \times 10^{-4}$  with the cosine learning rate scheduler. The matting network is trained for 150 epochs, while the uncertainty estimation decoder and the refinement network are trained for 75 epochs.

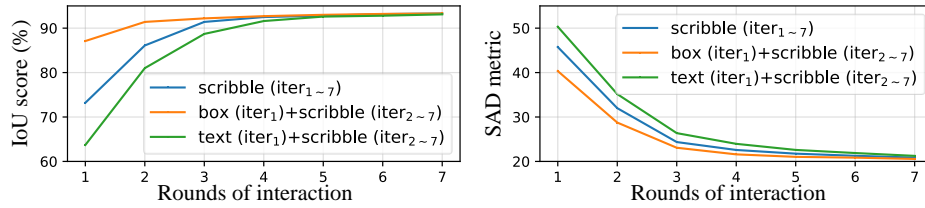
### 5.2 Dataset and Evaluation Protocol

We compare our method with existing interactive matting methods [5–7, 17, 32, 35, 36] on the SMPMat dataset and the HIM2K (natural) dataset [30]. All the methods are trained on the training set of SMPMat, and evaluated on both the validation set of SMPMat and HIM2K (natural). All the used metrics (the smaller, the better) follow previous works. We train different methods under their supported input type. For a fair comparison, we first train our DFIMat under the same protocol as the existing method (only one type of input) to make the comparison under each supported type. Then, we also train our DFIMat using mixed types of user input to show its full performance. We design some rules to imitate human behavior and simulate the user input during training and testing. Rules are as follows.

**Click & scribble input.** Since they usually conduct in a multi-round interaction fashion, we set a 5-round interaction loop, for both training and testing.



**Fig. 5:** Performance comparison under different rounds of interaction on SMPMat.



**Fig. 6:** Effect of different user inputs on the SMPMat dataset.

Each round adds 3 clicks/2 scribbles. The first round of input is randomly generated by GT (but the same for different methods in the same input image for a fair comparison). Subsequent inputs are generated in the most significant area calculated from previous prediction results and GT (see the supplementary material for the definition of the most significant area). For methods that do not consider multi-round interactions, we aggregate the inputs from rounds 1 to  $t$  and feed them together into the model as the input for round  $t$ .

**Box & text input.** A single round of interaction is built for both training and testing, where the input is obtained directly from GT.

**Mixed input.** For both training and testing, the interaction round is set to 5. We set the first round of input to a combination of text and any kind of visual input (click/scribble/box). The input types in 2 ~ 3 rounds are randomly selected from click and scribble, and clicks/scribbles are added to the most significant areas based on previous prediction results (same rule as the aforementioned single-type click or scribble input).

### 5.3 Comparison with the state-of-the-art methods

**Single-type user input.** We conducted comparisons of various models on the SMPMat validation set and the HIM2k natural subset, with results listed in Tab. 3. Experimental outcomes demonstrate that under single-type input settings, our approach consistently outperforms all state-of-the-art methods. To investigate how different models perform during multiple rounds of interaction, we present the SAD variation curves of various methods during the interactive process in Fig. 5. Notably, Our DFIMat also achieves more accurate prediction output with fewer interactions needed.

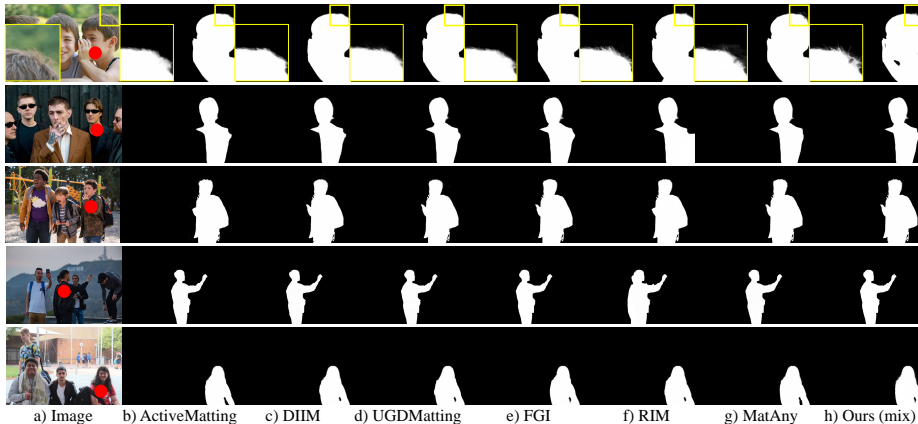
**Table 3:** Quantitative comparison on SMPMat validation set and HIM2K natural set. The MSE metrics are scaled by  $10^2$ .

Supported User Input	Method	SMPMat Validation				HIM2K Natural			
		SAD	MSE	GRAD	CONN	SAD	MSE	GRAD	CONN
Click	ActiveMatting [36]	36.56	0.93	17.85	37.46	16.69	0.49	7.85	16.73
	InteractiveMatting [32]	30.47	0.70	15.82	31.11	14.06	0.41	7.02	14.11
	DIIM [6]	33.49	0.81	16.37	32.04	15.87	0.43	7.36	15.89
	MatAny [37]	30.49	0.71	15.89	31.44	14.01	0.40	7.00	14.02
	DFIMat(click trained)	28.63	0.67	15.2	28.49	13.74	0.39	6.49	13.77
	DFIMat(mix trained)	28.01	0.66	16.84	27.36	13.59	0.38	6.47	13.62
Scribble	SmartScribbles [35]	26.59	0.58	16.01	26.77	12.42	0.41	6.27	12.26
	FGI [5]	27.34	0.61	16.85	26.95	13.63	0.39	6.51	14.01
	UGDMatting [7]	30.95	0.71	17.90	31.76	14.58	0.42	7.39	15.37
	MatAny [37]	26.37	0.58	15.97	26.82	12.40	0.41	6.25	12.51
	DFIMat(scribble trained)	24.14	0.50	15.82	24.19	12.22	0.38	6.25	12.23
	DFIMat(mix trained)	23.86	0.49	15.79	23.80	12.07	0.37	6.18	11.99
Box	MatAny [37]	50.31	2.36	30.16	50.52	21.85	0.93	10.79	22.44
	DFIMat(box trained)	47.44	2.17	28.13	47.35	20.46	0.88	9.88	21.38
	DFIMat(mix trained)	46.29	1.85	28.01	46.24	19.79	0.85	8.74	20.54
Text	RIM [17]	52.49	2.94	31.46	52.87	22.89	0.97	11.30	23.33
	DFIMat(text trained)	54.86	3.31	21.19	54.79	23.53	1.05	12.06	23.69
	DFIMat(mix trained)	50.32	2.83	30.70	50.24	22.45	0.92	11.19	22.66
Mix	DFIMat(mix trained)	22.89	0.47	15.54	22.73	11.77	0.36	5.98	11.80

**Multi-type user input.** From the test results in Tab. 3, the following conclusions can be drawn: (1) Unlike previous methods that only support one type of input, our model enables mixed types of inputs. Experiments show that when trained under mixed user inputs, the performance of our model can be further enhanced, even with the same single-type input inference. This verifies the benefit of multi-type user input for model training, as they can give more complementary information. (2) As in the bottom line of Tab. 3, it can be seen that when applying mixed type inputs during inference, the performance of our model can be further improved. This further verifies the benefit of multi-type user inference, and it also makes user interaction more flexible.

**Analysis on user input choice.** Here we investigate the roles of different input types, aiming to provide valuable principles for users on more effective interaction. We assume only one input per interaction. We use IoU to measure the coarse-grained instance capture accuracy, and SAD to measure the fine-grained matting accuracy. As in Fig. 6, box input can give a good start as its effectiveness for instance localization, scribble (we observe a similar role but slightly lower performance on click) is useful to refine local details, text is usually not-efficient. As a result, box at round 1 and scribble at the remaining iteration is an optimal choice for efficient user interaction.

**Qualitative comparison.** We follow the same protocol in Sec. 5.2 to conduct the experiment, as in Fig. 7. The red points in (a) indicate the target instances. For our DFIMat, we only give the result from our full model (i.e., mix trained & inference) in (h) due to the space limitation. More comparisons with other



**Fig. 7:** Qualitative comparisons among different methods. The red points in (a) indicate the target instances.

**Table 4:** Analysis of the decoupled design.

Setting	Complexity GFLOPs	Semantic Capture NoC @ 90%	Matting SAD MSE
Coupled Network	0.2186	7.43	25.19 0.53
Coupled Training	0.223	6.82	23.62 0.48
DFIMat	0.223	<b>6.51</b>	<b>22.89 0.47</b>

**Table 5:** Effectiveness of CRM.

Setting	Semantic Capture NoC @ 90%	Matting SAD MSE
w/o CRM	7.93	25.03 0.52
w/ CRM	<b>6.51</b>	<b>22.99 0.47</b>

variants of DFIMat refer to our supplementary material. From Fig. 7, it can be seen that DFIMat can accurately localize the target instance. More importantly, it shows superiority at perceiving fine-grained details: (1) Hair regions across all examples. (2) Other hollow body areas like the fingers or arm in row 2-5.

#### 5.4 Quantitative analysis on data synthesis method

Here we give some quantitative evaluation of our proposed data synthesis pipeline. We compare it with the existing methods [30]. We separately apply it and our method to synthesize a same amount of data (i.e., 45k instances) from model training. Then, we apply 3 different methods (i.e., MG [38], MatteFormer [25], and MRN) to train on the synthesized data, and evaluate their testing performance on the HIM2K natural dataset. The result is listed in Tab. 7, it can be observed that when trained on the data synthesized by our method, the performance of different models is significantly and consistently better than trained on the data generated by existing method [30], which further verifies the superiority of our data synthesis pipeline.

#### 5.5 Ablation studies

Here component effects are studied on the SMPMat dataset.

**Table 6:** Ablation study on the encoder setting of MRN.

Branch	SAD	MSE
Global	24.79	0.51
Local	23.35	0.48
Hybrid	<b>22.89</b>	<b>0.47</b>

**Table 7:** Matting performance comparison under datasets using different data generation schemes.

Method	MG [38]		MatteFormer [25]		MRN	
	SAD	MSE	SAD	MSE	SAD	MSE
Sys method in [30]	17.23	0.51	19.74	0.85	15.82	0.46
Ours	<b>12.48</b>	<b>0.43</b>	<b>13.59</b>	<b>0.38</b>	<b>11.77</b>	<b>0.36</b>

**Effect of the decoupled design.** Here we evaluate the following settings: **a) Coupled Network** (ISCN with matting head); **b) Coupled Training** (ISCN + MRN but trained jointly); **c) DFIMat** (Decoupled in both network and training). Tab. 4 shows that both decoupled network and decoupled training have obvious performance gains (in both semantic capture and matting that align with our insight). Besides, the extra complexity from our decoupled design is neglectable (2% in GFLOPS), which validates the effectiveness of our design. We think the reason is that it can make the 2 independent tasks more focused and easier to be optimized, thus leading to a clear performance gain.

**Effect of the contrastive reasoning module (CRM).** From Tab. 5, it can be seen that with CRM, both semantic capture and matting performance improve by a noticeable margin, which shows the advantages of our design.

**Design choices of the matting refinement network (MRN).** Tab. 6 shows that our dual-stream design can enhance the performance.

## 6 Conclusion and Limitations

In this paper, we propose DFIMat, a decoupled framework that enables flexible interactive matting in multi-person scenarios, which consists of two modules, the interactive semantic capture network and the matting refinement network. DFIMat enables flexible and multi-type user input by encoding different inputs into a unified visual-semantic space, resulting in a more effective and user-friendly matting experience. Concerning the multi-round interaction requirement for practical usage, we also design a contrastive reasoning module to enhance cross-round refinement. To address the limitation from the perspective of data, we introduce a new synthetic data generation pipeline that can generate much more realistic samples than previous arts. A new large-scale dataset SMPMat is subsequently established. Extensive experiments verify the significant superiority of DFIMat while also providing valuable principles for efficient interaction. Despite its effectiveness, our method produces less accurate results when only coarse user input (e.g., box, text) is provided, and similar phenomena can also be observed in other methods. Besides, our SMP-Mat dataset does not include crowd scenes due to unrealistic generation results from base diffusion model [27] in such cases. We will tackle those limitations in future works.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Cai, S., Zhang, X., Fan, H., Huang, H., Liu, J., Liu, J., Liu, J., Wang, J., Sun, J.: Disentangled image matting. In: CVPR. pp. 8819–8828 (2019)
3. Chen, Q., Ge, T., Xu, Y., Zhang, Z., Yang, X., Gai, K.: Semantic human matting. In: ACM MM. pp. 618–626 (2018)
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
5. Cheng, H., Xu, S., Jiang, X., Wang, R.: Deep image matting with flexible guidance input. arXiv preprint arXiv:2110.10898 (2021)
6. Ding, H., Zhang, H., Liu, C., Jiang, X.: Deep interactive image matting with feature propagation. *IEEE Transactions on Image Processing* **31**, 2421–2432 (2022)
7. Fang, X., Zhang, S.H., Chen, T., Wu, X., Shamir, A., Hu, S.M.: User-guided deep human image matting using arbitrary trimaps. *IEEE Transactions on Image Processing* **31**, 2040–2052 (2022)
8. Hou, Q., Liu, F.: Context-aware image matting for simultaneous foreground and alpha estimation. In: CVPR. pp. 4130–4139 (2019)
9. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetrm: modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021)
10. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: AAAI. vol. 36, pp. 1140–1147 (2022)
11. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
14. Li, J., Ma, S., Zhang, J., Tao, D.: Privacy-preserving portrait matting. In: ACM MM. pp. 3501–3509 (2021)
15. Li, J., Zhang, J., Maybank, S.J., Tao, D.: Bridging composite and real: towards end-to-end deep image matting. *IJCV* **130**(2), 246–266 (2022)
16. Li, J., Zhang, J., Tao, D.: Deep automatic natural image matting. arXiv preprint arXiv:2107.07235 (2021)
17. Li, J., Zhang, J., Tao, D.: Referring image matting. In: CVPR. pp. 22448–22457 (2023)
18. Li, Y., Lu, H.: Natural image matting via guided contextual attention. In: AAAI. pp. 11450–11457 (2020)
19. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)

20. Liu, J., Yao, Y., Hou, W., Cui, M., Xie, X., Zhang, C., Hua, X.s.: Boosting semantic human matting with coarse annotations. In: CVPR. pp. 8563–8572 (2020)
21. Liu, Y., Xie, J., Shi, X., Qiao, Y., Huang, Y., Tang, Y., Yang, X.: Tripartite information mining and integration for image matting. In: ICCV. pp. 7555–7564 (2021)
22. Lu, H., Dai, Y., Shen, C., Xu, S.: Indices matter: Learning to index for deep image matting. In: ICCV. pp. 3266–3275 (2019)
23. Lutz, S., Amplianitis, K., Smolic, A.: Alphagan: Generative adversarial networks for natural image matting. arXiv preprint arXiv:1807.10088 (2018)
24. Ma, S., Li, J., Zhang, J., Zhang, H., Tao, D.: Rethinking portrait matting with privacy preserving. IJCV pp. 1–26 (2023)
25. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: Transformer-based image matting via prior-tokens. In: CVPR. pp. 11696–11706 (2022)
26. Qiao, Y., Liu, Y., Yang, X., Zhou, D., Xu, M., Zhang, Q., Wei, X.: Attention-guided hierarchical structure aggregation for image matting. In: CVPR. pp. 13676–13685 (2020)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)
29. Shahrian, E., Rajan, D., Price, B., Cohen, S.: Improving image matting using comprehensive sampling sets. In: CVPR. pp. 636–643 (2013)
30. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. In: CVPR. pp. 2647–2656 (2022)
31. Tang, J., Aksoy, Y., Oztireli, C., Gross, M., Aydin, T.O.: Learning-based sampling for natural image matting. In: CVPR. pp. 3055–3063 (2019)
32. Wei, T., Chen, D., Zhou, W., Liao, J., Zhao, H., Zhang, W., Yu, N.: Improved image matting via real-time user clicks and uncertainty estimation. In: CVPR. pp. 15374–15383 (2021)
33. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. arXiv preprint arXiv:2308.06160 (2023)
34. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR. pp. 2970–2979 (2017)
35. Yang, X., Qiao, Y., Chen, S., He, S., Yin, B., Zhang, Q., Wei, X., Lau, R.W.: Smart scribbles for image matting. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **16**(4), 1–21 (2020)
36. Yang, X., Xu, K., Chen, S., He, S., Yin, B.Y., Lau, R.: Active matting. *Advances in Neural Information Processing Systems* **31** (2018)
37. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing* p. 105067 (2024)
38. Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., Yuille, A.: Mask guided matting via progressive refinement network. In: CVPR. pp. 1154–1163 (2021)
39. Zhang, Y., Gong, L., Fan, L., Ren, P., Huang, Q., Bao, H., Xu, W.: A late fusion cnn for digital matting. In: CVPR. pp. 7469–7478 (2019)



40. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)
41. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)