





# Supplementary Material: 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

## 1 Overview

We provide additional details and results in this Appendix. In Sec. 2.1, more qualitative comparisons with other variants of DFIMat is given. We further explore the fine-grained level performance, robustness, and generalization analysis of DFIMat in Sec. 2.2 to Sec. 2.4. In Sec. 2.5, we provide the definition of the most salient area (mentioned in L-382 of the manuscript). Examples of different user input types are given in Sec. 2.6. Sec. 3.1 provides additional user studies on the generated dataset, demonstrating its excellent visual authenticity. In Sec. 3.2 we give more statistics (dataset split) about SMPMat. We introduced how to obtain text prompt for image synthesis in Sec. 3.3. In Sec. 3.4, we give the details about the instance-level text description annotations in the proposed SMPMat dataset.

## 2 More Details of Experiments

### 2.1 More Qualitative Comparison with other variants of DFIMat

Since only the full model (i.e., mix trained & inference) of DFIMat has been given in the manuscript, here we also give the qualitative result of other variants of DFIMat. As shown in Fig. A1, Our method also consistently outperforms all existing methods under different input settings, and the matting quality achieves the best under our full model.

### 2.2 Fine-Grained Analysis

Tab. A1 substantiates our claim of achieving fine-grained superiority, particularly evident in the meticulous handling of intricate details and transitional zones within our model’s output. To further validate this assertion, we have incorporated the SAD in transition areas (SAD-T) metric into our evaluation framework. This metric is specifically designed to evaluate the errors of transition regions, which contains crucial fine details in the matting task.

---

<sup>\*</sup> These authors contributed equally to this work.

<sup>†</sup> Corresponding author.

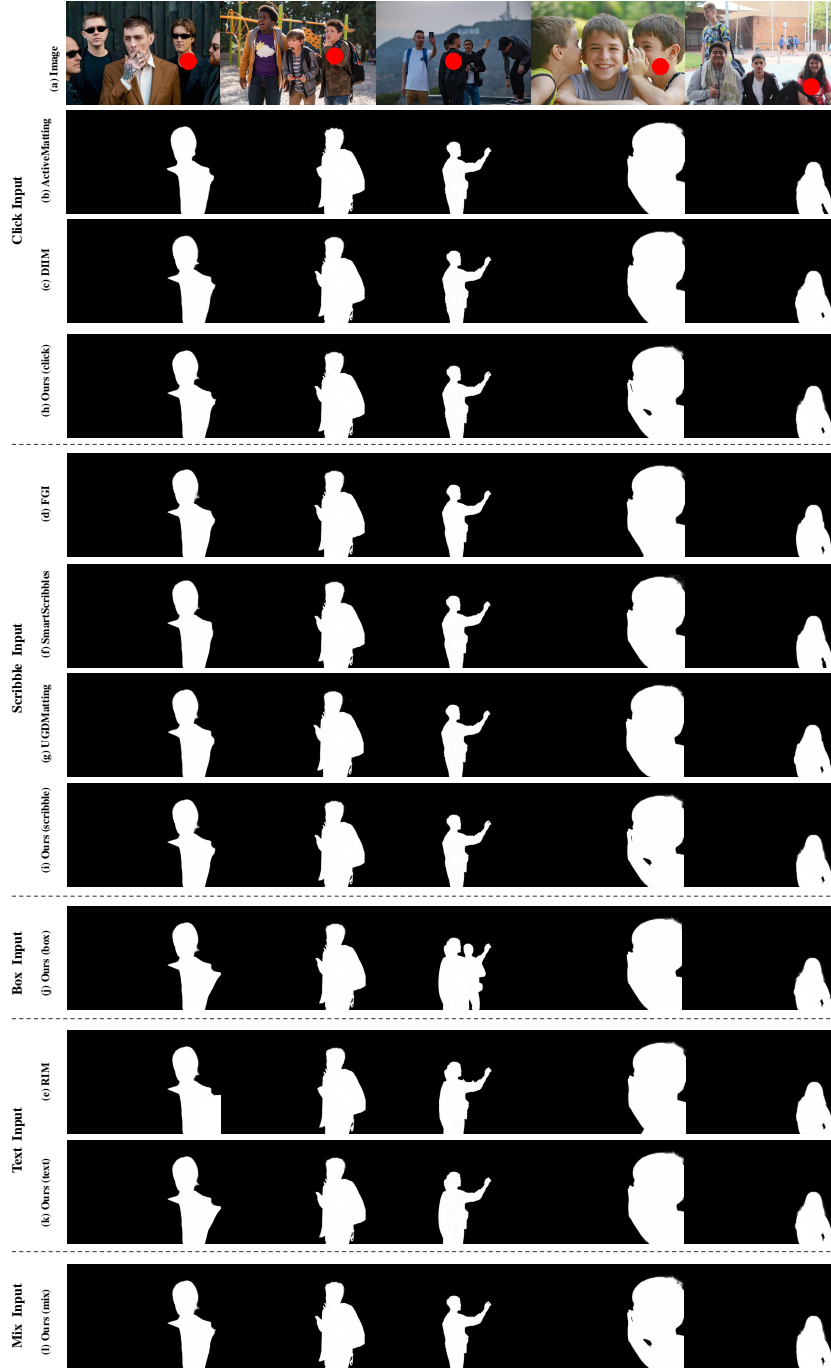


Fig. A1: More comparisons with other variants of DFIMat.

By including SAD-T in our analysis, we demonstrate a marked improvement in managing these critical areas, thus confirming our model’s capability to excel in fine-grained processing.

**Table A1:** Fine-grained analysis on SMPMat.

Method	SAD(↓)	SAD-T(↓)
Prev. SOTA(MatAny)	25.73	11.82
DFIMat	<b>22.89</b>	<b>9.52</b>
DFIMat-S	24.94	10.3

### 2.3 MRN’s Robustness

In order to thoroughly test the resilience and adaptability of our proposed Matting Refinement Network (MRN), we subjected the predicted masks obtained from the Interactive Semantic Capture Network (ISCN) to a series of rigorous tests involving the introduction of noise. This was achieved by applying morphological processing techniques, following the established protocol delineated in prior work [4], to systematically perturb the initial mask predictions.

The purpose of this experiment was twofold: first, to emulate real-world conditions where masks may be inherently imperfect due to various factors such as image quality, lighting conditions, or inherent limitations of the initial segmentation algorithm; second, to assess the robustness of MRN in the face of such inaccuracies, which are common in practical applications.

The results presented in Table A2 thus underscore the reliability of MRN as a post-processing step in image matting, particularly when dealing with noisy or imperfect input masks. This robustness ensures that MRN can be seamlessly integrated into workflows where the initial segmentation might not be perfect due to various constraints or limitations.

**Table A2:** Robust to noisy MRN input.

Input	SAD(↓)	SAD-T(↓)
dilate	23.36	9.88
erode	23.59	10.01
mix	23.41	9.95
none	<b>22.89</b>	<b>9.52</b>



**Fig. A2:** Examples of different input type, where red and green respectively represent the clicks (scribbles) belonging to the foreground/background categories, the yellow rectangles are used as box input, and the blue rectangles denote the most significant area calculated based on the previous prediction results.

## 2.4 Generality

To further probe into the versatility and broad applicability of our proposed model, we conducted extensive experiments on the Refmatte dataset [2]—a collection encompassing a wide array of foreground categories.

Our model, in this rigorous testing ground, not only met but surpassed expectations, demonstrating a remarkable superiority over the leading competitor, MatAny [3]. Specifically, in terms of the SAD metric, our model exhibited a notable 2.07 advantage, achieving an impressive SAD of 6.17 compared to MatAny’s 8.24. This significant margin of victory is indicative of our model’s superior adaptability and robustness, enabling it to handle the intricacies and variances of different foreground categories with finesse.

## 2.5 Definition of the Most Significant Area

In the multi-rounds interactive image matting, to simulate a realistic interaction process, we add new user input in the area with the most significant error in the previous prediction result. Specifically, we evenly divide the input image into a grid of  $4 \times 4$  regular regions, and calculate the sum of the alpha differences between the previous prediction result and the ground truth (GT) in each region. The region with the largest difference is regarded as the most significant area. Such setup ensures that the simulated inputs are in consistent alignment with real-world interactions.

## 2.6 Examples of Different Input Type

We give some examples of different input types in Fig. Å2. Each column of click, scribble, and mixed input types represents newly added input in the next round of interaction. Box type input allows the user to not completely align the edges of foregrounds. Mixed inputs can compensate for the limitations of a single type of input in foreground determination, especially for text input.

## 3 More Details about SMPMat

### 3.1 User Study

Sec. 5.4 of the manuscript presents quantitative experiments demonstrating that our SMPMat dataset outperforms other synthetic datasets in enhancing model performance. In order to further reflect the image quality of the generated dataset, we conducted a user study on the SMPMat dataset. The user study selected 1,000 images each from the SMPMat dataset and a subset of generated images from HIM2K. Fifty volunteers were invited to compare images from both datasets and select those they deemed more visually realistic. Specifically, each volunteer was given 50 images from each dataset and asked to pick out the 30 most realistic-looking images among the total of 100. To minimize the impact of individual bias on experimental fairness, each image was assessed by multiple volunteers.

The statistical outcome revealed that, out of 1,500 images judged to be closer to real images (excluding duplicates), 96.2% originated from SMPMat, while only 3.8% came from the synthetic images of HIM2K. The reason some SMPMat samples ranked lower in quality was due to deformations in human fingers and limbs, an issue requiring optimization of the stable diffusion model. However, overall, SMPMat had a clear edge in foreground-background style consistency and distribution, with content more closely adhering to real-world image patterns. The statistical results from the subjective evaluation further confirmed that the SMPMat dataset effectively narrowed the visual gap with real images.

### 3.2 More Statistics of SMPMat

We present more details about the statistics of SMPMat in Tab. Å3. Images were randomly split, with 60% going into training set and 40% into validation set.

**Table Å3:** The number of images and instances of different subsets

	Image	Instance
Training set	36,000	128,273
Validation set	4,000	14,084

### 3.3 Text Prompt Generation for Image Synthesis

In the image synthesis stage of SMPMat, we use GPT-4 [1] to obtain diverse and rich prompts and input them into our data generation pipeline.

Firstly, specific positive prompt prefix and negative prompt are set to enhance image quality and detail richness:

**Positive Prompt Prefix:** best quality, ultra detailed, masterpiece.

**Negative Prompt:** lowres, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, mutated hands, poorly drawn hands, poorly drawn face, cropped, worst quality, low quality, normal quality, artifacts, signature, watermark, blurry, extra arms, extra legs, missing arms, missing legs, long neck, humpbacked, bad feet, nsfw

Then, we divide the mainly used content-description prompt words into 10 categories, and refer to GPT-4 [1] and existing AI painting prompt generation websites (e.g., prompttool) to gather candidate prompt words for each category, creating a comprehensive vocabulary repository. In Tab. A4 we list the number of attributes (including foreground instance number) and the statistics for some attributes under each category, based on the text prompts used for image generation. For each image generation, we randomly sample prompt words from the vocabulary repository and combine them with commas, as shown in Algorithm 1. The style for image is set as photographic.

**Table A4:** Prompt vocabulary library.

	Category	Num.	Representative attribute (counts)
<b>Back-ground</b>	Environment	70	in autumn (64), fireworks (48), ...
	Weather	28	drizzle (58), hail (40), ...
	Scene	586	grove (61), castle (45), ...
<b>Fore-ground</b>	Number	4	2 (11806), 3 (11095), 4 (9775), 5 (7324)
	Eyes	43	brown eyes (1462), crazy eyes (1210), ...
	Hair	51	long hair (1551), pony-tail (1079), ...
	Identity	97	doctor (1668), teacher (1493), ...
<b>Action</b>	Hand	26	stretch (102), hands up (171), ...
	Leg	23	crossed legs (115), leg lift (108), ...
	Basic	20	grovel (126), squat (185), ...
	Compound	48	hug (172), princess carry (146), ...

### 3.4 Instance-level Text Description Annotation in the proposed SMPMat Dataset

Instance-level text description annotation is available in the proposed SMPMat dataset, which provides textual descriptions of each human instance and thus can be used for referring image matting. We use an automatic generation pipeline to

---

**Algorithm 1** Algorithm of Prompt generation

---

**Output:** positive prompt  $P_p$ , negative prompt  $P_n$

- 1: background = Sampler({Environment, Weather, Scene}, 1)
- 2: foreground\_num = random\_int(2,5)
- 3: **if** random\_float(0,1)>0.5 **then**
- 4:   identities = Sampler(Identity, Consistent\_flag=True, foreground\_num)
- 5: **else**
- 6:   identities = Sampler(Identity, Consistent\_flag=False, foreground\_num)
- 7: **end if**
- 8: action = Sampler({Hand Action, Leg Action, Compound Action}, 1)
- 9: face = Sampler({Eyes, Hair}, 1)
- 10:  $P_{pp}$  = constant\_positive
- 11:  $P_p$  = Concat( $P_{pp}$ , background, identities, action, face, ", ")
- 12:  $P_n$  = constant\_negative
- 13: **return** Outputs

---

generate those annotations. Specifically, we follow the expression generation engine in [2] to generate text annotations for SMPMat. We generate three types of text annotations for each foreground: basic expression, absolute position expression, and relative position expression. The basic expression describes the target foreground solely based on its attributes. In the absolute position expression, the target foreground is characterized by both its attributes and its absolute position within the image. The relative position expression details the target foreground by considering its attributes and establishing its position in relation to another foreground. Each type of expression has 2 fixed formats:

**Basic expression:**

the/a  $\langle att_0 \rangle \langle att_1 \rangle \dots \langle ide_0 \rangle$   
 the/a  $\langle ide_0 \rangle$  who/that is  $\langle att_0 \rangle \langle att_1 \rangle$ , and  $\langle att_2 \rangle$ .

**Absolute position expression:**

the/a  $\langle att_0 \rangle \langle att_1 \rangle \dots \langle ide_0 \rangle \langle rel_0 \rangle$  the photo/image/picture  
 the/a  $\langle ide_0 \rangle$  who/that is  $\langle att_0 \rangle \langle att_1 \rangle \langle rel_0 \rangle$  the photo/image/picture.

**Relative position expression:**

the/a  $\langle att_0 \rangle \langle att_1 \rangle \dots \langle ide_0 \rangle \langle rel_0 \rangle$  the/a  $\langle att_2 \rangle \langle att_3 \rangle \dots \langle ide_1 \rangle$   
 the/a  $\langle ide_0 \rangle$  who/that is  $\langle att_0 \rangle \langle att_1 \rangle \langle rel_0 \rangle$  the/a  $\langle ide_1 \rangle$  who/that is  $\langle att_2 \rangle \langle att_3 \rangle$ .

where  $\langle att_i \rangle$  represents attributes (e.g., color and clothing type predicted by off-the-shelf model),  $\langle ide_0 \rangle$  denotes identity (referencing the prompt during image synthesis).  $\langle rel_i \rangle$  delineates the relationship between the target foreground and the associated foreground  $\langle ide_i \rangle$ . This relationship is determined by comparing the positions of the geometric centers of the foregrounds, with four categories: *left*, *right*, *in front of*, and *behind*. In the annotation generation stage, we randomly select one format for each type of description.

## 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. Li, J., Zhang, J., Tao, D.: Referring image matting. In: CVPR. pp. 22448–22457 (2023)
  3. 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)
  4. 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)