Referring Image Matting
(Supplementary Material)

Jizhizi Li, Jing Zhang, and Dacheng Tao
The University of Sydney, Sydney, Australia
jili8515@uni.sydney.edu.au, jing.zhang1@sydney.edu.au, dacheng.tao@gmail.com

1. The Significance of RIM

(a) The original image

(b) highlight the **cat with white and black fur** in the image

(c) Top: Make the **beautiful cat** standing on the table Bottom: grab the **animal** out and paste on a green background

(d) **The man with a camera taking photo** of the cityscape

Figure 1. A case study of utilizing RIM for interactive image editing. The text inputs guide the interactive editing while the bold font indicates the given language descriptions.

Referring Image Matting (RIM) refers to extracting the soft and accurate foregrounds from the images based on given language descriptions. As a new task, we discuss the significance of RIM from the aspects of the industry impacts and academic impacts.

RIM, as a dense prediction problem that needs the collaboration of language and vision, provides a pathway to controllable image matting. Compared with conventional automatic image matting methods that can only predict a fixed set of pre-defined categories, e.g, people [6], animal [7], classes [8] or all the salient objects at one time [11, 17], RIM is able to predict the alpha matte of the specific foregrounds under language guidance, serving as a fundamental task for various downstream applications such as interactive image editing, human-machine interaction, virtual reality, augmented reality, film production, e-commerce promotion, etc. Compared with the other auxiliary user input-based image matting methods, e.g., trimap-based [16] and scribble-based [5], RIM provides much more freedom for the users as the inputs are simple and straightforward language descriptions.

As a novel task that has not been explored yet, RIM opens up lots of new research directions in the area of image matting. For example, how to align the visual and text features to better exploit both semantic and details information for such a dense level problem, how to bridge the domain gap between the typical composite training data and real-world testing data, how to distinguish between the same type objects, etc. Our proposed RefMatte, along with the baseline method CLIPMat, can serve as an initial starting point for such studies. Specifically, we discuss a case study and the differences with a related task Referring Image Segmentation (RIS) as follows to further emphasize the significance of RIM.

A case study of RIM. Here we provide a case study of RIM as a concrete example of utilizing RIM in downstream applications, e.g., interactive image editing. As can be seen from Figure 1, RIM is able to provide various interactive image editing results based on customized user text input, including highlighting any objects of interest and pasting the objects of interest to a reasonable background or a pure color. Different from the previous image matting methods that can only extract fixed-type foregrounds or all the salient foregrounds, RIM can easily separate the adherent foregrounds and focus on the one that best matches the language description. In this way, users can flexibly perform post-processing on either the *cat with white and black fur* or *the man with a camera taking photo* or any other customized language they like.

The differences with RIS. RIM is a very different task compared with RIS from the following several aspects:

1. RIS is only able to predict the coarse contour shape of the foreground from a low-resolution image, while RIM can predict the very fine details of the foregrounds from a high-resolution image.
2. RIS is a binary classification problem, while RIM is a pixel-level regression problem, requiring the datasets and methods to be very different from the two problems.
3. Datasets designed for RIS [4, 15] usually have low-resolution images and objects with very coarse shapes, while datasets for RIM should be all high-resolution images that preserve as many details as possible, as in our RefMatte.
4. Methods designed for RIS usually focus on extracting semantic features [3, 9] while RIM methods require features at both the (global) semantic level and (local) details level, making it more challenging, as in our proposed CLIPMat.

2. More Details about RefMatte

In this section, we provide more details about RefMatte, including the distribution of matting entities, linguistic details, and more statistics and visual results.

2.1. The Distribution of Matting Entities

Dataset	Category	Split	#Entities	#Categories	#Attrs. per Entity	#Entities in RefMatte train	#Entities in RefMatte test
AM-2k [7]	animal	train	1800	20	3	1800	-
		test	200	20		-	200
P3M-10k [6]	human	train	9186	1	6	9186	-
		test-1	485	1		-	485
		test-2	492	1		-	492
AIM-500 [8]	objects	test	200	93	3	95	105
SIM [13]	objects	train	271	82	3	271	-
		test	41	27		2	39
DIM [16]	objects	train	224	75	3	224	-
		test	38	27		7	31
HATT [11]	objects	train	210	58	3	210	-
		test	40	30		4	36
RefMatte (ours)	all-types	train	11799	230	3/6	11700	-
		test	1388	66		-	1388

Table 1. Statistics of the matting entities in our RefMatte which come from previous matting datasets.

Here, we present more details about the previous matting datasets that we used to retrieve the entities in our RefMatte, including the number of entities, categories, and attributes we generate per entity and the proportion of each dataset in our RefMatte train and test split. The results can be seen from Table 1. As can be seen, AM-2k [7] contributes all images of the animal categories, P3M-10k [6] contributes all images of the human category, and the others contribute the objects types like smog, plant, transparent glasses, etc. The number of attributes per entity is 3 for animals and objects, and 6 for humans. The proportion of each dataset in regard to the train and test split is shown in the table too. As can be seen, we try to reserve the original split in each matting dataset except for migrating the long-tailed categories to the RefMatte train set. However, the distribution of entities is unbalanced since most of them are human or animal. Thus, we duplicate some entities to form a balanced proportion of *human, animals, and objects* as *5:1:1*. The details are in the main paper.

2.2. Linguistic Details in RefMatte

In this subsection, we provide the linguistic details we used for constructing RefMatte, including the candidate words for the attributes and synonyms of the human type, the definitions of transparent and salient entities, syntax templates in the basic expressions, and the relationship words in the absolute/relative position expressions.

Candidate words for the human type To generate more precise synonyms, we define the basic synonyms for human type as `human being, citizenry, person, individual, mankind, mortal`. In addition to them, depending on the age, age group, and gender we defined, we provide more candidate words to serve as reasonable synonyms, which are shown in Table 2. All the candidate words are used to form the expression randomly.

Age	Age Group	Synonyms for Female	Synonyms for Male
0-2 4-6 8-12	child	`baby girl, little girl, girl`	`baby boy, little boy, boy`
15-20	youth	`girl, teenager, adolescent, miss, missy, young lady, young woman`	`boy, teenager, adolescent`
25-32 38-43 48-53	adult	`woman,lady`	`man`
60-100	senior	`old woman, senior citizen, pensioner`	`old man, senior citizen, pensioner`

Table 2. Synonyms of female and male at different age groups.

Definitions of transparent and salient entities Following Li et al. [8], we add the attributes `transparent` and `salient` for all the entities in RefMatte. We define the entities with category name synonyms including *smoke, glass, water, gauze, lace, ice, bubble wrap, plastic bag, net, fire, flame, cloth, mesh bag, mesh, wine glass, ice cube, spider web, wine, cloud smog, veil, wedding dress, fishing net, cloth net, light, water drop, drip, dew, crystal stone, beer* as `transparent` ones. Those entities with synonyms like *smoke, water, gauze, lace, fire, flame, net, leaves, spider web, mesh, wine, smog, light, water spray* as `non-salient` ones. For all the other entities, we can easily define them as `non-transparent` ones or `salient` ones, e.g., *human* and *animal* are both `salient` and `non-transparent`.

Syntax templates in the basic expressions We generate the basic expressions following the syntax templates as shown in Table 3, the templates are different for the human type and others since a human has six attributes and others have only three.

type	attributes	syntax template
human	$<att_0>$: `gender` $<att_1>$: `age` $<att_2>$: `non-transparent` $<att_3>$: `salient` $<att_4>$: `color` $<att_5>$: `clothes type`	`the` $<att_{0-3}>$ $<obj_0>$ `with the` $<att_{4-5}>$ `the` $<att_{0-3}>$ $<obj_0>$ `wearing the` $<att_{4-5}>$ `the` $<att_{0-3}>$ $<obj_0>$ `in the` $<att_{4-5}>$ `the` $<att_{0-3}>$ $<obj_0>$ `who is dressed in` $<att_{4-5}>$
others	$<att_0>$: `color` $<att_1>$: `non-/ transparent` $<att_2>$: `non-/ salient`	`the` $<att_{0-2}>$ $<obj_0>$ `the` $<obj_0>$ `which is` $<att_{0-2}>$

Table 3. Syntax template in the basic expression.

Relationship words in the absolute/relative position expressions As discussed in the paper, the syntax templates for the absolute position expressions are `the/a` $<att_0>$ $<att_1>$`...`$<obj_0>$ $<rel_0>$ `the photo/image/picture` or `the/a` $<obj_0>$ `which/that is` $<att_0>$ $<att_1>$ $<rel_0>$ `the photo/image/picture`. The syntax templates for the relative position expression are `the/a` $<att_0>$ $<att_1>$`...`$<obj_0>$ $<rel_0>$ `the/a` $<att_2>$ $<att_3>$`...`$<obj_1>$ or `the/a` $<obj_0>$ `which/that is` $<att_0>$ $<att_1>$ $<rel_0>$ `the/a` $<obj_1>$ `which/that is` $<att_2>$ $<att_3>$. Here we provide the candidate prepositional phrases for the relationship words $<rel_0>$ in Table 4 for each position relationship. Please note that the relationship `middle` is only used in the absolute position expressions.

Position Relationship	$< rel_0 >$ in absolute position expression	$< rel_0 >$ in relative position expression

left	`at the most left side of,`
	`on the far left of,`
	`at the leftmost edge of,`
	`farthest to the left of`
		`to the left of,`
		`on the left side of,`
		`at the left side of,beside,`
		`next to,close to,near`

right	`at the most right side of,`
	`on the far right of,`
	`at the rightmost edge of,`
	`farthest to the right of`
		`to the right of,`
		`on the right side of,`
		`at the right side of,beside,`
		`next to,close to,near`

middle	`in the middle of,` `n the center of`	-
top	`on top of,` `in the upper part of`	`above,over,on top of,on`
bottom	`below,in the lower part of`	`below,under,underneath`
in front of	`in front of`	`in front of`
behind	`behind,in the back of,` `at the back of`	`behind,in the back of,` `at the back of`

Table 4. Relationship words in the absolute/relative position expressions.

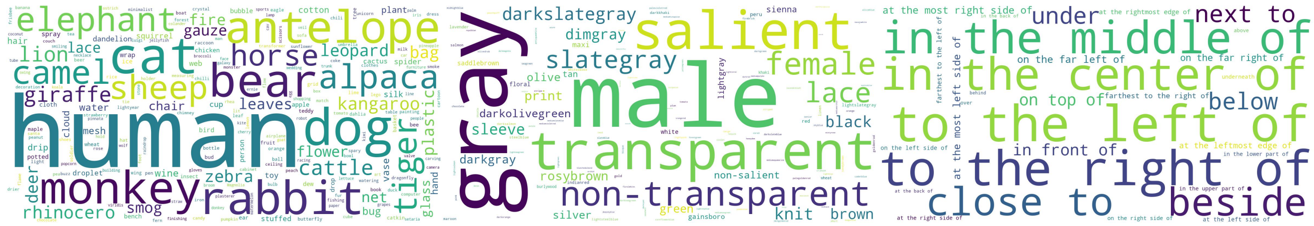
2.3. The Statistics of RefMatte

We present more details about the statistics of RefMatte in Table 5. For keyword-setting, since the text description is the entry-level category name, we remove the images with multiple entities belonging to the same category to avoid semantic ambiguity. Consequently, we have 30,391 images in the training set and 1,602 images in the test set in this setting. The numbers of alpha mattes, text descriptions, categories, attributes, and relationships are shown in the following columns, respectively. The average text length in the keyword-based setting is about 1, since there is usually a single word for each category, while it is much larger in the expression-based setting, i.e., about 17 in RefMatte and 12 in RefMatte-RW100.

Dataset	Split	Image Num.	Matte Num.	Text Num.	Category Num.	Attrs. Num.	Rels. Num.	Text Length
RefMatte Keyword	train	30,391	77,849	77,849	230	-	-	1.06
	test	1,602	4,085	4,085	66	-	-	1.04
RefMatte Expression	train	45,000	112,506	449,624	230	132	31	16.86
	test	2,500	6,243	24,972	66	102	31	16.80
RefMatte-RW100	test	100	221	884	29	135	34	12.01

Table 5. Statistics of RefMatte and RefMatte-RW100 regarding to the number of images, alpha mattes, text descriptions, categories, attributes, relationship words, and the average length of texts.

We also generate the word cloud of the keywords and attributes of the entities as well as the relationships between the entities in RefMatte in Figure 2. As can be seen, the dataset has a large portion of humans and animals since they are much more common in the image matting task. The most frequent attributes in RefMatte are *male, gray, transparent, and salient*, while the relationship words are more balanced, containing all kinds of relationships.



(a) (b) (c)

Figure 2. The word cloud of the keywords (a), attributes (b), and relationships (c) in RefMatte.

2.4. More Examples of RefMatte

We show more examples from our RefMatte training set and test set including their composition relations, keywords, basic expressions, absolute position expressions, and relative position expressions in Figure 4. We also show more examples from our RefMatte-RW100 test set, including their basic expression, absolute position expressions, relative position expressions, and free expressions in Figure 5. The green dots in both figures indicate the target objects.

3. More Details of CLIPMat

In this section, we present more details of our proposed baseline method CLIPMat, including more details about the three modules, *i.e.*, CP (context-embedded prompt), TSP (text-driven semantic pop-up), and MDE (multi-level details extractor). We also show the network structure of CLIPMat in Table 11.

Matting-related prefix templates We use a bag of words to serve as the matting-related prefix templates, aiming to reduce the gap between the long sentence used during pre-training CLIP [12] and the "single" word in the keyword-setting in RefMatte. Specifically, the templates in the bags of words are *"*$\{keyword\}$*"*, *"a photo of a* $\{keyword\}$*"*, *"a photograph of a* $\{keyword\}$*"*, *"an image of a* $\{keyword\}$*"*, *"a photo of the* $\{keyword\}$*"*, *"the foreground of the* $\{keyword\}$*"*, *"the mask of the* $\{keyword\}$*"*, *"the alpha matte of the* $\{keyword\}$*"*, *"to extract the* $\{keyword\}$*"*. The experiments have proved the effectiveness of using them in enhancing the ability of the pre-trained CLIP text encoder for the image matting task.

TSP details With the input of TSP as visual feature from CLIP image encoder $x_v \in \mathbb{R}^{(N+1)\times D_v}$ and text feature from CLIP text encoder $x_t \in \mathbb{R}^{L\times D_t}$, where $N = HW/P^2$ stands for the number of patches (tokens) in the ViT transformer [1], the additional one dimension denotes the class token which is not involved during feature reshaping, and L stands for the total length of the text and embedding context. We show the details of TSP as follows. First, both x_v and x_t pass through a layer norm, a linear layer, and another layer norm to align the feature dimension as D. Thus we have x'_v and x'_t with the same dimension D. We then pop up the semantic information from the visual feature by guiding it with the text feature through the cross-attention mechanism in transformer [14], thus we have $x_f \in \mathbb{R}^{(N+1)\times D}$. Furthermore, we adopt a self-attention mechanism to refine x_f and we adopt the residual connection to ease optimization. Finally, we pass x_f through a layer norm and a multilayer perception and then reshape it to $\mathbb{R}^{D'\times h\times w}$, where $h = \frac{H}{P}$ and $W = \frac{W}{P}$. This process can be formulated as follows:

$$x'_v = LN(Linear(LN(x_v))), \qquad x'_v \in \mathbb{R}^{(N+1)\times D}, \tag{1}$$

$$x'_t = LN(Linear(LN(x_t))), \qquad x_t \in \mathbb{R}^{L\times D}, \tag{2}$$

$$x_f = corss\_attn(x'_v, x'_t, x'_t), \qquad x_f \in \mathbb{R}^{(N+1)\times D}, \tag{3}$$

$$x_f = x_f + self\_attn(x_f, x_f, x_f), \qquad x_f \in \mathbb{R}^{(N+1)\times D}, \tag{4}$$

$$x_f = reshape(MLP(LN(x_f))), \qquad x_f \in \mathbb{R}^{D'\times h\times w}. \tag{5}$$

MDE details For MDE, the input feature is one of the four transformer blocks in the CLIP image encoder and the original image, denoted as x_v^i where $i \in \{1, 2, 3, 4\}$ and $X_m \in \mathbb{R}^{3\times H\times W}$, respectively. We show the details of MDE as follows. First, we reshape x_v^i and then normalize it by a 1×1 convolution layer, resulting in $x_v^i \in \mathbb{R}^{\frac{D_v}{2}\times\frac{H}{2^i}\times\frac{W}{2^i}}$. For X_m, we first normalize it by a 1×1 convolution layer and then down-sample it to the same size as x_v^i via max pooling, resulting in $x_m \in \mathbb{R}^{\frac{D_v}{2}\times\frac{H}{2^i}\times\frac{W}{2^i}}$. We then concatenate x_v^i and x_m to form $x_f^i \in \mathbb{R}^{D_v\times\frac{H}{2^i}\times\frac{W}{2^i}}$, which will be fed into a convolution layer, a batch norm layer, and a ReLU activation layer, results in the final output $x_f^i \in \mathbb{R}^{D_i\times\frac{H}{2^i}\times\frac{W}{2^i}}$. Finally, x_f^i is used as the input to the corresponding decoder layer at each level via a residual connection, which can preserve the details. This process can be formulated as follows:

$$x_v^i = norm(reshape(x_v^i)), \qquad x_v^i \in \mathbb{R}^{\frac{D_v}{2} \times \frac{H}{2^i} \times \frac{W}{2^i}}, \tag{6}$$

$$x_m = maxpool(norm(x_m)), \qquad x_m \in \mathbb{R}^{\frac{D_v}{2} \times \frac{H}{2^i} \times \frac{W}{2^i}}, \tag{7}$$

$$x_f^i = concat(x_v^i, x_m), \qquad x_f^i \in \mathbb{R}^{D_v \times \frac{H}{2^i} \times \frac{W}{2^i}}, \tag{8}$$

$$x_f^i = relu(bn(conv(x_f^i))), \qquad x_f^i \in \mathbb{R}^{D_i \times \frac{H}{2^i} \times \frac{W}{2^i}}. \tag{9}$$

4. More Details of Experiments

4.1. More Details of Experiment Settings

To customize the RIS methods [3, 9] for the newly proposed RIM task, we made slight changes to the existing methods for a fair comparison. For CLIPSeg [9], we choose CLIP [12] pre-trained **ViT-B/16** [1] as the image encoder and set the projection dimension of the decoder as 64 (D=64). We add one sigmoid layer on the output to normalize it to standard matting output. For MDETR [3], we choose ResNet-101 [2] as the image encoder, use the mask head as **smallconv**, and choose the prediction mask with the highest probability as the final output. For our CLIPMat, the channel numbers in matting semantic and details decoders are 768, 384, 192, and 96, respectively. We choose the CLIP [12] pre-trained **ViT-B/16** and **ViT-L/14** as the image encoder, respectively. For **ViT-L/14**, we change the kernel size and stride of the patch embedding layer from 14 to 16 for a fair comparison.

Both the CLIPSeg and MDETR use the weights that are further fine-tuned on VGPhraseCut [15]. However, CLIPMat only uses the CLIP pre-trained weights directly without further fine-tuning on VGPhraseCut [15] as we find it is unnecessary, which has also validated the value of our proposed RefMatte. For the parameters of position embedding that have a different shape from the pre-trained one, we reshape them by interpolation. The input size for all the methods is 512×512, and we choose the largest batch size for each model, which is 32 for CLIPSeg, 8 for MDETR, 12 for CLIPMat(ViT/B-16), and 4 for CLIPMat(ViT/L-14). All the learning rates are fixed as $1e-4$. For CLIPSeg and MDETR, the image and text encoders are all frozen to follow the original design in their papers. For CLIPMat with **ViT/B-16**, the learning rates of the image encoder and the text encoder are $1e-6$ and $1e-7$, respectively. For CLIPMat with **ViT/L-14**, the learning rates of the image encoder and the text encoder are all $1e-6$. All the methods are trained 50 epochs on two NVIDIA A100 GPUs, which takes about 1 day for CLIPMat(ViT/B-16) and 3 days for CLIPMat(ViT/L-14).

As for the optional matting refiner, we adopt the state-of-the-art automatic image matting model P3M from the work [6, 10], modifying the input from a single image to the image with a coarse map. We train the refiner on RefMatte with the settings following the original paper to serve as an optional post-refiner in our case. We present both the visual results of CLIPMat with or without the refiner in the following section, showing that CLIPMat already performs very well without the matting refiner, although better results can be achieved with its help.

4.2. Further Evaluation of the Main Results

To provide a comprehensive evaluation of the results, besides the conventional evaluation metrics SAD (sum of absolute differences), MSE (mean squared error), and MAD (mean absolute difference), we also calculate the average SAD, MSE, and MAD for all the entities in each image and average them over the test set, denoted as **SAD(E)**, **MSE(E)**, and **MAD(E)**, respectively. Please note that *SAD, MSE, MAD* are calculated on the basis of the foreground entities, which indicate the average error as per all foregrounds. However, *SAD(E), MSE(E), MAD(E)* are calculated on the basis of the images, indicating the average error as per all images. Take the *SAD* and *SAD(E)* as examples, we show the details of them in Eq. (10) and Eq. (11), where N stands for the number of images and M stands for the number of entities in one image (*e.g.*, $M = 2$ if the image contains a human and a dog). G stands for the ground truth label and P stands for the prediction. We simplify the SAD of each image as $|G - P|$. Thus, SAD(E), MSE(E), and MAD(E) reflect the models' ability to distinguish ambiguous foregrounds in the same image, serving as a more strict evaluation metric.

$$SAD = \frac{1}{(N \times M)} \times \Sigma_1^n (\Sigma_1^m |G - P|), \tag{10}$$

$$SAD(E) = \frac{1}{N} \times \Sigma_1^n (\frac{1}{M} \Sigma_1^m |G - P|). \tag{11}$$

Here, we provide all the results of MDETR [3], CLIPSeg [9], and our proposed CLIPMat with or without the post-matting refiner in Table 6. As can be seen, CLIPMat achieves the best results with both two backbones, where the larger backbone and the post-matting refiner improve the performance further. It is also noteworthy that the post matting refiner improves the results of MDETR and CLIPSeg by large margins, *i.e.*, 32.27 to 27.33 for MDETR in keyword-based setting, 17.75 to 12.17 for CLIPSeg, but only improves a little bit for CLIPMat, *i.e.*, 9.91 to 9.13 or 8.51 to 8.29. It owes to the excellent ability of CLIPMat to preserve details. Besides, for almost all the methods, SAD(E), MSE(E), and MAD(E) are larger than SAD, MSE, and MAD since they evaluate the ability of matting models to distinguish individual foreground in the same image, which is more challenging. However, CLIPMat's results on RefMatte-RW100 are even better in terms of SAD(E), MSE(E), and MAD(E), showing that CLIPMat has a good ability to extracting the correct targets. Furthermore, we provide more visual results to subjectively compare MDETR, CLIPSeg, and our proposed CLIPMat on the RefMatte test set and RefMatte-RW100 in both keyword and expression settings. The results are shown in the Figure 6 and Figure 7. As can be seen, CLIPMat performs very well in all the settings and outperforms all the other methods.

Setting	Method	Backbone	Refiner	SAD	MSE	MAD	SAD(E)	MSE(E)	MAD(E)
Keyword setting	MDETR [3]	ResNet-101 [2]	-	32.27	0.0137	0.0183	33.52	0.0141	0.0190
	MDETR [3]	ResNet-101 [2]	yes	27.33	0.0123	0.0155	28.22	0.0126	0.0160
	CLIPSeg [9]	ViT-B/16 [1]	-	17.75	0.0064	0.0101	18.69	0.0067	0.0106
	CLIPSeg [9]	ViT-B/16 [1]	yes	12.17	0.0042	0.0069	12.75	0.0044	0.0073
	CLIPMat	ViT-B/16	-	9.91	0.0028	0.0057	10.41	0.0029	0.0059
	CLIPMat	ViT-B/16	yes	9.13	0.0026	0.0052	9.56	0.0027	0.0055
	CLIPMat	ViT-L/14	-	8.51	0.0022	0.0049	8.98	0.0023	0.0051
	CLIPMat	ViT-L/14	yes	**8.29**	**0.0022**	**0.0027**	**8.72**	**0.0023**	**0.0050**
Expression setting	MDETR [3]	ResNet-101 [2]	-	84.70	0.0434	0.0482	90.45	0.0463	0.0515
	MDETR [3]	ResNet-101 [2]	yes	80.48	0.0424	0.0458	85.83	0.0452	0.0488
	CLIPSeg [9]	ViT-B/16 [1]	-	69.13	0.0358	0.0394	73.53	0.0381	0.0419
	CLIPSeg [9]	ViT-B/16 [1]	yes	64.48	0.0341	0.0367	68.56	0.0364	0.0391
	CLIPMat	ViT-B/16	-	47.97	0.0245	0.0273	50.84	0.0260	0.0273
	CLIPMat	ViT-B/16	yes	46.38	0.0239	0.0264	49.11	0.0253	0.0279
	CLIPMat	ViT-L/14	-	42.05	0.0212	0.0238	44.77	0.0226	0.0254
	CLIPMat	ViT-L/14	yes	**40.37**	**0.0205**	**0.0229**	**43.03**	**0.0218**	**0.0244**
RefMatte-RW100	MDETR [3]	ResNet-101 [2]	-	131.58	0.0675	0.0751	136.59	0.0700	0.0779
	MDETR [3]	ResNet-101 [2]	yes	125.78	0.0669	0.0717	130.72	0.0697	0.0744
	CLIPSeg [9]	ViT-B/16 [1]	-	211.86	0.1178	0.1222	222.37	0.1236	0.1282
	CLIPSeg [9]	ViT-B/16 [1]	yes	207.04	0.1166	0.1195	216.93	0.1222	0.1252
	CLIPMat	ViT-B/16	-	110.66	0.0614	0.0636	110.63	0.0612	0.0635
	CLIPMat	ViT-B/16	yes	107.81	0.0595	0.0620	107.23	0.0591	0.0616
	CLIPMat	ViT-L/14	-	88.52	0.0488	0.0510	87.92	0.0483	0.0505
	CLIPMat	ViT-L/14	yes	**85.83**	**0.0474**	**0.0495**	**84.93**	**0.0468**	**0.0488**

Table 6. Results on the RefMatte test set in two settings and the RefMatte-RW100 test set, a.k.a the complete version of Table 2 in the paper.

4.3. More Ablation Studies

4.3.1 Error Bars from Multiple Runs with Different Seeds

To calculate the error bars of our proposed method CLIPMat on RefMatte, we run the experiments on different random seeds and test them on both the keyword-based and expression-based settings of the RefMatte test set and RefMatte-RW100. We then report all the results in Table 7 by calculating the Standard Deviation (Std.) and mean value of the SAD. As can be seen, CLIPMat performs very stably on both the RefMatte test set and RefMatte-RW100.

4.3.2 Impact of Input Texts

Impact of prompt templates We trained CLIPMat with our proposed matting-related pre-embedding context to enhance the robustness of different input prompt templates. Here We investigate the impact of different prompt templates of CLIPMat(ViT-B/16) and show the results in Table 8. The default setting for our previous experiments in the main results is *a photo of*

Dataset	Setting	SAD	MSE	MAD	SAD-E	MSE-E	MAD-E
RefMatte test set	keyword	9.91	0.0028	0.0057	10.41	0.0029	0.0059
		9.82	0.0028	0.0056	10.40	0.0029	0.0059
		10.06	0.0030	0.0057	10.71	0.0031	0.0061
		Error bar of SAD	*Mean:*	*9.93*	*Std.:*	*0.1212*	
RefMatte test set	expression	47.97	0.0245	0.0273	50.84	0.0260	0.0273
		42.52	0.0215	0.0241	45.50	0.0231	0.0258
		50.35	0.0259	0.0287	53.09	0.0273	0.0303
		Error bar of SAD	*Mean*	*46.95*	*Std.:*	*4.0141*	
RefMatte -RW100	expression	110.66	0.0614	0.0636	110.63	0.0612	0.0635
		121.21	0.0676	0.0698	119.65	0.0667	0.0690
		117.79	0.0657	0.0677	120.46	0.0670	0.0691
		Error bar of SAD	*Mean:*	*116.55*	*Std.:*	*5.3826*	

Table 7. Results of CLIPMat with different random seeds on the RefMatte test set and RefMatte-RW100.

a $\{keyword\}$. As can be seen, CLIPMat is robust to all kinds of prompt templates while achieving the best results in the setting *an image of a* $\{keyword\}$, and also performs very well on matting-related prompt templates, *e.g.*, *the foreground of the* $\{keyword\}$, *to extract* $\{keyword\}$ and so on. We believe the success is owing to our careful design of matting-related pre-context embedded prompts. More efforts on prompt augmentation could be made in future work to further improve the performance.

prompt template	SAD	MSE	MAD	SAD(E)	MSE(E)	MAD(E)
$\{keyword\}$	9.95	0.0029	0.0057	10.45	0.0030	0.0060
a photo of a $\{keyword\}$	9.91	0.0028	0.0057	10.41	0.0029	0.0059
a photograph of a $\{keyword\}$	9.82	0.0028	0.0056	10.32	0.0029	0.0059
an image of a $\{keyword\}$	9.84	0.0028	0.0056	10.34	0.0029	0.0059
the foreground of the $\{keyword\}$	9.87	0.0028	0.0056	10.36	0.0029	0.0059
the mask of the $\{keyword\}$	9.90	0.0028	0.0057	10.39	0.0029	0.0059
the alpha matte of the $\{keyword\}$	10.01	0.0029	0.0057	10.50	0.0030	0.0060
to extract $\{keyword\}$	9.88	0.0028	0.0056	10.35	0.0029	0.0059

Table 8. Results of CLIPMat with different prompt templates on the RefMatte test set in the keyword-based setting.

Impact of different expressions Since we have introduced different types of expressions in our task, it is interesting to investigate the influence of each type on matting performance. As shown in Table 9, we evaluate CLIPMat(ViT-B/16) on the RefMatte test set and RefMatte-RW100. As can be seen, the relative position expression is more informative (or easy to understand) than others, leading to the best performance on both the synthetic test set as well as the real-world one. Among them, *RPE-1*, which is shorter and more straightforward compared with *RPE-2* achieves the best result in the expression-based setting. On the other hand, the absolute position expression has worse performance compared with relative ones in the expression-based setting but has comparable performance in RefMatte-RW100, probably due to manually labeled annotations being more straightforward and meaningful. Another interesting finding is that CLIPMat performs worse on basic expression than position-based expression, which is counter-intuitive. We believe the reason is that the model has been trained towards emphasizing the relationship between entities, resulting in a relatively poor ability to understand the entity and its own attributes. For the RefMatte-RW100 dataset, *FREE* prompts, which are labeled by human annotators following their preferred style, lead to the worst results, mainly due to the significant diversity of logic, grammar, and words used to describe the entities. More efforts could be made to study the most effective expressions that matter in automatic applications as well as improve the generalization ability of RIM models to deal with diverse expressions in human-machine interaction applications.

Setting	prompt template	SAD	MSE	MAD	SAD(E)	MSE(E)	MAD(E)
Expression-based setting	*BE*	59.20	0.0308	0.0337	62.81	0.0327	0.0357
	APE	59.41	0.0310	0.0339	62.81	0.0328	0.0358
	RPE-1	25.98	0.0120	0.0147	28.14	0.0131	0.0159
	RPE-2	46.29	0.0236	0.0264	48.33	0.0246	0.0275
RefMatte-RW100 dataset	*BE*	152.11	0.0857	0.0880	156.44	0.0881	0.0905
	APE	86.01	0.0469	0.0490	84.49	0.0458	0.0480
	RPE	85.08	0.0458	0.0479	83.59	0.0446	0.0469
	FREE	157.34	0.0883	0.0905	161.29	0.0905	0.0928

Table 9. Expression-based RIM results on RefMatte and RefMatte-RW100. BE: basic expression. APE: absolute position expression. RPE: relative position expression. FE: free expression.

4.3.3 The Impact of Pre-training and Freezing

We further investigate the impact of pre-training models and freezing the parameters by conducting the ablation studies on the keyword-based setting of RefMatte with CLIPMat(ViT-B/16) and presenting the results in Table 10. As can be seen, freezing the parameters of CLIP [12] pre-trained image encoder and text encoder results in SAD of 20.05, while fine-tuning the parameters as our proposed method results in SAD of 9.91. We hypothesize the benefits come from reducing the visual and text gap between the RefMatte and the CLIP dataset through fine-tuning. We set the learning rate of the image encoder as 0.01 times of matting decoders, and the learning rate of the text encoder as 0.001 times of matting decoders. On the other hand, pre-training CLIPMat on VGPhraseCut [15] does not provide performance improvement for CLIPMat, *i.e.*, SAD of 11.33 v.s. SAD of 9.91. Such results have also confirmed the value of RefMatte since training on it directly can achieve even better results, implying that the proposed CLIPMat can serve as a simple and strong baseline for referring image matting.

CLIP-Pretrain	CLIP-Freeze	Phrasecut-Pretrain	SAD	MSE	MAD	SAD(E)	MSE(E)	MAD(E)
✓	✓		20.05	0.0084	0.0115	21.39	0.0089	0.0122
		✓	11.33	0.0036	0.0065	11.86	0.0037	0.0068
✓			9.91	0.0028	0.0057	10.41	0.0029	0.0059

Table 10. Ablation studies of freezing the pre-trained CLIP [12] parameters and using VGPhraseCut [15] pre-trained weights.

4.4. Failure Cases

Figure 3. Some failure cases of CLIPMat, which is trained on RefMatte and tested on RefMatte-RW100. (a) Incorrect foreground instance. (b) Incomplete foreground details.

Although our proposed CLIPMat shows good performance on both the synthetic images and the real-world images after training on the RefMatte dataset, it still encounters some failure cases. We present some failure cases in Figure 3 (a) and (b). As shown in the figure, CLIPMat fails to locate the accurate foreground under some complex or ambiguous expression guidance. For example, as in (a), CLIPMat may not understand the word *feman* correctly and focus more on *looking at the camera*. In some other cases, like shown in (b) in the figure, CLIPMat fails to extract all the details of the foreground, which indicates the ability to preserve local details can be further improved. Such failure cases can be improved by 1) enhancing CLIPMat's abilities in understanding complex expressions and segmenting the foregrounds out with detailed boundaries, especially for those which have occlusions with other entities; 2) reducing the domain gap between synthetic and real-world images/expressions. We leave them as future work.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 5, 6, 7

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6, 7

[3] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2, 6, 7, 14, 15

[4] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 2

[5] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *TPAMI*, 30(2):228–242, 2007. 1

[6] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACM MM*, pages 3501–3509, 2021. 1, 2, 6, 17

[7] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *IJCV*, pages 1–21, 2022. 1, 2, 17

[8] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *IJCAI-21*, pages 800–806, 8 2021. 1, 2, 3, 17

[9] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, pages 7086–7096, June 2022. 2, 6, 7, 14, 15

[10] Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. Rethinking portrait matting with privacy preserving. *arXiv preprint arXiv:2203.16828*, 2022. 6

[11] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, 2020. 1, 2, 17

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6, 9

[13] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *CVPR*, pages 11120–11129, 2021. 2, 17

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 5

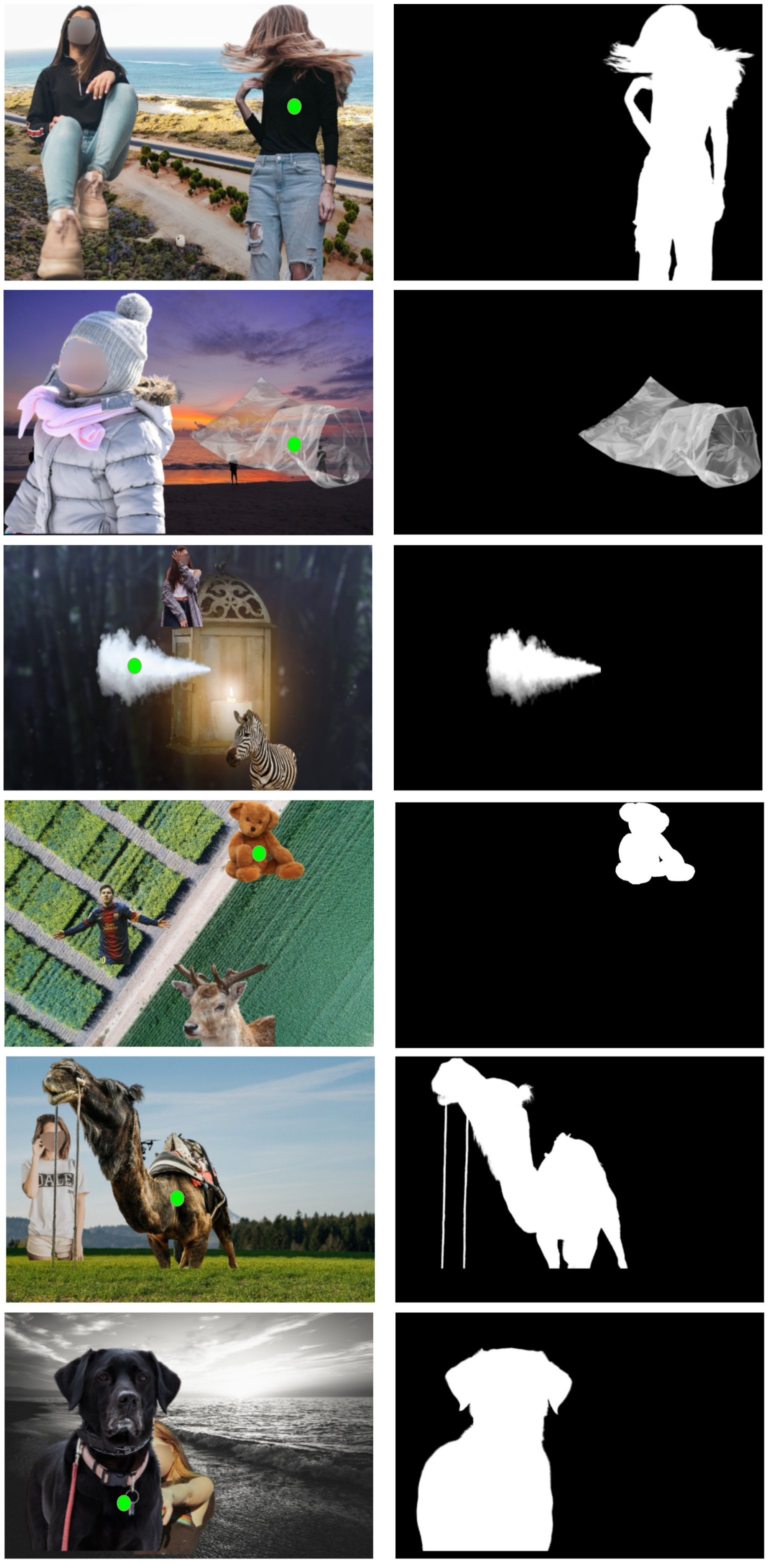
[15] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, pages 10216–10225, 2020. 2, 6, 9

[16] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, pages 2970–2979, 2017. 1, 2, 17

[17] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, pages 7469–7478, 2019. 1

Block name	Output size	Detail
CLIP - Text Encoder		
$embed$	$N \times 22(77) \times 512$	concatenate (text, context) + position embedding
T_1	$N \times 22(77) \times 512$	transformer block (heads 8, width 512)
CLIP - Image Encoder		
$conv1$	$N \times 768 \times 32 \times 32$	conv (3×3, 768, stride 16) + BN + ReLU
$embed$	$1025 \times N \times 768$	+ class embedding + position embeding
T_1	$1025 \times N \times 768$	transformer block (heads 12, width 768) $\times$ 3
T_2	$1025 \times N \times 768$	transformer block (heads 12, width 768) $\times$ 3
T_3	$1025 \times N \times 768$	transformer block (heads 12, width 768) $\times$ 3
T_4	$1025 \times N \times 768$	transformer block (heads 12, width 768) $\times$ 3
TSP		
$norm\_visual$	$N \times 1025 \times 256$	LN + Linear(256) + LN
$norm\_text$	$N \times 22(77) \times 256$	LN + Linear(256) + LN
$cross\_attn$	$N \times 1025 \times 256$	attention (256, heads=4, dropout=0.1)
$self\_attn$	$N \times 1025 \times 256$	attention (256, heads=4, dropout=0.1)
out	$N \times 64 \times 32 \times 32$	conv (1×1, 64, stride 1)
Matting Semantic Decoder		
D_2	$N \times 32 \times 128 \times 128$	[conv (3×3, 32, stride 3) + BN + ReLU] $\times$ 2 upsample(4)
D_1	$N \times 32 \times 512 \times 512$	[conv (3×3, 32, stride 3) + BN + ReLU] $\times$ 2 upsample(4)
D_0	$N \times 3 \times 512 \times 512$	conv (3×3, 3, stride 1)
MDE		
$norm\_visual$	$N \times 384 \times \frac{H}{2^i} \times \frac{H}{2^i}$	conv (1×1, 384, stride 1) upsample(2^{4-i})
$norm\_img$	$N \times 384 \times \frac{H}{2^i} \times \frac{H}{2^i}$	conv (1×1, 384, stride 1) maxpool (2^i)
out	$N \times D_i \times \frac{H}{2^i} \times \frac{H}{2^i}$	concatenate (visual, img) conv (3×3, D_i, stride 1) + BN + ReLU
Matting Details Decoder		
D_4	$N \times 384 \times 64 \times 64$	[conv (3×3, 384, stride 3) + BN + ReLU] $\times$ 2 upsample(2)
D_3	$N \times 192 \times 128 \times 128$	[conv (3×3, 192, stride 3) + BN + ReLU] $\times$ 2 upsample(2)
D_2	$N \times 96 \times 256 \times 256$	[conv (3×3, 96, stride 3) + BN + ReLU] $\times$ 2 upsample(2)
D_1	$N \times 32 \times 512 \times 512$	[conv (3×3, 32, stride 3) + BN + ReLU] $\times$ 2 upsample(2)
D_0	$N \times 1 \times 512 \times 512$	conv (3×3, 1, stride 1)
Collaborative Matting		
CM	$N \times 1 \times 512 \times 512$	pixel-wise multiply for output from two matting decoders output .

Table 11. Network structure of our proposed CLIPMat, where N stands for batch size. The input of CLIPMat is a batch of images of the size $N \times 3 \times 512 \times 512$, and a batch of text descriptions of size $N \times 14$ for keyword-setting and $N \times 69$ for expression-setting as well as the learnable context with size $N \times 8 \times 512$.



Composition relation: left/right
Keyword: human
Basic expression:
the female person who is dressed in black knit
Absolute position expression:
the non-transparent female lady with the black lace at the most right side of the picture
Relative position expression:
the salient female people with the black knit close to the female citizenry with the gray knit

Composition relation: left/right
Keyword: plastic bag
Basic expression:
the lightgray and salient plastic bag
Absolute position expression:
the plastic bag which is silver and salient at the rightmost edge of the image
Relative position expression:
the plastic bag which is lightgray and salient beside the non-transparent female individual with the thistle print

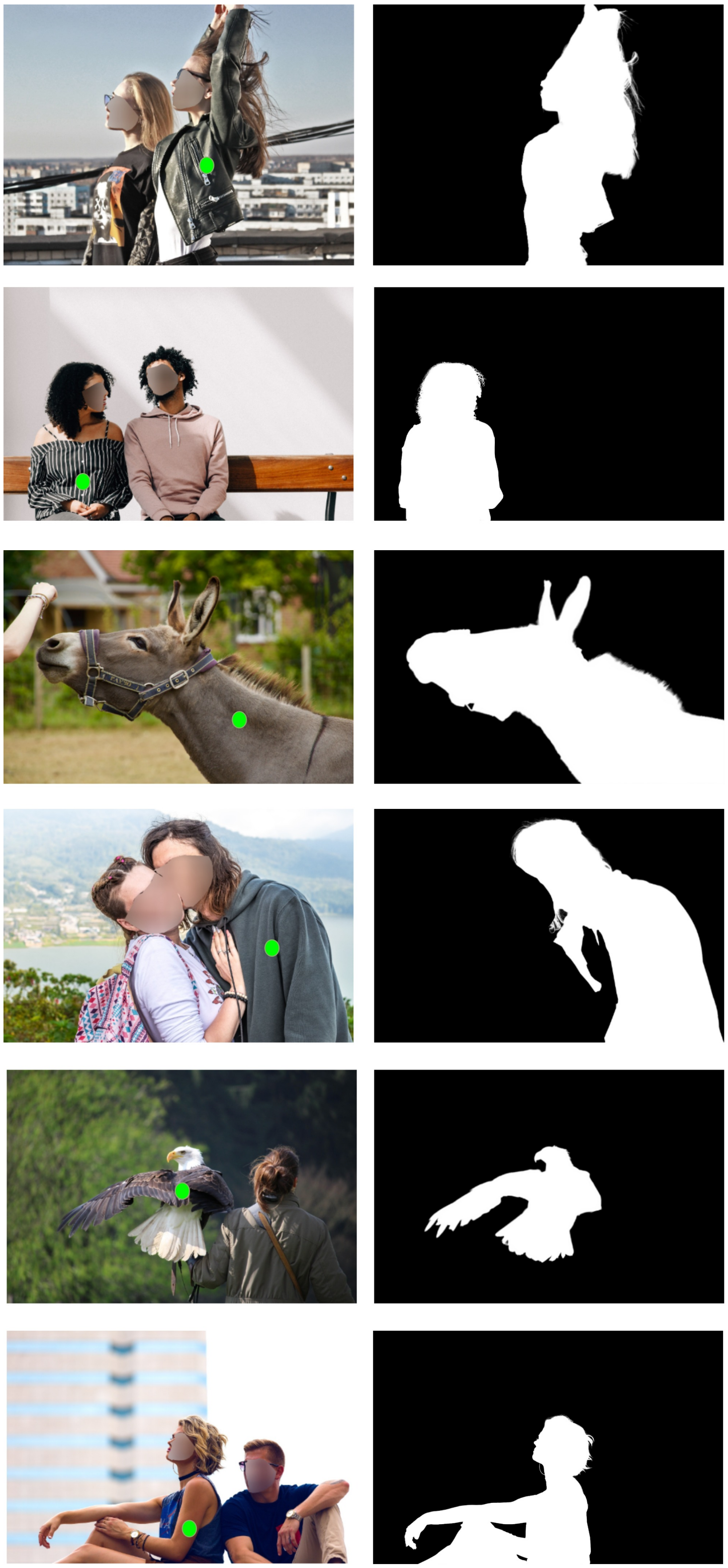
Composition relation: top/bottom
Keyword: smog
Basic expression:
the whitesmoke and non-salient smogginess
Absolute position expression:
the whitesmoke and transparent smogginess in the middle of the picture
Relative position expression:
the smoke which is gainsboro underneath the female people who is dressed in darkgray print

Composition relation: top/bottom
Keyword: teddy bear
Basic expression:
the teddy bear which is saddlebrown
Absolute position expression:
the teddy bear which is peru and non-transparent in the upper part of the image
Relative position expression:
the peru and salient and non-transparent teddy on the non-transparent male mortal with the darkslategray print

Composition relation: in front of/behind
Keyword: camel
Basic expression:
the animate being which is black and non-transparent
Absolute position expression:
the rosybrown and non-transparent creature in front of the photo
Relative position expression:
the black and salient animal in front of the non-transparent female individual with the white knit

Composition relation: in front of/behind
Keyword: dog
Basic expression:
the brute which is black
Absolute position expression:
the black and salient creature in front of the picture
Relative position expression:
the animal which is black and salient in front of the female people wearing the sienna lace

Figure 4. More examples from our RefMatte dataset. The first column shows the composite images with different foreground instances, and the second column and the third column show the ground truth alpha mattes and the natural language descriptions corresponding to the specific instances indicated by the green dots, respectively.



Basic expression:
the girl in a leather jacket wearing a pair of sunglasses
Absolute position expression:
the female human on the right side of the image
Relative position expression:
the beautiful girl on the right side of the short-hair girl
Free expression:
the girl who is enjoying the breeze blowing

Basic expression:
the woman with curly hair wearing a striped camisole
Absolute position expression:
the curly-haired female human-being on the left part of the picture
Relative position expression:
the female who is sitting to the left of the male
Free expression:
the lady smiles and looks at the man

Basic expression:
a gray donkey that has been equipped with a dark-blue bridle
Absolute position expression:
the donkey that dominates the image with its head and body centered at the image
Relative position expression:
the gook-looking animal on the right-hand side of the human arm
Free expression:
the peaceful donkey with beautiful eys and hairs, and trying to reach the human arm

Basic expression:
a long-haired man in a gray shirt
Absolute position expression:
a long-haired man in a gray shirt located at the right part of the picture
Relative position expression:
a long-haired man in a gray shirt standing to the left of the man
Free expression:
A long-haired man in a gray top hugging a woman in white with his face against the woman

Basic expression:
an eagle that has black feathers on the wings and white feathers on the body
Absolute position expression:
the handsome bird located onthe middle of the image
Relative position expression:
the beautiful eagle that is on the left part of the women
Free expression:
the handsome eagle that is looking back and about to spread its wings

Basic expression:
he woman with two big earrings
Absolute position expression:
the woman on the left half of the image
Relative position expression:
the woman on the left side of the man
Free expression:
the woman with long and yellow hair

Figure 5. More examples from our RefMatte-RW100. The first column shows real-world images with different foreground instances, and the second column and the third column show the ground truth alpha mattes and the natural language descriptions corresponding to the specific instances indicated by the green dots, respectively.

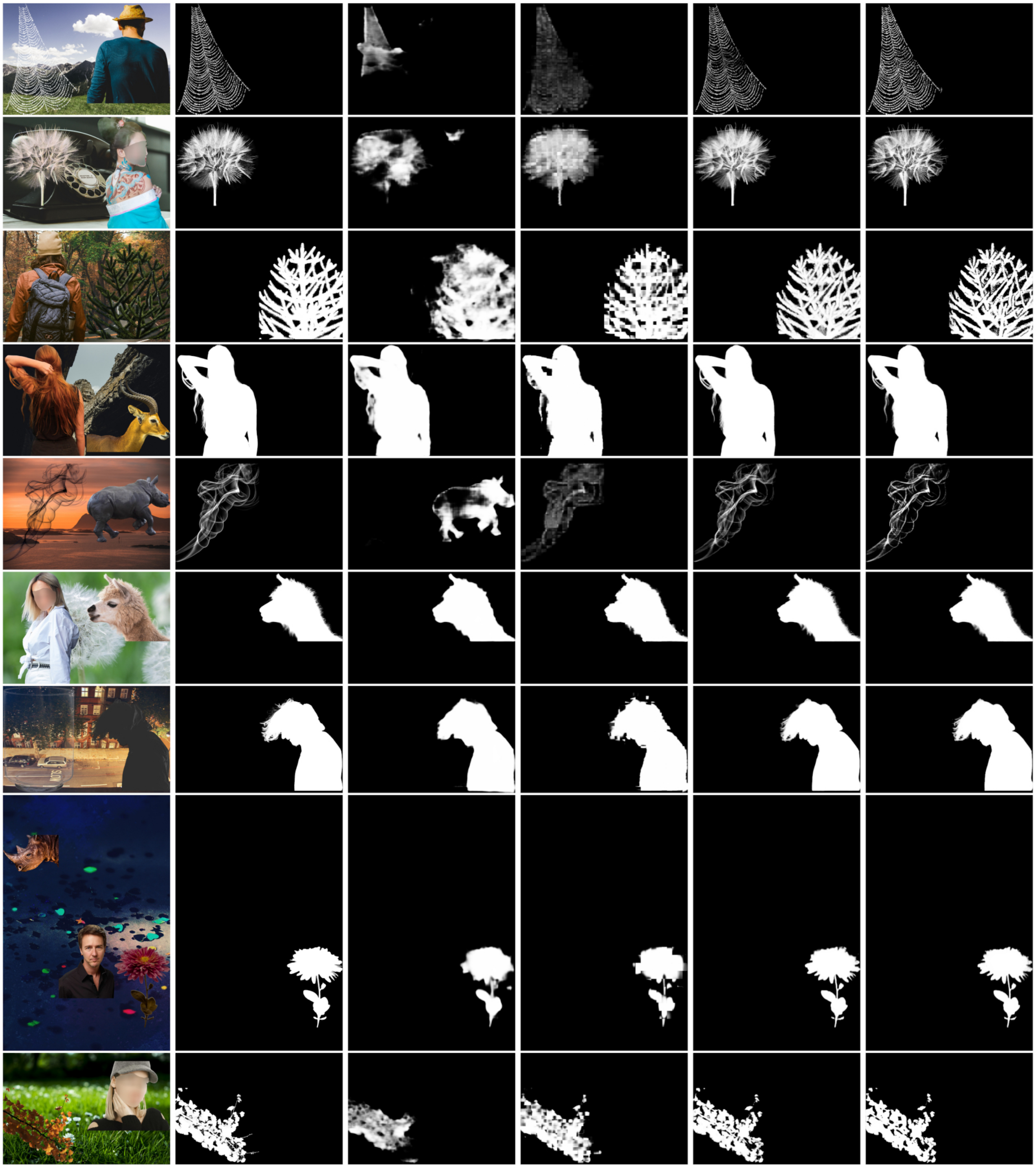


Figure 6. More subjective comparisons of different methods on the RefMatte under the keyword-based setting. From left to right: the original image, the ground truth, MDETR [3], CLIPSeg [9], our proposed CLIPMat, and CLIPMat with the matting refiner. The text inputs from the top to the bottom are: 1) net; 2) dandelion; 3) leaves; 4) human; 5) smog; 6) alpaca; 7) human; 8) flower; 9) leaves. We recommend zooming in for more details.

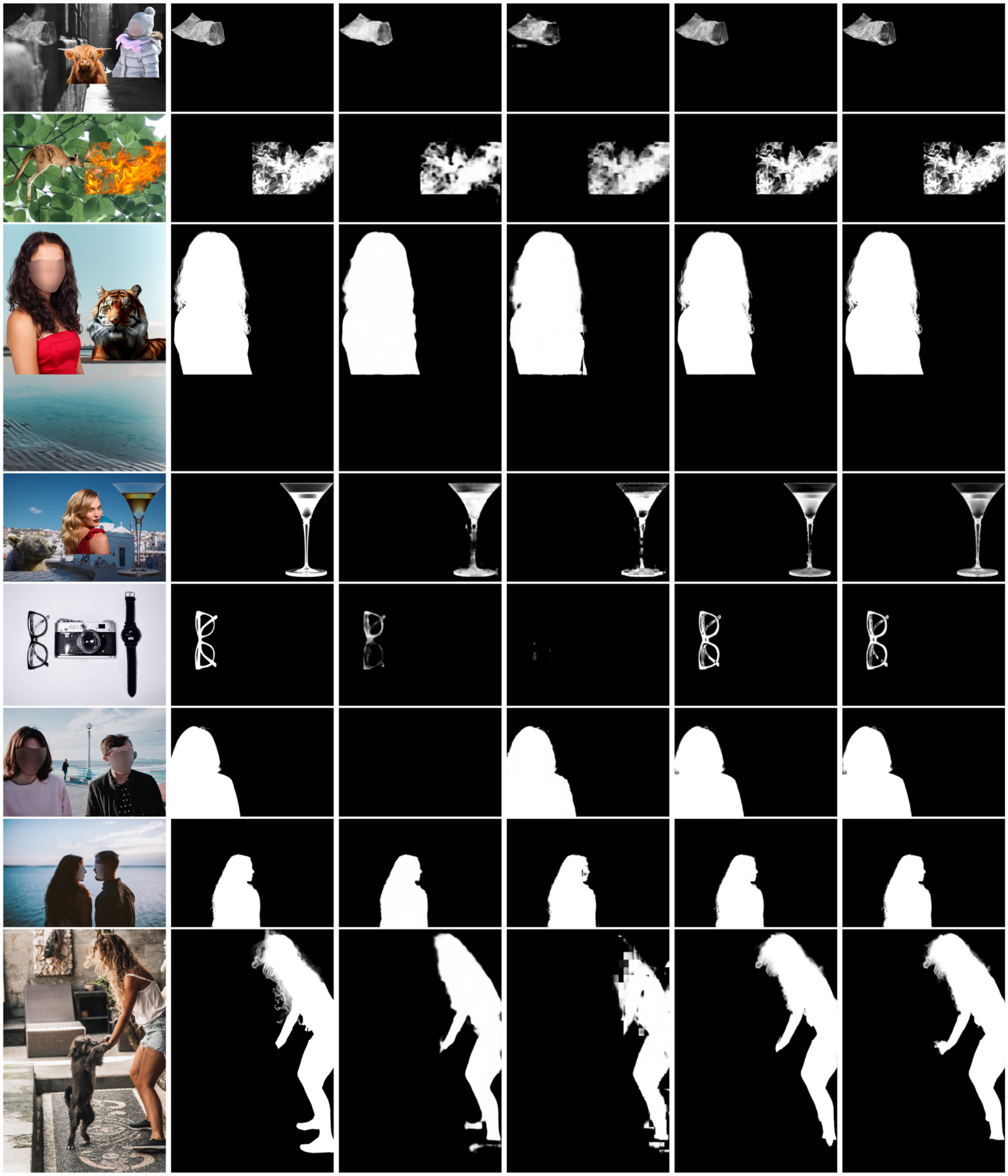


Figure 7. More subjective comparisons of different methods on the RefMatte under the expression-based setting and the RefMatte-RW100. From left to right: the original image, the ground truth, MDETR [3], CLIPSeg [9], our proposed CLIPMat, and CLIPMat with the matting refiner. The text inputs from the top to the bottom are: 1) the plastic bag which is lightgray and transparent; 2) the fire which is wheat at the most right side of the picture; 3) the woman in the crimson print on the left side of the rosybrown and salient brute; 4) the gray and salient vase at the most right side of the picture; 5) the good-looking glass on the left part of the photo; 6) a long-haired man in a pink shirt and sunglasses smiling; 7) a woman with long hair; 8) the lady in white clothes in the right of the picture. We recommend zooming in for more details.

A. Datasheet of RefMatte

A.1. Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A1: RefMatte is created to facilitate the study of a new task: referring image matting (RIM). RIM is first introduced in this paper as extracting the meticulous foreground in the image with linguistic keyword or expression as an auxiliary input. However, prevalent visual grounding methods are all limited to the segmentation level, probably due to the lack of high-quality datasets for RIM. To fill the gap, we establish the first large-scale challenging dataset **RefMatte** by designing a comprehensive image composition and expression generation engine to produce synthetic images on top of current public high-quality matting foregrounds with flexible logics and re-labelled diverse attributes.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: RefMatte is created by the authors.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A3: This study was supported by Australian Research Council Projects in part by FL170100117 and IH180100002.

A.2. Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances(e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A1: The RefMatte dataset consists of images covering 230 categories which are very popular in the field of image matting, and expressions describe each entity. Some typical types are humans, animals, plants, spiders web and so on. All the personally identifiable information has been preserved for privacy consent.

2. How many instances are there in total (of each type, if appropriate)?

A2: The RefMatte dataset contains 230 object categories, 47,500 images, 118,749 expression-region entities with high-quality alpha matte, and 474,996 expressions.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A3: The RefMatte itself is a large set that contains a large number of instances. It is large enough to be used for training deep models. However, the composition and expression engines we designed make it possible and easy to extend the dataset to a larger scale.

4. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

A4: Each instance consists of a high-resolution synthetic image generated by our composition engine, the high-quality alpha matte of the specific entity, the keyword and the expressions that used to describe the specific entity.

5. Is there a label or target associated with each instance? If so, please provide a description.

A5: Yes. Each instance is associated with a label, including an alpha matte, a keyword name, and several expressions. Some examples can be seen from the Figure 4 and Figure 5.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A6: No.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A7: Yes. We keep a JSON file to store the relationships between individual instances, i.e., information such as whether or not multiple instances are located on the same image.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A8: Yes. We split the dataset RefMatte into training and test sets manually. We keep all the long-tail categories in the training set only. More details about the data splits can be found in Section 3.3 of the paper.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: Although we have manually checked the annotation information very carefully, there may be some minor inaccurate expression labels. However, since linguistic expression generated by real human is also subjective and may contains some error. We believe this might be a source of noise to improve the generalization ability of trained models.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A10: The RefMatte dataset is comprised of the publicly available datasets, including AM-2k [7], P3M-10k [6], AIM-500 [8], SIM [13], DIM [16], and HATT [11]. These datasets are publicly available and can be downloaded from their websites. We appreciate the significant contribution of the authors to the research community. We show the details of each matting dataset in Section 2.1.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

A11: No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No.

A.3. Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: The data associated with each instance are generated through a semi-automatic style by combining the attribute information predicted by pretrained models and manual annotations. We report the details in Section 3.1 of the paper.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: The matting entities in the dataset RefMatte are collected from publicly available datasets described above, which can be directly downloaded from their websites. The final images are generated by our own proposed composition engine with the details described in Section 3.2 of the paper.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A3: No.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A4: The first author of this paper.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

A5: It took about 30 days to collect the data and about 2 months to complete organization and annotation.

A.4. Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A1: Yes. First, we collect and clean the data from currently available matting datasets to serve as matting entities, then we use our proposed composition and expression engine to generate the synthetic images with linguistic labels. The details can be seen in Section 3 of the paper.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

A2: N/A.

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A3: No. We process all the data with our code, which will be released.

A.5. Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A2: N/A.

3. What (other) tasks could the dataset be used for?

A3: The RefMatte dataset can be used for referring image matting studies. In addition, it can be used for machine learning topics like domain adaptation, referring image localization, and one-shot/zero-shot referring image matting.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A4: No.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

A5: No.

A.6. Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A1: Yes. The dataset will be made publicly available to the research community.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A2: It will be publicly available on the project website at https://github.com/JizhiziLi/RIM.

3. When will the dataset be distributed?

A3: The dataset will be distributed once the paper is accepted after peer-review.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A4: It will be distributed under the CC BY-NC license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A5: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A6: No.

A.7. Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

A1: The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: They can be contacted via email available on the project website.

3. Is there an erratum? If so, please provide a link or other access point.

A3: No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A4: No.

5. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

A5: N/A.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

A6: N/A.