THE UNIVERSITY OF SYDNEY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

# Referring Image Matting

*Jizhizi Li, Jing Zhang and Dacheng Tao*

THE UNIVERSITY OF SYDNEY, AUSTRALIA

**June, 2023**

Paper Tag:
THU-PM-176

# Summary

Different from conventional matting methods, which either requires user-defined input to extract a specific object or extracts all objects at once, **Referring Image Matting (RIM)**, is able to extract the alpha matte of the object that best matches the language description, enabling a more natural and simpler process.

We establish a large-scale dataset **RefMatte** by designing a generation engine to produce high-quality images with diverse text attributes, consisting of *47k images*, and *475k expressions*. We also construct 100 real-world images and manually annotated phrases to evaluate the out-of-domain generalization abilities.

Furthermore, we present a novel method **CLIPMat**, including a context-embedded prompt, a text-driven semantic pop-up, and a multi-level details extractor. Extensive experiments validated the superiority of CLIPMat. Please scan the QR code for code and datasets.

# Table of Contents


1. **Motivation**
2. **Dataset: RefMatte**
    1. Preparation of Matting Entities
    2. Image Composition and Expression Generation
    3. Dataset Split and Task Settings
3. **Method: CLIPMat**
    1. Overview
    2. CP: Context-embedded Prompt
    3. TSP: Text-driven Semantic Pop-up
    4. MDE: Multi-level Details Extractor
4. **Experiments**
    1. Objective Results
    2. Subjective Results
    3. Ablation Studies
5. **Conclusion**
6. **References**

# Motivation

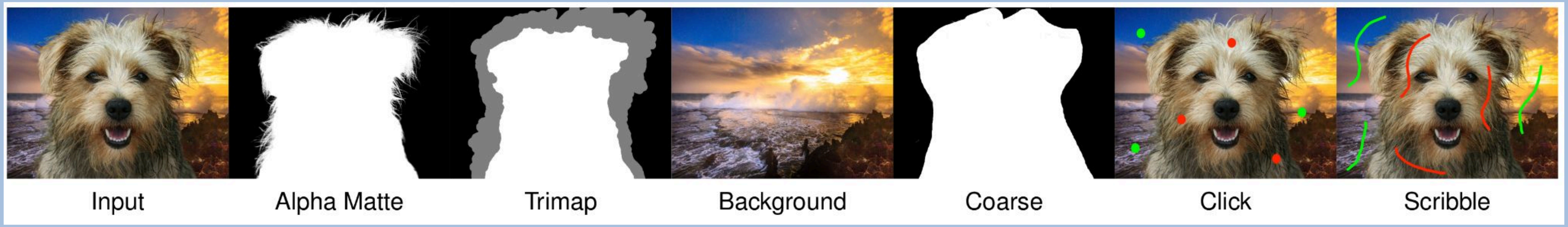



Image matting refers to extracting the foregrounds from arbitrary images with meticulous boundaries, i.e., removing the backgrounds.

Conventional image matting methods, either requires user-defined auxiliary input to extract a specific foreground object [1], or directly extracts all the foreground objects in the image discriminatively [2].

We aim to design a more controllable task by taking human's language as guidance, named as **Referring Image Matting (RIM)**.

# Motivation

(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

RIM can be used in a lot of downstream applications.

As show in the figure left, RIM is able to provide various interactive image editing results based on customised user text input, including highlighting any objects of interest and pasting the objects of interest to a reasonable background or a pure colour.

# Motivation

Here we show some results of our proposed method CLIPMat on our established dataset RefMatte. As can be seen, by defining a simple keyword like *"cattle"* or a complex expression like *"the man in a gold suit smiling sideways towards the camera"*, CLIPMat is able to predict the meticulous alpha matte of the specific object that best matches the language description, giving users large degree of freedom.

# Dataset: RefMatte

To provide support of RIM, we establish a large-scale challenging dataset RefMatte by designing a comprehensive image composition and expression generation engine to automatically produce high-quality images along with diverse text attributes based on public datasets.

First, We manually label each entity's category and annotate the attributes by leveraging off-the-shelf deep learning models. We show the details including the distribution of matting entities in the following table. We end up with 13,187 foreground entities and large amounts of background images from BG-20k [3].

| 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.

# Dataset: RefMatte

To present reasonably looking composite images with semantically clear, grammatically correct, as well as abundant and fancy expression, how to arrange the candidate entities and build up the language descriptions is the key to constructing RefMatte.

We then define six types of position relationships: *left, right, top, bottom, in front of* and *behind* and three types of expressions for each entity regarding different logic forms as shown in the right.

1. ***Basic expression*** This is the expression that describes the target entity with as many attributes as one can, e.g, `the/a` $<att_0>$ $<att_1>$`...`$<obj_0>$ or `the/a` $<obj_0>$ `which/that is` $<att_0>$ $<att_1>$`, and` $<att_2>$. For example, as shown in Figure 3(a), the basic expression for the entity flower is '`the lightpink and salient flower`';
2. ***Absolute position expression*** This is the expression that describes the target entity with many attributes and its absolute position in the image, *e.g.*, `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`. For example, as shown in Figure 3(a), the absolute position expression for the flower is '`the plant which is lightpink and salient at the rightmost edge of the picture`';
3. ***Relative position expression*** This is the expression that describes the target entity with many attributes and its relative position with another entity, *e.g.*, `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` $<att_2>$ $<att_3>$. For example, as shown in Figure 3(a), the relative position expression for the flower is '`the flower which is lightpink at the right side of the cat which is dimgray and non-transparent`'.

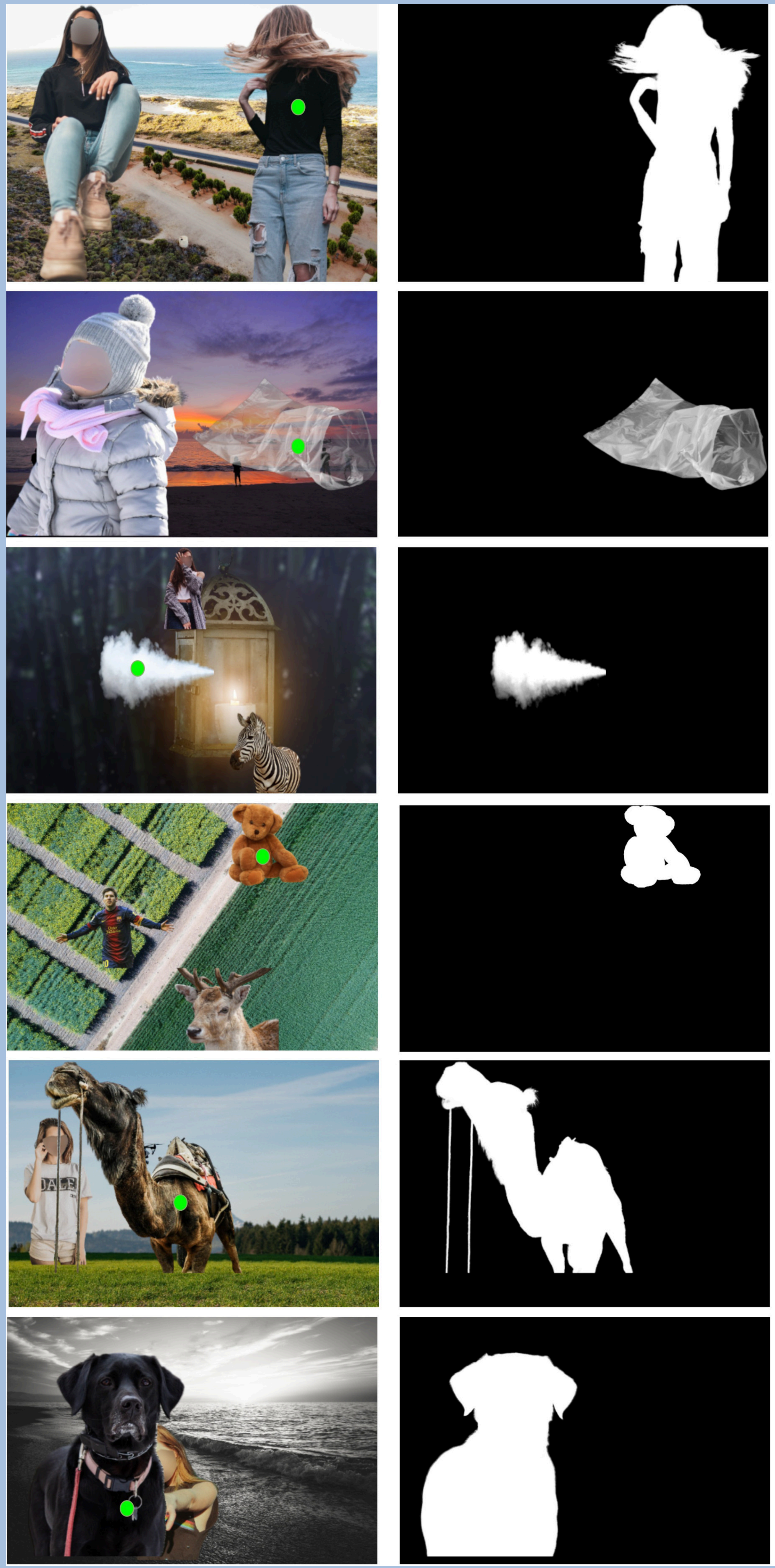
# Dataset: RefMatte

In total, we have 13,187 matting entities and split 11,799 for training and 1,388 for testing. We duplicate some samples to modify the proportion of humans, animals, and objects as 5:1:1. We then pick 5 humans, 1 animal and 1 object as one group to composite 20 images with various backgrounds. The final statistics is shown in the right.

We set up two settings as keyword and expression upon RefMatte to benchmark different language descriptions.

| Dataset | Image Num. | Matte Num. | Text Num. | Cate Num. | Text Length |
|---|---|---|---|---|---|
| RefMatte-keyword | 31,993 | 81,934 | 81,934 | 230 | 1.06 |
| RefMatte-Expression | 47,500 | 118,749 | 474,996 | 230 | 16.86 |
| RefMatte-RW100 | 100 | 221 | 884 | 29 | 12.02 |

# RefMatte



**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

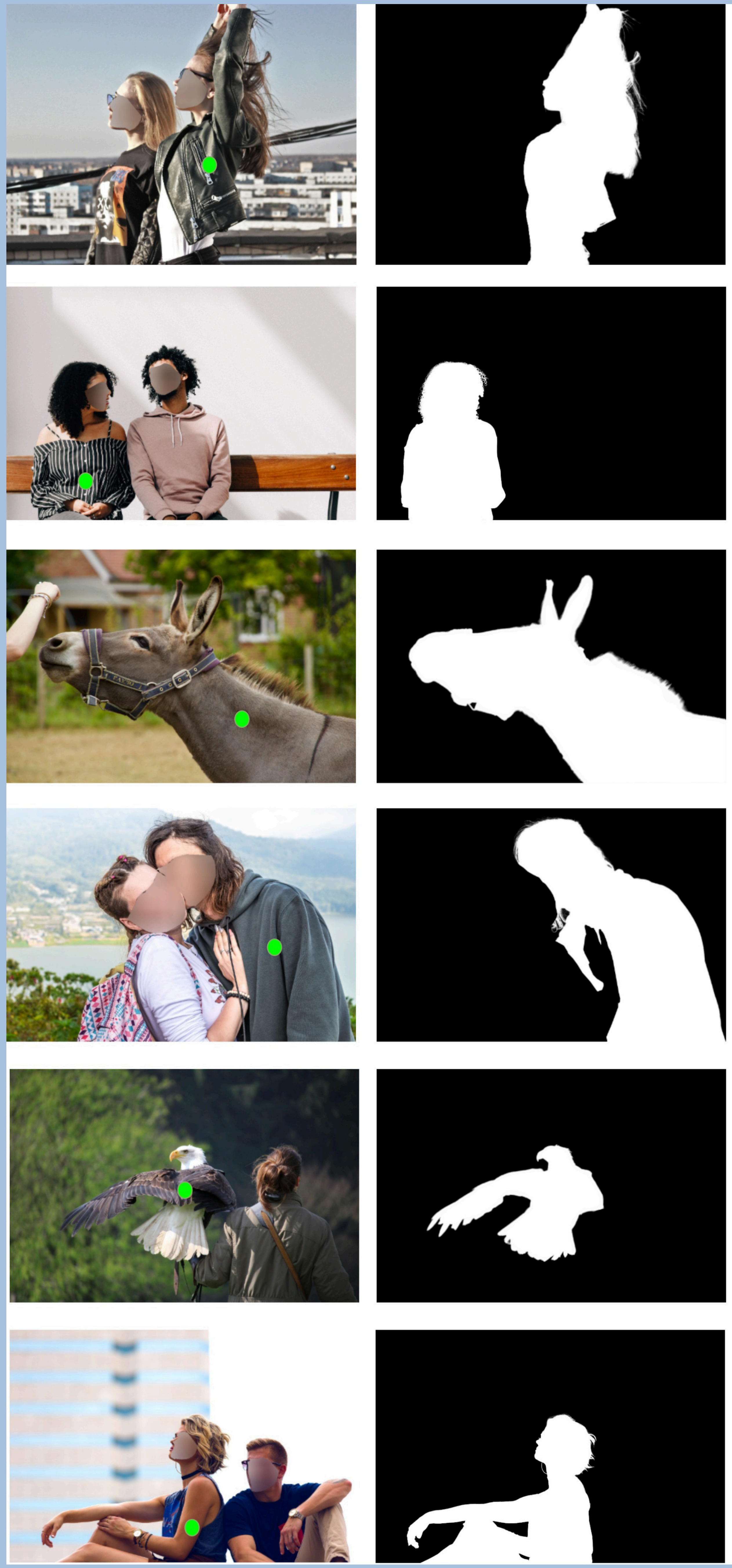
**Composition relation:** top/bottom
**Keyword:** smog
**Basic expression:**
the whitesmoke and non-salient smogginess
**Absolute position expression:**
the whitesmoke and transparent smogginess in the middle of the picture
**Relative position expression:**
the smoke which is gainsboro underneath the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

# RefMatte-RW100



**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

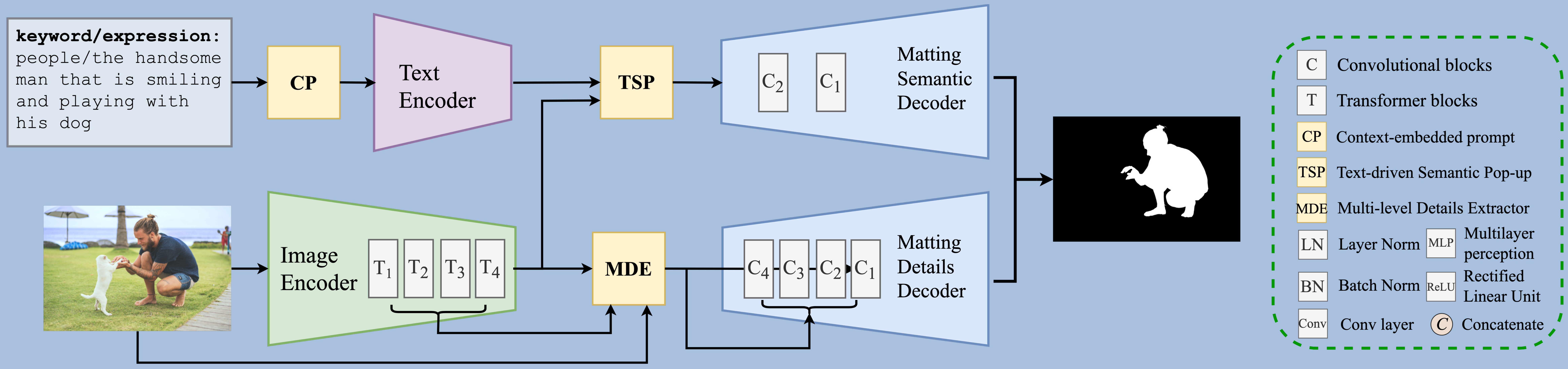
**Basic expression:**
an eagle that has black feathers on the wings and white feathers on the body
**Absolute position expression:**
the handsome bird located onthe middle of the image
**Relative position expression:**
the beautiful eagle that is on the left part of the women
**Free expression:**
the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

# Method: CLIPMat

## Overview

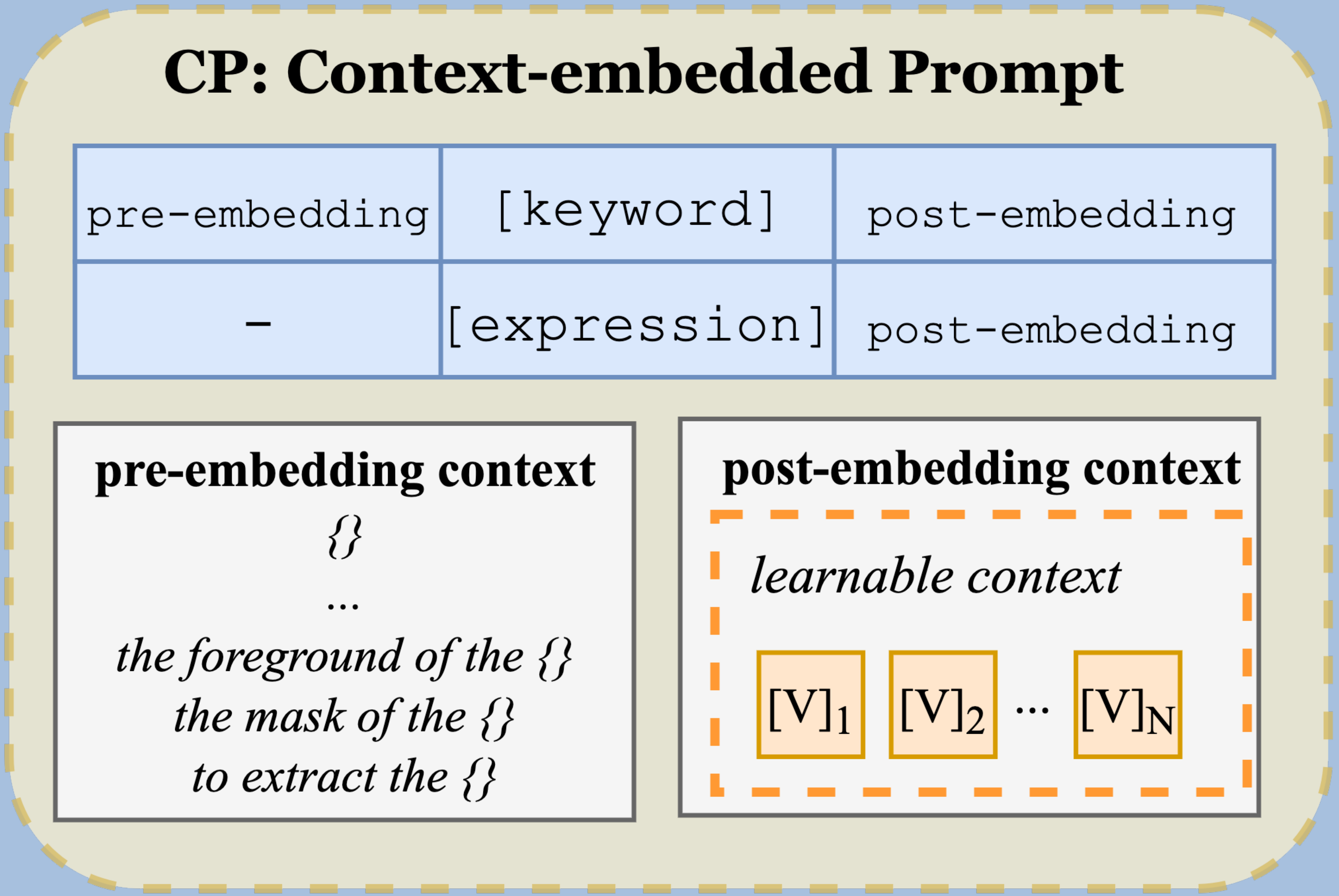
Motivated by the success of large-scale pre-trained vision language models like CLIP [4] and dual-decoder framework [3] from SOTA matting methods, we propose CLIPMat with ViT-B/16 or ViT-L/14 as image encoder, two decoders to predict trimap and the transition alpha respectively, before merging as the final alpha matte prediction. We show the overview framework of CLIPMat as follow.



# Method: CLIPMat

## CP: Context-embedded Prompt

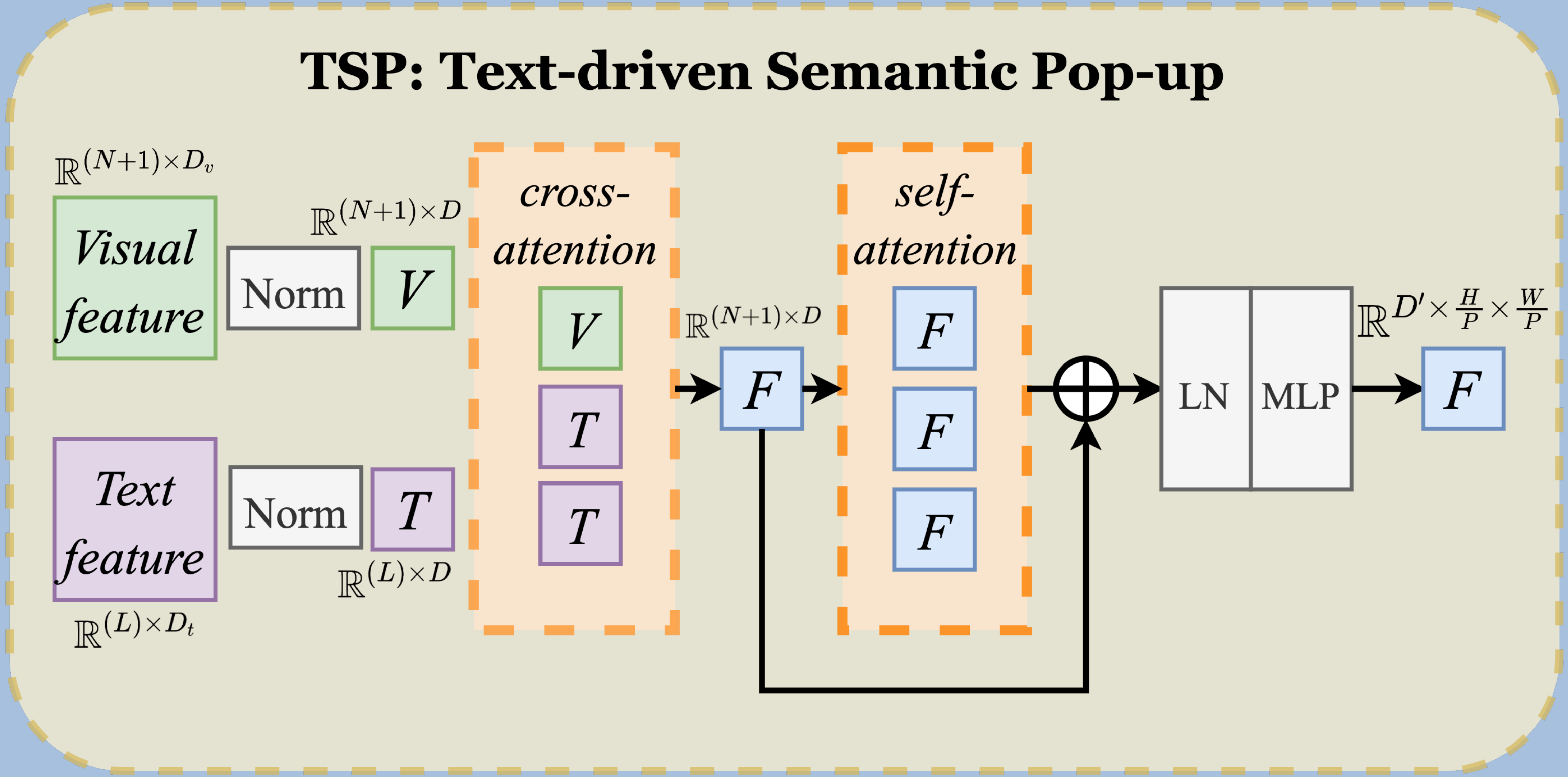
To enhance the understanding ability of the text input, we design two kinds of contexts to be embedding in the original prompt, i.e., pre-embedding context and post-embedding context. We Utilise customised matting context as prefix templates, e.g., the foreground of the {}, the mask of the {}, and to extract the {}. We use 14 and 69 for text length in keyword and expression setting, and 8 for the fixed length of learnable context.



# Method: CLIPMat

## TSP: Text-driven Semantic Pop-up

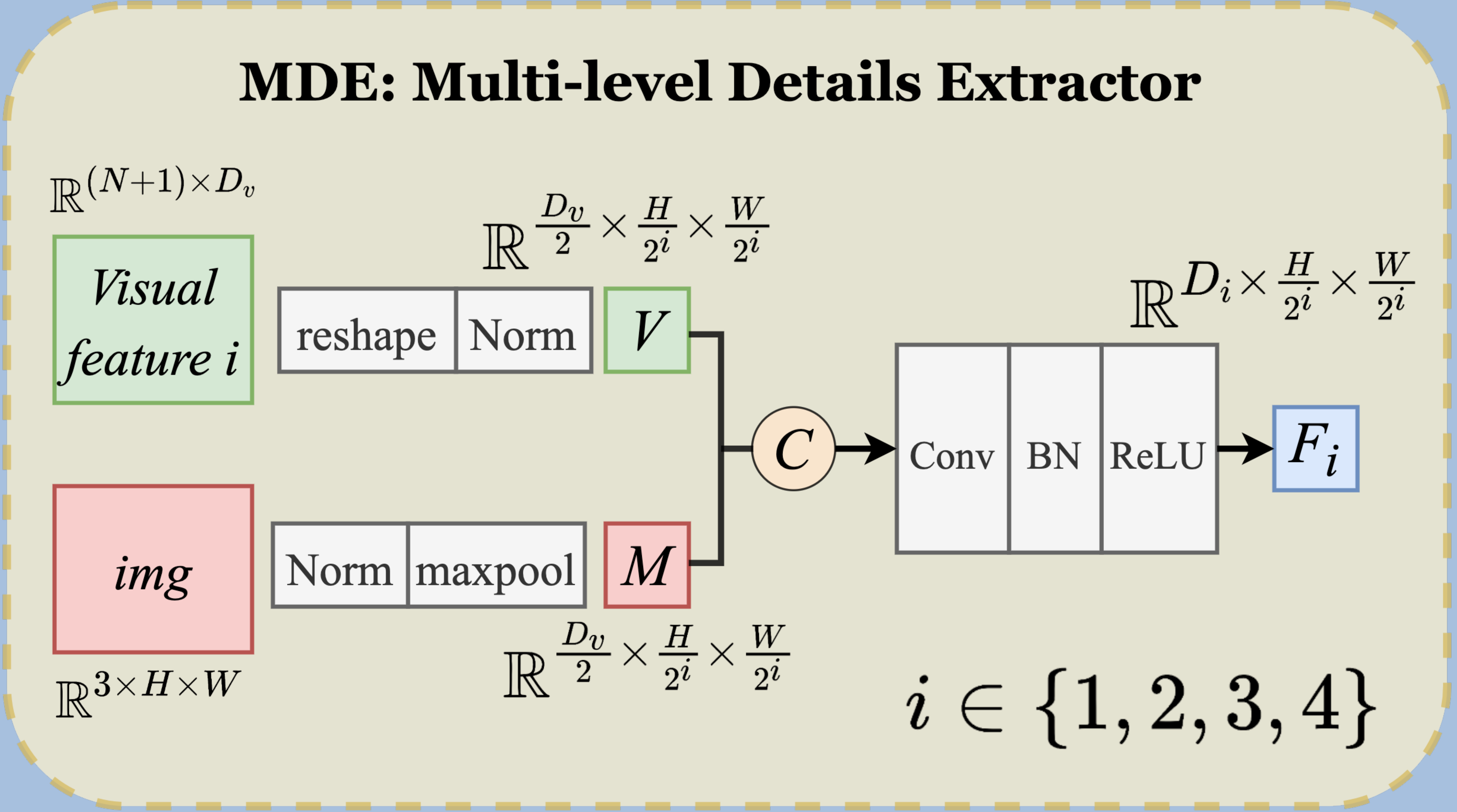
To ensure the text feature from the text encoder can provide better guidance on dense-level visual semantic perception, we propose TSP to process the text and visual features before the matting semantic decoder through linear projection, cross-attention, and self-attention.



# Method: CLIPMat

## MDE: Multi-level Details Extractor

To keep as much meticulous details in the final prediction, we propose MDE to extract useful local details from both the original image and multi-level features from the image encoder through reshaping, normalisation and a convolution layer.



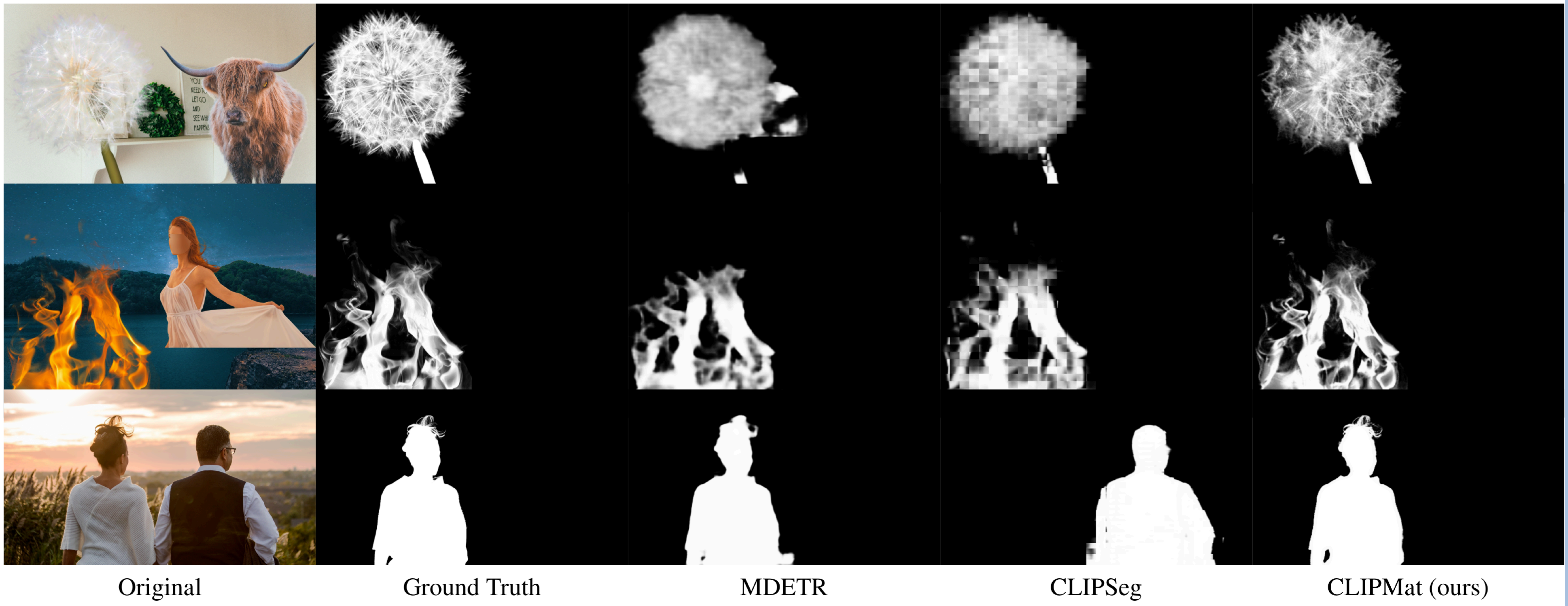
# Experiments

We evaluate MDETR [5], CLIPSeg [6] and CLIPMat on RefMatte keyword-setting, expression-setting, and RefMatte-RW100. We also utilise a coarse map-based matting method as an optional post-refiner [7] to further improve the results. The objective results are as follows. As can be seen, our proposed CLIPMat is able to outperform all SOTA methods from all metrics.

| Method | Backbone | Refiner | Keyword-setting | | | Expression-setting | | | RefMatte-RW100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SAD | MSE | MAD | SAD | MSE | MAD | SAD | MSE | MAD |
| MDETR[4] | ResNet-101 | - | 32.27 | 0.0137 | 0.0183 | 84.70 | 0.0434 | 0.0482 | 131.58 | 0.0675 | 0.0751 |
| CLIPSeg[5] | ViT-B/16 | - | 17.75 | 0.0064 | 0.0101 | 69.13 | 0.0358 | 0.0394 | 211.86 | 0.1178 | 0.1222 |
| CLIPMat | ViT-B/16 | - | 9.91 | 0.0028 | 0.0057 | 47.97 | 0.0245 | 0.0273 | 110.66 | 0.0614 | 0.0636 |
| CLIPMat | ViT-B/16 | Yes | 9.13 | 0.0026 | 0.0052 | 46.38 | 0.0239 | 0.0264 | 107.81 | 0.0595 | 0.0620 |
| CLIPMat | ViT-L/14 | - | 8.51 | 0.0022 | 0.0049 | 42.05 | 0.0212 | 0.0238 | 88.52 | 0.0488 | 0.0510 |
| CLIPMat | ViT-L/14 | Yes | 8.29 | 0.0022 | 0.0027 | 40.37 | 0.0205 | 0.0229 | 85.83 | 0.0474 | 0.0495 |

# Experiments

We also show subjective comparisons as follows, the text inputs from the top to the bottom are: 1) *dandelion*; 2) *the flame which is light salmon and non-salient*; 3) *the woman who is with her back to the camera*. CLIPMat also achieves better results compared with others.



# Experiments

We conduct ablation studies to validate the effectiveness of our proposed modules. The experiments are carried out in the keyword-based setting of RefMatte. We show the results in the right table.

| TSP | MDE | Pre-CP | Post-CP | SAD | MSE | MAD |
|---|---|---|---|---|---|---|
| | | | | 22.88 | 0.0097 | 0.0131 |
| ✓ | | | | 18.28 | 0.0068 | 0.0105 |
| ✓ | ✓ | | | 14.55 | 0.0050 | 0.0083 |
| ✓ | ✓ | ✓ | | 11.48 | 0.0036 | 0.0065 |
| ✓ | ✓ | | ✓ | 12.96 | 0.0045 | 0.0074 |
| ✓ | ✓ | ✓ | ✓ | 9.91 | 0.0028 | 0.0057 |

# Conclusion

In this paper, we define a novel task named referring image matting (RIM), establish a large-scale dataset RefMatte, and provide a baseline method CLIPMat.

RefMatte provides a suitable test bed for the study of RIM, thanks to its large scale, high-quality images, and abundant annotations, as well as two well-defined experiment settings. Together with the RefMatte-RW100, they can be used for both in-domain and out-of-domain generalization evaluation.

Besides, the CLIPMat shows the value of special designs for the RIM task and serves as a valuable reference to the model design. We hope this study could provide useful insights to the image matting community and inspire more follow-up research.

# References


[1] Xu, Ning, Brian Price, Scott Cohen, and Thomas Huang. "Deep image matting." CVPR, 2017.

[2] Li, Jizhizi, Jing Zhang, and Dacheng Tao. "Deep automatic natural image matting." *IJCAI, 2021*

[3] Li, Jizhizi, Jing Zhang, Stephen J. Maybank, and Dacheng Tao. "Bridging composite and real: towards end-to-end deep image matting." *IJCV,* 2022.

[4] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, ICML, 2021.

[5] Kamath, Aishwarya, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. "Mdetr-modulated detection for end-to-end multi-modal understanding." CVPR. 2021.

[6] Lüddecke, Timo, and Alexander Ecker. "Image segmentation using text and image prompts." CVPR, 2022.

[7] Li, Jizhizi, Sihan Ma, Jing Zhang, and Dacheng Tao. “Privacy-Preserving Portrait Matting.” ACM MM. 2021.

# Thanks!