MAD: Makeup All-in-One with Cross-Domain Diffusion Model

Supplementary Material

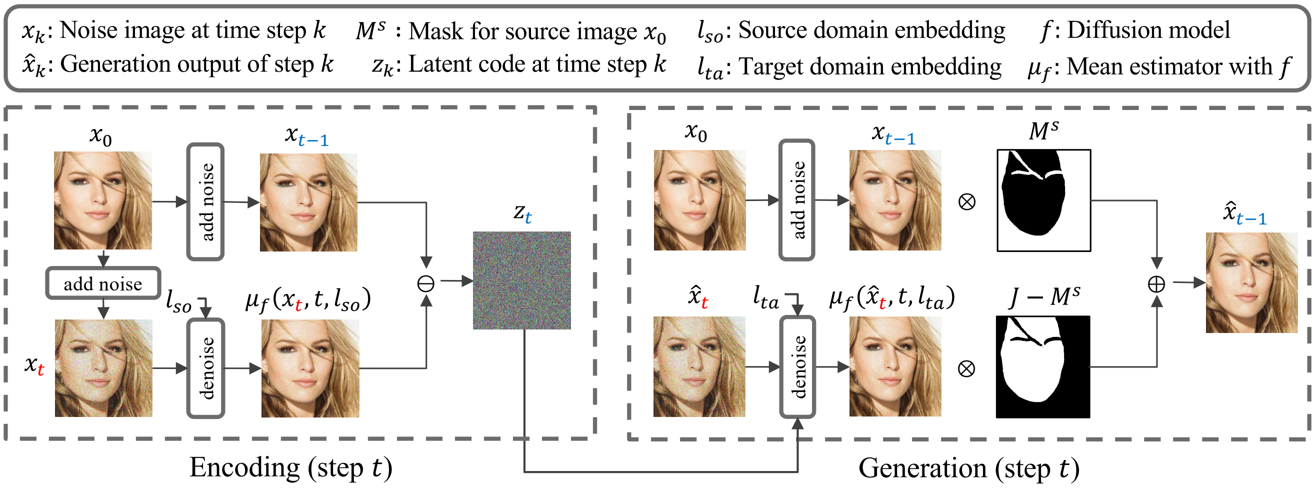

x_k: Noise image at time step k
M^s : Mask for source image x_0
l_{so}: Source domain embedding
f: Diffusion model
$\hat{x}_k$: Generation output of step k
z_k: Latent code at time step k
l_{ta}: Target domain embedding
μ_f: Mean estimator with f
x_0
add noise
x_{t-1}
z_t
add noise
l_{so}
$\mu_f(x_t, t, l_{so})$
x_t
denoise
$\ominus$
Encoding (step t)
x_0
add noise
x_{t-1}
M^s
$\otimes$
$\hat{x}_{t-1}$
$\hat{x}_t$
l_{ta}
$\mu_f(\hat{x}_t, t, l_{ta})$
$J - M^s$
$\oplus$
denoise
$\otimes$
Generation (step t)

Figure 8. Illustration of the cross-domain diffusion pipeline for time step t. Initially, the pipeline generates a latent code representing the source domain, which is subsequently used in the target domain generation to ensure detail preservation. During the generation phase, a preserved mask can be applied to maintain non-facial regions or to modify specific components.

A. Details of Makeup Transfer

Algorithm 1 outlines the procedure for single makeup transfer. For multi-makeup transfer, we replace the blending equation in Eq. 9 with Eq. 10. Furthermore, Fig. 8 offers a comprehensive visualization of the cross-domain translation process within our framework, as described in Sec. 3.

B. Annotation of MT-Text Dataset

The detailed annotation process for the MT-Text dataset is illustrated in Fig. 9. Using prompt and input images, the GPT-4v agent generates an initial labeling for three facial regions: "eyes," "lips," and "face," as depicted on the right.

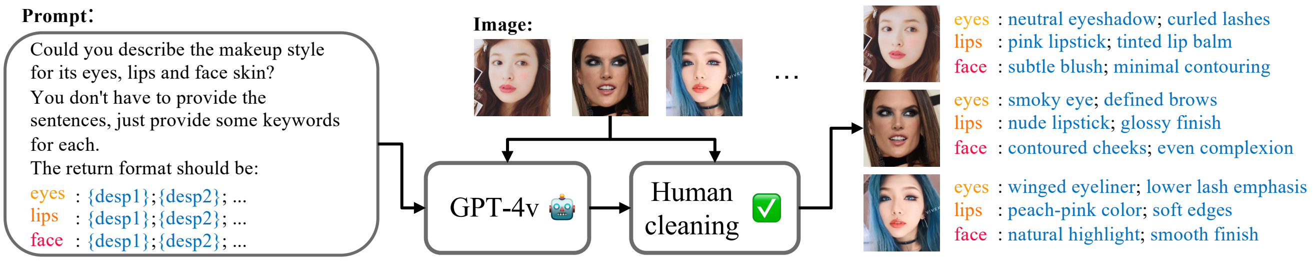
C. Visual Comparison

Makeup Removal. A visual comparison of makeup removal is provided in Fig. 10. Our method, using only the removal embedding, offers a good approximation of the subject's original appearance. However, incorporating the reference style significantly enhances the accuracy of the results. This improvement is quantifiable, as reflected by higher PSNR values when using the reference style compared to the embedding alone. Using only the embedding may not consistently align with the original features. Notably, when compared to other methods that also use reference styles, our approach achieves more precise removal of makeup on eyebrows, lips, and skin color.

Algorithm 1: Makeup Transfer

```
Input: step $K$, component set $C$, component masks
       $\{M^c | c \in C\}$ and starting time $\{t_c | c \in C\}$,
       source image $x_0$ and mask $M^s$, reference
       image $y_0$, model $f$, source and target
       embedding $l_{so}$ and $l_{ta}$, and scale $\alpha$
Output: denoised output $\hat{x}_0$
Obtain source and reference facial mesh $R_s$ and $R_r$ ;
Compute warping function $F_{\text{warp}}$ from $R_s$ and $R_r$ ;
$x'_0 \leftarrow (J - \alpha)x_0 + \alpha F_{\text{warp}}(y_0)$ ;   // Perform Blending
$\hat{x}_K \sim q(\hat{x}_K | x'_0)$ ;   // Obtain noisy input
for $t \leftarrow K$ to $1$ do
    /* Obtain source facial latent code */
    $x_{t-1} \sim q(x_{t-1}|x_0)$ ; $z_t = \frac{(x_{t-1} - \mu_f(x_t, t, l_{so}))}{\sigma_t}$ ;
    /* Component Asynchronous Masking */
    $M^s_t = \sum_{c \in C} \mathbb{1}_{\{t > t_c\}} \cdot M^c$ ;
    /* Perform masking for preservation */
    $\hat{x}'_{t-1} \leftarrow \mu_f(\hat{x}_t, t, l_{ta}) + \sigma_t z_t$ ;
    $\hat{x}_{t-1} \leftarrow M^s_t \cdot x_{t-1} + (J - M^s_t) \cdot \hat{x}'_{t-1}$;
return $\hat{x}_0$ ;
```

Text-to-Makeup. We provide the visual comparison in Fig. 11. Our approach can obviously preserve the identity better and still provide the correct makeup style. Other approaches tend to fit only the prompt without considering the



Could you describe the makeup style for its eyes, lips and face skin?
You don't have to provide the sentences, just provide some keywords for each.
The return format should be:
eyes : {desp1};{desp2}; ...
lips : {desp1};{desp2}; ...
face : {desp1};{desp2}; ...

eyes : neutral eyeshadow; curled lashes
lips : pink lipstick; tinted lip balm
face : subtle blush; minimal contouring

eyes : smoky eye; defined brows
lips : nude lipstick; glossy finish
face : contoured cheeks; even complexion

eyes : winged eyeliner; lower lash emphasis
lips : peach-pink color; soft edges
face : natural highlight; smooth finish

Figure 9. Illustration of our text labeling process. We first apply the GPT-4V model to provide a rough description of each area and clean the output to provide the correct labeling results.

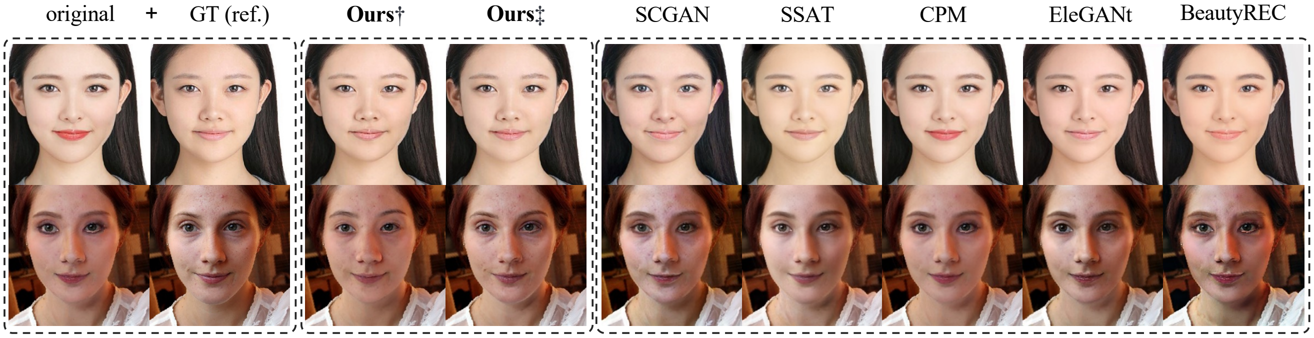
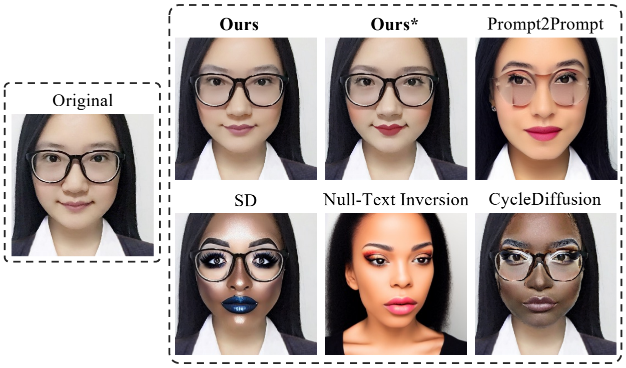


Figure 10. Visual comparison for makeup removal. The reference style of the first row is from the MT dataset, and the second is from the BF dataset. † represents makeup removal with an embedding, and † represents the makeup removal with the reference style.



"Makeup with winged eyeliner, well-defined lip line, and highlighted cheekbones"

Figure 11. Visual comparison for text editing with the prompt at the bottom. * represents the finetuned version with MT-Text.

original appearance. Compared with the non-finetuned version, the finetuned model can provide a more explicit style, such as a better illustration of Cupid's Bow (with a defined lip line), highlighted cheekbones with brighter color, and using eyelashes to accentuate the winged eyeliner.

Makeup Transfer. In Fig. 12, we illustrate the effectiveness of our makeup transfer method in various scenarios. Our blending approach selectively transfers makeup styles without incorporating undesired elements, such as dark tones (first row), from the reference images, showcasing our method's precision in capturing and applying only relevant styles. Additionally, our method is not severely affected by new skin tone (third row) due to utilizing original encoding information during generation. Our approach also enables adaptability to slightly different head sizes (second row) and different poses (third row) due to utilizing warping for accurate alignment of makeup elements before generation, ensuring consistency across various facial orientations.

Ablation Study. A comparative visual demonstration for last-K vs. DDIM is provided in Fig. 13 as evidence. Additionally, visual comparisons in Fig. 13 highlight CAM's ability to capture finer details for small areas, thereby enhancing the fidelity and precision of the makeup representation. Additional examples in Fig. 14 further demonstrate CAM's effectiveness in dealing with small areas to improve the realism and precision of makeup application.

D. Analysis of Last-K Step Denoising

We examine the results across different values of K and compare them with our component, asynchronous masking, as illustrated in Fig. 16. As K increases, style matching performance decreases, as indicated by higher KID scores. In contrast, larger values of K yield better identity preservation. Among all methods, ours achieves the best identity quality while maintaining competitive style matching.

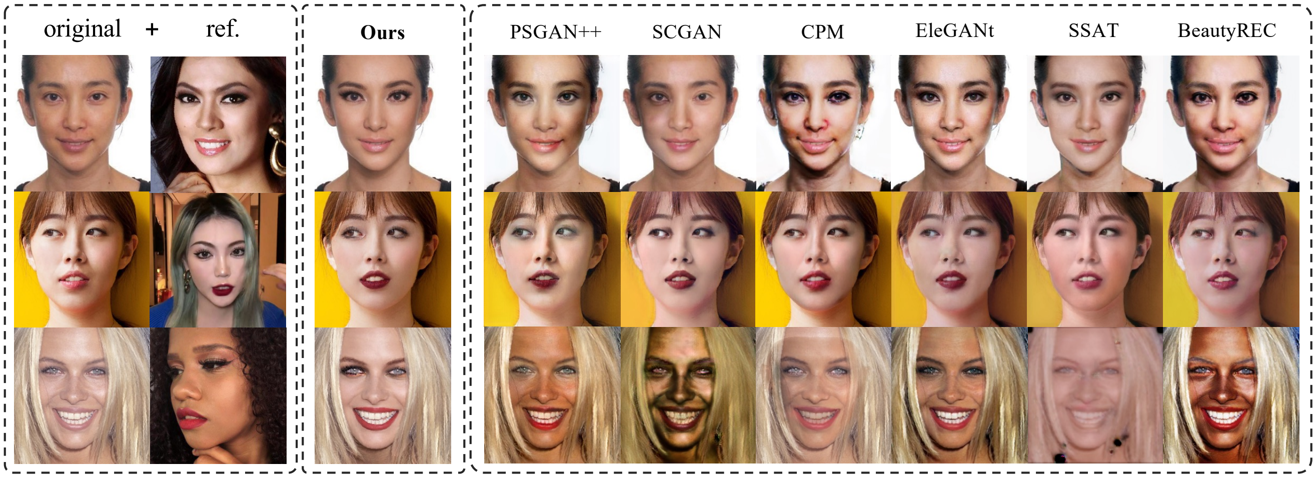


Figure 12. Visual comparison for makeup transfer. The reference image of the first row is from the MT dataset, the second row from the BF dataset, and the third row from the Wild dataset.

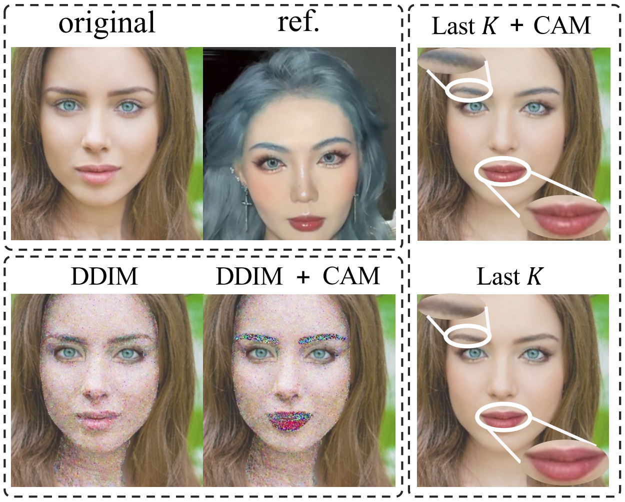


Figure 13. Visual comparison for ablation study. This figure contrasts the DDIM with our last-K step approach and shows the impact of component asynchronous masking (CAM) on the results.

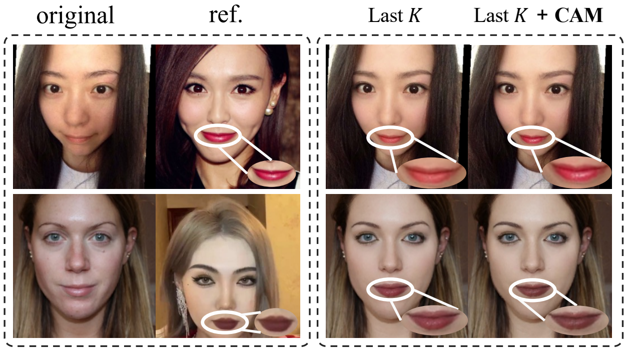


Figure 14. More examples to demonstrate that with component asynchronous masking (CAM), we can better transfer the styles, including luminance and color, for small areas.

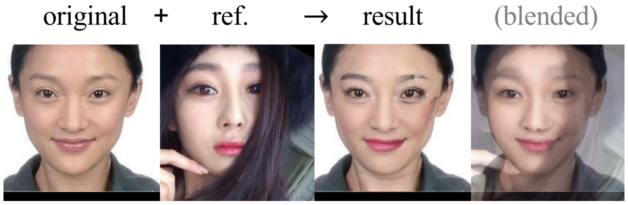


Figure 15. Illustration of the failure case due to facial occlusion. We provide the "blended" output on the rightmost to illustrate how the hair occlusion affects the transferred results.

E. Limitation and Failure Analysis

Here, we discuss the limitations and present failure cases of our model, along with possible reasons for these failures. This analysis is intended to guide future research directions in the makeup field based on our work.

Slow Generation Speed A notable challenge associated with our diffusion model is its relatively slow generation speed, with each image taking approximately 20 seconds to process, attributed primarily to the diffusion model's complexity. Although the process is optimized by focusing on the last K steps, the method still requires hundreds of steps to achieve the desired results. Future improvements could aim at improving the generation speed, potentially through the adoption of a latent diffusion model approach, as suggested by [19], or employing our idea with the Rectified Flow [11] for cross-domain translation.

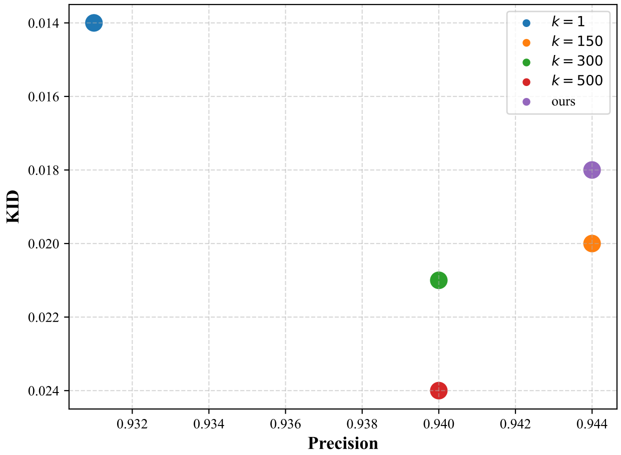


Figure 16. Illustration of different k for last-step k.

Facial Occlusion Challenges. Facial alignment, a critical step in our makeup transfer approach, can be problematic when faces are obscured by hair or other accessories. The style or the color for the occlusion part can be inconsistent with other areas, making the transfer process fail, as shown in Fig. 15. Since the left face of the reference image is blocked by its hair, we can observe some artifacts on the left face of the transferred result. A promising area for future research based on our work could be mixing in feature space instead of directly applying blending for pure images. This approach has the potential to eliminate visible artifacts, thereby preventing inconsistent or weird results.

Unnatural Text-to-Makeup Styles Although our approach supports text-to-makeup generation, it may produce unnatural makeup styles, even when accurately following the prompts. An example of this can be observed in the orange eyeshadow shown in Fig. 5. To address this issue, we plan to collect more contemporary makeup data and incorporate novel or bold makeup styles to achieve natural and creative results.