

# DreamMakeup: Face Makeup Customization using Latent Diffusion Models

Geon Yeong Park\*<sup>1</sup>  
Heechan Jeon<sup>2</sup>

Inhwa Han\*<sup>1</sup>  
Myeongjin Goh<sup>2</sup>

Serin Yang\*<sup>1</sup>  
Sung Won Yi<sup>2</sup>  
<sup>1</sup>KAIST    <sup>2</sup>Amorepacific  
\*Equal contribution

Yeobin Hong\*<sup>1</sup>  
Jin Nam<sup>2</sup>

Seongmin Jeong<sup>2</sup>  
Jong Chul Ye<sup>1</sup>

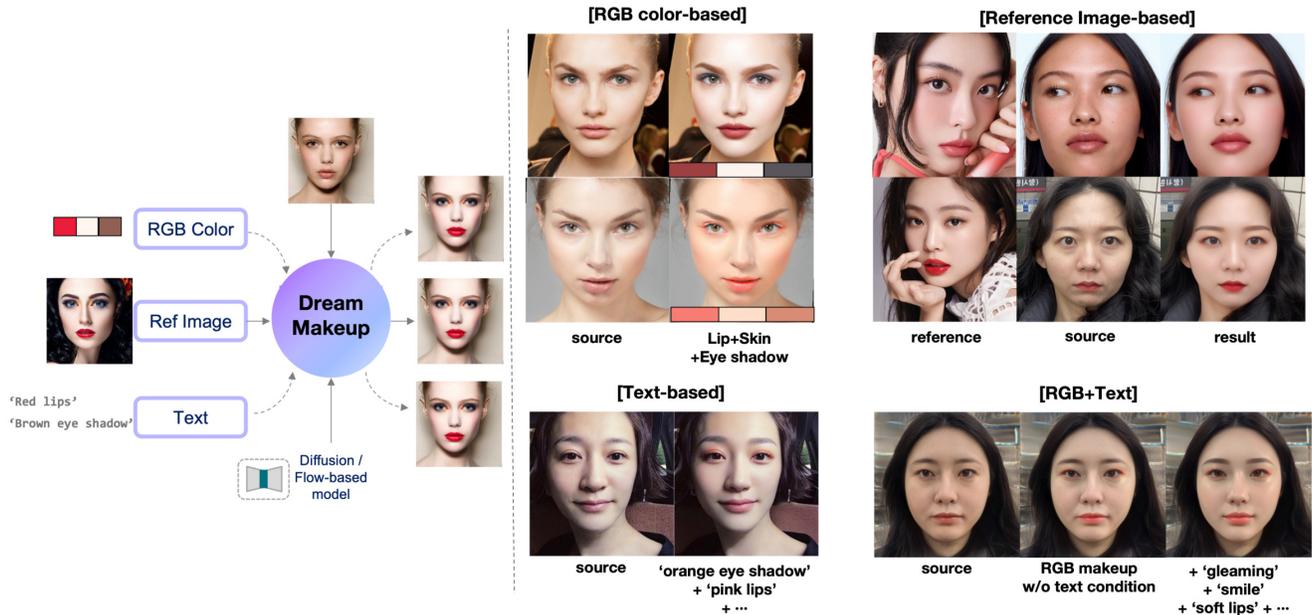


Figure 1. We present **DreamMakeup**, a training-free diffusion framework that generates high-fidelity makeup results by integrating diverse user inputs such as RGB colors, reference images, and text prompts. Our method produces high-quality, customized makeup while preserving facial identity, without requiring any fine-tuning. Please zoom in for detailed inspection.

## Abstract

The exponential growth of the global makeup market has paralleled advancements in virtual makeup simulation technology. Despite the progress led by GANs, their application still encounters significant challenges, including training instability and limited customization capabilities. Addressing these challenges, we introduce DreamMakup – a novel training-free Diffusion model based Makeup Customization method, leveraging the inherent advantages of diffusion models for superior controllability and precise real-image editing. DreamMakeup employs early-stopped DDIM inversion to preserve the facial structure and identity while enabling extensive customization through various conditioning inputs such as reference images, specific RGB colors, and textual descriptions. Our model demonstrates notable improvements over existing GAN-based and recent diffusion-based frameworks – improved customization, color-matching capabil-

ities, identity preservation and compatibility with textual descriptions or LLMs with affordable computational costs.

## 1. Introduction

The global makeup market size is valued at billions of dollars, and virtual makeup simulation technology is considered to be a rapidly growing sector within the beauty industry. Besides its industrial importance, face makeup customization is also an interesting problem in terms of generative modeling and editing. Specifically, one may have to disentangle and stylize each facial attribute in their independent style, while its composition should be well harmonized.

So far, virtual face makeup modeling is mainly driven by generative adversarial networks (GANs) [1, 7, 9, 10, 22]. Despite its advancements, GAN-based frameworks face inherent instability in adversarial training, leading to several limitations. Furthermore, existing GAN-based methods are

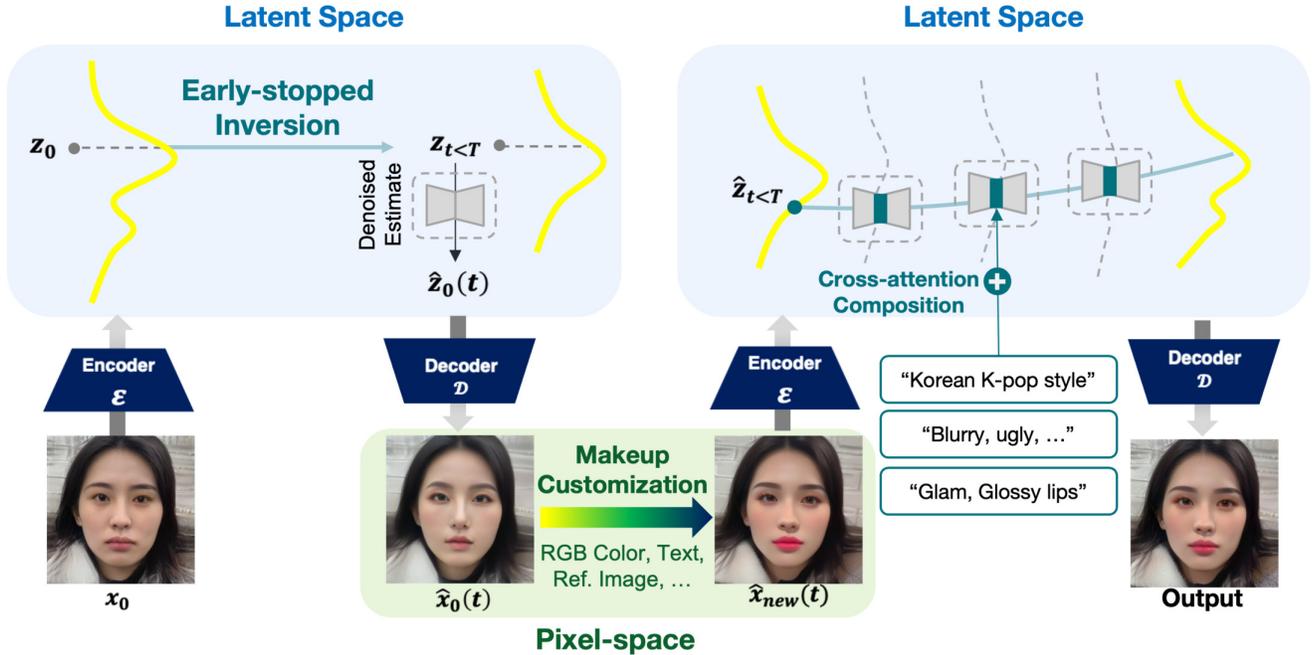


Figure 2. **Overview of DreamMakeup pipeline.** The key principle of our framework is to apply fine-grained guidance in high-dimensional pixel-domain during reverse sampling. After local makeup customization in pixel space, text prompts are leveraged to harmonize such local variations with a consistent global style in latent cross-attention space.

not fully customizable and lack controllability. Specifically, most of these frameworks *only* support makeup transfer tasks, inherently requiring reference target images. In many business contexts, users may seek to simulate facial makeup with a more degree of freedom, e.g. test with specific RGB colors of new cosmetic products, or linguistic descriptions such as “Glam makeup style”, etc.

In response to these challenges, this paper explores the adoption of diffusion models, recognized for their superior controllability and real-image editing capabilities. Diffusion models offer several advantages for facial beauty simulation. For instance, we can control the reverse sampling process using the enriched text-conditions. Moreover, it supports various style customization using LoRAs [6] supported by the vibrant user community. By employing techniques such as DDIM inversion [13, 16], diffusion models well preserve the overall structure and subject identity of given facial images, while retaining rich editing capabilities.

However, straightforward application of generic diffusion-based editing algorithms is inadequate for customized makeup, often failing to perform accurate color-matching and natural makeup in specified facial regions (Fig. 14). Consequently, while several specialized diffusion-based makeup frameworks have been introduced [8, 11, 17, 24], many of them still face similar constraints with GANs; many remain limited to reference-based makeup transfer, failing to leverage more diverse conditioning inputs. Additionally, they require expensive training/fine-tuning, potentially restricting

the generative capacity of powerful prior models, and often lose structural identity (Stable-Makeup in Fig. 6).

To this end, we introduce DreamMakeup, a novel *training-free* diffusion-based makeup customization distinguished by its advanced customization and superior identity preservation. Dreammakeup is fully compatible with a variety of conditionings to steer the makeup process, ranging from reference images and specific RGB colors to textual descriptions of desired makeup looks. As shown in Fig. 2, given pre-trained latent diffusion models (LDM), we commence by inverting facial images  $x_0$  into latents  $z_t$  through early-stopped DDIM inversion. Subsequently, we approximate the denoised estimate  $\hat{z}_0(t) := \mathbb{E}[z_0|z_t]$  and decode it back into the pixel space, preserving the facial structure attributed to the inversion process. Then, we stylize these facial representations in pixel space through transformations such as histogram matching, RGB color matching, or warping, toward a targeted makeup style. Finally, resuming the sampling process from these transformed representations, with advanced cross-attention control and interpolation-guided sampling, yields harmonized and coherent makeup outcomes. Our contributions are summarized as follows:

- We introduce *DreamMakeup*, a novel diffusion-based human face makeup framework that caters to a wide range of user preferences including text descriptions, colors, and reference images.
- DreamMakeup is computationally affordable as it does not fundamentally require fine-tuning. Moreover, we early-

stop DDIM inversion process to preserve the facial structures which further accelerates inference ( $< 4$  seconds for color transfer w/ SD v1.5 [15], GeForce RTX 4090).

- DreamMakeup outperforms real-world global AI makeup services in color makeup task, and state-of-the-art Diffusion/GAN-based frameworks in makeup transfer tasks. Furthermore, we demonstrate that our framework can be easily integrated with other foundational models, including Large Language Models (LLMs), facial classifiers, LoRAs or even other makeup diffusion models.

## 2. Related Works

Makeup transfer aims to modify a facial image to reflect a chosen makeup style, with numerous approaches developed using Generative Adversarial Networks (GANs). BeautyGAN [9] employed histogram matching to preserve color from the reference image. LADN [4] used local discriminators for heavy makeup. PSGAN [7] enhanced style controllability through matrices and addressed pose misalignments with attention mechanisms. SCGAN [1] tackled pose issues with style codes. RamGAN [20] improved makeup transfer with regional attention; EleGANt [22] controlled arbitrary regions using attention mechanisms, reducing computations.

However, GAN-based methods typically require large datasets of makeup and no-makeup images for training and reference images for inference, limiting application diversity and customizability. This is in line with recent diffusion-based methods [8, 11, 24] which often focus on adapting diffusion models for reference-based makeup with fine-tuning. Moreover, due to the stochastic generative nature of diffusion models, these works often fail to preserve the structural identity of input images (e.g., Stable Makeup in Fig. 6).

DreamMakeup overcomes these limitations by utilizing the powerful prior of foundational diffusion models, DDIM inversion and open-source LoRAs for image generation, enabling replication of makeup styles from references, manipulation via RGB or text prompts. This approach enhances controllability without relying on large datasets or fine-tuning.

## 3. Preliminary

Diffusion models aim to generate samples from the Gaussian noise through iterative denoising processes. Since pixel-space diffusion models are computationally heavy, the latent diffusion model (LDM) [15] operates the diffusion process on latent space instead of pixel space. Given a pixel-space clean sample  $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ , [15] leverages an autoencoder

$$\mathcal{E} : \mathbb{R}^d \rightarrow \mathbb{R}^k, \mathcal{D} : \mathbb{R}^k \rightarrow \mathbb{R}^d, \mathbf{x} \simeq \mathcal{D}(\mathcal{E}(\mathbf{x})), \forall \mathbf{x} \sim p_{\text{data}}(\mathbf{x}), \quad (1)$$

where  $\mathcal{E}$  is the encoder,  $\mathcal{D}$  is the decoder, and dimension of the latent space  $k < d$ . After training  $\mathcal{E}$ ,  $\mathcal{D}$ , forward and reverse diffusion process can be defined within the latent space  $\mathbf{z} = \mathcal{E}(\mathbf{x})$ .

The forward process is defined as a Markov chain, characterized by forward conditional densities:

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t | \beta_t \mathbf{x}_{t-1}, (1 - \beta_t)I), \quad (2)$$

$$p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t)I), \quad (3)$$

with  $\mathbf{z}_t \in \mathbb{R}^k$  representing the noisy latent variable at a timestep  $t \leq T$  that has the same dimension as  $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$  for  $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ , and  $\beta_t$  denotes an increasing sequence of noise schedule where  $\alpha_t := 1 - \beta_t$  and  $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$ . The goal of training LDM is to obtain a residual denoiser  $\epsilon_{\theta^*}$ :

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{\mathcal{E}(\mathbf{x}_0), t, \epsilon \sim \mathcal{N}(0, I)} [\|\epsilon_{\theta}(\mathbf{z}_t, t) - \epsilon\|]. \quad (4)$$

The reverse sampling from  $q(\mathbf{z}_{t-1} | \mathbf{z}_t, \epsilon_{\theta^*}(\mathbf{z}_t, t))$  is then

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta^*}(\mathbf{z}_t, t) \right) + \tilde{\beta}_t \epsilon, \quad (5)$$

where  $\epsilon \sim \mathcal{N}(0, I)$  and  $\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$ . For simplicity, we will omit  $*$  in  $\theta^*$ . After reverse sampling, the generated latent  $\tilde{\mathbf{z}}_0$  is decoded to the pixel space as  $\tilde{\mathbf{x}}_0 = \mathcal{D}(\tilde{\mathbf{z}}_0)$ .

To accelerate sampling, DDIM [16] proposes an alternative sampling method:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{z}}_0(t) + \sqrt{1 - \bar{\alpha}_{t-1} - \eta^2 \tilde{\beta}_t^2} \epsilon_{\theta}(\mathbf{z}_t, t) + \eta \tilde{\beta}_t \epsilon, \quad (6)$$

where  $\eta \in [0, 1]$  is a stochasticity parameter, and  $\hat{\mathbf{z}}_0(t)$  is the denoised estimate which can be equivalently derived using Tweedie’s formula [2]:

$$\hat{\mathbf{z}}_0(t) := \frac{1}{\sqrt{\alpha_t}} (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_{\theta}(\mathbf{z}_t, t)). \quad (7)$$

For text-guided sampling, we train the diffusion model with textual embedding  $c$ . We will often omit  $c$  from  $\epsilon_{\theta}(\mathbf{x}_t, t, c)$  to avoid notational complexity.

## 4. DreamMakeup

Our *primary* focus is on daily, realistic and aesthetic makeup (Fig. 1a), rather than complex and excessive makeup styles/transfer. Specifically, given an input non-makeup image  $\mathbf{x}_0$ , our main goal is to (a) customize the makeup style with coarse (e.g. RGB color) to fine (e.g. reference makeup image) level information, while (b) preserving the overall facial structure and subject identity to the greatest extent. Furthermore, we aim for applications beyond its primary scope. It can be extended to extreme makeup styles (Fig. 17), and its capacity to handle customized user inputs enables diverse applications such as hair dyeing (Fig. 18). Finally, we are interested in integrating with LLMs or facial classifiers, paving a new path for virtual makeup pipeline design.

To achieve this, one of our primary contributions is to integrate a pixel-space makeup customization during reverse

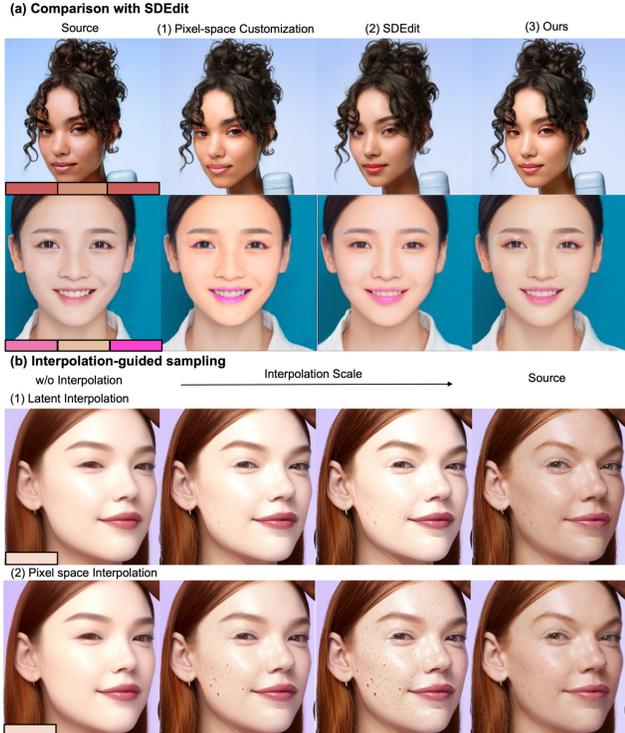


Figure 3. (a) Direct pixel-space customization results in color inconsistencies, while SDEdit (strength=0.2) degrades the subject’s identity. In contrast, our method applies the makeup faithfully while preserving identity. (b) The impact of the interpolation domain during reverse sampling. Latent-space interpolation (b-1) effectively preserves fine-grained facial details, whereas pixel-space interpolation (b-2) introduces significant visual artifacts. Further details are provided in Sec. 4.3.

	SDEdit	Ours	SDEdit	Ours
LPIPS (↓)	0.069	<b>0.040</b>	0.868	<b>0.921</b>
CLIP-I (↑)				

Table 1. Quantitative evaluation on preservation of facial structure.

sampling process given the decoded intermediate estimates  $\hat{x}_0(t) = \mathcal{D}(\hat{z}_0(t))$ . Furthermore, *DreamMakeup* leverages various user preferences for conditional guidance, e.g. target color, reference image, and textual make-up description. As shown in Fig. 2, the customization process consists of three main phases: (1) early-stopped DDIM inversion to impose structural consistency, (2) pixel-space customization to guide the sampling process towards target makeup style, and (3) reverse sampling with cross attention composition and interpolation guidance to accommodate complex textual makeup descriptions simultaneously.

#### 4.1. Early-stopped Inversion

While pixel-space editing achieves fine-grained control, the resulting edits must be seamlessly integrated without compromising the subject’s identity. Standard harmonization techniques that rely on the *stochastic* nature of the reverse diffusion process, often fail at this task by degrading the fa-

cial structure and identity. For instance, performing primary pixel-space makeup customization (Sec. 4.2) followed by noise addition and subsequent denoising refinement (SDEdit [12]) results in significant identity degradation (Fig. 3a-2, Tab. 1). For successful makeup customization, the sampling trajectory must be constrained to preserve the original facial identity, as humans are highly sensitive to subtle facial changes. To faithfully maintain the identity of the original face during the sampling process, we instead leverage *deterministic* DDIM inversion which is an iterative reverse simulation of the ODE flow in the limit of small steps. By setting  $\eta = 0$  in Eq. (6), the DDIM inversion [13, 16] for  $z_t$  is defined as:

$$z_t = a_t z_{t-1} - b_t \epsilon_\theta(z_{t-1}, t, c), \quad (8)$$

$$\text{where } a_t = \frac{\sqrt{\alpha_t}}{\sqrt{\alpha_{t-1}}}, b_t = \sqrt{\alpha_t} \left( \sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right).$$

This relies on the linearization approximation assuming  $\epsilon_\theta(z_{t-1}, t, c) \approx \epsilon_\theta(z_t, t, c)$ . However, the approximation error may accumulate following inversion steps, potentially causing identity loss. To prevent this, we terminate the inversion at  $t^* \leq T$  to reduce computational burdens and ensure structural consistency, in contrast to the conventional process inverts  $z_0$  to  $z_T$ . Fig. 2 shows that the denoised  $\hat{z}_0(t^*)$  from early-stopped  $z_{t^*}$  may decode faithfully the original sample  $x_0$ . This allows us to directly guide  $\hat{x}_0(t^*) = \mathcal{D}(\hat{z}_0(t^*))$  in a pixel space to enforce the target style. To further guarantee the faithful preservation, we propose interpolation-guided sampling which will be discussed in Sec. 4.3.

#### 4.2. Pixel-space makeup customization

Our next goal is to transform  $\hat{x}_0(t^*)$  in a manner that accurately emulates the desired makeup appearance indicated by reference image or a target RGB color. To this end, we introduce a pixel-space transformation  $\mathcal{T}(\cdot, \cdot) : \mathbb{R}^{H \times W \times 3} \times X \rightarrow \mathbb{R}^{H \times W \times 3}$ , offering multiple variants of  $\mathcal{T}$  to enable fine-grained makeup customization. Here  $X$  varies for different references, e.g. target RGB color, reference image, etc.

##### 4.2.1. Makeup transformation with RGB color

We first delineate an intuitive color transfer function that imposes the color characteristics of the reference makeup palette color on the source image. Let  $\mu_{src}(\hat{x}_0(t^*))$ ,  $\sigma_{src}(\hat{x}_0(t^*))$  represents the RGB mean and standard deviation of  $\hat{x}_0(t^*)$  computed across spatial dimensions. Given a reference color  $\mu_{tgt}$  and respective standard deviation  $\sigma_{tgt}$ , the color transfer function  $\mathcal{T}_{RGB}$  is defined as

$$\begin{aligned} \hat{x}_{new}(t^*) &= \mathcal{T}_{RGB}(\mu_{src}(\hat{x}_0(t^*)), \mu_{tgt}; \alpha) \\ &= \frac{\sigma_{src}(\hat{x}_0(t^*))}{\sigma_{tgt}} \left( \hat{x}_0(t^*) - \alpha [\mu_{src}(\hat{x}_0(t^*)) - \mu_{tgt}] \right), \end{aligned} \quad (9)$$

where  $0 \leq \alpha \leq 1$  represents a transfer scale. For simplicity, we empirically set  $\sigma_{src}(\hat{x}_0(t^*)) = \sigma_{tgt}$ .

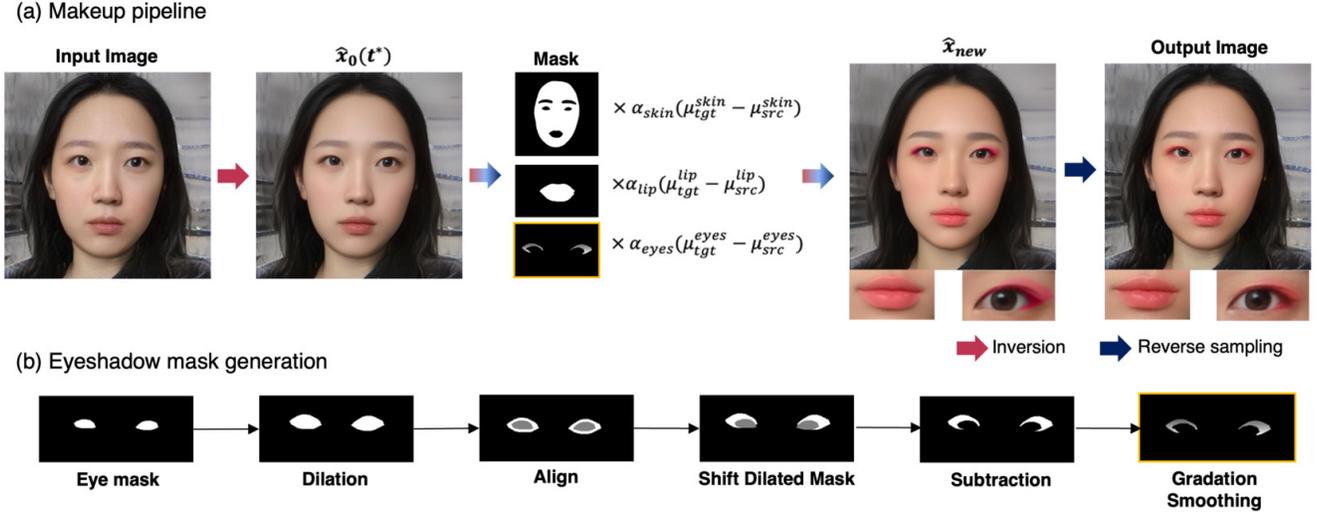


Figure 4. (a) Color-based makeup transformation. Mean RGB values within the masked area are adjusted with a scale  $\alpha$  to match the target RGB values. Output image is generated by reverse sampling from  $\hat{x}_{new}$ . (b) Eyeshadow mask is reproduced from eye mask manipulation.

The proposed color transfer supports attribute compositions. Specifically, we may transfer different  $\mu_{tgt}$  for lips, eye shadow, skin foundation, etc. For this, we segment each interested facial attribute using a segmentation model [23] pre-trained in pixel domain. We observed that the inverted image  $\hat{x}_0(t^*)$  is well segmented by the pre-trained model, owing to its high similarity with the original image  $x_0$ .

To prevent potential color discontinuities and ensure seamless integration, the eyeshadow and lip masks undergo further refinement. The eyeshadow mask is first reconstructed from the initial eye mask using transformations such as dilation and shifting, after which its edges are smoothed to create a natural blend (Fig. 4). Specifically, this smoothing is achieved by applying a gradient mask whose weights increase progressively from the inner to the outer boundary. The lip masks are also refined using a similar smoothing process. This allows fine control, enabling users to adjust the gradient decay rate for a more natural appearance. Any artifacts from this color transfer process are further refined through reverse diffusion sampling (More details in Sec. 4.3).

#### 4.2.2. Makeup transfer with reference image

The transformation  $\mathcal{T}$  can be varied depending on the downstream task. To demonstrate its universality, we consider conventional makeup transfer tasks [9]. Specifically, we simulate the makeup style of reference image through warping and histogram matching transformations.

First, histogram matching aligns the color distributions of the lip, eye shadow, and skin with reference. Then, the eyes of the source face are aligned with the reference through a series of warping transformations, including segmentation, dilation, affine, and diffeomorphic transformations, to ensure precise registration. This ensures that every pixel within the dilated mask area of the reference image corresponds to the

appropriate region on the source image. These steps enable the seamless adoption of tones and styles of the reference image, ensuring a natural makeup transfer.

#### 4.3. Makeup harmonization in reverse sampling

After pixel-space guidance, the output  $\hat{x}_{new}$  coarsely follows the desired makeup style in the local facial attributes, e.g. lips, eye shadow, etc, but could suffer from unnatural appearances ( $\hat{x}_{new}$  as in Fig. 4), requiring further refinement. Thus, our next goal is to (a) stylize  $\hat{x}_{new}$  with *multiple* text descriptions to harmonize such local variations with a consistent aesthetic style, e.g. “Korean K-Pop style”, “fair skin”, etc, while (b) maximally preserving the facial identity. This refinement begins with  $z_t = \sqrt{\bar{\alpha}_t}\mathcal{E}(\hat{x}_{new}(t^*)) + \sqrt{1 - \bar{\alpha}_t}\epsilon(z_{t^*}, t^*, c)$ , followed by subsequent denoising.

(a): **Cross attention composition.** Precise control over diverse makeup demands and facial attributes is challenging due to the need to balance multiple prompt strengths. To address this, we refine the image by integrating semantic makeup features directly into the cross-attention layer. Specifically, let  $Q_{t,l} \in \mathbb{R}^{P_l^2 \times d_l}$  represent the spatial query in  $l$ -th cross attention layer of U-Net. Given context vectors  $C \in \mathbb{R}^{N \times d_c}$ , let  $K, V \in \mathbb{R}^{d_c \times d_l}$  denote key and value matrices, respectively, where  $N$  refers to number of tokens,  $K = CW_{K,l}$  and  $V = CW_{V,l}$  with linear maps  $W_{K,l}, W_{V,l} \in \mathbb{R}^{d_c \times d_l}$ . Then, let  $K_s$  represents a  $s$ -th makeup concept key  $K_s$ , where  $K_{main}$  comes from the prompt used in inversion, i.e. “a photo of a woman”. Define  $V_s, V_{main}$  similarly. Then, we update the spatial query as:

$$Q^{new} = \text{softmax}\left(\frac{QK_{main}^T}{\sqrt{d}}\right)V_{main} + \quad (10)$$

$$\frac{1}{M} \sum_{s=1}^M \alpha_s \text{softmax}\left(\frac{QK_s^T}{\sqrt{d}}\right)V_s. \quad (11)$$

The degree and direction of  $s$ -th makeup concept can be controlled *individually* with  $\alpha_s$ , where  $\alpha_s < 0$  for negative makeup prompts, e.g. ugly, blurry, low-res. This linear combination incorporates detailed makeup descriptions independently, allowing fine-grained control of prompt strengths.

**(b): Interpolation-guided sampling.** While cross-attention composition effectively steers ODE sampling toward desired makeup styles, facial identity may degrade due to accumulated errors in DDIM inversion. As shown in Fig. 3b, DDIM inversion may fail to preserve fine-grained skin texture (e.g., pores, micro-blemishes), leading to an unnatural, *plastic* skin—an issue widely recognized in the SD community. To mitigate this, we *regularize* the sampling trajectory to remain closer to the original facial latents:

$$\begin{aligned} \tilde{z}_0(t) &= \arg \min_z \|z - \hat{z}_0(t)\|^2 + \frac{\lambda}{1-\lambda} \|z - \mathcal{E}(\mathcal{T}(x_0))\|^2 \\ &= (1-\lambda)\hat{z}_0(t) + \lambda\mathcal{E}(\mathcal{T}(x_0)), \\ z_{t-1} &= \sqrt{\bar{\alpha}_t}\tilde{z}_0(t) + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, c), \end{aligned} \quad (12)$$

where  $z'_0 = \mathcal{E}(\mathcal{T}(x_0))$  represents the makeup-transformed input latent. Serving as a pivotal reference,  $z'_0$  helps retain photorealistic skin details, ensuring both realism and identity preservation (Fig. 3b-1) while maintaining the intended makeup from  $\mathcal{T}(\cdot)$ . Conversely, performing this interpolation directly in the pixel space via  $(1-\lambda)\mathcal{D}(\hat{z}_0(t)) + \lambda\mathcal{T}(x_0)$  introduces undesirable visual artifacts, as shown in Fig. 3b-2. We observe that the latent space is a more suitable domain for this operation, facilitating a seamless interpolation between the denoised estimate and the transformed source. This approach strikes a superior balance between preserving the source’s identity and applying the custom makeup requirements. Notably, applying this regularization in just 1–2 early sampling steps significantly enhances identity preservation, with a moderate level of  $\lambda = 0.15$ .

## 5. Experimental Results

### 5.1. Experiment Settings

During inference, we used the Makeup Transfer (MT) dataset [9] for both source and reference images. Additionally, we employed artificially generated images for Asian women as a non-makeup source image. We used Stable Diffusion (SD) v1.5 and SDXL [14] as our base model, and further leveraged additional public LoRA weights and pre-trained models released in CivitAI, an open-source generative AI community. For facial segmentation, we utilized BiSeNet [23]. For cross



Figure 5. The virtual skin, lip, eye shadow makeup, and their combination by DreamMakeup (SD 1.5).

attention composition, we set the scaling factor for each makeup concept  $\alpha_s$  ranging from 0.1 to 0.7. For comparison, we tested state-of-the-art GAN-based makeup transfer methods, PSGAN [7], SCGAN [1], EleGANt [22], and CSD-MT [19], alongside SHMT [18] and Stable Makeup [24], which are recent diffusion-based makeup transfer frameworks. More experimental details are in appendix.

Method	LPIPS ↓	CLIP-I ↑	Makeup Artists		Non Artists	
			Detail ↑	Quality ↑	Detail ↑	Quality ↑
PSGAN	0.1879	0.7421	1.49	1.74	2.64	2.82
SCGAN	<u>0.0819</u>	0.7253	2.00	2.11	3.03	2.95
EleGANt	0.1877	<u>0.7662</u>	<u>3.04</u>	<u>3.12</u>	<u>3.67</u>	<u>3.72</u>
Ours	<b>0.0667</b>	<b>0.7694</b>	<b>3.42</b>	<b>3.42</b>	<b>3.95</b>	<b>4.00</b>

(a) Quantitative results on makeup transfer tasks.

Method	Beauty score	Makeup Artists		Non Artists	
		Detail ↑	Quality ↑	Detail ↑	Quality ↑
Service A	<u>2.90</u>	<u>4.08</u>	1.97	<u>3.69</u>	2.77
Service B	3.27	3.75	<u>2.61</u>	3.14	<u>2.83</u>
Ours	<b>3.38</b>	<b>4.22</b>	<b>4.19</b>	<b>3.93</b>	<b>4.27</b>

(b) Comparisons with global AI makeup services.

Table 2. Quantitative comparisons on makeup transfer task and color-based makeup transformation.

## 5.2. Results

### 5.2.1. Qualitative Results

Fig. 5 illustrates the RGB color-based makeup transformations applied to both synthetic and natural images. First three rows reflect the application of the eye shadow, skin, and lip colors indicated in the bottom right corner. Bottom row presents the combined results for each column’s corresponding colors. DreamMakeup effectively applies the



Figure 6. Qualitative comparison of reference image based makeup using references with diverse poses and makeup styles. GAN-based models exhibit artifacts and cropping issues (row 1,3,4). SHMT incorrectly transfers reference color and Stable-Makeup struggles with identity preservation, likely due to the stochastic nature of diffusion models.



Figure 7. DreamMakeup integrated with Stable Makeup [24].

specified RGB colors to the respective facial regions.

Qualitative comparison with reference-based makeup baselines is presented in Fig. 6. Our approach demonstrates superior performance over competing models. Most GAN-based methods - PSGAN, SCGAN, and CSD-MT produce noticeable artifacts, including identity loss and color bleeding. While EleGANT performs competitively, it fails on challenging cases, exhibiting dark artifacts in occluded regions (row 1) or generating weaker eye makeup. Among diffusion-based methods, SHMT struggles with inaccurate color transfer, and Stable Makeup exhibits poor identity preservation. Our method consistently produces clean, artifact-free results that faithfully replicate the reference makeup style.

As a training-free framework, DreamMakeup can be seamlessly integrated with other generative models. We demonstrate this modularity by incorporating our core components such as inversion, pixel-space customization, and interpolation guidance, into two backbones: the diffusion-based makeup transfer framework Stable Makeup [24] and the recent flow-based Stable Diffusion 3.0 [3]. In Fig. 7, inte-



Figure 8. DreamMakeup integrated with Flow-based models.

grating DreamMakeup improves structural consistency and reduces artifacts compared to the original Stable Makeup model. Furthermore, Fig. 8 illustrates that our framework is compatible with Diffusion Transformer architectures, highlighting its potential for synergistic applications across different model families.

### 5.2.2. Quantitative Comparison

We evaluate DreamMakeup against baselines on both color-based and reference image-based makeup tasks. For reference based makeup, we assess identity preservation using LPIPS and CLIP image similarity. Additionally, 10 expert makeup artists and 24 non-experts ranked 10 random outputs based on style accuracy and overall quality (scale: 1–5), where DreamMakeup consistently excelled (Tab. 2a).

For color-based makeup, we further compare against two global AI makeup services (50M+ downloads) while ensuring anonymity. Despite their widespread use, these services exhibit limitations in customization, such as restricted color

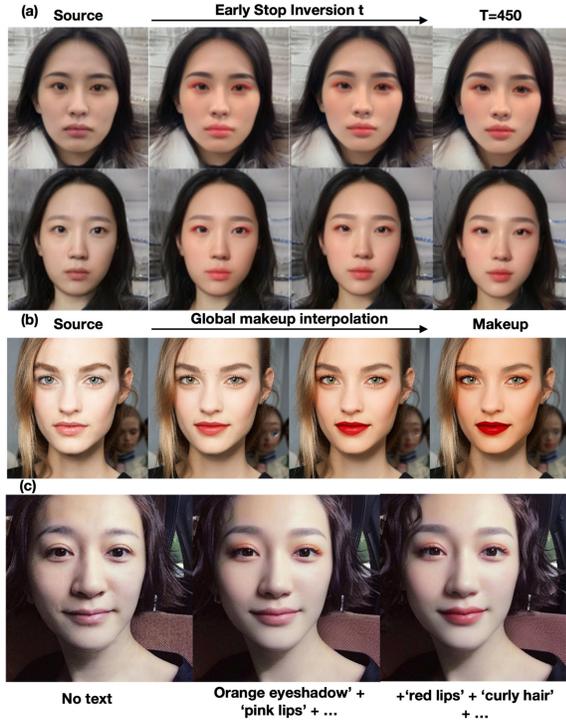


Figure 9. Ablation study for early-stop timestep  $t$ , interpolation scale  $\alpha$  and text guidance.

presets. Using 100 images per service, we evaluate beauty scores [21] and conduct a broader user study (300 images). DreamMakeup consistently outperforms others (Tab. 2b).

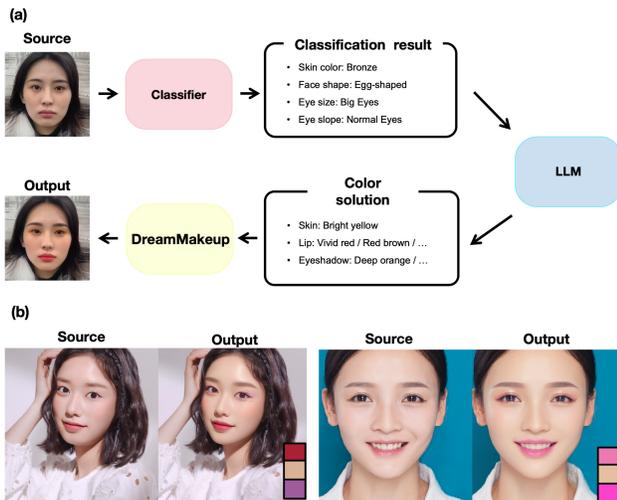


Figure 10. (a) DreamMakeup with integration with a classifier and an LLM. (b) From the source image, we can apply makeup based on the solution provided by the classifier and LLM. The color chips represent the color of lips, skin, and eye shadow from top to bottom.

### 5.2.3. Ablation Study

We conducted an ablation study on the key hyperparameters, with results shown in Fig. 9. Varying the early-stop timestep

$t^*$  for DDIM inversion allows for a balance between faithful identity preservation and stylistic variation. The transfer scale  $\alpha$ , defined in (9), governs the intensity of the applied makeup. As  $\alpha$  approaches 0, the output progressively resembles the original source image, effectively reducing the makeup’s opacity (Fig. 9(b)). Notably, this parameter can be adjusted independently for distinct facial regions (e.g., lips, eyes), enabling fine-grained control over the final look. The effect of textual guidance, implemented via cross-attention composition, is shown in Fig. 9(c). The inclusion of descriptive prompts harmonizes the local makeup applications with a cohesive global aesthetic, demonstrating the significance of text conditioning in achieving a polished result.

### 5.3. Integration with LLM

DreamMakeup can demonstrate improved performance by integrating with Large Language Models (LLMs) for personalized makeup recommendations. By harnessing the exceptional inference capabilities of LLMs, DreamMakeup selects makeup colors that are harmonious with the characteristics of the source image. We provide a pipeline in Fig. 10. This pipeline involves an initial extraction of facial attributes such as skin tone and facial structure from the source image via a classifier. This information is then conveyed to the LLM, which determines the most appropriate makeup colors for various facial regions, including the skin, eyes, and lips. DreamMakeup subsequently utilizes these recommendations to generate the final makeup-enhanced image.

These components (classifier, LLM, and DreamMakeup) operate sequentially during inference but are trained independently. The classifier was trained using ResNet50 on a dataset of 1,000 artificially generated images of Asian women, annotated with facial information. The LLM is trained on the dolly-v2-3b model with a specialized QnA dataset from beauty professionals. As illustrated in Fig. 10(b), the LLM adeptly matches skin, eye shadow, and lip colors to the source image, facilitating the application of these colors by DreamMakeup. The process ensures that the selected colors are well-suited to the source image, leading to an effectively applied makeup look. This demonstrates the potential of integrating classifiers and LLMs in DreamMakeup to provide customized makeup solutions based on in-depth analysis of facial features, thereby enhancing the personalization and effectiveness of makeup applications.

## 6. Conclusion

We introduced DreamMakeup, a novel training-free makeup framework utilizing powerful priors. Our method ensure precise makeup customization with rich conditions. We demonstrated DreamMakeup’s global effectiveness and efficiency in various tasks and verify compatibility with other frameworks.

## References

- [1] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He. Spatially-invariant style-codes controlled makeup transfer. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 6549–6557, 2021. 1, 3, 6
- [2] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 3
- [3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 7
- [4] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladt: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10481–10490, 2019. 3
- [5] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladt: Local adversarial disentangling network for facial makeup and de-makeup, 2019. 11
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [7] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiaoshi Feng, and Shuicheng Yan. Psgan: Pose-robust spatialaware gan for customizable makeup transfer. *arXiv preprint arXiv:1909.06956*, 3, 2019. 1, 3, 6
- [8] Qiaoqiao Jin, Xuanhong Chen, Meiguang Jin, Ying Chen, Rui Shi, Yucheng Zheng, Yupeng Zhu, and Bingbing Ni. Toward tiny and high-quality facial makeup with data amplify learning. In *European Conference on Computer Vision*, pages 340–356. Springer, 2025. 2, 3
- [9] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018. 1, 3, 5, 6
- [10] Yunfan Liu, Qi Li, Qiyao Deng, Zhenan Sun, and Ming-Hsuan Yang. Gan-based facial attribute manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1
- [11] Xiongbo Lu, Feng Liu, Yi Rong, Yaxiong Chen, and Shengwu Xiong. Makeupdiffuse: a double image-controlled diffusion model for exquisite makeup transfer. *The Visual Computer*, pages 1–17, 2024. 2, 3
- [12] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 4
- [13] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 4
- [14] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6
- [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3
- [16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4
- [17] Yuhao Sun, Lingyun Yu, Hongtao Xie, Jiaming Li, and Yongdong Zhang. Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24584–24594, 2024. 2
- [18] Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, Fei Du, Weihua Chen, Fan Wang, and Yi Rong. Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models. *Advances in Neural Information Processing Systems*, 37:16016–16042, 2024. 6
- [19] Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, and Yi Rong. Content-style decoupling for unsupervised makeup transfer without generating pseudo ground truth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7601–7610, 2024. 6
- [20] Jianfeng Xiang, Junliang Chen, Wenshuang Liu, Xianxu Hou, and Linlin Shen. Ramgan: Region attentive morphing gan for region-level makeup transfer. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022. 3
- [21] Lu Xu, Heng Fan, and Jinhai Xiang. Hierarchical multi-task network for race, gender and facial attractiveness recognition. In *2019 IEEE International conference on image processing (ICIP)*, pages 3861–3865. IEEE, 2019. 8
- [22] Chenyu Yang, Wanrong He, Yingqing Xu, and Yang Gao. Elegant: Exquisite and locally editable gan for makeup transfer. In *European Conference on Computer Vision*, pages 737–754. Springer, 2022. 1, 3, 6
- [23] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5, 6, 12
- [24] Yuxuan Zhang, Lifu Wei, Qing Zhang, Yiren Song, Jiaming Liu, Huaxia Li, Xu Tang, Yao Hu, and Haibo Zhao. Stablemakeup: When real-world makeup transfer meets diffusion model. *arXiv preprint arXiv:2403.07764*, 2024. 2, 3, 6, 7

## A. Pseudo Code

We provide the pseudo-code of DreamMakeup for RGB and textual guidance in Algorithm 1. Makeup transfer based on a reference image can be easily implemented using the same method, substituting the transformation  $\mathcal{T}_{RGB}$  with  $\mathcal{T}_{ref}$  and employing warping and histogram matching algorithms instead of RGB matching.

---

### Algorithm 1 DreamMakeup with RGB and text guidance

**Input:** Source image  $\mathbf{x}_0$ , early-stop timestep  $t^* \leq T$ , RGB scaling coefficient  $0 \leq \alpha \leq 1$ , target RGB color  $\mu_{tgt}$ , reference makeup image  $\mathbf{x}_{ref}$ , textual prompts for (inversion, editing)  $C_{inv}$ ,  $\{C_{edit,s}\}_{s=1}^N$ , degree of composition  $\{\alpha_s\}_{s=1}^N$ .

**Output:** Image with makeup transformation  $\tilde{\mathbf{x}}_0$ .

```

1:  $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ 
2:
3: 1. Early-stopped DDIM inversion
4: for  $t = 1$  to  $t^*$  do
5:    $\mathbf{z}_t = \text{DDIM-Inv}(\mathbf{z}_{t-1}, t, c)$  (Sec. 4.1)
6: end for
7:  $\hat{\mathbf{z}}_0(t^*) = \frac{1}{\sqrt{\bar{\alpha}_{t^*}}}(\mathbf{z}_{t^*} - \sqrt{1 - \bar{\alpha}_{t^*}}\epsilon_\theta(\mathbf{z}_{t^*}, t^*))$ .
8:  $\hat{\mathbf{x}}_0(t^*) = \mathcal{D}(\hat{\mathbf{z}}_0(t^*))$ 
9:
10: 2. Pixel-domain Diffusion Guidance
11:  $\hat{\mathbf{x}}_{new} = \mathcal{T}_{RGB}(\mu_{src}(\hat{\mathbf{x}}_0(t^*)), \mu_{tgt}; \alpha)$ 
12:  $\tilde{\mathbf{z}}_{t^*} = \sqrt{\bar{\alpha}_{t^*}}\mathcal{E}(\hat{\mathbf{x}}_{new}) + \sqrt{1 - \bar{\alpha}_{t^*}}\epsilon(\mathbf{z}_{t^*}, t^*, c)$ 
13:
14: 3. Reverse sampling with cross attention composition
15: for  $t = t^*$  to 1 do
16:    $\tilde{\mathbf{z}}_{t-1} \leftarrow \text{ReverseDDIM}(\tilde{\mathbf{z}}_t; t, \text{Composition}(\{\alpha_s\}_{s=1}^N, \{C_{edit,s}\}_{s=1}^N))$ 
17: end for
18:  $\tilde{\mathbf{x}}_0 = \mathcal{D}(\tilde{\mathbf{z}}_0)$ 

```

---

## B. Additional analysis

We provide in-depth analysis on the core components of DreamMakeup. Additional qualitative results are also provided in Fig. 19.

### B.1. Ablation studies

**Effects of gradation smoothing.** In the generating process of eye mask, gradation smoothing is essential. Without gradation smoothing, the edges of eye masks are accentuated, resulting in an unnatural outcome. Fig. 11 demonstrates that gradation smoothing makes the edge of the eye shadow natural and realistic.



Figure 11. The effect of the gradation smoothing of eyeshadow mask. Please zoom in for detailed inspection.



Figure 12. Ablation study on DDIM inversion, coloring, and reverse sampling. Text prompts guide the harmonization of unnatural color transitions into a cohesive aesthetic style during sampling.

### B.2. LoRA variation

We mainly utilized Dreamshaper<sup>1</sup>, ArienMixXL<sup>2</sup>, and BKG1<sup>3</sup> LoRA weights. Fig. 13 shows the experimental results of using other LoRA weights. For comparison, asian beauty v2<sup>4</sup>, Korean Alike<sup>5</sup>, Asian Cute Face<sup>6</sup>, koreanDoll-Likeness v15<sup>7</sup>, PMN 2<sup>8</sup> are used. The results are made with only text guidance, where the prompts are "deep red lip" and "heavy eye makeup". The results demonstrate how diverse makeup styles can be achieved by varying LoRA weights. In this paper, we mainly leverage BKG1 LoRA which shows better identity preservation and semantic alignment.

### B.3. Additional Results

To demonstrate the robustness of our method, we test its performance on a variety of challenging conditions, including images with low resolution, occlusions, and dark lighting, etc with results shown in Fig. 16. Although our primary focus is on generating natural, daily makeup, the framework's flexibility allows it to be applied to more extreme and artistic makeup styles as well (Fig. 17). Furthermore, the user can customize the target region for color transformation, extending the application of DreamMakeup beyond facial makeup to related tasks such as hair, eyebrow, and pupil coloring (Fig. 18).

<sup>1</sup><https://civitai.com/models/4384/dreamshaper>

<sup>2</sup><https://civitai.com/models/118913/sdxl-10-arienmixxl-asian-portrait>

<sup>3</sup><https://civitai.com/models/203947/beautiful-korean-girl-bkgv1>

<sup>4</sup><https://civitai.com/models/76883/2731-pretty-asian-face-asian-beauty-faces>

<sup>5</sup><https://civitai.com/models/193777/korean-alike-by-noerman>

<sup>6</sup><https://civitai.com/models/26914?modelVersionId=32215>

<sup>7</sup><https://civitai.com/models/26124/koreandoll likeness-v20>

<sup>8</sup><https://civitai.com/models/106028/korean-beauty>

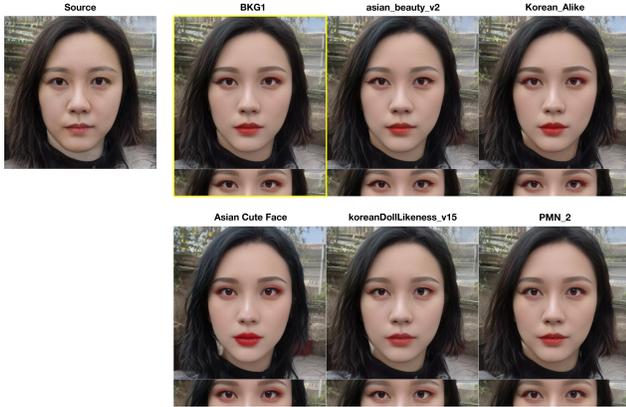


Figure 13. Results of using various LoRA weights.

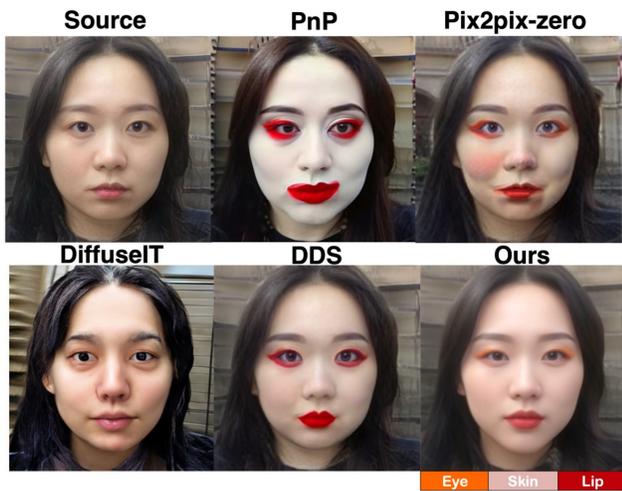


Figure 14. comparison with diffusion editing methods.



Figure 15. Comparisons of DreamMakeup with other global mobile AI makeup services.

### C. Experimental details

We use DDIM scheduler and set the early-stop inversion step ranging from  $t^* = 200$  to  $t^* = 400$ . The number of reverse steps is set to 30. LoRA scale  $s$  is set to 0.2. To

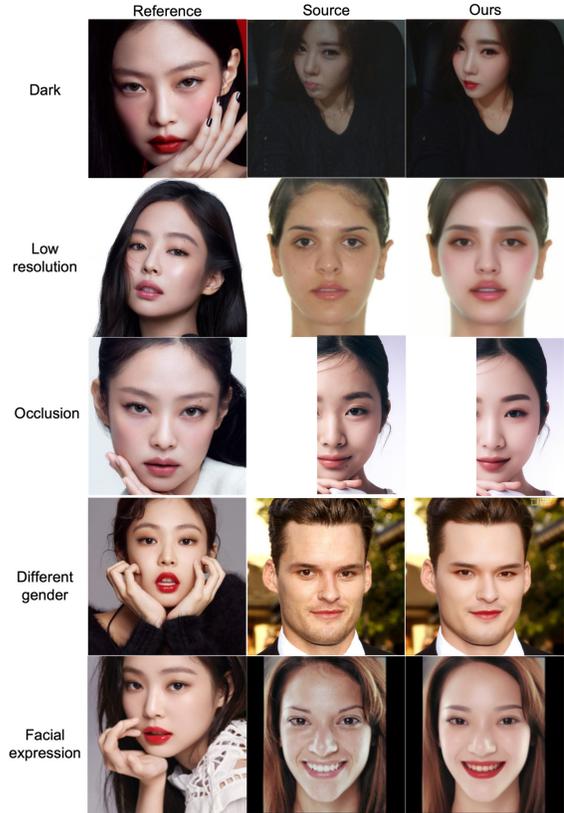


Figure 16. DreamMakeup results on extreme conditions.

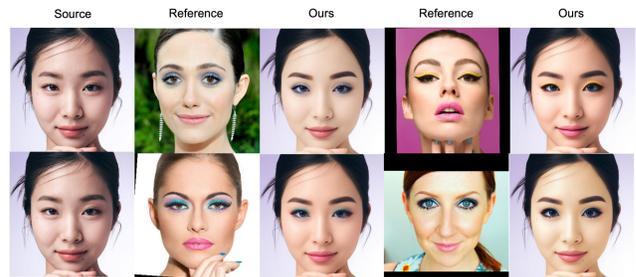


Figure 17. DreamMakeup results on extreme makeup. Reference images are from the LADN[5] dataset.

smooth eye shadow masks, we employed a cross-shaped kernel with the size of (12, 7) and performed 2 iterations of mask dilation. The textual prompts commonly used in cross attention composition are as follows:

- natural lips, natural makeup, fair skin, asian skin
- korean makeup, korean style, korean beauty, (A Classy and Cute Korean girl:1.3), cute, (Korean idol), K-pop, skm\_misoo, beautiful
- 32K, high-res, (masterpiece:1.3), best quality, 8K.HDR, smooth face, 1 girl,close up face, (photorealistic:1.6), [(detailed face:1.2):0.3]
- (Glossy lips:1.6), Gleaming lips, (fair skin:1.4), sharp focus, blusher
- (Goddess smile:1.3)

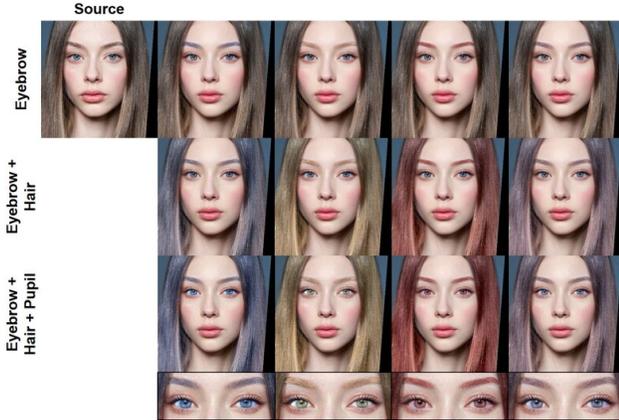


Figure 18. DreamMakeup results on RGB matching for eyebrows, pupils, and hair. We transform the color of each facial attribute within segmentation mask area.

- (worst\_quality:2.0) low quality, blur, deformed ugly, pixelated, cgi, illustration, cartoon, deformed, distorted, disfigured, poorly drawn

The directional degree of  $s$ -th composition,  $\alpha_s \leq 0$ , is assigned 0.1, 0.1, 0.3, 0.7, 0.1,  $-0.1$  for each prompt.

### C.1. LLM

To train the language model, we constructed a QnA dataset containing information matching makeup and facial attributes. The dataset consists of 460,000 pairs of questions and answers. Below is an example of the makeup dataset.

### Instruction: Which lip colors are suitable for women with the following condition? \nbronze skin, square face, angular jaw

### Response: deep red or vivid red or dark red.

As a base model, we utilized dolly-v2-3b<sup>9</sup> and fine-tuned the model for 3 epochs using the makeup dataset. To prevent the model from forgetting language proficiency during fine-tuning, we also incorporated the natural language dataset used to train this base language model. The training objective is to generate the subsequent tokens based on the tokenized instructions in an autoregressive manner.

### D. Limitations

While our proposed method offers an efficient, training-free framework for face makeup application via early-stop DDIM inversion, it requires multiple sampling timesteps to generate the final output. Since our method utilizes BiSeNet [23] for facial segmentation and a pre-trained diffusion model for the generative process, our approach may inherit the intrinsic limitations of these foundational models.

<sup>9</sup>Databricks, Free dolly: Introducing the world’s first truly open instruction-tuned llm, <https://github.com/databrickslabs/dolly>, 2023.

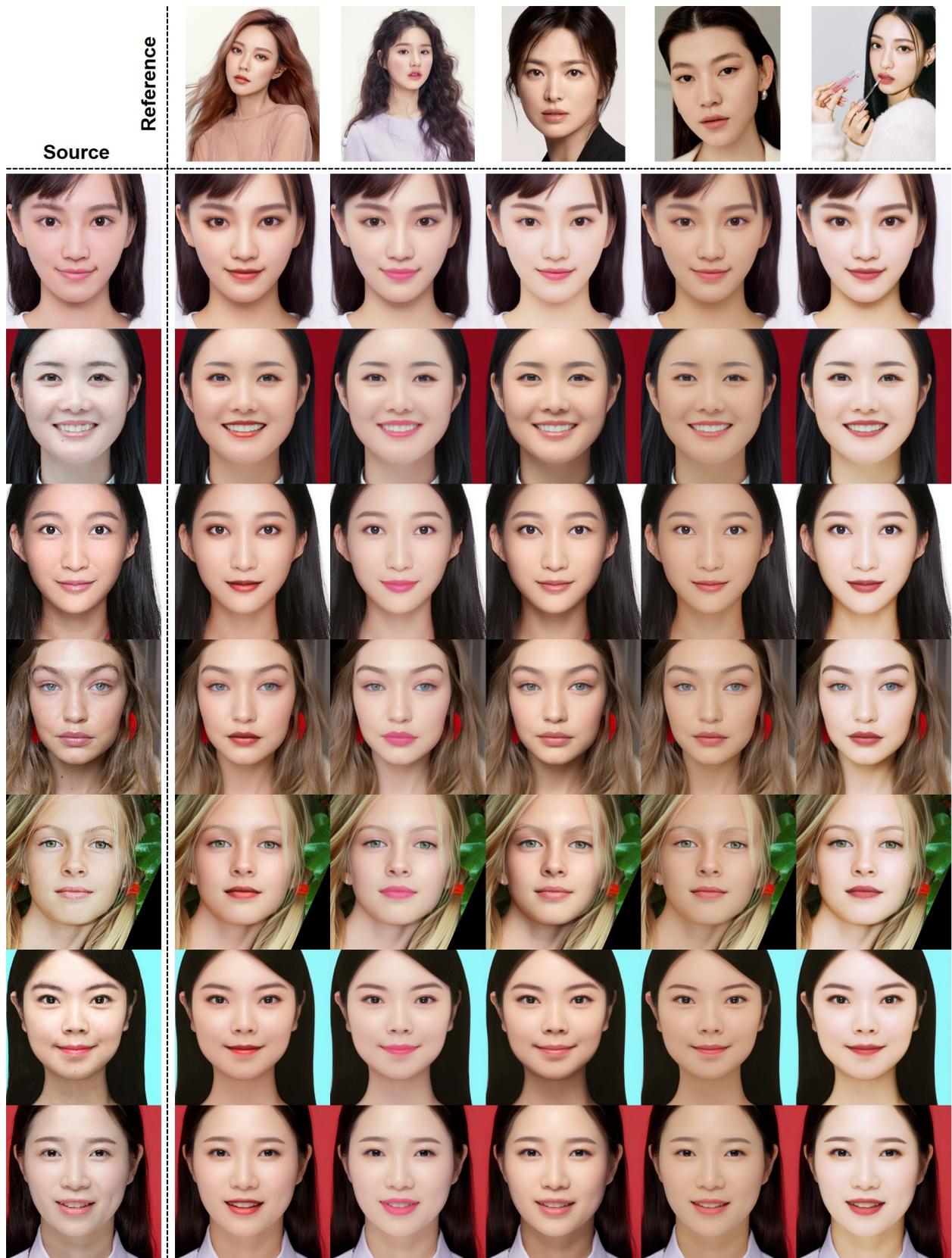


Figure 19. Addition results on makeup transfer.