MakeupMirror: Improving Facial Attribute Preservation in Diffusion Models for Makeup Transfer

Nefeli Andreou, Angel Martínez-González, Sabine Sternig, Matthieu Guillaumin,
Epameinondas Antonakos and Michael Opitz
Amazon
{nandreou, gonamang, ssternig, matthieg, antonak, micopitz}@amazon.de

Abstract

Makeup transfer models enable fun augmented reality (AR) experiences as well as virtual try-on (VTO) for online makeup shopping. While recent state-of-the-art diffusion-based solutions such as Stable-Makeup [45] dramatically improve the accuracy and realism of makeup transfer, they still face limitations in identity and skin color preservation, making production-level VTO for makeup shopping unrealistic. In this work, we propose MakeupMirror, a diffusion-based approach to makeup transfer that makes significant progress towards preserving facial features and skin tone. We introduce several technical innovations over Stable-Makeup: (1) integration of facial geometry conditioning with ControlNets to maintain facial fidelity; (2) region-specific makeup transfer control to enable precise makeup application across facial regions such as skin, eyes and lips; (3) skin tone-based makeup transfer modulation that prevent skin tone alteration in cross-subject transfer scenarios; and (4) integration of a Levenberg-Marquardt Langevin sampler to speed up inference while maintaining generation quality. Our experiments on CPM-Real, Makeup Wild, and (herein newly collected, more diverse) MakeupSelfies datasets show that MakeupMirror improves relative facial recognition similarity by $+60\%$, reduces relative skin tone difference by -50% over Stable-Makeup, with a latency of 0.7s, while achieving expert acceptance rate of 94% across core facial identity preservation criteria.

1. Introduction

Makeup transfer is the task of virtually applying the visible set of makeup products of a reference face image to a source face, while preserving its identity, see Fig. 1. Recent progress in the field has enabled several real-world applications, such as entertaining augmented reality (AR) functions, and is narrowing the gap to hyper-realistic virtual try-on (VTO) systems for beauty e-commerce, with which customers could faithfully visualize how products will render on their own unique features before they buy products. To fully close that gap, makeup transfer models still face challenges in appearance understanding and generation: makeup properties and application styles must be effectively disentangled from facial identity, transferred across subjects with different lighting, pose, expression, skin tones and facial geometries, and applied with region specific control while maintaining photorealistic quality.

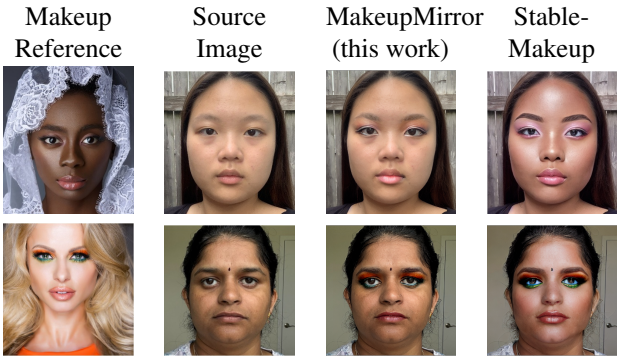


Figure 1. **Makeup transfer, qualitative comparison.** Makeup transfer is the task of applying a reference makeup (left image) to a source input image (second column). While both methods achieve photorealism and faithful makeup transfer, MakeupMirror (third column) better preserves facial attributes and skin tone of the source image compared to Stable-Makeup [45] (right image).

Traditional physically-based rendering approaches to makeup transfer [6, 10, 13, 33, 40] are fast and offer precise control but struggle to achieve photorealistic results, particularly for complex makeup looks involving multiple products of various textures. Early works in the generative AI space explored GAN-based approaches [9, 41] which offer improved quality but face inherent instability during training and lack full customization and controllability in makeup application. Following the image generation and editing domains, recent advances in makeup transfer lever-

age diffusion models [23, 30, 45], given their better controllability and improved visual fidelity. However, existing diffusion-based methods still face important limitations: unintended modifications to facial features and skin tones, as well as exaggerated and unfaithful transfer, see Fig. 1. These technical shortcomings – in particular unintentional skin tone, eye or nose shape modifications – render current approaches unsuitable for real-world e-commerce applications where fidelity, photorealism, speed, and controllability are all essential.

In this work, we introduce MakeupMirror, a diffusion-based model making significant progress in preserving facial attributes of the source image. Our improvements enable faithful makeup transfer, allowing users to apply looks with photorealistic quality, yet also speed, controllability, as well as convincing facial feature and skin tone preservation. Our approach builds upon Stable-Makeup [45], the state-of-the-art diffusion-based architecture for makeup transfer and makes the following contributions:

1. **Enhanced Facial Fidelity through Geometric Conditioning:** We integrate ControlNets [44] for Depth-Anything [42] estimation and Canny edge [2] detection maps to maintain facial structure, low-level details and overall identity during makeup transfer;
2. **Region-Specific Makeup Strength Control:** We implement adaptive control mechanisms which reduce classifier-free guidance and diffusion steps for skin regions while maintaining larger transfer for lips and eyes. This enables precise makeup application in facial regions where it is prominent while preventing facial feature transfer in other regions;
3. **Adaptive Skin Tone Preservation:** We incorporate skin tone difference detection between the reference and source images so as to automatically modulate makeup transfer intensity. When significant variations exist, this modulation prevents unwanted skin tone modifications while preserving makeup accuracy.
4. **Inference Acceleration:** We integrate a Levenberg–Marquardt Langevin sampler to accelerate inference, achieving a $2.8\times$ speed-up while maintaining makeup transfer quality, leading to a latency of 0.7s.

In the remainder of the paper, we first discuss related work (Sect. 2) then present our contributions in detail (Sect. 3). In our experiments (Sect. 4) on CPM-Real [20], Makeup Wild datasets [9], as well as on a newly collected one displaying larger diversity, which we denote as MakeupSelfies, we show that MakeupMirror improves relative facial recognition similarity by $+60\%$ and reduces relative skin tone difference by -50% compared to Stable-Makeup and leads to a 94% pass-rate in an audit conducted by beauty experts. We also conduct an ablation study and show how recent advances in diffusion sampling [36] speed-up MakeupMirror by $2.8\times$ without significant impact on quality.

2. Related Work

2.1. Diffusion Models

Diffusion models are a powerful family of generative models that synthesize high-quality images through an iterative denoising process. Following DDPM [8] which demonstrated the feasibility of recovering structured data from Gaussian noise, the extension to conditional diffusion enabled significant progress in numerous generation tasks including text-to-image [24, 26–28], text-to-video [1, 31, 32], inpainting [16], and image editing [4, 11, 34, 39, 46]. In this setting, the model predicts noise conditioned on an external signal, typically injected via cross-attention, introducing semantic control over the generative process while preserving the core denoising formulation.

Early models using conditional diffusion struggled with computational efficiency and accurate adherence to the controlling signals. Latent Diffusion Models (LDM) [27] address the high computational cost of pixel-space diffusion by performing the denoising process in a compressed latent representation of the image computed with image encoders. Classifier guidance [3] first introduced gradient-based conditioning using classifiers predictions from noisy images. Classifier-Free Guidance [7] later introduced a sampling-time technique that strengthens prompt adherence by linearly combining conditional and unconditional noise predictions without an external classifier. In particular, Stable-Diffusion [27] builds on the LDM design at scale integrating latent diffusion, cross-attention conditioning, and classifier-free guidance. Several other architectural extensions such as ControlNets [44], T2I-Adapters [19], and LoRA-based conditining [43] further introduced fine-grained structured spatial control, using signals such as pose, depth, edges, and layout, while preserving pre-trained priors. More recently, preference-based finetuning [35] and reinforcement learning [37] approaches further improved preference alignment beyond prompt application.

2.2. Makeup Transfer

Traditional makeup transfer relied on image processing and graphics techniques, including intrinsic decomposition and physically-based rendering [6, 10, 13, 33, 40]. Later, works in the space of generative AI removed explicit 3D world modeling and instead used a Generative Adversarial Networks (GAN) trained on large datasets of (*makeup*, *non-makeup*) image pairs to learn to transfer makeup [5, 9, 12, 38, 41]. Despite their large adoption, GAN-based methods lack diversity representation and often need facial region alignment, preventing their practical application in real-world scenarios and complex makeup compositions.

Building on recent image generation improvements, most works now leverage diffusion models for the makeup transfer task [14, 22, 23, 45, 47]. A core challenge ad-

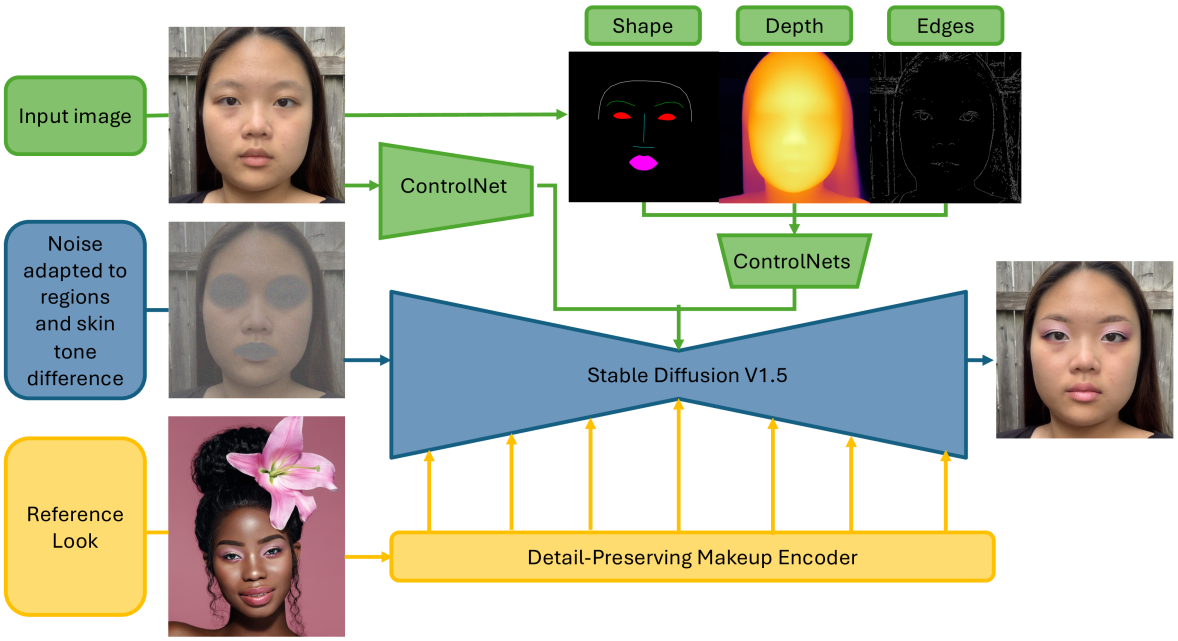


Figure 2. **Illustration of the MakeupMirror architecture.** Building upon Stable-Makeup, our approach enhances facial feature preservation by adding controller networks for Depth-Anything [42] and Canny edge [2] maps. We also adapt noise to facial segmentation regions and modulate the transfer strength based on the estimated skin tone difference between the source image and the reference look.

dressed in these methods is the disentanglement of person identity and global makeup representation. The approach presented in [14] integrates conditional diffusion with an adversarial discriminator that regularizes the denoising process, improving realism while preserving identity through spatially aligned makeup conditioning. In [30] the authors leverage 3D facial information to guide the diffusion process and improve facial geometric consistency while learning makeup representations in a self-supervised manner, addressing the lack of paired image data for training. Zhang *et al*. [45] introduced Stable-Makeup, a framework that performs makeup transfer by conditioning on facial pose to ensure preservation of facial identity while keeping makeup accuracy. In their method, makeup and pose representations are combined within the U-Net architecture using spatial-aware cross attention, allowing makeup features to be applied consistently across corresponding facial regions, e.g. eyes, lips, etc. More recently, FLUX-Makeup [47] alleviates the need for face-control modules by injecting region-aware makeup style information via LoRA modules trained on before-after makeup image pairs.

These methods exhibit undesirable failures in real-world cross-subject scenarios. In our experiments we observed that they often unintentionally alter facial features and skin tones. We argue that these limitations are likely inherited from their reliance on same-subject paired training data. Our work addresses these fundamental limitations by introducing a novel approach that preserves subject identity while achieving accurate makeup transfer. We incorporate facial geometric features and skin tone information into the diffusion process to explicitly address issues arising in cross-subject scenarios. Furthermore, we introduce region-specific strength control for precise makeup application, which not only enhances transfer accuracy but also enables efficient processing by reducing the diffusion steps.

3. Method

In this section, we present our enhanced makeup transfer approach, as illustrated in Fig. 2. We first describe the foundational architecture that builds upon Stable-Makeup [45] as the basis of our method (Sec. 3.1), then detail our technical innovations: geometric conditioning with ControlNets [44] (Sec. 3.2), region-specific makeup strength control (Sec. 3.3) and adaptive skin tone preservation (Sec. 3.4).

3.1. Background

Stable-Makeup [45] is the de facto state-of-the-art model for makeup transfer, itself building upon Stable Diffusion V1-5, which employs a LDM framework with a U-Net denoising architecture. Starting from noise $\mathbf{z_T} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ sampled from a standard Gaussian distribution, the model performs a series of T iterative denoising steps to recover a clean latent $\mathbf{z_0}$. At each timestep $t \in \{T, T -$

$1, \ldots, 1\}$, the denoising U-Net ϵ_θ predicts the noise component $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$, where $\mathbf{c}$ encodes the conditioning signals (input image and reference makeup). The sequential denoising process is guided via classifier-free guidance [7], which interpolates between conditional and unconditional noise predictions to control the fidelity–diversity tradeoff.

The Detail-Preserving (D-P) makeup encoder in Stable-Makeup extracts multi-scale makeup representations from reference images to capture intricate details such as eyelashes, eyebrows and fine lines. Formally, the encoder processes a reference makeup image I_m through a pretrained CLIP [25] visual backbone to produce multi-scale detailed makeup embeddings $E_m = \text{concat}_{k=0}^{K}(E_k, \text{dim} = 1)$, where E_k represents image embeddings at layer k in the CLIP visual backbone. The makeup embeddings are fed through a self-attention layer, which better preserves the multi-layer features compared to a linear layer, and incorporated into the U-Net via cross-attention layers. To maintain consistency with the source image, Stable-Makeup incorporates two adapted ControlNet [44] encoders: the content encoder processes the source image I_s to preserve pixel-level content and non-facial regions while the structural encoder uses dense colored lines based on facial keypoints to preserve facial structure.

3.2. Enhanced Geometric Conditioning

While the baseline Stable-Makeup architecture demonstrates effective makeup transfer capabilities, we observe undesired behaviors including unintentional skin tone modifications and facial feature alterations during the diffusion process (see Fig. 1). To address these limitations and maintain facial fidelity throughout makeup application, we employ two additional pretrained ControlNets that provide explicit and complementary geometric constraints to the diffusion U-Net. The depth conditioning utilizes the efficient Depth-Anything model [42] to predict depth maps F_D from the source image I_s, thus encoding three-dimensional facial structure information that preserves geometric relationships during the generation process at low computational cost. In parallel, a Canny edge [2] map F_E captures fine-grained contours including facial feature edges, jawlines, and significant anatomical landmarks. Both feature maps are injected into the corresponding layers of the main U-Net decoder through additive connections (*cf*. upper part of Fig. 2).

3.3. Region-Specific Strength Control

Existing makeup transfer methods using generative AI apply uniform transfer intensity across all facial regions, failing to disentangle makeup products from underlying facial characteristics, which results in over-application and unnatural skin tone modifications especially when the input and reference makeup image have significantly different skin tones. We implement region-specific makeup strength control that modulates transfer intensity based on facial segmentation, enabling differential treatment of skin regions versus feature regions (lips and eyes) within a single forward pass of the diffusion model. For segmentation, we handcraft a template mask and warp it on the input image.

Our approach operates by dynamically adjusting two key diffusion parameters: the classifier-free guidance scale w_c and the number of diffusion time-steps T_c. We define a template makeup transfer mask M which delineates facial regions where makeup should be applied, enabling spatial modulation of the diffusion process. The template mask is warped into the source image using a piecewise affine transformation computed from facial landmarks [15]. During the denoising process, the classifier-free guidance scale, w_c, controls the strength of makeup conditioning for skin regions $\tilde{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) = (1 + w_\text{c})\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}) - w_c\epsilon_\theta(\mathbf{z}_t, t, \emptyset)$ by interpolating between conditional and unconditional predictions [7], while the clamping time-steps parameter T_c constrains when makeup conditioning is applied. As seen in Fig. 2 the mask allows for higher noise to by applied on the lips and eyes, and introduces lower noise to the skin area. For time-steps $t > T_\text{c}$, the diffusion process operates without makeup conditioning on skin regions, effectively reducing transfer intensity while maintaining full conditioning for feature regions throughout all time-steps.

We define three discrete transfer strength levels $s \in \{\text{low}, \text{medium}, \text{high}\}$ with parameters $(w_c(s), T_c(s))$, where w_c denotes the classifier-free guidance scale for skin regions and T_c denotes the clamping time-steps:

$$(w_c(s), T_c(s)) = \begin{cases} (1.05, \lfloor 0.93 \times T \rfloor) & \text{if } s = \text{low} \\ (1.1, \lfloor 0.87 \times T \rfloor) & \text{if } s = \text{medium} \\ (1.6, \lfloor 0.73 \times T \rfloor) & \text{if } s = \text{high} \end{cases}$$

where T is the total number of inference steps.

3.4. Adaptive Skin Tone Preservation

Makeup transfer between individuals with significantly different skin tones presents a fundamental challenge: aggressive transfer can alter the recipient's natural skin tone, while conservative transfer may fail to capture the intended makeup aesthetic. We introduce adaptive skin tone difference detection that automatically modulates transfer intensity based on the perceptual skin tone difference between source and reference images.

Our system computes skin tone differences using the Monk scale [18], a perceptually-uniform 10-point skin tone classification system. The Monk scale values are extracted through a linear classifier trained on CLIP embeddings. Given source image I_s and reference makeup image I_m, we extract their respective Monk scale classifications μ_s and μ_m in the range $[1, \ldots, 10]$. The skin tone difference is computed as the absolute difference: $\Delta\mu = |\mu_s - \mu_m|$.

When significant skin tone variations are detected ($\Delta\mu \geq 4$), the system reduces the transfer strength by one level.

4. Experiments

In this section, we provide a thorough evaluation of our method. Public datasets, *e.g*. [9, 12, 20], tend to lack diversity in skin tone, raising bias and social fairness concerns. Consequently, we collect a new dataset with a wider variety of skin tones consisting of selfie images, serving as non-makeup source images, as well as curated stock images, serving as reference makeup images (Sec. 4.1). We show that our method achieves significantly better identity preservation compared to the state-of-the-art on public and our private benchmarks, while preserving the makeup on makeup-specific regions (Sec. 4.2). Next, we show a detailed ablation study of our design choices (Sec. 4.3). Finally, we show how we optimize the latency of our method for real-world deployment (Sec. 4.4).

4.1. MakeupSelfies Dataset Collection

To enable a fair and bias-free evaluation of our method which mimics an online shopping use-case, we curate two sets of imagery, which we denote as MakeupSelfies.

First, we recorded a set of images of faces without makeup, serving as input images. The goal is to recreate the in-store try-on experience, enabling customers to virtually try on products from the comfort of their own home. To this end, we collected about $1k$ frontal-view selfies from smartphone cameras in indoor and outdoor environments from a diverse group of people, *cf*. samples in Fig. 3a.

Second, we collected a set of visually elevated images with makeup, serving as makeup reference images. To this end, we leveraged high-quality licensed stock images. To collect these images, we used an LLM to generate ≈ 100 makeup related search queries. Next, we crawled stock websites with these search queries to get about $50k$ initial thumbnail images. However, these images can contain multiple humans, small faces, or non-frontal viewing faces. Consequently, we used heuristics to identify frontal-view face close-up images and remove images with multiple faces. Finally, as many stock images tend to show the same person, we used face-embeddings to de-duplicate identities, resulting in a shortlist of about $5.5k$ images. To ensure diversity, we used a skin tone prediction model, to group these images into monk scale skin tone buckets. Finally, human curators select a balanced set of 880 images from these skin tone buckets. Example images can be seen in Fig. 3b.

To highlight the diversity of our dataset, we ran the same skin tone analysis on public datasets [9, 12, 20] and compare the skin tone distribution (Fig. 4c) with that of the selfies (Fig. 4a) and reference looks (Fig. 4b). As shown in Fig. 4c, these public datasets tend to predominantly display individuals in lighter skin buckets, making a bias-free evaluation virtually impossible, while MakeupSelfies displays a more balanced distribution over skin tones in both subsets.

4.2. Comparison to the State-of-the-Art

In this section, we show that our method quantitatively and qualitatively outperforms the state-of-the-art Stable-Makeup [45] on two public datasets, *i.e*. CPM-Real [20], and Makeup Wild [9], as well as our MakeupSelfies dataset.

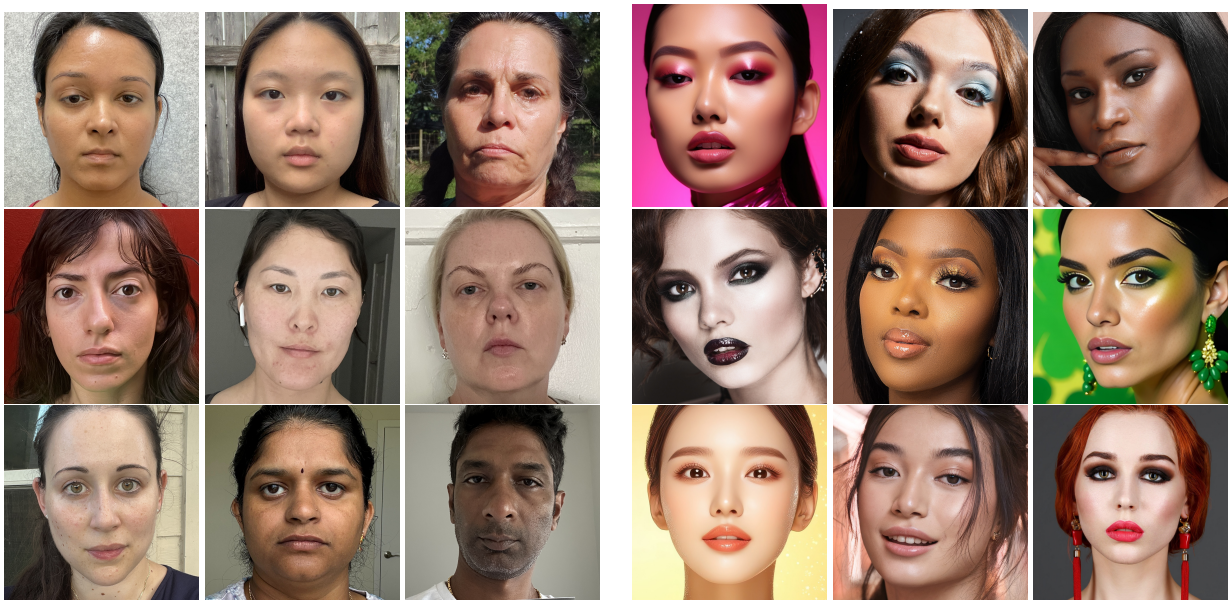
To this end, we follow existing evaluation protocols [45], and sample 2,000 makeup and non-makeup pairs for each of our datasets. As CPM-Real does not have non-makeup images, we use the non-makeup images of the Makeup Transfer [12] dataset to construct these pairs. To evaluate on MakeupSelfies, we follow a similar protocol, and sample non-makeup selfies with corresponding licensed stock makeup images for evaluation.

To measure identity preservation, we use a skin tone prediction model, as well as a face verification model [29]. More formally, given source image I_s, and transferred output image I_o, we extract their monk scale μ_s and μ_o and compute $\Delta\mu = |\mu_s - \mu_o|$. The lower the monk scale difference, the more accurately the transfer method preserves the skin tone. For face similarity, we extract FaceNet512 embeddings via DeepFace [29]. Similarly, the higher the similarity between the source image and transferred output image, the more we preserve the facial identity.

To measure makeup fidelity, we leverage DINOv2 [21] embeddings. As there is an inherent trade-off in preserving facial identity and makeup fidelity, we compute these features on the full face (Dv2-F), as well as on makeup-specific regions (Dv2-M), *i.e*. the eye region and mouth region. More specifically, we measure similarity between makeup reference image I_m and transferred makeup image I_o. To extract makeup-specific regions, we warp our makeup transfer mask (Sec. 3.3) on the makeup and output image, and compute DINOv2 features only on the non-masked regions (*i.e*. eyes, lips).

We summarize our results in Table 1. Compared to the baseline Stable-Makeup [45], our method preserves facial identity significantly better across all datasets – about -50% for ΔMonk, $+60\%$ for Face Sim, and flat for Dv2-M – while preserving makeup fidelity in makeup-specific regions. In regions such as cheeks, nose, forehead, *etc*., our method compares unfavorably to our baseline in terms of DINOv2 similarity (Dv2-F). We hypothesize that a high DINOv2 similarity in these regions typically also corresponds to saturated makeup and betrays unintentional skin tone change, and thus a loss in identity fidelity. Existing metrics conflate makeup coverage with accuracy, leaving precise, scalable makeup evaluation an open challenge.

Further, we illustrate qualitative results in Fig. 5. We see that our approach preserves the facial identity significantly better compared to Stable-Makeup. To quantify any



(a) Sample of selfies without makeup in MakeupSelfies.

(b) Sample of stock images used as reference looks in MakeupSelfies.

Figure 3. Sample images (source images and reference makeup image) from the herein collected MakeupSelfies dataset.

critical defects of our method, we also performed a manual human audit on about $2k$ samples pairs with 3 beauty expert annotators/sample. Specifically, we focus on detecting (1) defects on the lip region, (2) defects in the eye region (*e.g.* eye crease disappearing), (3) eyebrow quality (*e.g.* duplicated eyebrows), and (4) change of skin tone. Overall, we have found that our solution achieves a pass rate of 94% across these defects.

4.3. Ablation Study

Here, we show the effectiveness of each of our contributions quantitatively and qualitatively on our MakeupSelfies dataset. In Table 2 we see that introducing additional depth and Canny ControlNets already improves facial identity preservation (+0.177) as well as monk scale difference (-0.277), while maintaining makeup fidelity in makeup-specific regions. We achieve further improvements, when we add our novel skin tone preservation method, improving the face similarity (+0.203), and reducing the monk scale difference (-0.428).

As we qualitatively see in Fig. 5, our baseline, *i.e.* Stable-Makeup, tends to change the thickness of lips, and geometry of eyes and nose. While some of these changes might appear subtle, VTO applications are very personal experiences. People easily notice subtle changes to their own facial geometry, making these methods not suitable for real-world applications. By introducing depth- and edge-conditioned ControlNets, we significantly reduce these adverse qualitative effects. More prominently, the baseline tends to copy the skin tone of the makeup reference image on the source face, resulting in unnatural looks which people would not want to wear in real-life. Our second contribution addresses this problem, significantly reducing skin tone change and yielding more realistic makeup transfer. In a perceptual study, a low transfer threshold was generally found to produce the most favorable results. However, for makeup looks that include cheek-applied products such as blush, the adaptive skin tone transfer mechanism proposed in our work yields better perceptual outcomes. Notably, viewer tolerance for skin tone transfer intensity varies especially when input and output images are viewed side by side.

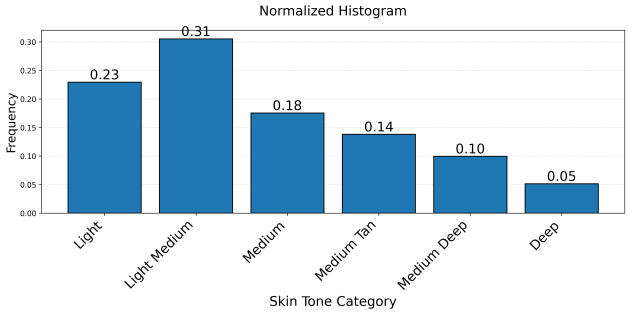
4.4. Reducing Latency for Production

Real-world VTO GenAI experiences for online shopping benefit from low latency and low operational cost. To this end, we reduce the inference time of our approach by leveraging recent advances in few-step inference approaches. More specifically, we evaluate a Latent Consistency Models LoRA (LCM) [17] and a few step Levenberg-Marquardt-Langevin (LML) sampler [36].

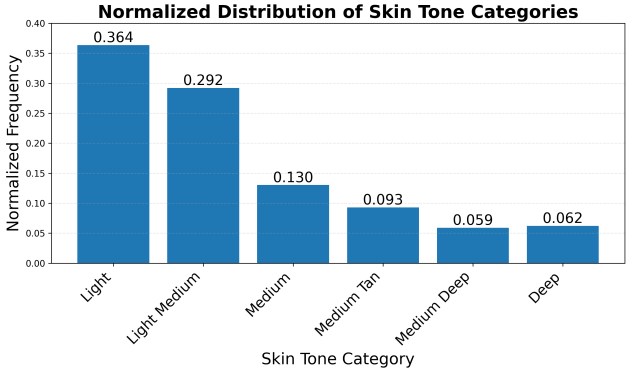
For a fair comparison, we used TensorRT (TRT) to speed up inference of the three variants of the model, and deploy them on NVIDIA L40S GPUs. We summarize the latency, and number of model invocations (NFEs) of all approaches in Table 3. TRT already reduces the latency from 3.2s to 2.0s for the base model. The LCM and LML optimizations bring significant speed-ups but also introduce approximations in inference and alter the outputs of the models. To compare the quality of our methods, we performed a manual

	CPM-Real				Makeup Wild				MakeupSelfies			
Method	Dv2 (M)	Dv2 (F)	Δ Monk	Face Sim.	Dv2 (M)	Dv2 (F)	Δ Monk	Face Sim.	Dv2 (M)	Dv2 (F)	Δ Monk	Face Sim.
StableMakeup [45]	**0.418**	**0.614**	0.701	0.523	0.249	**0.631**	0.741	0.590	**0.583**	**0.627**	1.132	0.487
MakeupMirror	0.416	0.392	**0.349**	**0.886**	**0.251**	0.503	**0.320**	**0.893**	0.5631	0.452	**0.427**	**0.867**

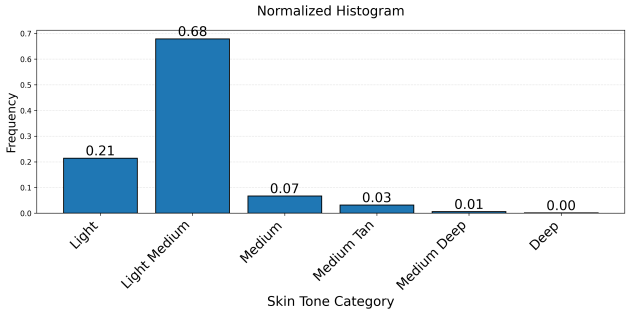
Table 1. Our method preserves facial identity significantly better compared to our baseline, as measured in monk scale difference (Δ Monk) and face similarity (Face Sim.). It preserves makeup on makeup-specific regions, as measured in DINOv2 similarity (Dv2 (M)).



(a) Skin tone distribution in our selfie image collection.



(b) Skin tone distribution in our licensed stock image collection.



(c) Skin tone distribution in public datasets [9, 12, 20].

Figure 4. Skin tone distributions of our datasets (top, middle) compared to existing public datasets (bottom). Our datasets have a more balanced skin tone distribution, enabling a fairer and bias-free evaluation.

human audit on 500 samples via pairwise preference scores, as well as measure their defects. More specifically, we compare each method with our MakeupMirror model and rate if its output is better (+1 points), equal (±0 points), or worse

Method	Dv2 (M)	Dv2 (F)	Δ Monk	Face Sim.
Baseline	**0.583**	**0.627**	1.132	0.487
+ ControlNets	0.574	0.584	0.856	0.664
+ Skin Tone	0.563	0.452	**0.427**	**0.867**

Table 2. Our contributions consistently improve makeup transfer in terms of face similarity and monk scale difference, while preserving the makeup accuracy in makeup-specific regions.

MakeupMirror	NFEs	Latency	Preference	Δ Defects
with TensorRT	30	2.0s	-	-
with LCM+TRT	4	**0.45s**	$-5/500$	+5/500
with LML+TRT	10	0.70s	$\pm\mathbf{0}/\mathbf{500}$	**+3/500**

Table 3. Evaluation of latency optimization approaches for our makeup transfer method. While not the fastest, using LML yields a 2.8× speed-up without compromising on output quality.

(-1 points) compared to MakeupMirror. Overall, the LML sampling method achieves slightly lower defects (3 / 500 vs 5 / 500) and higher preference compared to the LCM method (± 0 / 500 vs -5 / 500), with a significant 2.8× speed-up in inference time.

5. Conclusion

In this work, we present MakeupMirror, a diffusion-based method for makeup transfer that marks a significant improvement in preserving skin tone and facial features from source images over Stable-Makeup. These improvements are achieved by extending Stable-Makeup with three contributions: 1. additional geometric conditioning in the ControlNet branch, namely with depth and low-level edge maps; 2. region-specific makeup strength control by leveraging face segmentation and adjusting transfer weight and clamped time-steps based on the region; 3. modulating the transfer strength to take into account the difference of skin tone between the reference and the source images. 4. leveraging Levenberg-Marquardt-Langevin sampling to achieve 2.8× speed-up of the makeup transfer pipeline. Our experiments on public datasets and a newly collected one with more balanced skin tone distribution shows an improvement in relative facial similarity (+60%) and reduction of skin tone alteration (−60%) compared to Stable-Makeup, leading to a pass rate of 94% according to an audit by beauty experts. This high pass rate confirms the viability of the approach for a production-ready VTO experience for online beauty shopping.

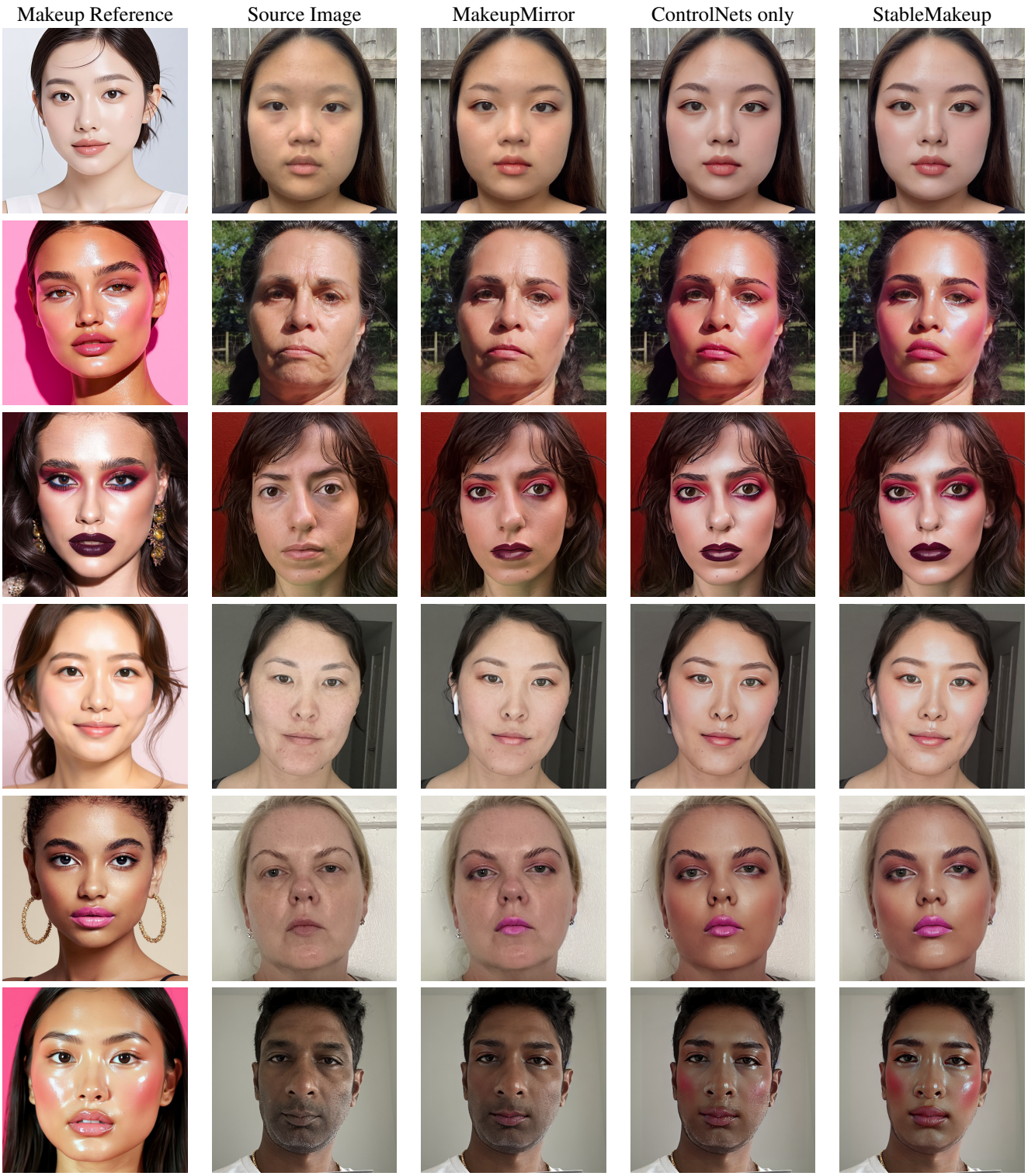


Figure 5. Qualitative results of each of our improvements compared to StableMakeup (column 5). StableMakeup suffers from changing facial geometry (*e.g.* lip thickness, nose, eye shape, *etc.*) and skin tone, especially when copying makeup between different ethnicity. Our contributions alleviate these effects in large part.

References

[1] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. 2

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 2, 3, 4

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[4] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 2

[5] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10481–10490, 2019. 2

[6] Dong Guo and Terence Sim. Digital face makeup by example. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–79, 2009. 1, 2

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 4

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[9] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5194–5202, 2020. 1, 2, 5, 7

[10] Chen Li, Kun Zhou, Hsiang-Tao Wu, and Stephen Lin. Physically-based simulation of cosmetics via intrinsic image decomposition with facial priors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(6):1455–1469, 2019. 1, 2

[11] Pengzhi Li, Qinxuan Huang, Yikang Ding, and Zhiheng Li. Layerdiffusion: Layered controlled image editing with diffusion models. In *SIGGRAPH Asia 2023 Technical Communications*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[12] 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*, page 645–653, New York, NY, USA, 2018. Association for Computing Machinery. 2, 5, 7

[13] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. "wow! you are so beautiful today!". *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(1s), 2014. 1, 2

[14] 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

[15] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. 4

[16] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022. 2

[17] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference, 2023. 6

[18] Ellis Monk. The monk skin tone scale. 2023. 4

[19] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Feng. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[20] Thao Nguyen, Anh Tran, and Minh Hoai. Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 7

[21] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 5

[22] Qihe Pan, Yiming Wu, Xing Zhao, Liang Xie, Guodao Sun, and Ronghua Liang. Supervised makeup transfer with a curated dataset: Decoupling identity and makeup features for enhanced transformation. *arXiv preprint arXiv:2602.00729*, 2026. 2

[23] Geon Yeong Park, Inhwa Han, Serin Yang, Yeobin Hong, Seongmin Jeong, Heechan Jeon, Myeongjin Goh, Sung Won Yi, Jin Nam, and Jong Chul Ye. Dreammakeup: Face makeup customization using latent diffusion models. *arXiv preprint arXiv:2510.10918*, 2025. 2

[24] 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. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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. 4

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image gener-

ation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[27] 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. 2

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[29] Sefik Ilkin Serengil and Alper Ozpinar. Boosted LightFace: A Hybrid DNN and GBM Model for Boosted Facial Recognition. *Gazi University Journal of Science*, 2026. 5

[30] Zhaoyang Sun, Shengwu Xiong, Yaxiong Chen, Fei Du, Weihua Chen, Fang Wang, and Yi Rong. Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models. *Advances in neural information processing systems*, 2024. 2, 3

[31] Kling Team. Kling-omni: A generalist generative framework for multimodal video synthesis. *arXiv preprint arXiv:2512.16776*, 2025. 2

[32] Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2406.09203*, 2024. 2

[33] Wai-Shun Tong, Chi-Keung Tang, M. S. Brown, and Ying-Qing Xu. Example-based cosmetic transfer. *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 211–218, 2007. 1, 2

[34] Linoy Tsaban and Apolinário Passos. Ledits: Real image editing with ddpm inversion and semantic guidance. *arXiv preprint arXiv:2307.00522*, 2023. 2

[35] Bram Wallace et al. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[36] Fangyikang Wang, Hubery Yin, Lei Qian, Yinan Li, Shaobin Zhuang, Huminhao Zhu, Yilin Zhang, Yanlong Tang, Chao Zhang, Hanbin Zhao, et al. Unleashing high-quality image generation in diffusion sampling using second-order levenberg-marquardt-langevin. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2, 6

[37] Yichi Wu, Hongming Zhang, Xiaohui Li, Nian Wen, Yingxue Gao, Yelong Shen, and Nan Duan. A unified framework for incorporating human feedback into text-to-image generation. *arXiv preprint arXiv:2305.06500*, 2023. 2

[38] Jianfeng Xiang, Junliang Chen, Wenshuang Liu, Xianxu Hou, and Linlin Shen. Ramgan: Region attentive morphing gan for region-level makeup transfer. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, page 719–735, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[39] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023. 2

[40] Lin Xu, Yangzhou Du, and Yimin Zhang. An automatic framework for example-based virtual makeup. In *Proceedings of the 20th IEEE International Conference on Image Processing*, pages 3206–3210, 2013. 1, 2

[41] 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, 2

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 2, 3, 4

[43] Haonan Yu, Xiangyu Chen, Kunhao Chen, Weiwei Shi, Xiaodong Xie, Yong Zhang, Tao Qin, and Tie-Yan Liu. Lora: Low-rank adaptation for fast text-to-image diffusion finetuning. *arXiv preprint arXiv:2307.02904*, 2023. 2

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 3, 4

[45] Yuxuan Zhang, Yirui Yuan, Yiren Song, and Jiaming Liu. Stablemakeup: When real-world makeup transfer meets diffusion model. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 1, 2, 3, 5, 7

[46] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6027–6037, 2023. 2

[47] Jian Zhu, Shanyuan Liu, Liuzhuozheng Li, Yue Gong, He Wang, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Dawei Leng, et al. Flux-makeup: High-fidelity, identity-consistent, and robust makeup transfer via diffusion transformer. *arXiv preprint arXiv:2508.05069*, 2025. 2, 3