From Synthetic to Real: Toward Identity-Consistent Makeup Transfer with Synthetic and Real Data

Yue Yu, Jiayu Wang, Jiajia Shi, Jingjing Chen, *Senior Member, IEEE*, and Yu-Gang Jiang, *Fellow, IEEE*

***Abstract*—Makeup transfer aims to apply the makeup style of a reference portrait to a source portrait while preserving identity and background. Early methods formulate this task as unsupervised image-to-image translation, relying on surrogate objectives and often yielding limited performance. Recent diffusion- and flow-based approaches instead exploit synthetic data for supervised training, leading to significant improvements. However, these methods still face two critical challenges: synthetic supervision frequently fails to faithfully preserve identity, and the domain gap between synthetic and real data limits generalization, resulting in degraded performance in complex real-world scenarios. To address these issues, this paper first proposes ConsistentBeauty, a novel data curation pipeline that ensures makeup fidelity and strict identity consistency within the synthesized data. Second, we propose RealBeauty, a synthetic-to-real post-training framework. Beyond supervised learning on curated synthetic data, we further adapt the model to real-world scenarios through reinforcement learning and design novel verifiable rewards tailored to the makeup transfer task. It allows the model to further benefit from real makeup patterns beyond synthetic supervision. In addition, we establish a new diverse benchmark for makeup transfer, covering a wide range of skin tones, ages, genders, poses, and makeup styles, thereby enabling a more comprehensive evaluation of model performance under diverse real-world conditions. Extensive experiments show that our method achieves state-of-the-art performance on multiple benchmarks and demonstrates clear advantages in identity preservation and performance on complex real-world cases.**

***Index Terms*—Makeup transfer, image editing, dataset curation, benchmark.**

I. Introduction

MAKEUP transfer aims to seamlessly render the makeup from a given portrait image onto a source portrait. It requires both accurate makeup transfer and faithful identity preservation, while keeping background regions unchanged. With the practical value in social media, e-commerce and the beauty industry, this task has attracted increasing attention in computer vision. However, although a non-makeup source portrait and a reference makeup image are readily available, the corresponding transferred result is generally unavailable for direct supervision. As a result, makeup transfer remains a challenging task.

Yue Yu, Jiayu Wang, and Jiajia Shi are with the College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China (e-mail: yuy24, jiayuwang25, 25213050337@m.fudan.edu.cn).

Jingjing Chen and Yu-Gang Jiang are with Institute of Trustworthy Embodied AI, Fudan University, Shanghai 200433, China (e-mail: chenjingjing, ygj@fudan.edu.cn).

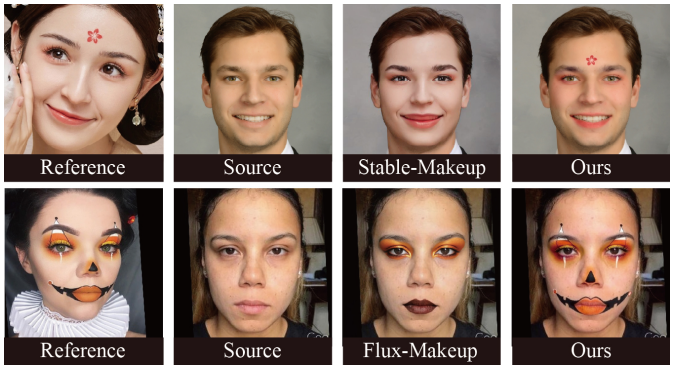

Fig. 1. Existing methods for makeup transfer usually suffer from two key limitations: **Top row**: failure to preserve source identity; **Bottom row**: degraded performance under complex makeup

Existing methods can be broadly divided into two categories. The first category treats makeup transfer as an unsupervised image-to-image translation task and commonly follows a CycleGAN-style [1] training scheme. These methods [2]–[4] adopt surrogate objectives, such as histogram-matching loss and perceptual loss, to guide model training. However, such objectives cannot effectively capture the desired makeup effects, resulting in limited makeup transfer capability. The second category focuses on synthesizing pseudo ground-truth images to enable supervised training. Early methods [5], [6] perform affine transformations and color matching to "paste" the reference makeup onto the source portrait, generating coarse-grained pseudo ground truth. With the rapid development of diffusion and flow-based generative models [7]–[9] as well as their strong performance in image editing and image-to-image translation tasks [10]–[12], more recent studies [13]–[16] employ image editing models to synthesize high-quality makeup and non-makeup pairs for supervised training. Benefiting from this, these methods achieve substantially improved transfer capability. Nevertheless, this line of work still suffers from two limitations, as shown in Fig. 1. First, these methods do not always faithfully preserve the source identity, as the models used to synthesize training data may introduce unintended changes to facial characteristics. Second, although these methods work well for simple and common makeup patterns, their performance degrades in complex real-world cases due to the gap between synthetic training data and real-world scenarios.

These observations suggest that effective makeup transfer requires addressing two issues simultaneously, i.e., (1) how to construct training data that preserves both makeup fidelity and identity consistency; and (2) how to bridge the gap between

synthetic training data and real-world scenarios. To this end, we first introduce **ConsistentBeauty**, a novel data curation pipeline. Specifically, we collect a substantial amount of non-makeup portraits and construct high-quality makeup layers covering light, heavy, and artistic styles. Rather than relying on image editing models, we employ a 3D triangulation-based method to seamlessly apply the same makeup layer to two non-makeup images with different identities. This process enables us to construct training triplets ⟨*non-makeup source, makeup reference, target image*⟩ for supervised training. In this way, our data construction pipeline ensures strict identity consistency and high makeup fidelity. To further reduce the gap between synthetic training data and real-world cases, we then propose **RealBeauty**, a synthetic-to-real post-training framework for makeup transfer. RealBeauty adopts a two-stage training strategy, where the model is first cold-started on our curated dataset and then further optimized via reinforcement learning on unpaired real-world makeup data. During the reinforcement learning stage, we design a new task-specific verifier that accurately measures the makeup fidelity. In this way, our post-training framework enables the model to jointly leverage synthetic supervision and real-world data, thereby improving real-world generalization.

In addition, we establish a new benchmark for the makeup transfer task. Existing studies mostly rely on earlier datasets [2], [3], [17] with randomly sampled pairs, which lack sufficient diversity. We curate a new benchmark containing 300 non-makeup portraits and 512 makeup images. The non-makeup images cover different skin tones, ages, genders, and poses, while the makeup images include diverse styles from light to heavy and from simple to complex. This benchmark enables a more comprehensive evaluation of models' transfer capability under various conditions.

To summarize, our key contributions are threefold:

- We propose a data curation pipeline and construct a large-scale dataset of high visual quality, strong makeup consistency and strict identity consistency for supervised training in makeup transfer. Additionally, we establish a diverse and challenging benchmark that enables comprehensive evaluation of makeup transfer models.
- We introduce a synthetic-to-real post-training framework for makeup transfer and design a novel verifiable reward, enabling the model to learn jointly from synthetic and real data. Our method achieves state-of-the-art performance on multiple benchmarks.
- Extensive experiments demonstrate the advantages of our curated dataset and the effectiveness of the proposed method.

II. Related Work

A. Makeup Transfer

Makeup transfer has attracted continuous attention owing to its practical value. Early approaches were based on GAN [18]. In the absence of real ground-truth as supervision, methods such as BeautyGAN [2], BeautyRec [4], and PSGAN [3] were trained in an unsupervised image-to-image translation manner. They typically adopted surrogate objectives such as histogram matching to compute makeup-related losses, thereby enforcing makeup consistency. EleGANt [5] and SSAT [6] constructed pseudo ground truth (PGT) using simple matching and affine transformation techniques as supervisory signals. With the paradigm shift in image generation models in recent years, more methods [13], [15], [16], [19] have adopted diffusion-based frameworks for makeup transfer. SHMT [19] decouples makeup information from human identity features and trains the model in a self-supervised manner. Stable-Makeup [13], Flux-Makeup [15], and EvoMakeup [16] employ image editing models to synthesize paired makeup and non-makeup samples for supervised training. In particular, Flux-Makeup and EvoMakeup are built upon flow-based diffusion models and leverage more powerful image editing models [10], [20] for data construction. Compared with previous methods, these models exhibit stronger makeup transfer capabilities. However, since the identity characteristics may be altered in the synthetic makeup images relative to the original non-makeup images, models trained on such data often struggle to preserve identity consistency, which limits the overall performance of the models.

B. Diffusion and Flow-based Generative Models

In recent years, diffusion models [21]–[25] have developed rapidly, demonstrating remarkable effectiveness across a wide range of generation tasks [26]–[28]. Compared with GAN-based image generation models [18], [29], [30], diffusion models exhibit better training stability and stronger generative capability. As defined in DDPM [22], diffusion models generate images by iteratively denoising a random noise, gradually transforming it into a meaningful image. Stable Diffusion [7] further advanced this paradigm by performing the noising and denoising processes in the latent space, significantly reducing computational costs. DiT [31] introduced a Transformer-based architecture that replaces the conventional U-Net used in previous diffusion models. Stable Diffusion 3 [9] and Flux integrate flow mechanisms [32], [33] and transformer-based diffusion models, further enhancing the models' image generation capabilities.

C. Reinforcement Learning for Diffusion Models

Reinforcement learning has emerged as a key technique in the post-training phase of large language models (LLMs). When equipped with appropriate rewards, it enables LLMs to better align with specific preferences or further enhance their capabilities in particular domains. In contrast to LLMs, reinforcement learning for diffusion models remains in an initial stage of development. Earlier works [34]–[36] directly adopted rewards as training objectives for supervised learning. Diffusion-DPO [37] formalizes the DPO [38] loss for diffusion models, enabling true reinforcement-style training for this paradigm. Inspired by the GRPO algorithm introduced in DeepSeekMath [39], subsequent works [40]–[42] begin exploring the application of GRPO for optimizing diffusion models. In this work, by performing reinforcement learning, the model can be trained on real data rather than solely on synthetic data.

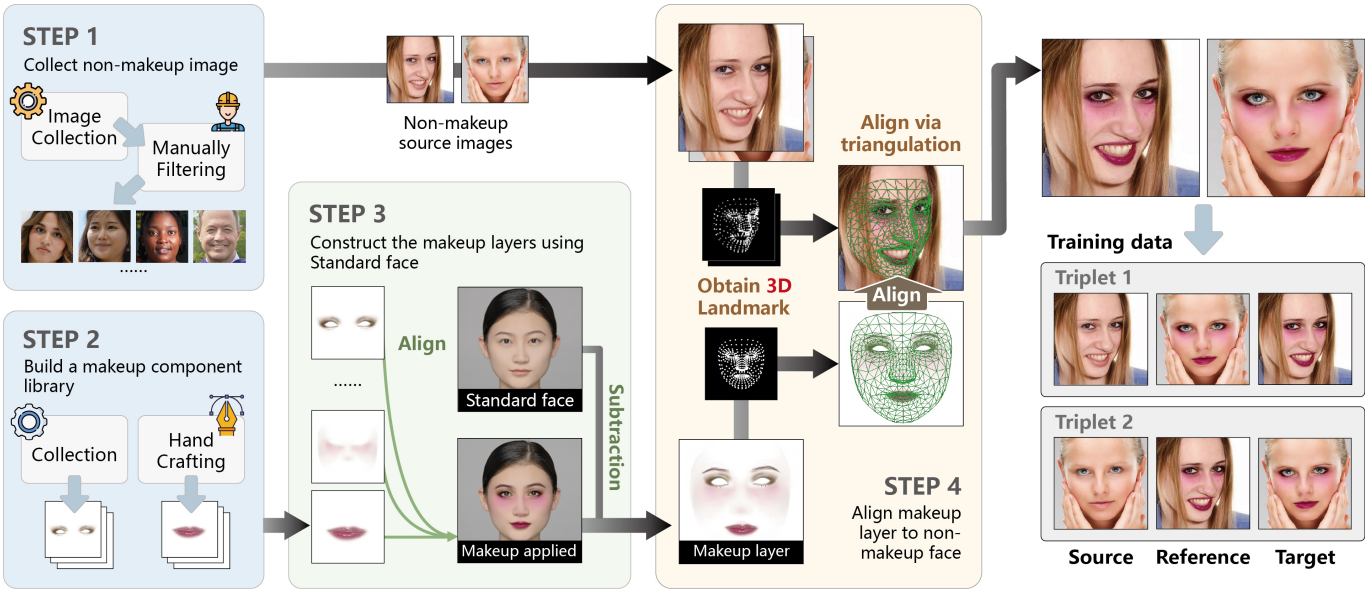

Fig. 2. Overview of the proposed ConsistentBeauty data curation pipeline. By constructing makeup layers on a standard face and transferring them to other non-makeup portraits via 3D triangulation, the proposed pipeline constructs high-quality triplets with strict identity consistency and high makeup fidelity, thereby providing reliable supervision for makeup transfer.

III. Method

In this section, we provide a detailed description of our data curation pipeline, dataset, synthetic-to-real post-training framework, and benchmark.

A. Preliminary

1) Rectified Flow: The rectified flow is a continuous normalizing flow defined by an ordinary differential equation (ODE). This ODE drives a trajectory from an initial noise distribution p_1 to the target data distribution p_0, denoted as

$$\mathrm{d}(z_t) = v_\pi(z_t, t)\mathrm{d}t, \tag{1}$$

where v_π denotes the learnable velocity field and z_t represents the intermediate state of a data point at time t. With this ODE, a data point $x_1 \sim p_1$ can be transformed into a corresponding point $x_0 \sim p_0$. For rectified flow model [32], z_t is defined as

$$z_t = (1-t)x_0 + t\epsilon, \tag{2}$$

where ϵ is standard Gaussian noise. Under this formulation, the training objective becomes

$$\mathcal{L} = \mathbb{E}_{t,p_t(x_0),p(\epsilon)}||v_\pi(z_t,t) - (\epsilon - x_0)||_2^2. \tag{3}$$

2) Reinforcement Learning for Flow-based Generative Models: As suggested in [36], the iterative denoising process of rectified flow models can be formulated as a Markov Decision Process (MDP). Since reinforcement learning requires the policy model to perform stochastic sampling, the deterministic ODE originally used in rectified flow must be reformulated. A stochastic differential equation (SDE) that preserves the marginal probability density of the original ODE throughout the entire trajectory is constructed as

$$\mathrm{d}z_t = \left(v_\pi(z_t,t) - \frac{\sigma_t^2}{2}\nabla \log p_t(z_t)\right)\mathrm{d}t + \sigma_t \mathrm{d}w, \tag{4}$$

where σ_t controls the stochasticity and $\mathrm{d}w$ represents the Wiener process.

B. Data Curation

1) Data Curation Pipeline: The goal of makeup transfer is to apply the makeup style of the reference image I_{ref} onto the person in the source image I_{src}. The generated results should faithfully reflect the reference makeup style while preserving the identity of I_{src}. To ensure both makeup fidelity and identity consistency in synthetic data, we propose a data curation pipeline, **ConsistentBeauty**, that produces high-quality triplets $\langle I_{src}, I_{ref}, I_{tgt}\rangle$, where I_{tgt} denotes the target result of the transfer, serving as a supervision signal. In each triplet, I_{tgt} shares the same identity as I_{src} and the same makeup style as I_{ref}. The curation pipeline is illustrated in Fig. 2 and consists of four steps as follows.

Step 1. We first collect non-makeup face images from the Internet and the publicly available FFHQ dataset [43]. These images cover diverse skin tones, genders, ages, lighting conditions, and poses. After manually filtering out samples with unfavorable pose, severe occlusions, or unsatisfactory image quality, we obtain 19,639 high-quality non-makeup portraits.

Step 2. We curate a high-quality makeup component library by collecting and manually crafting, consisting of seven categories: eyebrows, eyeliners, eyelashes, eyeshadows, blushes, lipsticks, and complex facial patterns.

Step 3. We construct the makeup layers. The collected makeup components are randomly combined and aligned to a non-makeup standard face I_{non}^{std} using 2D facial landmarks, resulting in a makeup-applied standard face I_{make}^{std}. The makeup layer layer_{make} is then defined as the pixel-wise residual between the makeup-applied and non-makeup standard faces in the color space, i.e., $\mathrm{layer}_{make} = I_{make}^{std} - I_{non}^{std}$. By

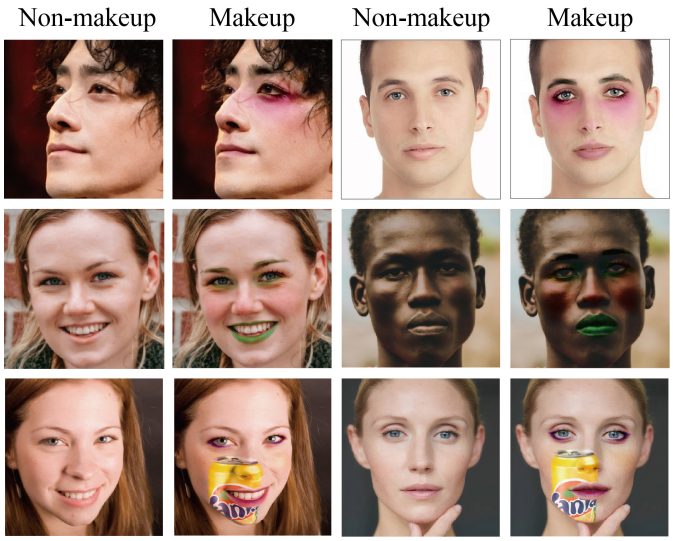

Fig. 3. Examples from our ConsistentBeauty dataset. Different non-makeup images within the same row are applied with the same makeup style, demonstrating strict identity consistency and high makeup fidelity in the constructed triplets. In particular, the sticker-style data shown in the third row is useful for improving the model's ability to transfer complex facial patterns.

leveraging the 3D facial geometry of the standard face I^{std}, the resulting makeup layer can be transferred to other non-makeup faces, producing more natural and seamless results.

Step 4. Using 3D triangulation, the constructed makeup layer is aligned to non-makeup portrait I_{src} in 3D space. Finally, two source images with the same makeup layer can serve as references for each other, forming two triplets of $\langle I_{src}, I_{ref}, I_{tgt} \rangle$.

2) Constructed Dataset: Building upon our non-makeup portrait collection and the makeup component library, we construct a total of 290,144 triplets to form the **ConsistentBeauty** dataset. Examples of the makeup-applied images we construct are shown in Fig. 3. Compared with existing data construction methods and datasets, our dataset offers the following advantages: **Strict identity consistency and high makeup fidelity**. Our approach does not alter the original identity from the source image or distort the makeup appearance. In each triplet, the identities in I_{src} and I_{tgt} are strictly consistent, while I_{ref} and I_{tgt} share the same makeup appearance. **High scalability**. Different makeup components are disentangled from each other as well as from human identities. These components can be randomly combined to generate new makeup layers, and different layers can be applied to various identities. This design allows our dataset to be efficiently expanded. **Clearer Supervision.** Compared to methods that construct only image pairs, our triplet-based formulation is more suitable for makeup transfer, as it provides clearer supervision by explicitly separating the identity and makeup information, enabling more effective learning of the task.

3) Ethics Statement: The Institutional Review Board has approved the construction of the ConsistentBeauty dataset. The dataset is intended exclusively for academic research. Access is restricted to scholarly use and requires formal application.

C. Synthetic-to-Real Post-Training Framework

To better align the model with the domain of real data and real-world scenarios, thereby further enhancing its makeup transfer capability, we propose **RealBeauty**, a synthetic-to-real post-training framework. As part of this framework, we also design a novel verifier that provides a verifiable reward for the makeup transfer task.

1) Training Procedure: Inspired by the post-training paradigm of LLMs, we first perform supervised fine-tuning (SFT) as a cold start using the proposed ConsistentBeauty dataset to obtain the model's basic capability for makeup transfer. Subsequently, reinforcement learning is applied to align the model with the domain of real data, thereby further improving its performance. This training procedure is applied to flow-based diffusion model, enabling it to learn jointly from both synthetic and real data. The model is optimized in the SFT stage with the rectified flow-matching loss

$$\begin{aligned}\mathcal{L}_{SFT} =& \mathbb{E}_{t\sim p(t), x_{tgt}, x_{src}, x_{ref}}, \\ & [||\pi(z_t, t, x_{src}, x_{ref}) - (\epsilon - x_{tgt})||_2^2]\end{aligned} \tag{5}$$

where π represents the model, t denotes the timestep, and x_{tgt}, x_{src} , x_{ref} are the latents encoded from I_{tgt}, I_{src} , I_{ref}, respectively. ϵ is standard Gaussian noise and $z_t = (1-t)x_{tgt} + t\epsilon$. After SFT cold start, reinforcement learning is performed on real data using Group Relative Policy Optimization (GRPO) [39]. As demonstrated in Flow-GRPO [40], the inference process of flow models can be formulated as a Markov Decision Process (MDP) for reinforcement learning. In the proposed framework, we utilize three rewards. To be specific, we use the proposed makeup perceptual verifier to assess makeup fidelity, and face similarity [44] to measure human identity consistency. Both of these are verifiable rewards. In addition, ImageReward [45] is adopted to assess the aesthetic quality aligned with human preference. The three scores are combined to obtain the final reward, which is used to compute the GRPO advantage A, formulated as:

$$r = \lambda_1 r_{makeup} + \lambda_2 r_{face} + \lambda_3 r_{IR}, \tag{6}$$

$$A_i = \frac{r_i - \text{mean}(r_1, r_2, \cdots, r_G)}{\text{std}(r_1, r_2, \cdots, r_G)}. \tag{7}$$

In Eq.(6), r_{makeup}, r_{face} and r_{IR} denote the scores from the makeup perceptual verifier, face similarity, and ImageReward, respectively, λ are their weights. In Eq.(7), the index i denotes the i-th sampled image within a group of G samples.

2) Makeup Perceptual Verifier: The effectiveness of reinforcement learning heavily depends on the quality of the verifiable reward. For the makeup transfer task, the key is accurately measuring the consistency between the makeup in the generated result and the reference, which serves as the reward signal. Existing works commonly evaluate this consistency by computing the cosine similarity between the CLIP [46] (or DINO [47]) embeddings of the reference and generated images. However, such metrics are often affected by non-makeup information, including human facial identity, skin tone, and background, resulting in vulnerable assessments. To address this issue, we design a task-specific makeup perceptual verifier that mitigates the influence of makeup-irrelevant

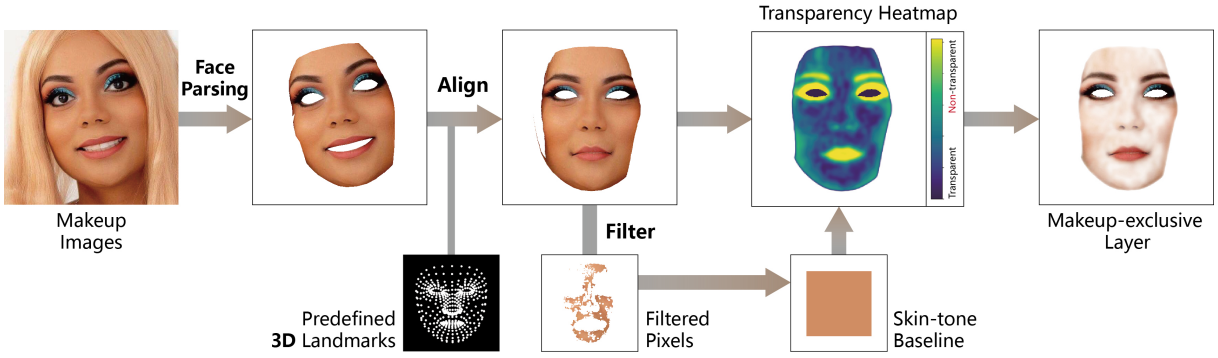

Fig. 4. Overview of the proposed Makeup Perceptual Verifier for extracting a makeup-exclusive layer from the input image. By removing background regions and reducing the influence of identity and skin tone, the verifier extracts makeup-focused representations for more accurate evaluation of makeup transfer quality.

information, thereby providing a more accurate assessment of the transfer quality.

To eliminate the influence of background area, we first perform face parsing to remove non-facial regions, obtaining R_{face}. We then align the facial region to a predefined 3D landmark template by performing facial triangulation, thereby reducing the influence of human identity information. Finally, we mitigate the impact of skin tone in R_{face} by assigning transparency to non-makeup skin areas, which requires careful design. To achieve this, we introduce a skin-tone filtering algorithm. Specifically, we first apply a Gaussian filter to remove high-frequency details such as eyes, lips, and decorative makeup patterns, and retain the smooth low-frequency components corresponding to skin tone. For the filtered pixels, we further remove the boundary chroma and luminance values in their distribution to exclude potential smooth makeup regions, such as blush. Subsequently, we compute the average hue, chroma, and luminance of the remaining pixels to define a skin-tone baseline. We assume that pixels in R_{face} with larger deviations from the skin-tone baseline are more likely to belong to makeup regions. For all pixels (not only the filtered part) in R_{face}, we compute a soft transparency map that makes non-makeup regions more transparent while preserving makeup regions. The transparency for each pixel is calculated as follows:

$$\text{Transparency} = 1 - e^{-(\lambda_h \Delta H + \lambda_c \Delta C + \lambda_l \Delta L)} \tag{8}$$

where ΔH, ΔC and ΔL denote the hue, chroma, and luminance differences from the skin-tone baseline, respectively, and λ is their weights. After the above processing, we obtain a makeup-exclusive layer, in which irrelevant information is greatly reduced. The overall process is illustrated in Fig. 4. In reinforcement learning, we compute the cosine similarity between the DINO [47] embeddings extracted from the makeup-exclusive layers of the reference and generated images, which serves as the verifiable reward.

D. Benchmark

Existing benchmarks suffer from two limits. First, both the source and reference images primarily consist of frontal portraits of light-skinned women. This makes it difficult to evaluate model performance under diverse conditions, such as those involving different skin tones or profile faces. Furthermore, the reference images primarily feature simple makeup styles, restricting the evaluation of a model's ability to transfer complex makeup patterns. To address these issues, we introduce BeautyBench, a comprehensive and high-quality benchmark dataset for makeup transfer, to systematically evaluate a model's ability to transfer diverse makeup styles. To this end, we selected 50 source images for each group defined by a combination of skin tone and gender from FFHQ [43], with 70% of the images in each group being frontal and 30% being profile. We collected reference images from both internet data and CPM-Real [48], and manually categorized them by makeup type, resulting in 176 light-makeup images, 164 heavy-makeup images, and 172 images featuring complex facial painting patterns, for a total of 512 makeup images. We then randomly paired images between the source and reference sets to construct a 1,000-sample transfer test set that covers a wide range of makeup styles and facial characteristics.

IV. Experiments

A. Experimental settings

Implementation Details. We take the FLUX.1-dev as pretrained model and trained the model with a LoRA [49] of 512 rank throughout the entire experiment. During supervised fine-tuning, the model is trained on the ConsistentBeauty dataset for 20,000 steps using 4 A100 GPUs, with a batch size of 8 per GPU and a learning rate of 1e-5. In the reinforcement learning stage, we utilize real makeup images collected from the internet and combine the reward signals from makeup perceptual verifier, face similarity, and image reward [45] as the final reward, with equal weights assigned to each. The model undergoes 240 iterations of training on 8 A100 GPUs with a batch size of 1 per GPU and a learning rate of 2e-6. Inference was performed with a step size of 25 and a guidance scale of 4.

Evaluation Datasets. In addition to evaluating on BeautyBench, we assess the methods on two makeup transfer datasets—MT [2] and LADN [17] following existing works. For each dataset, 1,000 source-reference image pairs were randomly selected to perform makeup transfer, resulting in

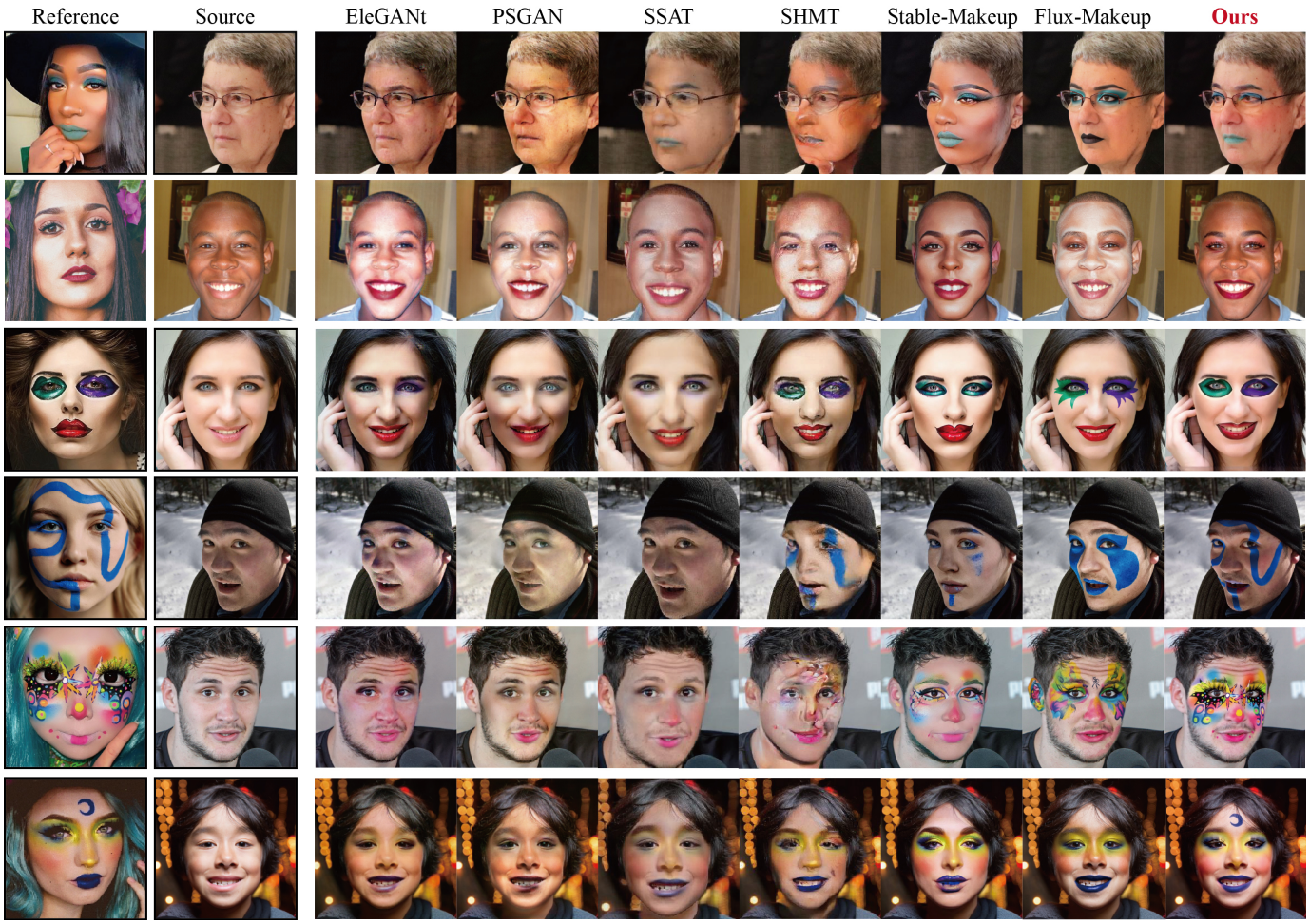

Fig. 5. Qualitative comparison with representative makeup transfer methods. Existing methods mainly exhibit two limitations: (1) insufficient identity consistency with the source image. For example, in the first three rows, some methods show noticeable identity shift toward the reference. (2) limited transfer capability in complex cases. In the last three rows, most methods struggle to accurately reproduce complex makeup patterns. In contrast, our method achieves better identity consistency and higher makeup fidelity.

a total of 3,000 transfer results per method for comprehensive evaluation.

Evaluation Metrics. Following previous works, we first employ several standard automatic metrics to evaluate the performance of makeup transfer methods from different perspectives. To assess makeup fidelity, we compute the cosine similarity between the generated and reference images within the CLIP [46] embedding space, referred to as the CLIP-I score. Following PhotoMaker [50], we utilize Face similarity metric to measure the ID consistency, which calculates the cosine similarity of face embedding [44] between source and transferred results. In addition, we use L2M to measure background preservation over non-facial regions defined by facial parsing [51].

Since makeup transfer requires both makeup fidelity and identity consistency, achieving high scores in one metric while failing in the other cannot be considered a successful transfer. Therefore, in addition to the aforementioned metrics, we further use the product of CLIP-I and Face Sim, denoted by $C \times F$, for a more comprehensive evaluation. As both CLIP-I and Face Similarity are cosine-based similarity measures, they share the same scale, allowing their product to be directly used as a unified score. A high score can only be achieved when both makeup fidelity and identity consistency are well preserved.

B. Qualitative Analyses

We compare our method with six representative makeup transfer methods, as shown in Fig. 5, including EleGANt [5], PSGAN [3], SSAT [6], SHMT [19], Stable-Makeup [13], and FLUX-Makeup [15]. The three GAN-based models are largely limited to transferring basic makeup styles and perform weakly on artistic, face-painting-style makeup. In addition, their background preservation is generally unsatisfactory. The diffusion-based models exhibit stronger transfer capability than the GAN-based approaches: they show some ability to handle face-painting-style makeup but still fail to accurately generate the reference painting patterns. They are also inadequate in preserving identity consistency. Specifically, Stable-Makeup tends to shift generated facial features and expressions toward the reference image, resulting in weak identity consistency with the source. SHMT produces severely distorted faces when there are significant differences between the source

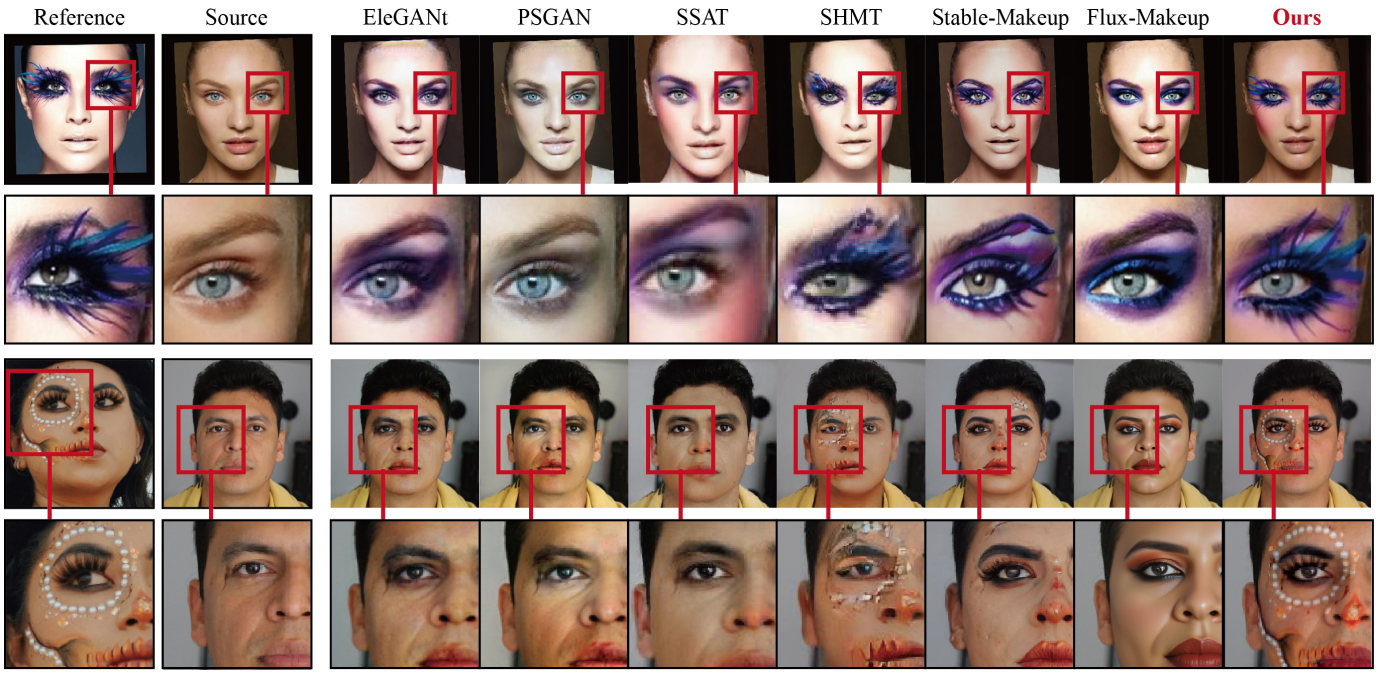

Fig. 6. The zoomed-in regions highlight fine-grained details, such as decorative elements and intricate makeup patterns. The close-up comparisons show that our method produces realistic and high-quality details. These details are highly consistent with the reference makeup.

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT MAKEUP TRANSFER METHODS ACROSS THREE DATASETS. **RED BOLD** VALUES INDICATE THE BEST RESULTS, AND BLUE UNDERLINED VALUES INDICATE THE SECOND-BEST RESULTS. C AND F REFER TO THE CLIP-I SCORE AND FACE SIMILARITY. $C \times F$ DENOTES THE PRODUCT OF THESE TWO METRICS.

	MT				LADN				BeautyBench			
Method	**L2M↓**	**C↑**	**F↑**	**C×F ↑**	**L2M↓**	**C↑**	**F↑**	**C×F ↑**	**L2M↓**	**C↑**	**F↑**	**C×F ↑**
CPM	0.077	0.618	0.621	0.384	0.136	0.647	0.589	0.381	0.177	0.503	0.509	0.256
EleGANT	0.117	0.603	0.861	0.519	0.067	0.582	0.839	0.488	0.090	0.440	0.833	0.367
PSGAN	0.114	0.596	0.836	0.498	0.061	0.565	0.830	0.469	0.082	0.431	0.824	0.355
SSAT	0.183	0.592	0.788	0.466	0.218	0.609	0.769	0.468	0.190	0.471	0.731	0.344
SHMT	0.060	0.643	0.493	0.317	0.060	0.702	0.416	0.292	0.066	0.508	0.363	0.184
Stable-Makeup	0.046	**0.708**	0.579	0.410	0.048	**0.773**	0.528	0.408	0.054	**0.656**	0.374	0.245
Flux-Makeup	0.033	0.632	**0.901**	0.569	**0.035**	0.695	**0.859**	0.597	0.040	0.531	0.784	0.416
Ours	**0.026**	0.644	**0.901**	**0.580**	0.062	0.706	0.853	**0.602**	**0.028**	0.541	**0.869**	**0.470**

and reference, leading to visually unsatisfactory results. Flux-Makeup exhibits inadequate transfer capability when faced with complex makeup references, failing to preserve the structure of facial painting patterns, which leads to inaccurate transfer results. In contrast, our model preserves the facial features, expressions, and skin tone of the source image while faithfully transferring the makeup patterns from the reference, yielding realistic, natural results across diverse styles. To provide a more fine-grained comparison, we provide zoomed-in views of specific regions, as illustrated in Fig. 6. Compared with other methods, our method more faithfully captures the texture and fine details of the decorative elements around the eyes in the first row, and more accurately renders the intricate patterns of complex makeup in the second row.

C. Quantitative Evaluations

To comprehensively evaluate the makeup transfer performance of different methods, we compare them using four quantitative metrics: L2M, CLIP-I, Face Similarity, and $C \times F$. The quantitative results are reported in Table I. As shown in Table, our method achieves the highest $C \times F$ on all three benchmarks, demonstrating superior overall transfer performance. Additionally, we achieve the best L2M and Face Similarity on both MT and BeautyBench, which intuitively demonstrates that the proposed method excels in both identity consistency and background preservation. As for CLIP-I, the proposed method attains the second-highest score across all three benchmarks, further indicating strong makeup transfer capability.

Furthermore, to more intuitively illustrate the joint per-

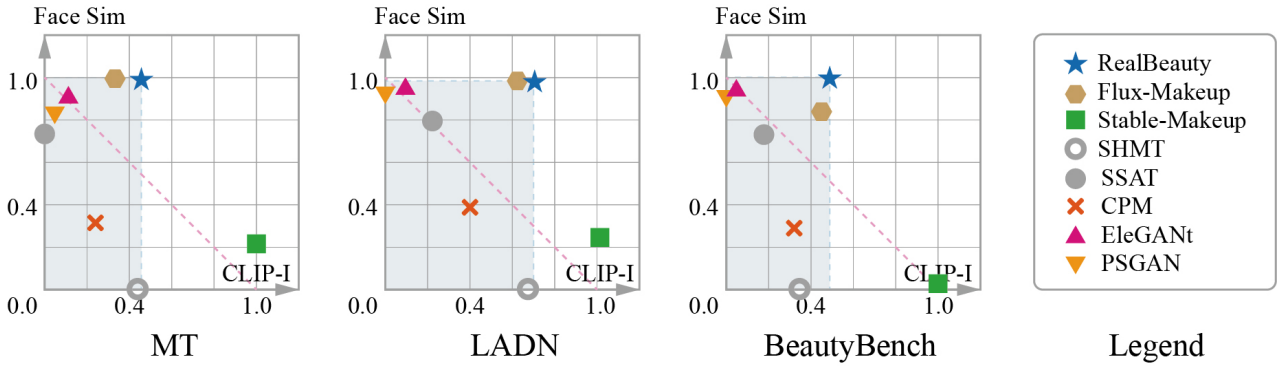

Fig. 7. Visualization of different makeup transfer methods in the two-dimensional space of min-max normalized CLIP-I and Face Similarity. For each metric, min-max normalization is applied across all methods to linearly map scores into $[0, 1]$ while preserving their relative positions. Methods with high values on both metrics occupy a larger rectangular area in this space, indicating stronger joint performance in makeup fidelity and identity consistency.

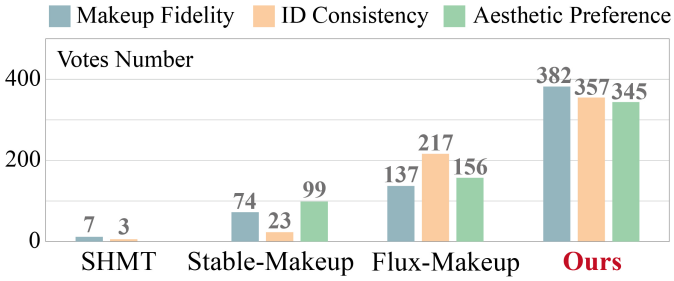

Fig. 8. The results of the user study comparing our method with SHMT, Stable-Makeup, and Flux-Makeup across three evaluation aspects: makeup fidelity, identity consistency, and aesthetic preference.

formance of different methods, we visualize the min-max normalized CLIP-I and Face Similarity in a two-dimensional space, as shown in Fig. 7. The distribution of all methods on the plot reveals three distinct patterns. The first pattern, including Stable-Makeup [13] and SHMT [19], lies closer to the x-axis and exhibits high CLIP-I but relatively low Face Similarity, indicating that these methods struggle to maintain identity consistency. The second pattern, including EleGANt [5], PSGAN [3] and SSAT [6], lies closer to the y-axis and shows higher Face Similarity but lower CLIP-I, suggesting that these methods make minimal changes to the original portrait and fail to achieve effective makeup transfer. The third pattern, including Flux-Makeup [15] and the proposed method, lies in the upper-right region across all benchmarks, suggesting that these methods achieve a more balanced performance, effectively preserving both makeup fidelity and identity consistency.

From both the table and the figure, it is evident that the proposed method and Flux-Makeup exhibit stronger performance. This is because both methods are built on the more powerful Flux.1 series [10]. The two methods show comparable performance on MT and LADN. However, on BeautyBench, which is more challenging, our method demonstrates a clear advantage in both Face Similarity and $C \times F$. This indicates that the proposed method performs better in handling more diverse and complex cases.

D. User study

In the user study, we compared our method with three representative makeup transfer approaches: SHMT [19], Stable-Makeup [13], and Flux-Makeup [15]. Participants were shown four results from each method for every transfer task and were asked to answer three questions: (1) Which image best matches the makeup style in the reference image? (2) Which image best preserves the identity of the source image? (3) Which image is the most aesthetically pleasing? These questions were designed to evaluate makeup fidelity, identity consistency, and aesthetic preference, respectively. To ensure objectivity, the method names were concealed, and the order of images was randomized.

A total of 30 participants assessed the results of the four methods on 20 randomly selected makeup transfer tasks from BeautyBench, resulting in 1,800 valid votes. Before participation, all participants were informed about the purpose and procedure of the user study, and their informed consent was obtained. As shown in Fig. 8, our method garnered the highest number of votes across all three aspects, significantly outperforming the other methods in every category. These results underscore the clear superiority of our method, demonstrating its ability to excel in all key aspects of makeup transfer.

E. Ablation Study

As shown in Table II, the fully trained model (SFT + RL) achieves overall comparable quantitative performance to the SFT-only model across all three benchmarks. Although no significant numerical improvement is observed, the RL training brings notable visual enhancements to the generated results, as shown in Fig. 9. In the first row, the fully trained model produces a more pronounced makeup effect, demonstrating better fidelity in the transfer of makeup details. The second and third rows show more complex cases. RL training helps the model avoid incomplete or failed edits, improving the transfer capability. This demonstrates the effectiveness of the proposed framework in further enhancing the model's ability for complex scenarios.

TABLE II
ABLATION STUDY OF THE REINFORCEMENT LEARNING STAGE ON MT, LADN, AND BEAUTYBENCH. L2M DENOTES BACKGROUND PRESERVATION, WHILE C AND F DENOTE THE CLIP-I SCORE AND FACE SIMILARITY, RESPECTIVELY. $C \times F$ REPRESENTS THE PRODUCT OF THESE TWO METRICS. THE RESULTS INDICATE THAT THE RL STAGE MAINTAINS PERFORMANCE COMPARABLE TO THE SFT-ONLY BASELINE ACROSS ALL THREE BENCHMARKS.

	MT				LADN				BeautyBench			
Method	**L2M↓**	**C↑**	**F↑**	**C×F ↑**	**L2M↓**	**C↑**	**F↑**	**C×F ↑**	**L2M↓**	**C↑**	**F↑**	**C×F ↑**
SFT	0.025	0.644	0.917	0.591	0.062	0.707	0.868	0.614	0.027	0.534	0.891	0.476
SFT+RL	0.026	0.644	0.901	0.580	0.062	0.706	0.853	0.602	0.028	0.541	0.869	0.470

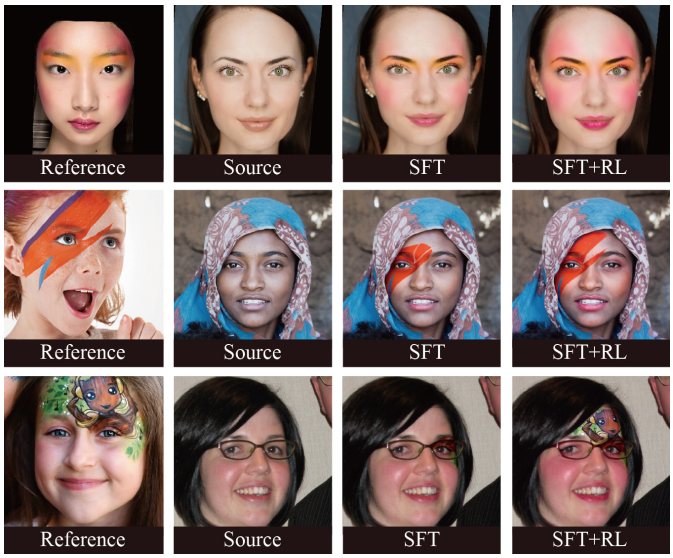

Fig. 9. Qualitative comparison between the SFT-only and SFT+RL models. The RL model mitigates the insufficient transfer issue of the SFT-only model, leading to improved visual quality in makeup transfer.

F. Benchmark Analysis for Makeup Transfer

Table III reports the benchmark statistics of MT, LADN, and BeautyBench with respect to skin tone, gender, and makeup style. In terms of skin-tone distribution, BeautyBench shows the highest skin-tone distribution entropy of 1.951 bits, indicating greater diversity. Regarding gender distribution, it presents a balanced split, with male and female samples each accounting for 50%, whereas the other two benchmarks are dominated by female subjects. In terms of makeup style, we categorize the makeup images into two types: normal makeup and complex makeup. Complex makeup includes intricate patterns, face painting, and facial decorations, while normal makeup refers to more conventional styles without such fancy designs. Among the three benchmarks, BeautyBench contains a considerably larger proportion of complex makeup samples. These statistics show that BeautyBench provides a more diverse and challenging benchmark for evaluating makeup transfer methods, allowing it to better reflect the real-world performance of makeup transfer methods.

Additionally, as illustrated in Table I, compared to MT and LADN, existing methods show noticeably lower scores in CLIP-I, Face Similarity, and $C \times F$ on the proposed benchmark. This further demonstrates that the benchmark presents more challenging tasks, pushing the model to its performance boundaries.

TABLE III
BENCHMARK STATISTICS OF MT, LADN, AND BEAUTYBENCH, SHOWCASING THE DIVERSITY AND COMPREHENSIVENESS OF OUR BENCHMARK.

Benchmarks		**MT**	**LADN**	**BeautyBench**
Skin Tone	Distribution Entropy↑	1.185 bits	1.449 bits	**1.951** bits
Gender	Male	22.4%	11.7%	50%
	Female	77.6%	88.3%	50%
Makeup Style	Normal	98.52%	80%	66.4%
	Complex	1.48%	20%	33.6%

V. CONCLUSION

In this work, we propose a novel data curation pipeline for producing high-quality supervised training data and construct a new dataset, ConsistentBeauty. We further introduce RealBeauty, a synthetic-to-real post-training framework for makeup transfer, enabling models to learn effectively from both synthetic and real data, thereby enhancing their capability. In addition, we establish a more challenging benchmark that provides a more comprehensive evaluation of this task. Our method achieves state-of-the-art performance on multiple datasets, and the experimental results validate the effectiveness of our approach.

REFERENCES

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
[2] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 645–653.
[3] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. 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*, 2020, pp. 5194–5202.
[4] Q. Yan, C. Guo, J. Zhao, Y. Dai, C. C. Loy, and C. Li, "Beautyrec: Robust, efficient, and component-specific makeup transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1102–1110.
[5] C. Yang, W. He, Y. Xu, and Y. Gao, "Elegant: Exquisite and locally editable gan for makeup transfer," in *European conference on computer vision*. Springer, 2022, pp. 737–754.
[6] Z. Sun, Y. Chen, and S. Xiong, "Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 36, no. 2, 2022, pp. 2325–2334.
[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[8] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[9] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[10] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser *et al.*, "Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space," *arXiv preprint arXiv:2506.15742*, 2025.

[11] S. Zou, Y. Huang, R. Yi, C. Zhu, and K. Xu, "Cyclediff: Cycle diffusion models for unpaired image-to-image translation," *IEEE Transactions on Image Processing*, 2026.

[12] T. Xia, Y. Zhang, T. Liu, and L. Zhang, "Consistent image layout editing with diffusion models," *IEEE Transactions on Image Processing*, vol. 34, pp. 6978–6992, 2025.

[13] Y. Zhang, Y. Yuan, Y. Song, and J. 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*, 2025, pp. 1–9.

[14] X. Yang, S. Ueda, Y. Huang, T. Akiyama, and T. Taketomi, "Ffhq-makeup: Paired synthetic makeup dataset with facial consistency across multiple styles," *arXiv preprint arXiv:2508.03241*, 2025.

[15] J. Zhu, S. Liu, L. Li, Y. Gong, H. Wang, B. Cheng, Y. Ma, L. Wu, X. Wu, D. Leng *et al.*, "Flux-makeup: High-fidelity, identity-consistent, and robust makeup transfer via diffusion transformer," *arXiv preprint arXiv:2508.05069*, 2025.

[16] H. Wu, Y. Fu, Y. Li, Y. Gao, and K. Du, "Evomakeup: High-fidelity and controllable makeup editing with makeupquad," *arXiv preprint arXiv:2508.05994*, 2025.

[17] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "Ladn: Local adversarial disentangling network for facial makeup and de-makeup," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2019, pp. 10 481–10 490.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[19] Z. Sun, S. Xiong, Y. Chen, F. Du, W. Chen, F. Wang, and Y. Rong, "Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 16 016–16 042, 2024.

[20] P. Wang, Y. Shi, X. Lian, Z. Zhai, X. Xia, X. Xiao, W. Huang, and J. Yang, "Seededit 3.0: Fast and high-quality generative image editing," *arXiv preprint arXiv:2506.05083*, 2025.

[21] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[24] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[26] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2308.06721

[27] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[28] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 392–18 402.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.

[31] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[32] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[33] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[34] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, "Aligning text-to-image models using human feedback," *arXiv preprint arXiv:2302.12192*, 2023.

[35] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee, "Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79 858–79 885, 2023.

[36] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," *arXiv preprint arXiv:2305.13301*, 2023.

[37] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, "Diffusion model alignment using direct preference optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8228–8238.

[38] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[39] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[40] J. Liu, G. Liu, J. Liang, Y. Li, J. Liu, X. Wang, P. Wan, D. Zhang, and W. Ouyang, "Flow-grpo: Training flow matching models via online rl," *arXiv preprint arXiv:2505.05470*, 2025.

[41] Z. Xue, J. Wu, Y. Gao, F. Kong, L. Zhu, M. Chen, Z. Liu, W. Liu, Q. Guo, W. Huang *et al.*, "Dancegrpo: Unleashing grpo on visual generation," *arXiv preprint arXiv:2505.07818*, 2025.

[42] J. Li, Y. Cui, T. Huang, Y. Ma, C. Fan, M. Yang, and Z. Zhong, "Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde," *arXiv preprint arXiv:2507.21802*, 2025.

[43] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[44] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[45] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15 903–15 935, 2023.

[46] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[47] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[48] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 13 305–13 314.

[49] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[50] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 8640–8650.

[51] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

VI. Biography Section

Yue Yu received the Bachelor's degree and M.S. degree from the Chinese University of Hong Kong, Hong Kong, China, in 2019 and Fudan University, Shanghai, China, in 2024. He is currently pursuing his Ph.D. degree in the College of Computer Science and Artificial Intelligence, Fudan University. His research interests include image generation and image editing.

Jiayu Wang Jiayu Wang received the B.E. degree from Northwest University, Xi'an, China, in 2022. She is currently pursuing her Ph.D. degree in the College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China. Her research interests include image generation and image editing.

Jiajia Shi received the B.S. degree in computer science and technology from Shandong University, Qingdao, China, in 2025. She is currently pursuing the master's degree in the College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China. Her research interests include multimedia forensics and security, with a specific focus on multimodal forgery detection and deepfake identification.

Jingjing Chen (Senior Member, IEEE) is now an Associate Professor at the Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China. Before joining Fudan University, she was a postdoc research fellow at the School of Computing in the National University of Singapore. She received her Ph.D. degree in Computer Science from the City University of Hong Kong in 2018. Her research interests lie in the areas of robust AI, multimedia content analysis, and deep learning. She received multiple prestigious recognitions, including the IEEE Multimedia Rising Star Runner Up Award (2023), the ACM SIGMM Rising Star Award (2024).

Yu-Gang Jiang (Fellow, IEEE) received the Ph.D. degree in Computer Science from City University of Hong Kong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York, during 2009-2011. He is currently a Distinguished Professor at Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision, embodied AI and trustworthy AI. His research has led to the development of innovative AI tools that have been used in many practical applications like defect detection for high-speed railway infrastructures. His open-source video analysis toolkits and datasets such as CU-VIREO374, CCV, THUMOS, FCVID and WildDeepfake have been widely used in both academia and industry. He currently serves as Chair of ACM Shanghai Chapter and Associate Editor of several international journals. For contributions to large-scale and trustworthy video analysis, he was elected to Fellow of IEEE, IAPR and CCF.