IPMT: An Identity-Preserving Makeup Transfer Model

Haodong Song[1[0009-0007-5492-0995]],Yinglin Zheng[1[0000-0003-4671-6111]],Yuxin Lin[1[0009-0003-0252-7789]], Pingping Gu[2[0009-0005-7626-3845]],Wangzheng Shi[1[0009-0008-4049-3539]],Wentao Chen[3*[0009-0009-9913-3279]],Ruolong Ma[3[0009-0007-0815-5348]],Jianan Lin[4[0009-0002-0981-6877]] and Ming Zeng[1[0000-0002-5056-0706]]

[1] School of Informatics, Xiamen University, China,
[2] Xiamen University Tan Kah Kee College, China
[3] China Academy of Information and Communications Technology, China
[4] Xiamen Aerospace Siert Robot System Company Limited, China
chenwentao@caict.ac.cn

Abstract. Makeup transfer serves as an important application in the field of image processing.Makeup transfer aims to transfer a reference makeup style to a source image, thereby synthesizing a high-fidelity facial image. However, when processing heavy or artistic makeup, existing generative methods frequently corrupt the identity-specific geometric features of the face. To address this issue, we propose an Identity-Preserving Makeup Transfer (IPMT) model, which is implemented via an end-to-end self-supervised framework. Specifically, we construct a multi-stream encoding pipeline and integrate a Dynamic Fusion Block to perform pixel-level adaptive fusion, achieving rigorous alignment between makeup semantics and target spatial geometry. Furthermore, to effectively prevent structural deformation without relying on pseudo-paired data, a one-step denoising approximation strategy is introduced during the training phase. By incorporating a parsing consistency loss and a 3D vertex loss during the early timesteps of generation, our method imposes explicit constraints on both the 2D planar layout and the 3D topological structure of the face. Extensive experiments demonstrate that the proposed IPMT significantly mitigates the identity inconsistency problem while ensuring high-fidelity makeup transfer.

Keywords: Makeup Transfer,Diffusion Model,Identity Preservation,Computer Vision.

1 Introduction

Makeup transfer represents a highly challenging conditional image synthesis task within the field of image processing. It can be formulated as a controllable generation task where the objective is to synthesize high-fidelity facial images that faithfully preserve the structural identity of the source input while effectively injecting the makeup semantics from a reference image.

The technique of makeup transfer has found significant applications in various fields, such as virtual try-on, social media filters, and film production, demonstrating vast developmental potential.In recent years, the continuous evolution of generative models has provided strong momentum for the progress of makeup transfer

technologies. From early Generative Adversarial Network (GAN)-based methods such as BeautyGAN [9] and PSGAN [8], to mainstream diffusion models like SHMT [17] and Stable-Makeup [24], related approaches have achieved remarkable improvements in both the realism of generated images and the accuracy of makeup restoration. Relying on their powerful generative capabilities and stable training processes, diffusion models have emerged as the mainstream approach for makeup transfer, significantly outperforming traditional GANs in handling complex texture details.

Despite substantial progress in related research, existing methods consistently struggle to balance makeup integrity and identity consistency when processing heavy or artistic makeup. The core bottleneck of this issue lies in the fact that, in the absence of pixel-level ground-truth supervision, heavy makeup scenarios easily trigger a deep entanglement between the identity and makeup feature spaces. Current mainstream diffusion-based makeup transfer methods predominantly rely on global condition injection mechanisms to transfer makeup features. These mechanisms fail to precisely delineate the spatial boundaries between the source identity features and the reference makeup features, which easily leads to spatial feature leakage. Consequently, models are unable to effectively disentangle identity and makeup features. Furthermore, existing methods generally lack explicit geometric prior anchoring designs and fail to impose targeted rigid constraints on core facial structures. Ultimately, when restoring complex makeup details, these models inevitably introduce unintended modifications to the intrinsic geometric structure of the source face, causing identity feature distortion and failing to achieve the core objective of makeup transfer.

To address the aforementioned issues, this paper proposes an Identity-Preserving Makeup Transfer (IPMT) model, constructing an end-to-end makeup synthesis network. This method discards the reliance on pseudo-paired data and guides the U-Net denoising generation process through multi-branch feature extraction and dynamic alignment mechanisms. Simultaneously, it introduces a parsing consistency loss and a 3D vertex loss to construct dual geometric constraints. These constraints anchor the core facial identity from both the 2D semantic layout and 3D topological structure dimensions, achieving a robust disentanglement between makeup styles and facial features. This allows for the precise restoration of intricate makeup details while strictly preserving the intrinsic geometric structure of the source face.

The main contributions of this paper are summarized as follows:

• We propose an end-to-end Identity-Preserving Makeup Transfer (IPMT) network, which achieves robust disentanglement of makeup style and facial identity features, enabling the model to learn a generalizable and abstract makeup representation.

• To strictly preserve the facial identity of the source image, we design a comprehensive geometric constraint mechanism integrating a Parsing Consistency Loss and a 3D Vertex Loss. This imposes rigid constraints on core facial geometric features from the dual dimensions of 2D planar layout and 3D topological structure.

• Extensive qualitative and quantitative experimental results demonstrate that the IPMT model effectively mitigates the identity inconsistency problem in makeup transfer. It not only achieves high-quality, fine-grained makeup transfer but also demonstrates stronger identity preservation while maintaining competitive makeup fidelity compared to current leading methods.

2 Related Work

2.1 Diffusion Models

Diffusion models have recently emerged as powerful competitors to generative adversarial networks (GANs) [3] by generating realistic images through an iterative inverse denoising process. The feasibility of recovering realistic images from random Gaussian noise was first demonstrated by DDPM [7]. Following this, DDIM [15] transformed the sampling process into a non-Markovian process, enabling fast and deterministic inference. To mitigate computational complexity while preserving high synthesis quality and flexibility, Latent Diffusion Models (LDM) [13] perform the diffusion process within the latent space of powerful pre-trained autoencoders.

Furthermore, diffusion models have demonstrated remarkable progress in controllable generation and image editing. ControlNet [23] integrates varied spatial conditions into the generation process through a bypass structure with a trainable copy network. Similarly, adapters like T2I-Adapter [10] and IP-Adapter [21] have been developed to leverage the controllable capabilities of pre-trained T2I models. Concurrently, diffusion-based image editing methods, including training-free manipulations [19], have significantly broadened the scope of image control. These advancements in precise spatial conditioning and feature manipulation establish a robust foundation for complex conditional synthesis tasks. In this work, we build upon these controllable generative priors to specifically address the stringent geometric and semantic alignment requirements inherent in identity-preserving makeup transfer.

2.2 Makeup Transfer

The evolution of makeup transfer has progressed from pixel-level statistics to deep generative disentanglement. Early approaches, primarily based on Generative Adversarial Networks (GANs), formulated the task as domain translation. BeautyGAN [9] pioneered this by combining global histogram matching with a CycleGAN architecture. To overcome spatial rigidity, PSGAN [8] introduced an Attentive Makeup Morphing (AMM) module to adaptively warp reference makeup to the source face, effectively handling pose variations. Subsequently, SSAT [16] incorporated semantic correspondence via transformers to ensure makeup was transferred to semantically correct regions.

Beyond spatial alignment, recent works have focused on unsupervised content-style disentanglement to address data scarcity. CSD-MT [18] posits that makeup style and facial identity can be decoupled in the frequency domain, extracting makeup information from low-frequency components without requiring pseudo-ground truth. However, these methods often struggle with training instability or fail to synthesize the high-frequency realistic details, characteristic of complex cosmetics.The emergence of Denoising Diffusion Probabilistic Models (DDPMs) has shifted the paradigm toward high-fidelity texture synthesis [7]. Leveraging the strong generative priors of LDM, recent approaches have achieved superior visual quality in self-supervised settings [13]. For instance, Stable-Makeup [24] injects makeup style guidance via cross-attention layers to achieve semantic alignment between the source and reference images. Despite

their success in texture generation, current diffusion-based methods primarily rely on implicit style injection or cross-attention mechanisms. This implicit guidance often lacks precise geometric constraints, resulting in inadequate structure preservation. The generative model thus inadvertently alters facial features and gives rise to identity inconsistency when transferring heavy or artistic makeup styles.

3 Methods

3.1 Preliminary

Latent Diffusion Models (LDMs) operate in the compressed latent space of a pretrained autoencoder, effectively reducing computational complexity while preserving high-fidelity image synthesis capabilities. Given an image$x \in R^{H\times W\times 3}$,the encoder $\mathcal{E}$ maps it into a low-dimensional latent representation $\mathrm{z}_0 = \mathcal{E}(\mathrm{x})$. The forward diffusion process gradually injects Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ into z_0 over a sequence of timesteps t, producing the noisy latent z_t. In the reverse denoising process, a time-conditional UNet, denoted as ϵ_θ, is trained to predict the added noise ϵ. This process is typically guided by additional conditioning inputs c, such as text prompts or reference images. The overall objective of the standard LDM is optimized via the following simplified noise-prediction loss:

$$L_{diffusion} = E_{z_{\mathbb{0}},c,\epsilon\sim\mathbb{N}(\mathbb{0},\mathbb{1}),t}\left[||\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(c)\right)||_2^2\right] \quad (1)$$

where τ_θ serves as a domain-specific encoder that projects the conditioning input c into an intermediate representation space. In the context of makeup transfer, our framework builds upon the robust generative prior of LDMs. However, we observe that optimizing standard conditional LDMs via $L_{diffusion}$ alone is insufficient for this task.When dealing with heavy or artistic makeup, the vanilla reverse process tends to corrupt identity-specific features, leading to severe identity inconsistency. To address this, our proposed IPMT framework introduces explicit identity-preserving and topological constraints into the generation pipeline.

3.2 Disentangled Feature Encoding and Fusion

Our proposed IPMT framework is specifically designed to tackle the identity leakage and inconsistency problems in high-fidelity makeup transfer by strictly regularizing the generative process within a geometry-aware manifold. Leveraging the powerful generative prior of the pre-trained LDM [13], we formulate makeup transfer as a constrained conditional reconstruction task. Unlike previous approaches that rely on unstable pseudo-paired supervision, IPMT operates in a fully self-supervised manner to ensure that the intrinsic identity of the source is strictly preserved.

As illustrated in **Fig. 1.** The overall framework of our Identity-Preserving Makeup Transfer (IPMT) pipeline. The model extracts and aligns identity, makeup, and noise features to condition a Latent Diffusion Model, which generates the final output., the architecture effectively disentangles the entangled facial signal into two orthogonal

latent spaces: a rigid structure space governed by identity geometric priors, and a fluid style space representing makeup semantics. The final output is synthesized via a progressive denoising trajectory, conditioned jointly on these disentangled representations.

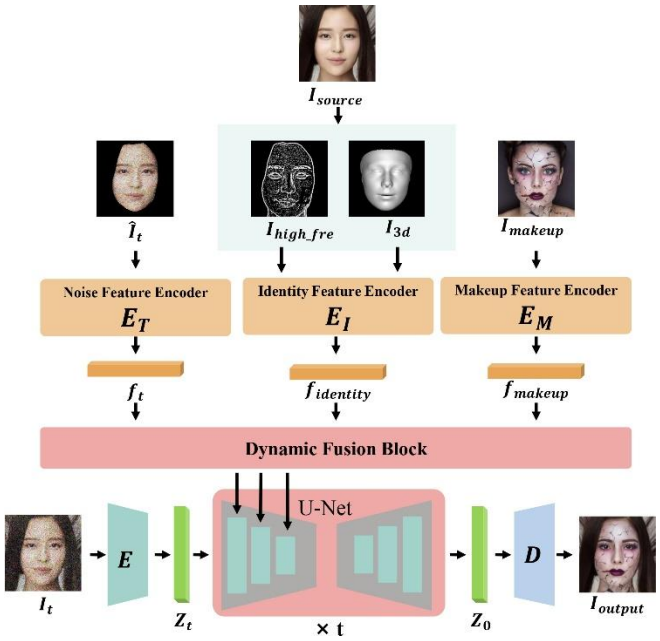

Fig. 1. The overall framework of our Identity-Preserving Makeup Transfer (IPMT) pipeline. The model extracts and aligns identity, makeup, and noise features to condition a Latent Diffusion Model, which generates the final output.

To provide precise guidance for the denoising process and maintain core identity features, we introduce a multi-stream encoding mechanism. Specifically, the identity preservation branch extracts a comprehensive structural identity feature, denoted as $f_{identity}$. To avoid the loss of fine-grained facial attributes, we compose $f_{identity}$ by fusing two complementary signals: (1) a dense 3D facial layout I_{3d} extracted via 3D Dense Face Alignment-V2[5]; and (2) a high-frequency residual map I_{high_fre} obtained through a Gaussian high-pass filter. These inputs are processed by the Identity Feature Encoder E_I, a hierarchical convolutional network designed to extract multi-scale spatial feature maps that encapsulate robust topological constraints.Simultaneously, the makeup style is encoded into a latent makeup feature vector f_{makeup} from the reference image I_{makeup} utilizing a pre-trained CLIP image encoder as our Makeup Feature Encoder E_M. Specifically, E_M projects the visual input into a global semantic space, extracting abstract chromatic and textural representations that are inherently invariant to spatial layouts.

Furthermore, to capture the dynamic temporal dependency, we employ a Noise Feature Encoder E_T to extract the noise feature f_t from the noisy target foreground image $\hat{I}_t$,providing a dynamic spatial representation of the current generation state.

A critical challenge in diffusion-based transfer is the spatial misalignment between the source identity and the reference makeup, which often leads to structural deformation. To dynamically bridge this semantic-spatial gap without distorting the source geometry, we propose a novel Dynamic Fusion Block (DFB). Distinct from simple feature concatenation or implicit frequency mixing, our DFB leverages a time-aware cross-attention mechanism to achieve pixel-level adaptive fusion. Formally, DFB actively constructs a spatial query by concatenating the time-dependent noise feature f_t and the structural identity feature $f_{identity}$:

$$Q_{spatial} = W_Q \cdot Concat(f_t, f_{identity}) \tag{2}$$

where W_Q is a learnable projection weight. The abstract makeup feature f_{makeup} is mapped to Key and Value representations:

$$K_{makeup} = W_K \cdot f_{makeup}, V_{makeup} = W_V \cdot f_{makeup} \tag{3}$$

The dynamically fused feature f_{fusion} is then obtained via standard attention mechanisms, adapting the makeup semantics to the target spatial geometry:

$$f_{\text{fusion}} = \text{Softmax}\left(\frac{Q_{\text{spatial}} K_{\text{makeup}}^T}{\sqrt{d}}\right) \cdot V_{\text{makeup}} \tag{4}$$

This mathematical formulation ensures that the stylistic features are spatially modulated and semantically registered onto the source coordinates, actively preventing the reference style from bleeding into and altering the source identity.

Finally, to seamlessly inject these dynamically aligned control signals into the U-Net denoiser, we propose a multi-scale spatial injection strategy. As depicted in **Fig. 1**, the fused feature representations f_{fusion} are projected and injected into the intermediate layers of the U-Net at multiple resolutions via spatial cross-attention mechanisms. This multi-scale modulation ensures that the generative model is consistently guided by the precise facial geometry and the target makeup semantics across all feature hierarchies. Consequently, the model learns to reconstruct I_{output} by iteratively refining the noisy latents, achieving a seamless makeup fusion while strictly preserving the source identity.

3.3 Identity Preservation via Dual Geometric Constraints

As the core mechanism of our IPMT framework, we require a dedicated strategy to ensure that the learned makeup is accurately transferred to the source face without distorting the underlying facial geometry that defines the subject's identity. To this end, IPMT introduces a specialized loss function module that operates on an intermediate, one-step denoised prediction. This strategy ensures that the model maintains strict geometric and semantic consistency, particularly during the crucial early stages of generation. We adopt a self-supervised disentanglement-and-reconstruction training paradigm, employ a spatial warping strategy to construct unsupervised training samples, and introduce dual geometric constraint losses to strictly anchor the source identity, eliminating the reliance on pseudo-paired data, as illustrated in **Fig. 2**.

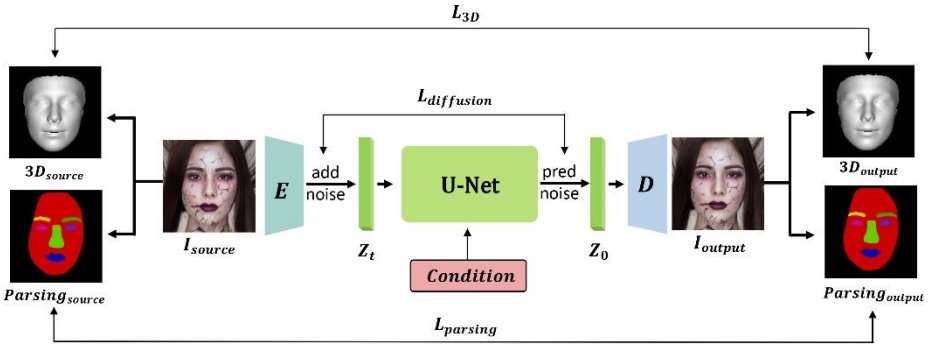

Fig. 2. Self-supervised training pipeline. The framework adds noise and reconstructs the source image (I_{source}) using a U-Net. It applies 3D shape and facial parsing losses (L_{3D} and $L_{Parsing}$) to strictly preserve the source identity.

One-Step Denoising Approximation. During the reverse diffusion process, the total number of noise injection steps is denoted as T (typically $T < 1000$). Directly predicting the clean latent from a state of near-pure noise is highly inaccurate. However, for a relatively small value of t bounded by a threshold R_t (where $t \leq R_t < T$), we can reliably estimate the denoised latent $\hat{z}_0$ directly from the noisy latent z_t and the predicted noise $\epsilon_\theta(z_t, t, C)$. Let x_0 be the ground-truth image and z_t be its latent representation after t steps of the forward diffusion process. Under this specific timestep condition, we approximate the clean latent $\hat{z}_0$ by reversing the diffusion in a single step. This conditional one-step reverse formula is formally defined as:

$$\hat{z}_0 = \frac{z_t - \sqrt{1-\overline{\alpha}_t}\,\epsilon_\theta(z_t, t, C)}{\sqrt{\overline{\alpha}_t}}, \quad t \leq R_t \tag{5}$$

where $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of the noise schedule coefficients, and C is the conditional information. After the reversal, the estimated latent $\hat{z}_0$ is passed through the VAE decoder $\mathcal{D}$ to yield the reconstructed spatial image, $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$.

Parsing Consistency Loss. To explicitly prevent geometric distortions and identity leakage, we apply a Parsing Consistency Loss ($L_{parsing}$) between the ground-truth image x_0 and the one-step prediction $\hat{x}_0$. It is well-established that makeup shouldn't have a significant impact on the face segmentation of the source image. This loss is calculated using a pre-trained face parsing network to enforce strict semantic similarity. The field has seen a proliferation of diverse network architectures for face parsing [11, 14, 25]. In this study, we adopt the pre-trained face parsing network Farl [25] for face parsing tasks. This network boasts both faster processing speed and higher parsing accuracy, enabling it to support efficient model training.

From the face parsing network, we obtain per-pixel softmax probability distributions for both the ground-truth image x_0 and the predicted image $\hat{x}_0$, denoted as Q and P respectively. Each distribution describes the probability of a pixel belonging to one of the 19 predefined semantic facial classes. To specifically enforce geometric consistency on crucial identity-defining regions, we construct a binary mask M that isolates key areas

such as the eyes, nose, and mouth. The subsequent loss calculation is then confined to these masked regions, ensuring their underlying shapes are rigidly preserved by constraining their respective probability distributions.

Crucially, this loss is applied conditionally, focusing only on a predetermined range of early timesteps ($t \leq R_t$, where R_t is a threshold). This encourages the model to establish the subject's core identity and structure early in the reverse diffusion process. The determination of R_t is strategically grounded in the temporal dynamics of diffusion models, which typically synthesize high-level geometric structures before refining high-frequency textural details. Consequently, enforcing the parsing constraint during the early phase secures the facial topology, while relaxing it in later stages grants the model sufficient flexibility to render intricate makeup styles without over-regularization. Extensive empirical evaluations demonstrate that setting R_t to approximately 25% of the total sampling steps (e.g., $R_t = 250$ for $T = 1000$) yields an optimal trade-off between structural rigidity and stylistic editability.Formally,the Parsing Consistency Loss is defined as:

$$L_{parsing} = \frac{1}{|M|} \sum_{i \in M} \left(\sum_{c=1}^{C} Q_{i,c} \log \left(\frac{Q_{i,c}}{P_{i,c}} \right) \right) \tag{6}$$

Here, $Q_{i,c}$ and $P_{i,c}$ are the probabilities that pixel i belongs to class c, C is the number of classes, and $|M|$ is the number of pixels in the mask. This loss provides a fundamental semantic constraint, forming the backbone of IPMT's identity-preserving capability, effectively regularizing the model's geometry to ensure absolute structural consistency without overly constraining the target makeup appearance.

3D Vertex Loss. To further explicitly enforce structural integrity and complement the 2D semantic constraints of the parsing map, we introduce a 3D Vertex Loss (L_{3D}). While $L_{diffusion}$ ensures the planar layout of facial components remains stable, it may not fully capture fine-grained 3D topological deformations. Therefore, we utilize 3DDFA-V2[5] to extract dense 3D facial vertices from both the ground-truth image x_0 and the predicted image $\hat{x}_0$. By calculating the L_1 distance between these corresponding vertex sets, we impose a rigid 3D geometric constraint.

Consistent with our parsing strategy, this loss is also applied exclusively during the early stages of generation to strictly constrain the facial topology.The 3D Vertex Loss is defined as:

$$L_{3D} = ||V(\hat{x}_0) - V(x_0)||_1 \tag{7}$$

where $V(\cdot)$ denotes the 3D vertex extraction operation. Together, $L_{parsing}$ and L_{3D} form a comprehensive geometry-aware manifold that rigorously prevents identity leakage.

Our final training objective for IPMT, L_{total}, integrates three core components carefully designed to balance high-fidelity makeup synthesis with strict identity preservation. The foundation is a Core Diffusion Loss ($L_{diffusion}$), which follows the standard LDM training paradigm [13] to learn the denoising process.

In summary, the total loss function is formulated as:

$$L_{total} = L_{diffusion} + \lambda_{parsing} L_{parsing} + \lambda_{3D} L_{3D} \tag{8}$$

where $\lambda_{parsing}$ and λ_{3D} are the respective balancing hyperparameters. Both the Parsing Consistency Loss ($L_{Parsing}$) and the 3D Vertex Loss (L_{3D}) are computed exclusively during the early stages of the denoising process ($t \leq R_t$).This temporal strategy allows them to serve as robust structural anchors for identity preservation, securing the facial topology without disrupting the model's capacity to render fine-grained makeup details in later stages.

4 Experiments

4.1 Datasets

We conduct comprehensive evaluations on three benchmark datasets to validate different aspects of our identity-preserving framework. Our primary training and quantitative analysis are performed on the MT dataset [9], utilizing a 60%/40% train-test split on its 3,334 high-quality images. To assess generalization capabilities on complex textures and avant-garde styles, scenarios where identity leakage is typically most severe, we employ the LADN dataset [4]. Furthermore, we use the challenging Wild-MT dataset [8] to verify the broad applicability and robust identity preservation capabilities of our proposed method under unconstrained scenarios with significant variations in pose and expression.

4.2 Implementation Details

Our IPMT model is implemented in PyTorch and trained on a single NVIDIA RTX A6000 GPU. The backbone is initialized with a pre-trained LDM. Training is conducted for 300k steps with a batch size of 2 using the Adam optimizer (learning rate 1×10^{-5}, $\beta_1 = 0.9, \beta_2 = 0.999$), which takes approximately 7 days. The balancing hyperparameters for the objective function are set to $\lambda_{parsing} = 0.15$ and $\lambda_{3D} = 0.1$. The Parsing Consistency Loss is applied to the one-step denoised approximation $\hat{x}_0$ exclusively at early timesteps ($t \leq R_t$, with $R_t = 250$). During inference, we adopt a 50-step procedure, yielding an average inference time of 1.5 seconds per 256×256 image. It should be noted that the 3DDFA-V2 and the additional face parsing network are exclusively utilized during the training phase to construct the geometric constraints. Consequently, they do not introduce any additional computational overhead or latency during inference.

4.3 Evaluation Metrics

To comprehensively evaluate our method, we employ five key metrics. Makeup fidelity is assessed via CLIP-I [12] for high-level semantic alignment, and DINO [22] for fine-grained textural detail comparison.

For our core objective of identity preservation, we utilize Parsing-acc to measure face segmentation accuracy, ensuring 2D geometric structural consistency. Specifically, distinct from the probabilistic parsing loss applied to intermediate latents during

training, Parsing-acc evaluates the final synthesized output by computing the pixel-level accuracy between the hard categorical segmentation masks of the source and generated images across all facial semantic regions. The metric is formulated as the overall proportion of correctly classified pixels over the entire face, thereby providing a comprehensive assessment of global geometric structure preservation.

Furthermore, traditional 2D face recognition networks (e.g., ArcFace [1]) often erroneously entangle complex makeup styles as inherent identity features. To mitigate this evaluation bias, assessing facial expressions within a 3D manifold offers a robust, macro-level evaluation of facial feature correctness that is largely invariant to 2D textural variations. Therefore, we introduce the Expression Error (EXP-error) to explicitly quantify dynamic structural consistency during the evaluation phase. Specifically, we leverage a pre-trained 3DDFA-V2 [5] to extract the decoupled 10-dimensional expression parameters from the generated and source images, and compute their L_1 distance. Finally, overall image realism is evaluated using the Fréchet Inception Distance (FID) [6]. For CLIP-I, DINO, and Parsing-acc, higher values indicate better performance, whereas lower values are preferred for EXP-error and FID.

4.4 Baselines

We conduct a comprehensive comparison of our method against several state-of-the-art models. Specifically, we evaluate four specialized makeup transfer methods: CSD-MT [18], SSAT [16], EleGANt [20], and Stable-Makeup [24]. Among these, Stable-Makeup is diffusion-model-based, while the others are GAN-based. Furthermore, to evaluate our method's competitiveness in broader scenarios, we compare it with the industry-leading general-purpose large model, Seedream [2]. By leveraging its powerful image generation capabilities via text prompting, Seedream serves as a crucial benchmark to measure the performance advantages of our proposed method against a top-tier, general-purpose generative model.

5 Results

5.1 Experimental Results

Quantitative Comparison. In the quantitative evaluation, we independently and randomly sample 1000 source-reference pairs from the MT, LADN and Wild-MT datasets respectively for the evaluation of all quantitative metrics. **Table 1** presents the comparative results of various models regarding makeup fidelity (evaluated via CLIP-I and DINO) and image realism (evaluated via FID). The experimental results indicate that the proposed IPMT model demonstrates exceptional makeup transfer capabilities across multiple datasets. Whether in standard scenarios, unconstrained environments with severe pose and expression variations, or when dealing with complex artistic makeup, IPMT effectively captures and transfers the stylistic paradigm and local textural details of the reference image, achieving leading or highly competitive performance. Furthermore, the consistently competitive FID scores prove that IPMT not only achieves high-fidelity makeup injection but also generates natural, artifact-free, and

high-quality images. Compared to existing general-purpose large models and diffusion-based specific methods, IPMT exhibits superior robustness and generalization capabilities while maintaining the intensity of style rendering.

Table 1. Quantitative results of DINO, CLIP-I, and FID on MT, Wild-MT, and LADN datasets.

Model	MT			Wild-MT			LADN		
	DINO↑	CLIP-I↑	FID↓	DINO↑	CLIP-I↑	FID↓	DINO↑	CLIP-I↑	FID↓
CSD-MT	0.762	0.728	60.99	0.786	0.696	35.36	0.816	0.723	59.76
SSAT	0.759	0.724	55.97	0.78	0.692	41.94	0.793	0.732	48.63
EleGANt	0.741	0.721	63.21	0.756	0.695	**33.71**	0.768	0.717	59.62
Stable-Makeup	0.784	0.723	57.96	**0.842**	0.717	90.58	0.862	0.744	47.21
Seedream	0.773	0.731	**46.38**	0.805	0.703	44.88	0.859	**0.751**	63.77
IPMT(Ours)	**0.788**	**0.737**	53.43	0.826	**0.721**	51.14	**0.864**	0.746	**42.24**

Table 2 evaluates the performance of various methods specifically regarding identity consistency. The results demonstrate that IPMT achieves excellent comprehensive performance on both Parsing-acc and EXP-error. Although some traditional generative models, such as SSAT and EleGANt, perform prominently on these metrics, this is directly caused by their weak makeup transfer capabilities, which impose virtually no substantive impact on the original facial structure. In contrast, baseline models equipped with strong style transfer capabilities, such as Stable-Makeup and Seedream, frequently suffer from severe identity leakage and structural deformation, leading to noticeably lower scores on these identity preservation metrics. Constrained by the comprehensive consistency loss, IPMT successfully disentangles makeup semantics from facial geometry, achieving an optimal preservation of the source identity structure while simultaneously applying high-fidelity makeup.

Table 2. Quantitative results of EXP-error and Parsing-acc on MT, Wild-MT, and LADN datasets.

Model	MT		Wild-MT		LADN	
	EXP-error↓	Parsing-acc↑	EXP-error↓	Parsing-acc↑	EXP-error↓	Parsing-acc↑
CSD-MT	0.125	0.905	0.105	0.886	0.121	0.888
SSAT	0.119	0.881	0.093	0.872	0.091	0.883
EleGANt	**0.065**	0.922	**0.078**	0.915	0.083	0.916
Stable-Makeup	0.170	0.911	0.218	0.924	0.181	0.892
Seedream	0.184	0.879	0.243	0.881	0.217	0.874
IPMT(Ours)	0.107	**0.931**	0.089	**0.932**	**0.075**	**0.928**

Fig. 3. Qualitative comparison of the results produced by different methods. Our IPMT explicitly prevents identity leakage and structural distortion, achieving superior makeup transfer while strictly preserving the source subject's identity.

Qualitative Comparison. As illustrated in **Fig. 3**, we conduct a comprehensive qualitative comparison between our Identity-Preserving Makeup Transfer (IPMT) model and other state-of-the-art methods. First, compared to GAN-based approaches such as SSAT and CSD-MT, our method demonstrates superior capability in capturing fine-grained makeup details, resulting in a much stronger stylistic consistency with the reference makeup. While SSAT and CSD-MT often struggle to replicate complex artistic styles or produce blurred textures, IPMT accurately synthesizes intricate local patterns and chromatic attributes. Furthermore, when compared to advanced diffusion-based and general-purpose large models like Stable-Makeup and Seedream, our approach exhibits significant advantages in facial feature preservation. Although these baseline models achieve high-quality makeup rendering, they frequently suffer from severe identity leakage. Specifically, they erroneously integrate facial attributes from the reference image into the synthesized output, causing noticeable structural deformations in core facial regions, such as the nose and lip shapes.

By explicitly decoupling the makeup style representation from the facial structure, our parsing-guided framework effectively mitigates the aforementioned identity leakage. Consequently, IPMT generates artifact-free, high-fidelity results that successfully transfer complex makeup semantics while strictly preserving the intrinsic geometric identity of the source subject.

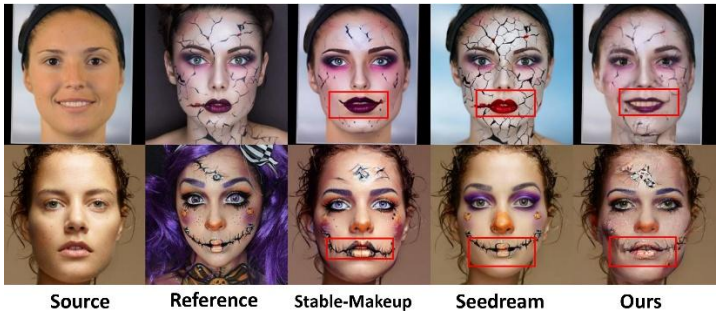

Fig. 4. Detailed comparison highlighting that our IPMT framework successfully preserves the source's fine-grained facial structure and completely avoids the identity leakage artifacts frequently seen in other generative models.

Furthermore, the detailed comparisons in **Fig. 4** highlight that our method demonstrates significant superiority in preserving identity features. The proposed method precisely reconstructs complex makeup effects without altering the facial geometric structure underlying the source identity (e.g., mouth and nose), thereby simultaneously achieving high stylistic fidelity and rigorous identity preservation.

User Study. To comprehensively evaluate the perceptual performance of IPMT, we conducted a user study comparing it against five state-of-the-art methods. We utilized 20 source-reference pairs and recruited 80 participants to blindly evaluate the generated results across three dimensions: makeup fidelity, identity preservation, and overall preference.

Table 3. User study results (preference ratio %).

Model	Makeup Fidelity ↑	Identity Preservation ↑	Overall Preference ↑
SSAT	4.2	11.5	3
EleGANt	7.4	10.8	3.7
CSD-MT	8.9	8.2	5.3
Stable-Makeup	13.1	9	4.9
Seedream	14.6	7.3	8.3
IPMT (Ours)	**51.8**	**53.2**	**74.8**

As shown in **Table 3**, IPMT not only leads in the overall preference ratio but also achieves the highest scores in individual metrics for identity preservation and makeup fidelity, significantly outperforming existing baseline models across all evaluation dimensions. These results validate the effectiveness of our proposed identity-preserving framework, which achieves high-fidelity makeup transfer while strictly maintaining the source identity. Future research will involve more comprehensive evaluations on a broader scale to further validate the robustness of our subjective conclusions.

5.2 Ablation Study

Ablation Study on Core Components. To evaluate the contribution of each core component in the IPMT framework, we conduct an ablation study on the LADN dataset.

We investigate three variant configurations: (1) w/o DFB, where the Dynamic Fusion Block is replaced by direct feature concatenation; (2) w/o 3D Guide, where the L_{3D} is omitted; and (3) w/o Parsing Guide, where the $L_{Parsing}$ is removed.

Table 4. Quantitative results of the ablation study on LADN dataset.

Methods	DINO ↑	CLIP-I ↑	FID ↓	Parsing-acc ↑	EXP-error ↓
IPMT w/o DFB	0.819	0.709	**39.46**	0.905	0.086
IPMT w/o 3D Guide	**0.871**	0.735	46.85	0.862	0.119
IPMT w/o Parsing Guide	0.851	0.741	44.21	0.887	0.131
IPMT (Full)	0.864	**0.746**	42.24	**0.928**	**0.075**

These variants are compared against the full IPMT model to validate the necessity of each module. The quantitative results presented in **Table 4** demonstrate that each component is integral to achieving high-performance makeup transfer. Specifically, the removal of the DFB leads to a significant decline in makeup fidelity metrics (DINO and CLIP-I), confirming that the DFB is essential for the effective injection of stylistic features. Conversely, excluding either the 3D or Parsing guides results in a marked degradation in structural consistency and expression preservation metrics. The full IPMT model achieves the optimal balance, maintaining superior identity preservation while simultaneously delivering high-fidelity style transfer.

Fig. 5. Visual comparison of the ablation study. From left to right: source face, reference makeup, ablated variants (w/o DFB, w/o 3D Guide, w/o Parsing Guide), and the full IPMT model. Our full model maintains the best identity consistency, particularly in preserving the source's facial structure.

As illustrated in **Fig. 5**, the visual comparison reinforces the quantitative findings. Variants lacking the Parsing or 3D guides frequently exhibit structural artifacts and identity leakage when dealing with heavy makeup. In contrast, the full IPMT model achieves high makeup fidelity while maintaining rigorous identity consistency. Notably, the shape and spatial layout of core facial features and fine-grained details, such as the nose, mouth position, and wrinkles, accurately conform to the source image rather than being distorted by the reference style. This demonstrates that our geometry-aware constraints effectively prevent structural deformation, ensuring that the synthesized output remains faithful to the source identity.

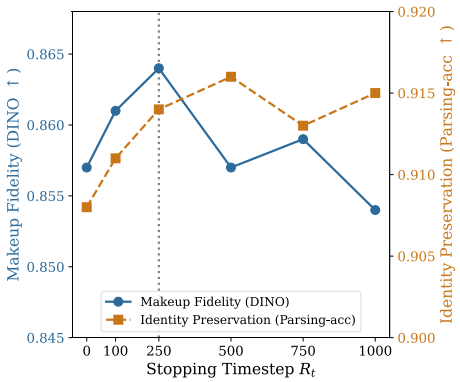

Fig. 6. Impact of the timestep threshold R_t. The chart illustrates the dynamic trade-off during the reverse diffusion process. Setting $R_t = 250$ achieves the optimal balance between maintaining structural identity and transferring makeup semantics, a critical factor in the identity-preserving framework.

Impact of the Timestep Threshold R_t. To investigate the effect of the timestep threshold R_t, which determines the duration of geometric constraints applied during the reverse diffusion process, we conduct an ablation study across various timesteps, as illustrated in **Fig. 6**.We randomly sample a total of 600 source-reference pairs across all datasets for testing. Specifically, we employ DINO to evaluate makeup fidelity and Parsing-acc to assess identity preservation. When R_t is set to a minimal value, the model fails to maintain the source identity. Increasing R_t initially improves both identity preservation and makeup fidelity, demonstrating that applying semantic constraints during the early timesteps successfully establishes the core structural identity of the subject. However, as R_t increases further, makeup fidelity exhibits a steady decline while the identity score plateaus. The temporal dynamics of diffusion models indicate that enforcing rigid parsing constraints during the later stages, in which high-frequency textural details are refined, results in over-regularization of the generation process and constrains the model's ability to synthesize intricate makeup styles. Consequently, defining R_t at an intermediate optimal value (e.g., $R_t = 250$) achieves the optimal trade-off between structural rigidity and stylistic editability, which is essential for the identity-preserving objective of IPMT.

6 Conclusion

In this paper, we propose an Identity-Preserving Makeup Transfer (IPMT) model to address the identity inconsistency problem in high-fidelity makeup generation.By adopting a self-supervised learning framework, our approach effectively eliminates the dependence on pseudo-paired data, enhances the generalization ability of the model, and achieves high-quality and robust makeup synthesis based on the disentanglement of makeup style and facial identity.To strictly maintain the consistency of facial features, we introduce the Parsing Consistency Loss and the 3D Vertex Loss to impose

explicit spatial constraints, thereby preventing facial structure deformation during the style transfer process.Extensive quantitative and qualitative evaluations demonstrate that the IPMT model effectively alleviates identity leakage, demonstrating stronger identity preservation while achieving competitive makeup fidelity compared to existing leading methods.

7 Limitations and Future Work

Despite its overall effectiveness, our framework exhibits certain limitations when handling extreme spatial discrepancies. Specifically, in scenarios involving severe pose variations between the source and reference images, the current alignment mechanism may encounter difficulties in achieving optimal semantic registration. Future work will investigate more robust 3D-aware strategies to effectively address these challenging variations.

References

1. Deng, J. et al.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019).
2. Gao, Y. et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346. (2025).
3. Goodfellow, I.J. et al.: Generative adversarial nets. Advances in neural information processing systems. 27, (2014).
4. Gu, Q. et al.: Ladn: Local adversarial disentangling network for facial makeup and demakeup. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 10481–10490 (2019).
5. Guo, J. et al.: Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision. pp. 152–168 Springer (2020).
6. Heusel, M. et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. 30, (2017).
7. Ho, J. et al.: Denoising diffusion probabilistic models. Advances in neural information processing systems. 33, 6840–6851 (2020).
8. Jiang, W. et al.: 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. pp. 5194–5202 (2020).
9. Li, T. et al.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 645–653 (2018).
10. Mou, C. et al.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. pp. 4296–4304 (2024).
11. Narayan, K. et al.: FaceXFormer: A Unified Transformer for Facial Analysis. arXiv preprint arXiv:2403.12960. (2024).
12. Radford, A. et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 PmLR (2021).

13. Rombach, R. et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022).
14. Serengil, S., Ozpinar, A.: A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. Journal of Information Technologies. 17, 2, 95–107 (2024). https://doi.org/10.17671/gazibtd.1399077.
15. Song, J. et al.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502. (2020).
16. Sun, Z. et al.: Content-style decoupling for unsupervised makeup transfer without generating pseudo ground truth. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7601–7610 (2024).
17. Sun, Z. et al.: Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models. Advances in Neural Information Processing Systems. 37, 16016–16042 (2024).
18. Sun, Z. et al.: Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. In: Proceedings of the AAAI Conference on artificial intelligence. pp. 2325–2334 (2022).
19. Tsaban, L., Passos, A.: Ledits: Real image editing with ddpm inversion and semantic guidance. arXiv preprint arXiv:2307.00522. (2023).
20. Yang, C. et al.: Elegant: Exquisite and locally editable gan for makeup transfer. In: European conference on computer vision. pp. 737–754 Springer (2022).
21. Ye, H. et al.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721. (2023).
22. Zhang, H. et al.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605. (2022).
23. Zhang, L. et al.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023).
24. Zhang, Y. et al.: 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. pp. 1–9 (2025).
25. Zheng, Y. et al.: General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18697–18709 (2022).