Anchoring on Reality: Breaking the Pseudo-Target Ceiling in Makeup Transfer

Bo Wei[1], Xianhui Lin[2†], Yi Dong[2], Zhongzhong Li[2], Zonghui Li[2]
Zirui Wang[2], Jiachen Yang[2], Xing Liu[2], Hong Gu[2]
Xiaoming Li[3] ✉, and Wangmeng Zuo[1] ✉

[1]Harbin Institute of Technology, China [3]Nanjing University, China
[2]vivo BlueImage Lab, vivo Mobile Communication Co., Ltd, China
csbowei@gmail.com, xhlin129@gmail.com, ydong@outlook.com,
csxmli@nju.edu.cn, wmzuo@hit.edu.cn
https://csbowei.github.io/ART/

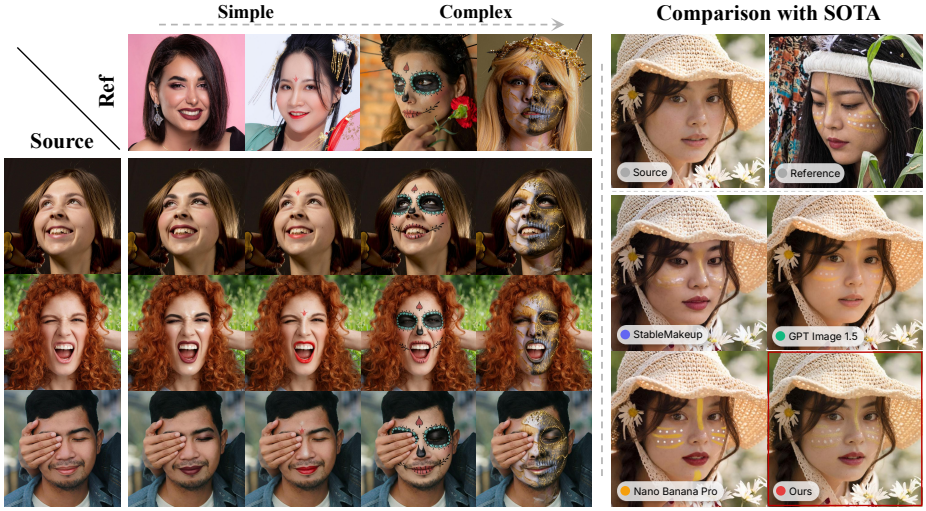

Fig. 1: ART enables high-fidelity transfer of diverse makeup styles (simple to complex) at up to 2K resolution, while preserving identity, geometry, and makeup placement under challenging expressions, occlusions, and cross-gender scenarios.

Abstract. Makeup transfer applies a reference cosmetic style to a source face while preserving its identity and geometry. However, this task is severely hindered by the lack of real paired training data. Current methods rely on either weak priors or synthetic pseudo-targets from large-scale editing models. These paradigms provide suboptimal guidance, often leading to degraded fine-grained details, synthetic artifacts, and identity drift. To this end, we propose **A**nchoring on **R**eality Makeup **T**ransfer (ART), a two-stage framework with a reality-anchored refinement cycle. In Stage I, the model is initialized with pseudo-targets to

†Project lead. ✉Corresponding author.

establish basic semantic alignment and global makeup placement. Crucially, Stage II shifts supervision from pseudo-targets to the real reference, reconstructing it from its bare-skin counterpart through a differentiable cycle that penalizes any omitted detail and overrides synthetic artifacts. Furthermore, we introduce MakeupFaces2K (MF2K), the first 2K-resolution in-the-wild makeup portrait dataset comprising 8,573 images. Extensive experiments demonstrate that our method achieves superior makeup fidelity, strong background stability, and robust identity preservation, especially for complex makeup styles.

Keywords: Makeup Transfer · Refinement Cycle · Diffusion Transformer

1 Introduction

The task of makeup transfer aims to synthesize a target portrait that faithfully adopts the cosmetic style of a reference image while strictly preserving the identity and facial geometry of a source image. Beyond consumer photo editing, the demand for high-fidelity transfer is increasing in professional applications, such as cinematic post-production and virtual avatars. In these settings, cosmetics often go beyond daily makeup to encompass complex styles, including heavy artistic patterns, face paintings, glitter, and dense stickers. Faithfully transferring these styles, especially across gender or ethnicity, remains a formidable challenge for existing methods.

Fundamentally, progress in complex makeup transfer is bottlenecked by the lack of real paired supervision. Many prior methods formulate makeup transfer as an unpaired image-to-image translation problem guided by weak priors, such as cycle consistency or domain adversarial losses [5, 11, 30, 36, 37, 50]. However, these approaches often struggle to achieve photorealistic fidelity and precise spatial placement when dense correspondence is ambiguous. Driven by the recent surge in large-scale generative models, state-of-the-art methods typically construct synthetic pseudo-targets using external editing engines [41, 48] and train models to regress their predictions to these generated pairs. While this paradigm improves overall realism, it imposes an inherent **pseudo-target ceiling**: the model systematically inherits the biases and artifacts of the constructed targets, such as identity drift, degraded fine-grained details, and unintended background edits. Such flaws are inevitable, since a model trained merely to imitate pseudo-targets can never surpass them.

To break this ceiling, we propose ART, a two-stage Diffusion Transformer (DiT) [27] framework that shifts the learning paradigm from synthetic regression to differentiable reality-anchored reconstruction. Our key insight is that, although the ideal transfer target is unavailable, the real reference itself provides all the authentic makeup details that supervision requires. Unlike traditional cycle consistency (e.g., CycleGAN [50]), which only enforces invertibility back to the source, our cycle reconstructs the real reference itself, providing direct supervision for the transferred makeup.

Specifically, Stage I performs pseudo-target initialization to establish basic semantic alignment and global makeup placement. In Stage II, we introduce a reality-anchored refinement cycle. We first predict a preliminary transfer output and retain it as a differentiable *makeup carrier* in the latent space. The DiT is then constrained to reconstruct the real reference image from its bare-skin counterpart (extracted via an auxiliary makeup remover), conditioned solely on this carrier. Crucially, because this carrier remains a differentiable latent rather than a decoded image, the reconstruction gradient flows back into the initial transfer prediction. This directly penalizes any omission of fine cosmetic detail, forcing the transfer prediction to reproduce the reference's true textures and overwrite the artifacts inherited from the pseudo-targets. To stabilize this refinement cycle, we introduce a controlled-noise bottleneck during the generation of the *makeup carrier*. By modulating the amount of pseudo-target prior retained under noise injection, this bottleneck preserves the coarse makeup placement as a structural foundation while affording the model sufficient flexibility to recover high-fidelity details from the real reference.

Progress in high-fidelity transfer is further hindered by the limitations of existing public datasets [11, 17, 20, 24, 44], which are typically low-resolution and underrepresent complex makeup styles (e.g., dense stickers, face paintings) as well as specific demographics (e.g., male portraits). To facilitate research in high-fidelity synthesis, we introduce MakeupFaces2K (MF2K), the first 2K-resolution in-the-wild makeup portrait dataset. Comprising 8,573 images, MF2K spans diverse makeup intensities ranging from bare skin to extreme artistic styles while maintaining broad demographic diversity. Based on this dataset, we construct training triplets and derive a curated evaluation set that specifically emphasizes complex, fine-grained patterns and cross-domain transfers.

In summary, our main contributions are:

- **Reality-Anchored Refinement.** We propose ART, a two-stage DiT framework that breaks the pseudo-target ceiling. By anchoring a refinement cycle to the real reference, our model effectively overrides pseudo-target artifacts while stabilizing training via a controlled-noise bottleneck.
- **A 2K-Resolution In-the-Wild Makeup Dataset.** We introduce MakeupFaces2K (MF2K), the first 2K-resolution makeup portrait dataset with 8,573 images, covering diverse makeup intensities and underrepresented demographics to support high-fidelity makeup transfer and related tasks.
- **State-of-the-Art Performance.** Our method achieves superior makeup fidelity across diverse styles, while consistently preserving fine-grained details and source identity, even under occlusions and extreme expressions.

2 Related Work

2.1 Unpaired and Weakly Supervised Transfer

Due to the scarcity of paired data, early frameworks formulated makeup transfer strictly as an unpaired image-to-image translation task. Driven primarily

by GANs [8], methods like CycleGAN [50], LADN [11], and CSD-MT [37] relied on weak priors, such as domain adversarial losses, to constrain the cosmetic mapping. To better handle localized cosmetics, subsequent approaches introduced spatial alignment and style-code controls for semantic correspondences [6, 19, 34, 35]. Recently, diffusion models [14, 27, 29, 32] have emerged as powerful generative backbones. Building on this, SHMT [36] proposes a self-supervised hierarchical strategy to decompose and progressively inject makeup textures, while MAD [30] treats makeup tasks as cross-domain translations governed by unified domain embeddings. Despite leveraging powerful generative priors, these paradigms lack a ground-truth transfer target and can therefore supervise the output only indirectly. When dense spatial correspondence is ambiguous, they often struggle to achieve precise placement and recover fine-grained details (e.g., dense stickers, glitter), exposing the inherent limitation of relying solely on weak unpaired supervision.

2.2 Surrogate Supervision via Pseudo-Targets

To inject stronger supervisory signals, a dominant line of research constructs synthetic paired data as surrogate ground-truth. Many methods enforce color and tone consistency via histogram matching [17,20,22,42,44], or explicitly model structured makeup patterns [24]. To further align spatial placement, other works construct pseudo-targets via geometric warping [4, 40, 45] or derive feature-level semantic targets [51].

Driven by large-scale foundation models, state-of-the-art diffusion frameworks now heavily rely on external editing engines to synthesize massive pseudo-paired datasets [12, 41, 48]. StableMakeup [48] and EvoMakeup [41] construct synthetic pseudo-targets via dedicated data-construction pipelines, directly regressing the model outputs to these generated targets. While such regression provides essential spatial guidance and improves global layout, minimizing deviation from these imperfect pseudo-targets inevitably propagates their biases and artifacts into the prediction. This pseudo-target ceiling motivates us to seek supervision beyond synthetic approximations.

2.3 Cycle Consistency in Generative Modeling

Cycle consistency [50] is a foundational mechanism widely adopted to regularize domain invertibility. Recently, this concept has also been retrofitted into diffusion pipelines (e.g., CycleNet [43], CycleDiff [52]) to stabilize unpaired manipulation and zero-shot editing.

However, existing cycle objectives primarily verify source recoverability, acting as domain-level regularizers rather than supervising the forward generation with instance-level detail. Consequently, while they improve overall translation stability, they remain insufficient for the precise, fine-grained texture recovery that complex cosmetics demand.

In contrast, our reality-anchored refinement cycle reconstructs the real reference itself, verifying whether the transfer prediction carries reference-specific

makeup cues. Rather than imposing a domain-level invertibility penalty, it provides a corrective, instance-level supervisory signal that drives the network to preserve fine-grained cosmetic details and override pseudo-target artifacts, bridging weak domain regularization and exact textural grounding.

3 Methodology

In this section, we first introduce the preliminaries (Sec. 3.1), then present the two-stage strategy with a focus on the reality-anchored refinement cycle (Sec. 3.2), and finally describe our high-resolution makeup dataset (Sec. 3.3).

3.1 Preliminaries

Given a source image I_{src} and a reference image I_{ref}, the goal of makeup transfer is to synthesize an output that preserves the identity and geometry of I_{src} while faithfully adopting the cosmetic style of I_{ref}. Since paired ground-truth is unavailable, we construct a synthetic pseudo-target I_{pseudo} and extract a bare-skin counterpart $I_{\text{bare\_ref}}$ of the reference via an auxiliary makeup remover for our refinement cycle. We operate in the latent space of a pretrained VAE [18], denoting the latent representation of any image I as $z = \mathcal{E}(I)$.

We train the conditional DiT using the Flow Matching (FM) objective [21]. For a target latent z and condition c, we sample noise $\epsilon \sim \mathcal{N}(0, I)$ and a timestep $\sigma \in [0, 1]$ to construct the noisy latent $z_\sigma = (1 - \sigma)z + \sigma\epsilon$. The model predicts a velocity field $v_\theta(z_\sigma, \sigma; c)$ and the objective is defined as:

$$\mathcal{L}_{\text{FM}}(\theta; z, c) = \mathbb{E}_{\epsilon,\sigma}\left[w(\sigma)\left\|v_\theta(z_\sigma, \sigma; c) - (\epsilon - z)\right\|_2^2\right], \tag{1}$$

where $w(\sigma)$ is a standard weighting function.

Following standard DiT conventions for image editing, condition injection is implemented via token concatenation [38, 39, 49]. The input sequence to the transformer is constructed by concatenating the position-encoded tokens of the noisy target, the source, and the reference latents.

3.2 Two-Stage Training Strategy

To break the pseudo-target ceiling imposed by I_{pseudo}, which is synthesized by a large-scale image editing model, we propose a two-stage training strategy that shifts the learning paradigm toward differentiable reality-anchored reconstruction. An overview is shown in Fig. 2.

Stage I: Pseudo-Target Initialization. In this stage, we establish the foundation for the subsequent refinement cycle by independently training the main transfer model and an auxiliary makeup remover $\mathcal{R}$.

Transfer Model Initialization. The transfer model is initialized to imitate the synthetic pseudo-target I_{pseudo}. We minimize the standard FM objective conditioned on the source and reference:

$$\mathcal{L}_{\text{init}} = \mathcal{L}_{\text{FM}}\big(\theta; z_{\text{pseudo}}, z_{\text{src}}, z_{\text{ref}}\big). \tag{2}$$

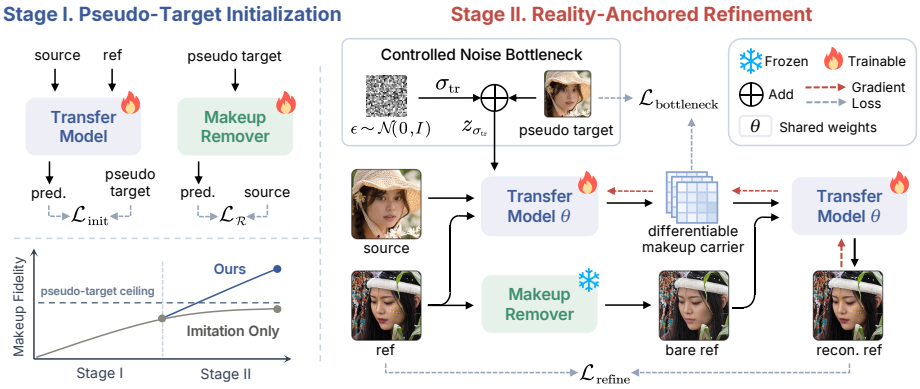

Fig. 2: Overview of ART. ART is a two-stage framework that shifts supervision from the synthetic pseudo-target to the real reference. **Stage I** performs pseudo-target initialization, training the transfer model for global makeup placement and an auxiliary makeup remover for bare-skin extraction. **Stage II** predicts a differentiable makeup carrier $\hat{z}$ at noise level σ_{tr}, then reconstructs the real reference from its bare-skin counterpart conditioned on $\hat{z}$; as $\hat{z}$ stays differentiable, the gradient of $\mathcal{L}_{\text{refine}}$ back-propagates into the transfer prediction to recover fine-grained makeup cues and override synthetic artifacts, while $\mathcal{L}_{\text{bottleneck}}$ preserves the global structure.

This establishes basic semantic alignment and global makeup placement. However, its performance is inherently bounded by the artifacts (e.g., incorrect details or identity drift) present in I_{pseudo}.

Bare-Skin Counterpart Extraction. In parallel, we train an auxiliary makeup remover $\mathcal{R}$ to recover a bare-skin face from its makeup counterpart, isolating the facial identity to anchor the Stage II refinement cycle. We implement $\mathcal{R}$ as a one-step denoising model and supervise it with the pair $(I_{\text{pseudo}}, I_{\text{src}})$, where I_{pseudo} is the source I_{src} rendered with the reference makeup. Its optimization is driven by a standard FM objective, complemented by auxiliary regularizers for perceptual fidelity, identity preservation, and structural consistency:

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_{\text{FM}}\big(\phi; z_{\text{src}}, z_{\text{pseudo}}\big) + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{lmk}}\mathcal{L}_{\text{lmk}}\,. \tag{3}$$

where $\mathcal{L}_{\text{lpips}}$ [47] improves perceptual quality, $\mathcal{L}_{\text{id}}$ enforces identity consistency via ArcFace embeddings [7], and $\mathcal{L}_{\text{lmk}}$ [3] preserves the facial geometric structure. Once trained, $\mathcal{R}$ is applied to the reference to obtain its bare-skin counterpart, $I_{\text{bare\_ref}} = \mathcal{R}(I_{\text{ref}})$.

Stage II: Reality-Anchored Refinement. This stage shifts supervision from the synthetic pseudo-target to the real reference, anchoring the model to authentic makeup detail rather than a synthetic approximation.

Starting from the pseudo-target latent z_{pseudo}, we perturb it with Gaussian noise at noise level σ_{tr} to obtain the noisy latent $z_{\sigma_{\text{tr}}}$. A one-step Euler update is then applied to predict the makeup carrier $\hat{z}$:

$$\hat{z} = z_{\sigma_{\text{tr}}} - \sigma_{\text{tr}}\, v_\theta(z_{\sigma_{\text{tr}}}, \sigma_{\text{tr}};\, z_{\text{src}}, z_{\text{ref}})\,. \tag{4}$$

Rather than decoding $\hat{z}$ into a pixel-space image, we retain it within the continuous computational graph as a differentiable *makeup carrier*.

Reality-Anchored Reconstruction. The same transfer model (with shared weights) is then repurposed to reconstruct the real reference latent z_{ref}, conditioned solely on the bare-skin counterpart $z_{\text{bare\_ref}}$ and the predicted carrier $\hat{z}$. The corresponding refinement loss $\mathcal{L}_{\text{refine}}$ is defined as:

$$\mathcal{L}_{\text{refine}} = \mathcal{L}_{\text{FM}}\big(\theta; z_{\text{ref}}, z_{\text{bare\_ref}}, \hat{z}\big)\,. \tag{5}$$

Notably, since the makeup carrier $\hat{z}$ remains differentiable, the gradient of $\mathcal{L}_{\text{refine}}$ back-propagates smoothly through the DiT blocks directly into the velocity field v_θ responsible for the initial transfer prediction (Eq. 4). This formulation explicitly penalizes the omission of fine-grained makeup cues, encouraging the transfer prediction to preserve the details necessary for faithfully reconstructing the real reference.

While $\mathcal{L}_{\text{refine}}$ provides a high-fidelity refinement constraint, optimizing it alone could lead to degenerate solutions, such as neglecting the source identity to trivialize the refinement objective (Suppl. Sec. B.2). To mitigate this and stabilize the global makeup placement, we introduce $\mathcal{L}_{\text{bottleneck}}$ as a structural regularizer. Specifically, it enforces consistency with the pseudo-target prior exclusively at the noise bottleneck σ_{tr}:

$$\mathcal{L}_{\text{bottleneck}} = \mathbb{E}_\epsilon\Big[w(\sigma_{\text{tr}})\,\big\|v_\theta(z_{\sigma_{\text{tr}}}, \sigma_{\text{tr}};\, z_{\text{src}}, z_{\text{ref}}) - (\epsilon - z_{\text{pseudo}})\big\|_2^2\Big]\,, \tag{6}$$

where $z_{\sigma_{\text{tr}}}$ is the noised latent defined in Eq. (4).

The overall objective for Stage II is a weighted combination of the refinement loss $\mathcal{L}_{\text{refine}}$ and the structural regularizer $\mathcal{L}_{\text{bottleneck}}$:

$$\mathcal{L}_{\text{s2}} = \lambda_{\text{refine}}\,\mathcal{L}_{\text{refine}} + \lambda_{\text{bot}}\,\mathcal{L}_{\text{bottleneck}}\,. \tag{7}$$

Here, the controlled noise level σ_{tr} serves as a semantic information bottleneck that modulates the generative trade-off. $\mathcal{L}_{\text{bottleneck}}$ preserves the coarse makeup placement as a structural foundation, while affording the model sufficient flexibility to recover high-fidelity texture via $\mathcal{L}_{\text{refine}}$. An appropriate σ_{tr} (ablated in Sec. 4.4) ensures that global structure is robustly preserved while overriding synthetic artifacts.

3.3 2K-Resolution Makeup Dataset

MF2K. Existing makeup transfer datasets are often limited in image resolution and provide insufficient coverage of challenging cases [11,17,20,24,44], which fundamentally hinders the learning and evaluation of localized, fine-grained details. To address these limitations, we introduce MakeupFaces2K (MF2K), the first 2K-resolution makeup portrait dataset comprising 8,573 images at 2048×2048 resolution. MF2K covers a wide range of makeup intensities, including bare skin (3,139), light makeup (2,063), heavy makeup (1,798), and artistic styles (1,573),

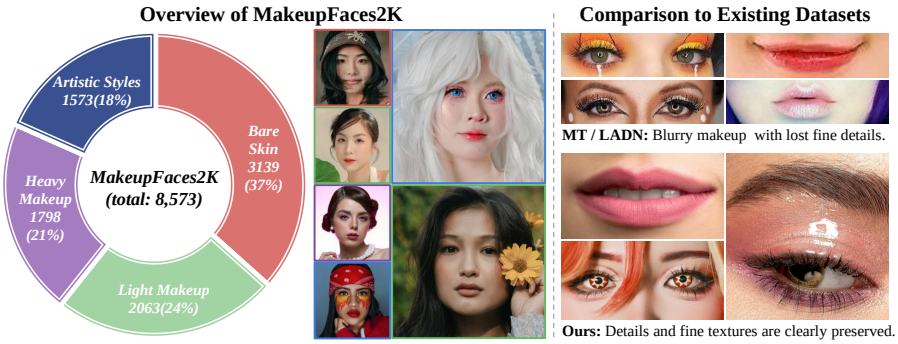

Fig. 3: MF2K, the first 2K-resolution in-the-wild makeup dataset. MF2K contains 8,573 images spanning four makeup intensities from bare skin to artistic styles, with broad demographic diversity. Compared with existing datasets, it captures fine-grained makeup textures and thus serves as a demanding benchmark for high-fidelity transfer. Zoom in to see the fine details.

and includes diverse demographics with improved coverage of male portraits. With its high-resolution and diverse styles, MF2K not only enables high-fidelity training but also constitutes a demanding benchmark for fine-grained makeup transfer. We summarize the dataset statistics and compare MF2K with existing public datasets in Fig. 3.

Pseudo-target triplets. As the MF2K dataset contains only bare-skin and makeup portraits and lacks paired ground-truth for makeup transfer, we construct the training triplets by sampling source–reference pairs $(I_{\text{src}}, I_{\text{ref}})$ and generating a corresponding pseudo-target I_{pseudo} using a large-scale image editing model $\mathcal{G}$:

$$I_{\text{pseudo}} = \mathcal{G}(I_{\text{src}}, I_{\text{ref}}). \tag{8}$$

To ensure balanced style coverage, we perform stratified sampling across makeup categories, yielding 2,000 source–reference pairs and their corresponding training triplets $\big(I_{\text{src}}, I_{\text{ref}}, I_{\text{pseudo}}\big)$. For the remaining images, which lack a pseudo-target, we predict the carrier from the source latent z_{src} in place of z_{pseudo}, and train the transfer model with the refinement loss alone, setting λ_{bot} to 0. More details about dataset statistics and sampling ratios are provided in Suppl. Sec. C.1.

4 Experiments

4.1 Experimental Settings

Implementation details. We employ the pre-trained FLUX.1-Kontext-dev [2] as our backbone and extend its Rotary Position Embedding (RoPE) with an explicit image-level index to support multi-image inputs. We fine-tune it via Low-Rank Adaptation (LoRA) [15] using the proposed two-stage training strategy. The LoRA rank and scaling factor are both set to 32. We optimize the model

using Prodigy [23] with a learning rate of 1. Training is conducted on 4 NVIDIA H20 GPUs, with a total batch size of 16 for 512×512 inputs and 4 for 2K-resolution inputs. Stage I is trained for 10k steps as pseudo-target initialization, followed by 10k steps of reality-anchored refinement in Stage II. During Stage II, the controlled noise level is fixed at $\sigma_{\text{tr}} = 0.6$ with loss weights $\lambda_{\text{refine}} = 1$ and $\lambda_{\text{bot}} = 0.2$. The remover $\mathcal{R}$ is pre-trained separately and remains frozen during the refinement stage. Other parameters are detailed in Suppl. Sec. B.3.

Baselines. We compare our method with representative approaches across three categories: GAN-based makeup transfer models, including PSGAN [17] and ElegANt [45]; diffusion-based methods, including StableMakeup [48], SHMT [36], and MAD [30]; and leading commercial image editing models, including Nano Banana Pro [10] and GPT Image 1.5 [25]. For all open-source baselines, we use the official implementations and follow the recommended settings.

4.2 Evaluation

Datasets. We evaluate our method on three public makeup datasets including MT [20], MT-Wild [17], and LADN [11], as well as our MF2K. For each dataset, we randomly sample 100 source–reference pairs with non-overlapping identities to form the evaluation set. In particular, to assess performance under complex scenarios, we sample the reference images in the MF2K evaluation set exclusively from its artistic-makeup subset. We enforce a strict separation between training and evaluation splits across all datasets to prevent data leakage.

Metrics. We evaluate makeup transfer quality from four dimensions: makeup similarity, identity preservation, background stability, and image quality. Makeup similarity is measured using a rubric-based Visual Language Model (VLM) evaluation, which assesses how accurately makeup textures and patterns from the reference are transferred to the output. Concretely, we use Gemini 3 Pro [9] and Qwen2.5-VL-72B-Instruct [1] as judges to assign a makeup similarity score on a 0–10 scale, denoted as MSim_G and MSim_Q, respectively. For fairness, all images fed into the VLM judges are resized to 512×512. Identity preservation is quantified by an ArcFace-based similarity metric (ID) [7], computed as the cosine similarity between ArcFace embeddings of the generated image and the source image. Background stability is evaluated using L2-M [48], which computes the pixel-averaged L_2 distance within the unedited background mask. We assess image quality with Fréchet Inception Distance (FID) [13], computed between generated outputs and real images from the evaluation set. Although prior work often employs feature-based distances (e.g., CLIP [28] and DINO [26, 31]) for makeup similarity, we find that these global metrics can be unreliable under certain makeup conditions. We analyze the failure modes and report the corresponding results together with qualitative evidence in Suppl. Sec. C.2.

4.3 Experiment Results

Qualitative results. Fig. 4 and Fig. 5 compare performance on simple daily styles and complex cosmetics (e.g., dense stickers, face paintings). GAN-based

Fig. 4: Qualitative comparison on simple makeup. Zoom in to see details.

Fig. 5: Qualitative comparison on complex makeup. Zoom in to see details.

methods (PSGAN, EleGANt) manage basic tones but produce unnatural blending on simple styles, and fail completely on complex geometries by merely applying global color tints. Previous diffusion methods (SHMT, StableMakeup) attempt to generate fine-grained patterns but often suffer from severe blurring,

Table 1: Quantitative comparison on four makeup transfer evaluation sets. Banana Pro and GPT 1.5 denote Nano Banana Pro and GPT Image 1.5. Best and second-best are in **bold** and underline.

Methods	MT Test Set [20]					LADN Test Set [11]				
	$MSim_G$↑	$MSim_Q$↑	ID↑	L2-M↓	FID↓	$MSim_G$↑	$MSim_Q$↑	ID↑	L2-M↓	FID↓
PSGAN [17]	6.13	5.26	.80	31.06	98.60	3.79	3.44	.80	17.32	108.21
EleGANt [45]	6.98	4.94	.83	34.95	95.56	5.64	4.82	.81	20.34	90.52
MAD [30]	6.60	5.10	.69	-	98.77	3.78	3.49	.60	-	96.02
SHMT [36]	5.87	4.22	.86	6.34	102.57	4.78	5.00	**.85**	5.57	91.58
StableMakeup [48]	8.34	7.30	.49	9.55	**82.70**	7.81	7.89	.47	8.91	**65.99**
Banana Pro [10]	8.34	7.07	.67	9.98	86.66	8.18	7.71	.67	38.40	70.75
GPT 1.5 [25]	8.69	**7.51**	.36	33.22	85.94	8.85	**8.13**	.35	43.59	69.14
Ours	**8.96**	7.12	**.87**	**3.94**	87.72	**9.14**	8.04	.81	**3.54**	68.67
Methods	**MT-Wild Test Set [17]**					**Our MF2K Test Set**				
	$MSim_G$↑	$MSim_Q$↑	ID↑	L2-M↓	FID↓	$MSim_G$↑	$MSim_Q$↑	ID↑	L2-M↓	FID↓
PSGAN [17]	3.65	2.71	.82	42.22	112.61	1.49	0.74	.79	45.23	166.79
EleGANt [45]	5.30	4.30	.82	43.86	113.24	2.60	1.87	.79	48.88	139.07
MAD [30]	3.93	2.85	.61	-	118.43	1.86	1.37	.57	-	156.41
SHMT [36]	3.11	2.45	**.88**	8.44	115.81	3.47	2.50	**.82**	6.71	133.97
StableMakeup [48]	8.21	**7.62**	.48	9.91	86.51	6.73	7.04	.43	9.26	**100.60**
Banana Pro [10]	8.45	7.39	.53	10.38	**80.48**	8.34	7.06	.65	13.27	104.45
GPT 1.5 [25]	8.74	7.53	.28	35.27	84.07	8.43	**7.69**	.35	28.32	102.94
Ours	**8.77**	7.00	.85	**4.37**	91.11	**9.22**	7.68	.74	**4.31**	103.27

skin tone shifts, and structural misregistration. MAD exhibits severe facial distortions, while commercial editing models frequently alter the intrinsic facial geometry and identity. In contrast, by leveraging the proposed reality-anchored refinement cycle, our method faithfully transfers the reference cosmetics while preserving sharp boundaries and source identity in both simple and complex settings.

Quantitative results. Tab. 1 shows that our method achieves the best overall balance across the four evaluation sets. It ranks first on $MSim_G$ across all four sets, by the largest margin on the artistic MF2K set (9.22 vs. 8.43), and remains competitive on $MSim_Q$; we provide results from more VLM judges in Suppl. Sec. C.2. For identity, ART obtains the highest ID on MT and the second-highest on LADN and MT-Wild, whereas methods that achieve higher identity, such as SHMT, do so at the expense of makeup fidelity. Background stability is likewise the strongest, with the lowest L2-M on every set, suppressing the unintended edits commonly inherited from synthetic pseudo-targets. On FID, our method scores higher than StableMakeup. FID measures global distributional similarity and is known to diverge from fine-grained perceptual quality [16, 33]; in our case, faithfully reproducing the reference's intricate artistic textures introduces patterns that are out-of-domain relative to the average-face statistics encoded by FID. Overall, these results demonstrate that the reality-anchored refinement cycle improves makeup fidelity without incurring the identity drift or background instability exhibited by competing methods.

User Study. To assess perceptual preferences, we collected 864 ratings from 21 participants, who scored the anonymized results in randomized order on a 1–5 Likert scale for Makeup Similarity, ID Consistency, and Image Quality; participant demographics and the rating rubrics are detailed in Suppl. Sec. C.3. As reported in Tab. 2, our method ranks first on all criteria. It achieves the highest Makeup Similarity (3.67) for faithful pattern transfer, and the best ID Consistency (4.03), indicating that improved makeup fidelity does not induce identity drift under human judgment. Meanwhile, our Image Quality (3.83) remains on par with the strongest commercial editors. Overall, human evaluations corroborate our quantitative findings and support the effectiveness of the reality-anchored refinement cycle.

Table 2: User study results.

Methods	Makeup Similarity ↑	ID Consistency ↑	Image Quality ↑
PSGAN [17]	1.61	3.00	2.07
EleGANt [45]	2.01	3.03	2.23
MAD [30]	1.51	1.97	1.51
SHMT [36]	1.73	<u>3.21</u>	2.19
StableMakeup [48]	2.75	2.28	2.73
Banana Pro [10]	<u>3.43</u>	3.12	3.81
GPT 1.5 [25]	3.31	2.15	<u>3.82</u>
Ours	**3.67**	**4.03**	**3.83**

4.4 Ablation Studies

We conduct ablation studies to isolate the contribution of each design, covering the two-stage training strategy, the controlled-noise bottleneck, and the pseudo-target supervision and data. Unless otherwise specified, all evaluations follow the settings detailed in Sec. 4.2.

Two-stage training. We evaluate the effectiveness of the proposed framework by comparing the full model (Stage I & II) against baselines trained solely via Stage I or Stage II. As reported in Tab. 3, training with Stage I alone establishes basic makeup placement but exhibits pronounced identity drift, since it merely imitates the imperfect pseudo-targets. Conversely, training with Stage II alone preserves identity but cannot transfer makeup faithfully: without the Stage I initialization, the makeup carrier is largely uninformative, so reconstructing the reference yields only a weak signal for learning the transfer. Our full framework combines the complementary strengths of both stages. It achieves the highest MSim_G (9.22 on MF2K), capturing fine-grained cosmetic details, while substantially improving identity preservation and attaining the lowest L2-M (4.31 on MF2K) to reduce unintended background modifications. Overall, these results validate that Stage I is crucial for the structural foundation, while the Stage II refinement cycle effectively breaks the pseudo-target ceiling to recover fine-grained patterns and mitigate synthetic artifacts.

Controlled-noise bottleneck. We analyze the sensitivity of the controlled noise level σ_{tr} in the Stage II refinement cycle on the MF2K evaluation set. Fig. 6 reveals a clear trade-off. A small σ_{tr} tightly constrains the refinement, so the model edits conservatively, preserving identity but recovering few fine-grained textures and thus lowering MSim. A larger σ_{tr} relaxes this constraint and raises MSim, but the weaker structural guidance lowers both identity similarity and

Table 3: Ablation of the two-stage training strategy.

Variants	MT Test Set [20]					LADN Test Set [11]				
	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓
Stage I only	7.81	**7.47**	.56	4.87	**84.93**	8.30	8.00	.60	11.52	**63.14**
Stage II only	6.96	5.35	**.93**	4.45	92.72	8.34	7.79	**.81**	4.38	70.64
Stage I&II	**8.96**	7.12	.87	**3.94**	87.72	**9.14**	**8.04**	**.81**	**3.54**	68.67
Variants	**MT-Wild Test Set [17]**					**Our MF2K Test Set**				
	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓
Stage I only	**8.77**	**7.78**	.46	4.59	**83.73**	8.82	**7.74**	.57	7.68	**96.85**
Stage II only	6.60	5.50	**.91**	5.02	101.58	8.08	6.72	.70	4.98	103.97
Stage I&II	**8.77**	7.00	.85	**4.37**	91.11	**9.22**	7.68	**.74**	**4.31**	103.27

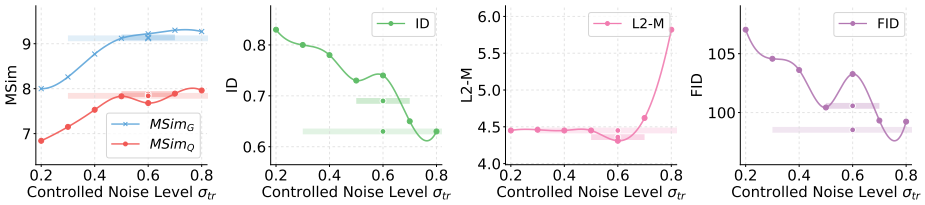

Fig. 6: Sensitivity to the controlled noise level σ_{tr}. Markers denote discrete, fixed σ_{tr} values, and shaded horizontal bands denote σ_{tr} sampled from the corresponding continuous interval.

background stability (higher L2-M). We therefore set $\sigma_{\text{tr}} = 0.6$ by default, which provides a strong balance between cosmetic fidelity and identity preservation while achieving the best background stability.

Impact of pseudo-target supervision and data. Tab. 4 and Fig. 7 analyze how the pseudo-target source and the training data affect transfer. We train our framework on pseudo-targets from either Nano Banana Pro (Ours_{NBP}, our default) or StableMakeup (Ours_{SM}). Although StableMakeup produces lower-quality pseudo-targets, Ours_{SM} still improves markedly over StableMakeup itself (8.56 vs. 6.73 in MF2K MSim_G) and nearly matches Ours_{NBP}, showing that the refinement re-anchors the model to real references rather than being capped by its initial pseudo-targets. We then retrain StableMakeup on our MF2K data ($\text{StableMakeup}_{\text{MF2K}}$) to isolate the contribution of the dataset itself. Its improvements are limited to simpler styles (MT, MT-Wild) and reverse on the complex makeup of LADN and MF2K, where StableMakeup's global CLIP conditioning cannot represent such fine-grained detail. Overall, these results confirm that, while higher-quality pseudo-targets help, the reality-anchored refinement is the key to surpassing them and breaking the pseudo-target ceiling.

4.5 Resolution Scaling

To explore high-resolution portrait makeup transfer, we perform experiments at spatial resolutions of 512, 1024, and 2048 on our MF2K dataset. To supervise

Table 4: Ablation of the pseudo-target source and training data. $\text{StableMakeup}_{\text{MF2K}}$ denotes StableMakeup trained on our MF2K data. Ours_{SM} and Ours_{NBP} denote our models trained with pseudo-targets synthesized by StableMakeup and Nano Banana Pro, respectively.

Methods	MT Test Set [20]					LADN Test Set [11]				
	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓
StableMakeup [48]	8.34	7.30	.49	9.55	**82.70**	7.81	7.89	.47	8.91	**65.99**
$\text{StableMakeup}_{\text{MF2K}}$	8.47	7.47	.60	30.59	88.22	7.47	7.75	.52	51.60	74.67
Ours_{SM}	8.83	**7.53**	.65	4.79	87.07	9.07	**8.46**	.61	4.24	67.68
Ours_{NBP}	**8.96**	7.12	**.87**	**3.94**	87.72	**9.14**	8.04	**.81**	**3.54**	68.67
Methods	**MT-Wild Test Set [17]**					**Our MF2K Test Set**				
	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓	$\text{MSim}_G\uparrow$	$\text{MSim}_Q\uparrow$	ID↑	L2-M↓	FID↓
StableMakeup [48]	8.21	**7.62**	.48	9.91	**86.51**	6.73	7.04	.43	9.26	100.60
$\text{StableMakeup}_{\text{MF2K}}$	8.66	7.62	.42	21.48	90.12	6.58	6.92	.51	28.17	100.67
Ours_{SM}	**8.99**	7.45	.75	5.46	88.60	8.56	**8.16**	.51	5.90	**95.84**
Ours_{NBP}	8.77	7.00	**.85**	**4.37**	91.11	**9.22**	7.68	**.74**	**4.31**	103.27

Fig. 7: Effect of the pseudo-target source and training data. Even from weaker pseudo-targets, Ours_{SM} matches Ours_{NBP} in recovering the fine-grained makeup, whereas StableMakeup fails even when retrained on MF2K ($\text{StableMakeup}_{\text{MF2K}}$), showing that the reality-anchored refinement breaks the pseudo-target ceiling.

fine details in high-resolution training, we add a wavelet-based loss [46] that uses a discrete wavelet transform (DWT) to emphasize high-frequency detail while preserving low-frequency structure. Formally, it is defined as $\mathcal{L}_{\text{WLF}} = \mathbb{E}\left[w_t \| f(v_\theta(z_t, t)) - f(\epsilon - x_0)\|^2\right]$, where w_t is the loss weight, and $f(\cdot)$ denotes the DWT. As shown in Fig. 8, increasing the resolution progressively recovers finer and sharper makeup details. To our knowledge, this is the first high-fidelity makeup transfer at 2K resolution, directly meeting the high-resolution demands of real-world applications.

4.6 Limitations and Failure Cases

Although our framework faithfully transfers visual appearance cues, it does not explicitly enforce physical lighting and reflectance consistency. Consequently, significant illumination mismatches can yield contradictory results, such as retaining intense specular highlights from a directionally lit reference within a

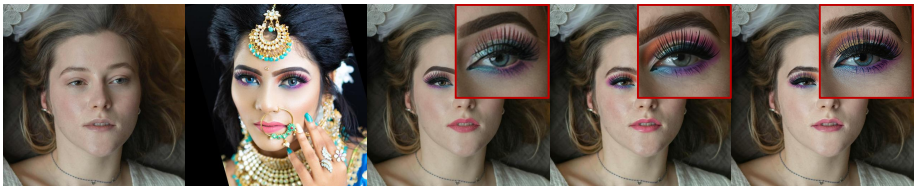

Fig. 8: Resolution scaling examples. We show 512/1024/2048 outputs with aligned zoom-in crops. Higher resolution recovers finer and sharper makeup details.

softly lit source environment. A detailed discussion of limitations is provided in Suppl. Sec. D.

5 Conclusion

In this paper, we presented ART, a two-stage framework for high-fidelity makeup transfer that breaks the pseudo-target ceiling. Through a reality-anchored refinement cycle, the transfer model recovers fine-grained makeup textures and precise placement by using the reconstruction of the real reference as its supervisory signal, while a controlled-noise bottleneck stabilizes the refinement and preserves global structure. We further introduced MakeupFaces2K, a 2K-resolution in-the-wild makeup portrait dataset that supports research on high-resolution makeup transfer and related facial editing tasks. Extensive experiments show that our method delivers the best makeup fidelity and background stability while robustly preserving identity. Beyond makeup transfer, the proposed reality-anchored refinement offers a general paradigm for DiT-based image editing.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant No. 2022YFA1004100.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
2. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)

3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision. pp. 1021–1030 (2017)
4. Chang, H., Lu, J., Yu, F., Finkelstein, A.: Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 40–48 (2018)
5. Chen, H.J., Hui, K.M., Wang, S.Y., Tsao, L.W., Shuai, H.H., Cheng, W.H.: Beautyglow: On-demand makeup transfer framework with reversible generative network. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 10042–10050 (2019)
6. Deng, H., Han, C., Cai, H., Han, G., He, S.: Spatially-invariant style-codes controlled makeup transfer. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 6549–6557 (2021)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: 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)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. Google DeepMind: Gemini 3 Pro. `https://deepmind.google/models/gemini/pro/` (2025), accessed: March 5, 2026
10. Google DeepMind: Gemini 3 Pro Image. `https://deepmind.google/models/gemini-image/pro/` (2025), accessed: March 5, 2026
11. Gu, Q., Wang, G., Chiu, M.T., Tai, Y.W., Tang, C.K.: Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
12. He, F., Li, H., Ning, X., Li, Q.: Beautydiffusion: Generative latent decomposition for makeup transfer via diffusion models. Information Fusion **123**, 103241 (2025)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
16. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9307–9315 (2024)
17. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: 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)
18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
19. Kips, R., Gori, P., Perrot, M., Bloch, I.: Ca-gan: Weakly supervised color aware gan for controllable makeup transfer. In: European conference on computer vision. pp. 280–296. Springer (2020)
20. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: 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)

21. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
22. Liu, S., Jiang, W., Gao, C., He, R., Feng, J., Li, B., Yan, S.: Psgan++: Robust detail-preserving makeup transfer and removal. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8538–8551 (2021)
23. Mishchenko, K., Defazio, A.: Prodigy: An expeditiously adaptive parameter-free learner. arXiv preprint arXiv:2306.06101 (2023)
24. Nguyen, T., Tran, A.T., Hoai, M.: 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 (CVPR). pp. 13305–13314 (June 2021)
25. OpenAI: GPT Image 1.5 Model. `https://developers.openai.com/api/docs/models/gpt-image-1.5` (2025), accessed: March 5, 2026
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
30. Ruan, B.K., Shuai, H.H.: Mad: Makeup all-in-one with cross-domain diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 749–758 (2025)
31. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
32. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
33. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. Advances in Neural Information Processing Systems **36**, 3732–3784 (2023)
34. Sun, Z., Chen, Y., Xiong, S.: Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. In: Proceedings of the AAAI Conference on artificial intelligence. vol. 36, pp. 2325–2334 (2022)
35. Sun, Z., Chen, Y., Xiong, S.: Ssat++: A semantic-aware and versatile makeup transfer network with local color consistency constraint. IEEE Transactions on Neural Networks and Learning Systems **36**(1), 1287–1301 (2023)
36. Sun, Z., Xiong, S., Chen, Y., Du, F., Chen, W., Wang, F., Rong, Y.: Shmt: Self-supervised hierarchical makeup transfer via latent diffusion models. Advances in Neural Information Processing Systems **37**, 16016–16042 (2024)
37. Sun, Z., Xiong, S., Chen, Y., Rong, Y.: 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)

38. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
39. Tan, Z., Xue, Q., Yang, X., Liu, S., Wang, X.: Ominicontrol2: Efficient conditioning for diffusion transformers. arXiv preprint arXiv:2503.08280 (2025)
40. Wan, Z., Chen, H., An, J., Jiang, W., Yao, C., Luo, J.: Facial attribute transformers for precise and robust makeup transfer. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1717–1726 (2022)
41. Wu, H., Fu, Y., Li, Y., Gao, Y., Du, K.: Evomakeup: High-fidelity and controllable makeup editing with makeupquad. arXiv preprint arXiv:2508.05994 (2025)
42. Xiang, J., Chen, J., Liu, W., Hou, X., Shen, L.: Ramgan: Region attentive morphing gan for region-level makeup transfer. In: European conference on computer vision. pp. 719–735. Springer (2022)
43. Xu, S., Ma, Z., Huang, Y., Lee, H., Chai, J.: Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. Advances in Neural Information Processing Systems **36**, 10359–10384 (2023)
44. Yan, Q., Guo, C., Zhao, J., Dai, Y., Loy, C.C., Li, C.: Beautyrec: Robust, efficient, and component-specific makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1102–1110 (June 2023)
45. Yang, C., He, W., Xu, Y., Gao, Y.: Elegant: Exquisite and locally editable gan for makeup transfer. In: European conference on computer vision. pp. 737–754. Springer (2022)
46. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23464–23473 (2025)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
48. Zhang, Y., Yuan, Y., Song, Y., Liu, J.: 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)
49. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19513–19524 (2025)
50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
51. Zhu, M., Yi, Y., Wang, N., Wang, X., Gao, X.: Semi-parametric makeup transfer via semantic-aware correspondence. arXiv preprint arXiv:2203.02286 (2022)
52. Zou, S., Huang, Y., Yi, R., Zhu, C., Xu, K.: Cyclediff: Cycle diffusion models for unpaired image-to-image translation. arXiv preprint arXiv:2508.06625 (2025)