

MagicMakeup: A Region-Controllable Diffusion Transformer for High-Fidelity Makeup-Transfer

Ziyi Wang^{1,2*}, Siming Zheng^{2*}, Yang Yang^{1,2}, Shusong Xu²,
Hao Zhang², Bo Li², Changqing Zou^{1†}, Peng-Tao Jiang^{2†}

¹ State Key Lab of CAD&CG, Zhejiang University, China

² vivo BlueImage Lab, vivo Mobile Communication Co., Ltd., China

<https://vivocameraresearch.github.io/magicmakeup>



Fig. 1: MagicMakeup: a diffusion transformer for high-fidelity, region-controllable makeup transfer. It supports mixing references for eyes and lips, single-region editing, and full-face makeup application.

Abstract. Makeup-transfer applies the reference makeup to the source face while preserving the source identity. Despite advances in full-face editing by diffusion-based methods, strong regional controllability, makeup fidelity, and identity preservation remain challenging. The reasons are (i) pixel-to-attention misalignment that causes spillover into non-target

* Equal contributions; † Corresponding author.

areas and weakens regional control; (ii) unclear transfer/preservation concept separation under two-image conditioning, leading to coupling between makeup attributes and identity; and (iii) the lack of a high-resolution dataset that is identity-consistent and region-labeled for fine-grained supervision. In this paper, we propose MagicMakeup, a diffusion transformer-based framework for region-controllable and high-fidelity makeup transfer, built on spatial constraints and concept disentanglement. To enable precise region-specific editing while preserving identity, we propose Token-Aligned Region Gating, which aligns pixel masks with attention and applies region-specific logit gating. To clarify the concepts of transfer and preservation, we further introduce Cross-Modal Perception Guidance, which aligns text and image features to enhance cross-modal concept perception. We also design a pipeline for the generation of 1024×1024 data pairs through region-specific makeup removal and establish a unified benchmark in synthetic and real settings. Extensive quantitative and qualitative experiments show that MagicMakeup improves regional controllability, makeup fidelity, and identity preservation, with strong robustness across styles, races, and poses.

Keywords: Makeup-Transfer · Diffusion Models · Image Editing

1 Introduction

Given a source face and a makeup reference, Makeup-transfer generates an output that preserves the source identity and facial geometry while transferring the reference makeup (color, texture, style), supporting full-face and regional editing. Compared with generic face editing, makeup-transfer places greater emphasis on identity preservation, region-specific control, and realistic high-resolution results, supporting applications such as virtual makeup try-on and personalized digital avatars.

Recently, several methods [31, 33, 35, 36, 44, 48] have been proposed for the makeup-transfer task. Most use pretrained diffusion models or Diffusion Transformers (DiT), integrate reference makeup through cross-attention or feature concatenation, and are trained on unpaired or synthetic pseudo-paired data. However, achieving regional controllability, makeup fidelity, and identity preservation simultaneously remains challenging for three reasons, as shown in Fig. 2:

(i) Limited spatial controllability. Many diffusion methods [33, 35, 44, 48] support only full-face transfer. Even with pixel masks [31], constraints misalign in the attention space, causing spillover and weakening region-specific control and precision. **(ii) Conceptual ambiguity.** In two-image conditioning with a source face and a makeup reference, current models often blur the transfer/preservation concepts. They fail to restrict transfer to makeup attributes while preserving source identity and facial geometry, which couples makeup and identity. As a result, the makeup details are not faithfully transferred, or the reference facial structure leaks into the output. **(iii) Limited data quality and availability.** High-quality real makeup/non-makeup pairs are scarce and often misaligned in expression, background, and lighting, which makes them

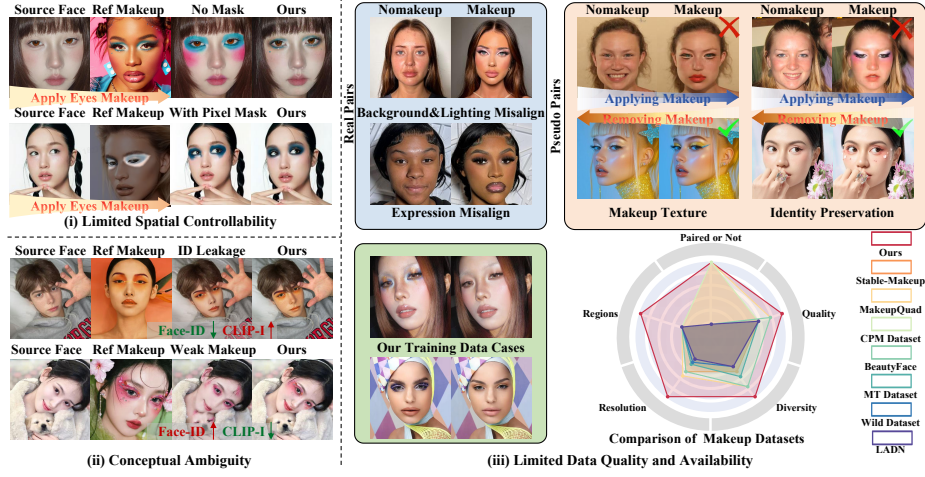


Fig. 2: Three main challenges in makeup transfer: (i) Limited Spatial Controllability, where pixel-to-attention misalignment prevents reliable region-specific control, causing makeup spillover into non-target areas and weakening precision in localized edits. (ii) Conceptual Ambiguity, where the separation between identity preservation and makeup transfer is unclear, leading to identity-attribute coupling. This causes either the reference facial structure to leak into the output or the makeup details to be inadequately transferred. (iii) Limited data quality and availability, where real makeup/non-makeup pairs are scarce and often misaligned in expression, background, and lighting, leading to reliance on makeup-applied pseudo-pairs that lack texture, compromise identity, and introduce distribution shift.

unsuitable for supervised training. As a result, most works use unpaired data [9, 16, 22, 31, 39, 41, 42, 47] or synthesize pseudo-pairs by applying makeup to the source via text prompts [26, 44, 48], but the generated results lack rich makeup texture and often fail to preserve identity, leading to a distribution shift from real data. In addition, existing benchmarks lack region annotations and are of low resolution [9, 16, 22, 41], which limits fine-grained, region-specific evaluation.

To mitigate these gaps, we propose MagicMakeup, a DiT-based framework for region-controllable, high-fidelity makeup transfer, built on token-aligned region constraints and transfer/preservation disentanglement. To impose region-specific constraints, we introduce Token-Aligned Region Gating (TARG), which maps pixel masks to tokens and applies region-specific logit gating. To clarify the transfer/preservation concepts, we introduce Cross-Modal Perception Guidance (CMPG), which strengthens cross-modal alignment between text and image features to decide which makeup attributes to transfer from the makeup reference and which identity features to preserve from the source image.

Furthermore, we develop an automated data pipeline to construct paired makeup and non-makeup images with region masks. We obtain pairs via region-specific makeup removal on 1024×1024 real makeup images rather than syn-

thesizing makeup. Makeup removal better preserves facial structure, producing non-makeup data closer to the real-world distribution, while the original makeup images retain diverse styles and textures, providing realistic, aligned supervision. Finally, we establish a unified high-resolution benchmark across synthetic and real settings, covering diverse demographics and fine-grained facial regions, with standardized evaluation of identity preservation, makeup fidelity, and regional control accuracy. Our contributions to the community are threefold:

- We propose MagicMakeup, a DiT-based, region-controllable framework achieving state-of-the-art performance across diverse styles and real-world data and showing strong robustness and generalization.
- We introduce TARG and CMPG to (i) enforce token-aligned, region-specific attention constraints and (ii) disentangle transfer from preservation, enabling precise regional makeup transfer with reduced spillover and improved identity consistency.
- We build an automated “real makeup to regional makeup removal” data construction pipeline that produces high-resolution, identity-consistent, texture-faithful pairs enabling realistic, well-aligned supervision. And we establish a unified high-resolution benchmark that spans synthetic and real settings for fair evaluation.

2 Related Work

2.1 Facial Makeup-Transfer

Makeup-transfer has advanced rapidly. Early GAN methods [6, 22, 41] preserved identity via alignment and identity/makeup disentanglement. PSGAN [16] captures local detail but needs hand-tuned boundaries; CPM [26] uses UV correspondence yet fails across identities and large style gaps; SCGAN [47] and EleGANt [42] introduce attention and region-wise semantics. Overall, these methods depend on strict alignment and delicate losses, and still suffer mask misalignment and boundary leakage under large style changes.

With the rise of diffusion models [7, 12, 27, 29, 30, 32, 43], newer methods address content boundaries more directly through stronger generation and semantic conditioning. DiffAM [33] enables attribute-conditioned edits, but sparse or ambiguous labels make the preserve/transfer boundary unstable. SHMT [35] builds pseudo-paired data via a 3D face model with Laplacian contours and stabilizes geometry using a hierarchical Transformer, but it still suffers from boundary noise and weak fusion in overlapping regions. Stable-Makeup [44] uses CLIP-guided text and image conditioning with face-control modules, but once alignment drifts, errors amplify and spread across regions. MAD [31] enables region-wise makeup application but still suffers from makeup spillover, concept ambiguity, and high computational cost at high resolutions. Despite progress, the field still lacks a testable criterion for deciding what to preserve and what to transfer, and a spatial control mechanism fully aligned with internal representations.

2.2 Makeup-Transfer Datasets

Given the difficulty of collecting high-quality paired makeup data, makeup transfer has long relied on unpaired training. MT [22] introduced this paradigm with surrogate losses such as color-histogram matching. LADN [9] expanded data diversity and keypoint annotations, and Makeup-Wild [16] further increased scale, but source-reference identity mismatch limited strict supervision. MT-HR [23], and BeautyFace [41] improved resolution and annotations, and MT-Text [31] added text-driven control. However, strongly paired makeup/non-makeup samples with the same identity, similar pose, and clear region masks remain rare, and available data are often low resolution (e.g., 256×256). These gaps hinder fine-grained, controllable, and distribution-faithful supervision.

With the rise of image editing models [1–3, 5, 8, 13, 14, 20, 21, 25, 38, 45, 46], many works turn to pseudo-paired data: SHMT [35] derives pairs from a single reference using a 3D face model [10] with Laplacian contours; Stable-Makeup [44], BeautyBank [24] and Flux-Makeup [48] apply makeup to source images to synthesize pairs at scale, but limited application quality and predefined styles yield makeup outputs with limited identity alignment, style diversity, and visual quality, causing distribution shift from real data. Moreover, support for region-wise editing is weak. Thus, building high-quality, style-diverse, and scalable strongly paired datasets is a key challenge for makeup-transfer.

3 Method

3.1 Overview

The overall pipeline of MagicMakeup is illustrated in Fig. 3. Given a source face $\mathbf{I}_s$ and a reference makeup image $\mathbf{I}_r$, the goal is to generate an output image $\mathbf{I}_o$ that transfers the fine-grained makeup details from $\mathbf{I}_r$ while preserving the identity and facial structure of $\mathbf{I}_s$. To achieve this, we first encode both $\mathbf{I}_s$ and $\mathbf{I}_r$ into token sequences $\mathbf{X}_s$ and $\mathbf{X}_r$, and sample a noise token $\mathbf{X}_n$ for the generation process. Simultaneously, the text prompt $\mathbf{T}$ is encoded into text embeddings $\mathbf{X}_T$. These visual and textual tokens are then processed jointly by our framework to iteratively denoise the latent token $\mathbf{X}_n$ and synthesize the target image tokens.

To ensure makeup is transferred only to designated regions of the face and to prevent spillover, we introduce TARG, which uses region masks $\mathbf{M}_s$ and $\mathbf{M}_r$ for the source and reference faces, respectively, to restrict attention within the target facial regions. Additionally, to clarify and strengthen the concepts of preservation and transfer, we propose CMPG, which modulates both the image and text features. CMPG aligns the preservation concept $\mathbf{C}_s$ with the source image $\mathbf{I}_s$ and the transfer concept $\mathbf{C}_r$ with the reference image $\mathbf{I}_r$, updating both image and text features accordingly. The updated features $\mathbf{X}_s^*$, $\mathbf{X}_r^*$, $\mathbf{X}_{C_s}^*$, and $\mathbf{X}_{C_r}^*$ are then integrated by MM-DiT. Finally, an image decoder reconstructs the denoised tokens into the final output image $\mathbf{I}_o$, successfully transferring the makeup of $\mathbf{I}_r$ and retaining the identity of $\mathbf{I}_s$.

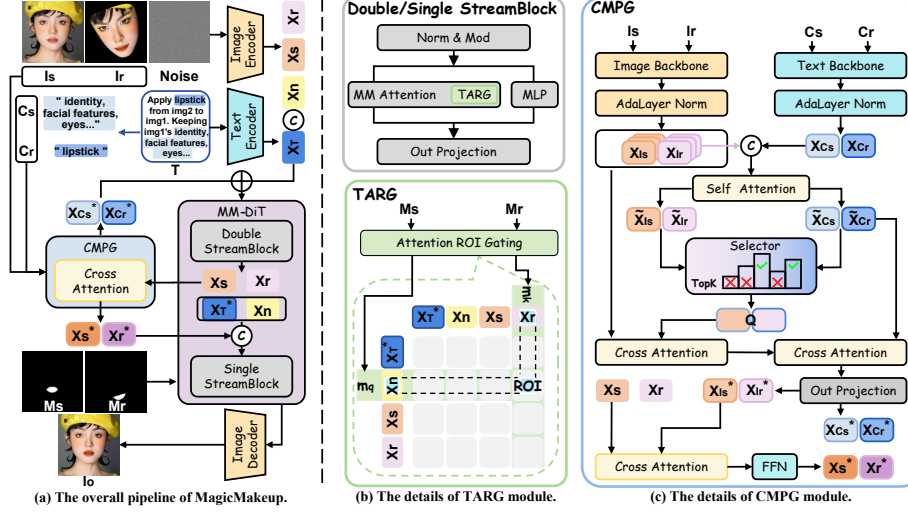


Fig. 3: The overview of our MagicMakeup pipeline, as shown in (a). We introduce Token-Aligned Region Gating within DiT’s multi-modal attention, which maps pixel masks to tokens and applies attention ROI gating to impose region-specific constraints, as shown in (b). We introduce Cross-Modal Perception Guidance, which strengthens cross-modal alignment between text and image features to clarify the preservation/transfer concepts, as shown in (c).

3.2 Token-Aligned Region Gating

Most existing diffusion-based makeup-transfer pipelines are optimized for full-face transfer, making region-specific edits on eyes and lips unreliable. To enable precise transfer on three target regions, namely face, eyes, and lips, we propose Token-Aligned Region Gating (TARG) using refined facial masks. A key challenge is that pixel masks do not reliably constrain attention-based information flow, since pixel masks are misaligned with token interactions, allowing makeup cues from the reference to leak outside the region of interest (ROI). TARG resolves this mismatch by projecting the ROI mask onto the token grid and enforcing the constraint directly in attention.

For each target region (face, eyes, or lips), we downsample and flatten the source and reference masks $\mathbf{M}_s$ and $\mathbf{M}_r$ to the token grids of MM-Attention, yielding a latent query gate $\mathbf{m}_q \in \{0, 1\}^{L_n}$ and a reference key gate $\mathbf{m}_k \in \{0, 1\}^{L_r}$, with lengths L_n and L_r for $\mathbf{X}_n$ and $\mathbf{I}_r$. We define an additive logit mask on latent-to-reference token pairs as

$$M_{i,j} = \begin{cases} 0, & \mathbf{m}_q(i) = 1 \text{ and } \mathbf{m}_k(j) = 1, \\ -\infty, & \text{otherwise,} \end{cases} \quad (1)$$

where i and j index latent-query and reference-key tokens, respectively. In practice, $-\infty$ is implemented with a sufficiently large negative constant. The masked

attention is computed as

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right) \mathbf{V}. \quad (2)$$

We apply $\mathbf{M}$ only to the latent-query and reference-key interactions and keep all other attention terms unchanged. As a result, latent tokens outside the ROI cannot attend to $\mathbf{I}_r$, suppressing cross-region leakage and enabling precise region-level makeup transfer.

3.3 Cross-Modal Perception Guidance

Unlike generic face editing, makeup-transfer requires conditioning on two images: the source face and the makeup reference. This makes it challenging to transfer makeup attributes while preserving identity geometry without mixing them. To address this, we propose Cross-Modal Perception Guidance (CMPG), which semantically decouples what to preserve and what to transfer, and jointly strengthens both the image and text representations for these two concepts.

From the instruction $\mathbf{T}$, we extract two concept phrases: a preservation concept $\mathbf{C}_s$ and a transfer concept $\mathbf{C}_r$. Intuitively, $\mathbf{C}_s$ describes identity-related cues (e.g., *identity*, *facial features*...), while $\mathbf{C}_r$ specifies the target makeup attributes in the edited region (e.g., *lipstick* for lip makeup or *eyeshadow* for eye makeup). At each denoising step (Fig. 3(c)), CMPG processes two concept-image pairs $(\mathbf{C}_s, \mathbf{I}_s)$ and $(\mathbf{C}_r, \mathbf{I}_r)$ and outputs concept-aware text modulations $\mathbf{X}_{C_s}^*, \mathbf{X}_{C_r}^*$ together with image modulations $\mathbf{X}_{I_s}^*, \mathbf{X}_{I_r}^*$. The goal is to make preservation/transfer cues more explicit in the text stream and to make each image stream respond to its corresponding concept without cross-contamination.

Concept-guided evidence selection. We extract concept tokens $\mathbf{X}_C$ from a frozen text backbone and image tokens $\mathbf{X}_I$ from the last layers of a frozen image backbone, where $(\mathbf{C}, \mathbf{I})$ denotes the current pair, i.e., $(\mathbf{C}_s, \mathbf{I}_s)$ or $(\mathbf{C}_r, \mathbf{I}_r)$. For each concept-image pair, we first co-update the two token sets with a joint self-attention block to align concept semantics with responsive image regions, yielding $\tilde{\mathbf{X}}_C$ and $\tilde{\mathbf{X}}_I$. To focus subsequent updates on concept-relevant evidence, we rank image tokens by their similarity to the concept tokens and select a compact set of top- k visual queries:

$$\mathbf{Q} = \tilde{\mathbf{X}}_I \left[\text{Top-}k(\text{Max}(\tilde{\mathbf{X}}_I \tilde{\mathbf{X}}_C^\top)) \right], \quad (3)$$

where $\text{Max}(\cdot)$ pools similarities over concept tokens. $\mathbf{Q}$ captures the most concept-aligned regions and serves as shared evidence for updating both modalities.

Updating image and text perceptions. Starting from $\mathbf{Q}$, we refine the queries with two cross-attentions. The first attends to a multi-layer visual context bank (from the last three backbone layers) of the corresponding image, and the

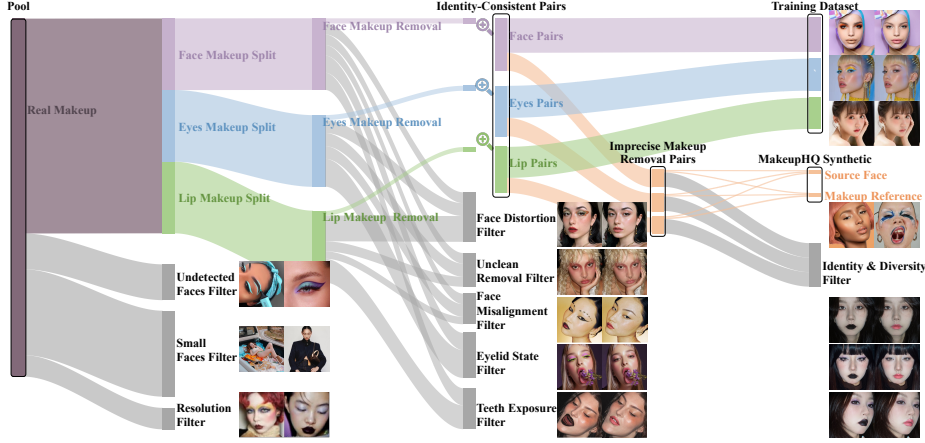


Fig. 4: Diagram of our training and synthetic dataset construction pipeline. It shows the filtering stages for face distortion, unclean removal, and imprecise regions, resulting in identity-consistent, region-labeled makeup/non-makeup pairs across face, eye, and lip splits.

second attends to the concept tokens. AdaLayerNorm heads map the refined queries to image modulation and span-aware text modulation.

We write the text modulation $\mathbf{X}_{C_s}^*$, $\mathbf{X}_{C_r}^*$ only on the token spans of $\mathbf{C}_s$ and $\mathbf{C}_r$ in the text sequence $\mathbf{X}_T$, which is then fed into MM-DiT as textual conditioning. The image modulation $\mathbf{X}_{I_s}^*$, $\mathbf{X}_{I_r}^*$ is injected into the corresponding image streams $\mathbf{X}_s, \mathbf{X}_r$ via a cross-attention adapter, and the modulated features are further processed by the DiT single-stream blocks.

By strengthening concept-aware representations for both text and images at each denoising step, CMPG provides timestep-conditioned and concept-disentangled guidance, reducing identity leakage and attribute drift while improving local makeup fidelity.

3.4 Dataset Construction Pipeline

Makeup/No-Makeup Pairs. In practice, we attempted to synthesize paired data by applying makeup with Flux-Kontext [20], GPT-4o [15], and Nano-Banana [37]. Both text-guided and reference-image-guided generation are unreliable: the former often yields unrealistic makeup with weak details and limited diversity, while the latter frequently distorts facial structure and still fails to faithfully transfer the reference makeup as shown in Fig. 2 (iii). We therefore construct pairs via region-specific makeup removal from real makeup photos. This is more stable because facial geometry is strongly constrained by the input and the model’s facial prior, whereas adding makeup requires inventing high-frequency, style-specific textures. Makeup removal acts like subtracting a local appearance layer, preserving identity and providing realistic, aligned supervision.

As illustrated in Fig. 4, we begin with a pool of 50,000 real 1024×1024 makeup faces spanning diverse styles and poses. After removing undetected, very small faces, and low-resolution samples, we crop and normalize the remaining images. This produces about 3,000 standardized images, which are then grouped into face, eyes, and lip makeup splits. For each split, FLUX-Kontext [20] performs region-specific makeup removal (full face, eyes, lips), generating multiple non-makeup candidates which are then matched with corresponding makeup pairs.

We then apply a cascade of filters. First, identity-consistent pairs are obtained by removing face distortion and misalignment: DFA-MobileNet [19] aligns faces, AdaFace [18] similarity removes unchanged or distorted results, and landmark-mask differences eliminate eyelid-state and teeth-exposure mismatches. Next, an unclean-removal filter ensures sufficient regional color-histogram change; among remaining candidates, the best pair per group is selected using a balanced score combining identity similarity and regional color difference. This yields around 10,000 clean, identity-consistent pairs. Finally, an imprecise-region filter enforces large change within the labeled region and small change elsewhere, leaving 6,772 region-labeled pairs (2,392 face, 2,224 eyes, 2,156 lip). During training, the non-makeup image serves as the source and the makeup image as the GT.

Compared with applying makeup to bare faces, removal better preserves facial structure for the non-makeup, while the makeup retains authentic texture and style diversity. This provides realistic and aligned supervision for training.

Reference Makeup. A challenge is to construct references that differ in identity from the makeup image while preserving the makeup appearance. Prior method [44] modifies only a small subset of facial landmarks, yielding insufficient deformation and causing structural leakage. We instead construct an identity-separated reference by choosing, for each makeup image, a face with a large identity difference as the deformation target and warping makeup to that geometry using 478 facial landmarks, Delaunay triangulation with piecewise-affine transforms, and edge smoothing. The resulting reference image preserves the makeup image’s makeup texture and style but adopts a new identity and facial layout, which better decouples identity from makeup.

Masks. We obtain a full-face mask from face parsing [40] and generate eye and lip masks from the MediaPipe Face Landmarker (468 points) [17] using predefined region landmark sets. Since eyeshadow is not directly available, we construct a soft eyeshadow mask by dilating and shifting the eye mask, multiplying it with a radial gradient, and smoothing the boundary.

3.5 MakeupHQ Benchmark

Current benchmarks are low-resolution, lack dedicated splits for full-face, eye, and lip makeup, and exhibit small identity gaps with limited style, pose, and expression diversity, which fails to support the fine-grained, region-specific makeup-transfer needed in real-world applications. To address this, we introduce MakeupHQ Bench as a supplementary benchmark for makeup transfer, consisting of

two test sets: (i) MakeupHQ-Synthetic, a synthetic cross-identity split selected from the remaining pool by identity and diversity filtering, and (ii) MakeupHQ-Real, an independent real-world split with real non-makeup sources and real makeup references. MakeupHQ-Real is image- and identity-disjoint from both the training data and MakeupHQ-Synthetic.

All source images and makeup references are standardized using the same preprocessing pipeline, including detection, center-cropping, normalization, parsing, and landmark-based mask extraction, followed by careful manual screening. MakeupHQ-Synthetic contains 267 makeup references, including 139 eye references and 128 face or lip references, together with 277 source images. MakeupHQ-Real contains 295 makeup references, including 174 eye references and 121 face or lip references, together with 243 source images. Style coverage includes Western, Japanese/Korean, East Asian, and neo-Chinese; poses range from frontal to profile across multiple viewpoints; expressions include open/closed eyes and mouth, with or without teeth. Compared with prior datasets, MakeupHQ Bench offers higher resolution, broader style diversity, clearer region labels, and a distribution closer to real data, enabling fine-grained and controllable evaluation.

4 Experiments

4.1 Datasets and Metrics

We evaluate MagicMakeup on our benchmark, which includes two high-resolution datasets at 1024×1024 : MakeupHQ-Synthetic (500 pairs) and MakeupHQ-Real (500 pairs). For comparability with prior work, we additionally report performance on the public Makeup-Wild dataset [16] (1000 pairs) with diverse poses at 256×256 .

We assess four aspects of quality: identity fidelity, makeup similarity, unedited-region consistency, and distribution quality. For identity fidelity, we extract AdaFace embeddings [18] after standard face alignment and compute cosine similarity, which offers stronger discriminativeness and makeup invariance than SSIM. Following Stable-Makeup [44], we use CLIP-I [28] and DINO-I [4] to measure cosine similarity between the makeup reference and the generated image. Unedited-region consistency is evaluated using L2-M, which computes the mean squared error on masked non-edited regions (background for full-face transfer) between the source and generated images. Finally, we report FID [11] between generated results and real makeup images in each test split using Inception features to assess overall distribution quality.

4.2 Implementation Details

We train our MagicMakeup on our training set using 8 NVIDIA A100 80GB GPUs with a batch size of 2 per GPU. We optimize with AdamW at a learning rate of $5e-5$ for 36K iterations. During training, the text prompt is restricted to three predefined classes and the LoRA rank is set to 256. At inference time, we use 28 sampling steps with a guidance scale of 2.5 to generate the final results.

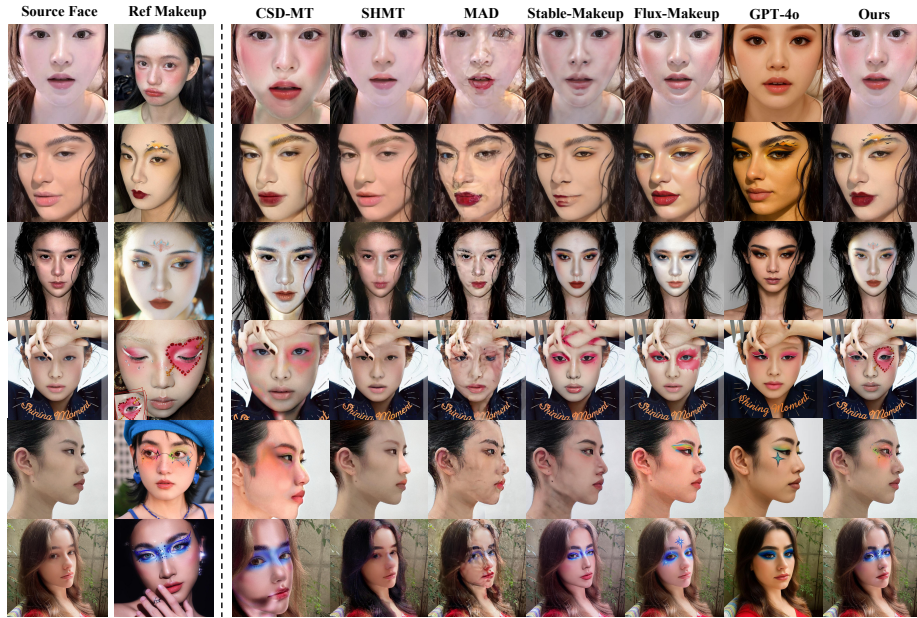


Fig. 5: Qualitative comparison of makeup-transfer results under synthetic and real settings, including diverse poses, backgrounds and makeup styles. Our method demonstrates superior identity preservation and makeup fidelity. Please zoom in for better visualization.

4.3 Comparison with SOTA Methods

We compare with representative makeup-transfer approaches, including GAN-based CPM [26], EleGANt [42], SSAT [34], CSD-MT [36], and diffusion-based SHMT [35], MAD [31], Stable-Makeup [44] and Flux-Makeup [48]. These methods serve as our baselines for quantitative and qualitative comparison, and all models are evaluated under the same settings on unified benchmarks.

Qualitative Comparison Fig. 5 shows side-by-side results on MakeupHQ-Synthetic and MakeupHQ-Real across varied poses, backgrounds, and a spectrum from light to heavy makeup. Prior methods often degrade identity consistency or bleed color outside facial regions, and we observe face rescaling and background stretching in several cases. Their transfers capture only coarse color cues and often miss fine details and full texture, resulting in incomplete makeup. In contrast, MagicMakeup preserves identity and scene geometry while delivering complete and well-aligned transfers. Light styles look clean and natural, and heavy styles retain the reference color distributions and texture details at the correct facial locations. Non-edited regions, including hair and background, remain intact without warping or deformation.

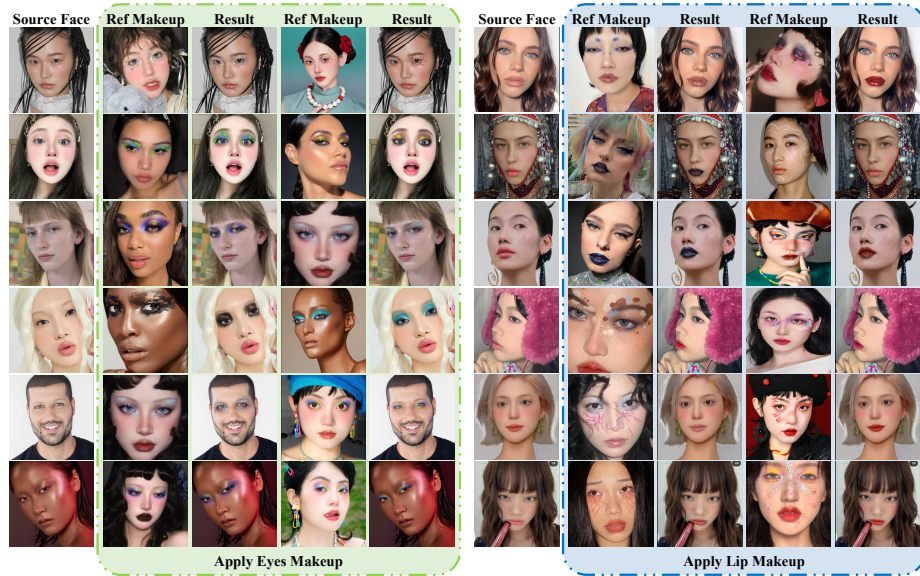


Fig. 6: Demonstration of eye and lip makeup transfer. Our method applies makeup precisely within the specified regions, producing clean boundaries and fine-grained cosmetic details. The results remain consistent under pose and expression changes and across diverse skin tones, while preserving non-edited facial areas and keeping the background free of spillover or geometric distortion.

Beyond full-face transfer, we also demonstrate eye and lip makeup application in Fig. 6. The results remain stable under noticeable pose and expression changes, and show consistent, fine-grained placement confined to the target region. Importantly, non-edited areas preserve their original structure and appearance, with no visible spillover or geometric distortion.

Quantitative Comparison As shown in Tab. 1, on the two 1024×1024 datasets MakeupHQ-Synthetic and MakeupHQ-Real, MagicMakeup achieves the highest DINO-I and CLIP-I and the lowest L2M and FID, while maintaining competitive Face-ID. These results indicate better transfer fidelity, image quality, and identity stability than prior methods. Since L2M is computed only on non-edited background regions, our much lower values further confirm that MagicMakeup preserves global geometry and avoids the face rescaling and background stretching observed in other approaches. On the public 256×256 Makeup-Wild dataset with diverse poses, MagicMakeup also leads in CLIP-I and L2M, and its FID is close to the best baseline, which is consistent with distribution differences at lower resolution rather than a transfer failure.

We note that faithful makeup transfer can lower Face-ID since makeup changes affect identity-related cues, while methods that transfer less makeup may report

Table 1: Quantitative comparison of different methods. Results are shown on MakeupHQ-Synthetic and MakeupHQ-Real at 1024×1024 resolution, and Makeup-Wild at 256×256 . Arrows indicate the preferred direction for each metric, and red values highlight the best scores.

Dataset	Metric	CPM [26]	EleGANt [42]	SSAT [34]	CSD- MT [36]	SHMT [29]	MAD [31]	Stable- Makeup [44]	Flux- Makeup [48]	Ours
MakeupHQ- Synthetic (1024×1024)	DINO-I $\uparrow$	0.625	0.520	0.461	0.576	0.488	0.500	0.612	0.576	0.660
	CLIP-I $\uparrow$	0.777	0.744	0.698	0.796	0.710	0.727	0.789	0.765	0.816
	Face-ID $\uparrow$	0.400	0.796	0.684	0.563	0.824	0.590	0.569	0.806	0.730
	L2M $\downarrow$	1610.938	2148.089	2816.613	3334.562	823.657	180.674	101.416	111.476	56.373
	FID $\downarrow$	72.191	86.841	120.193	83.896	92.601	91.372	69.733	64.119	59.238
MakeupHQ- Real (1024×1024)	DINO-I $\uparrow$	0.575	0.534	0.448	0.596	0.482	0.476	0.616	0.552	0.623
	CLIP-I $\uparrow$	0.722	0.733	0.665	0.772	0.682	0.703	0.774	0.745	0.799
	Face-ID $\uparrow$	0.361	0.755	0.641	0.522	0.777	0.539	0.525	0.792	0.706
	L2M $\downarrow$	2158.131	4038.093	3138.460	5572.246	593.029	183.091	163.979	104.464	49.768
	FID $\downarrow$	99.105	97.251	131.545	91.931	115.319	119.164	81.342	85.427	76.273
Makeup- Wild (256×256)	DINO-I $\uparrow$	0.608	0.587	0.569	0.602	0.564	0.545	0.619	0.568	0.581
	CLIP-I $\uparrow$	0.621	0.640	0.581	0.640	0.586	0.609	0.637	0.621	0.665
	Face-ID $\uparrow$	0.406	0.777	0.706	0.499	0.847	0.598	0.721	0.928	0.884
	L2M $\downarrow$	605.857	1540.744	2709.617	2151.729	1246.029	169.281	71.525	42.609	25.839
	FID $\downarrow$	86.588	94.202	109.184	94.705	103.265	105.794	85.132	88.708	92.924

higher Face-ID. Even so, MagicMakeup remains competitive on Face-ID while improving fidelity and preserving spatial stability across both high-resolution and in-the-wild settings, validating a better balance between transfer strength and identity preservation.

4.4 Ablation Study

To evaluate the contribution of each component, we conduct ablation studies on TARG and CMPG. All variants are trained for 10K iterations on our full dataset and tested on the MakeupHQ-Synthetic dataset. Quantitative results are shown in Tab. 2, and qualitative results in Fig. 7.

Starting with the base DiT model, adding Token-Aligned Region Gating (TARG) improves spatial precision: attention is concentrated on the intended regions, and cross-region leakage is effectively suppressed. As shown in Variant 2 of Fig. 7, compared to Variant 1, adding TARG reduces lip color changes in the eye-makeup task, leading to lower error in non-edited areas and higher identity-preservation scores across all datasets, as reported in Tab. 2. This improvement results from projecting pixel masks into the attention domain, blocking non-ROI interactions, and reducing cross-region leakage, which stabilizes background geometry and prevents face rescaling.

Building on TARG, adding Cross-Modal Perception Guidance (CMPG) further enhances semantic alignment with the reference while maintaining the spatial stability achieved by TARG. As shown in Variant 3 of Fig. 7, compared to

Table 2: Quantitative results of ablation study. The table shows the performance of different variants *w/o* TARG and CMPG.

Region	Variant	TARG	CMPG	Face-ID $\uparrow$	DINO-I $\uparrow$	CLIP-I $\uparrow$
Eyes	1	×	×	0.870	0.791	0.883
	2	✓	×	0.890	0.799	0.890
	3	✓	✓	0.922	0.801	0.904
Lip	1	×	×	0.879	0.653	0.877
	2	✓	×	0.885	0.659	0.870
	3	✓	✓	0.971	0.673	0.884
Face	1	×	×	0.805	0.733	0.833
	2	✓	×	0.857	0.712	0.811
	3	✓	✓	0.858	0.753	0.840

**Fig. 7: Qualitative results of ablation study.** Variant 2 with TARG improves region-specific accuracy, while Variant 3 with CMPG further enhances makeup fidelity.

Variant 2, the eyeshadow transfer becomes more complete and stable. Makeup-similarity and identity scores improve, and the background remains stable, as reported in Tab. 2. CMPG clarifies the preservation and transfer processes by aligning text and image cues during denoising, reducing content-boundary ambiguity, and further enhancing local style details and precision.

5 Limitation

Although MagicMakeup supports controllable full-face, eye, and lip makeup transfer, it does not yet provide finer-grained control over individual cosmetic components, such as blush or contour. In addition, performance may degrade under extreme side-view poses, where part of the makeup can be incompletely preserved, as illustrated in the fifth row of Fig. 5. This limitation is mainly due to the limited availability of real makeup data with extreme side-view poses, which remain rare in real-world datasets.

6 Conclusion

In this paper, we present MagicMakeup, a diffusion-transformer framework for high-fidelity, region-specific makeup transfer. Our system integrates Token-Aligned Region Gating to align pixel masks with attention and gate logits by region to prevent spillover, Cross-Modal Perception Guidance to align text and image features and make preservation/transfer decisions explicit, and an automated 1024×1024 makeup-removal pairing pipeline that supplies identity-consistent, region-labeled supervision. Additionally, we assemble a unified benchmark spanning synthetic and real data for standardized evaluation of regional controllability, makeup fidelity, and identity preservation. These components produce precise region-specific edits and achieve state-of-the-art results across diverse styles, races, and poses.

7 Ethical Considerations and Dataset Release

Our work uses real-face images for high-resolution facial makeup transfer, requiring careful consideration of privacy, consent, copyright, bias, and misuse risks. The benchmark is intended only for non-commercial academic research on cosmetic makeup transfer, not for identity recognition, face swapping, identity generation, or impersonation. To reduce these risks, we remove account identifiers, usernames, URLs, and unnecessary metadata and filter inappropriate, sensitive content during data curation. Restricted evaluation benchmark will be provided only through gated access under a Data Usage Agreement (DUA), which prohibits redistribution, identity-related use, and deceptive or malicious manipulation. We will also provide an opt-out/removal mechanism for individuals or content owners. We considered diversity in facial appearances and makeup styles during curation, while acknowledging that web-collected data may still reflect demographic, cultural, and style biases. Therefore, the benchmark should be used only within its intended research setting.

References

1. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)
2. Bodur, R., Gundogdu, E., Bhattarai, B., Kim, T.K., Donoser, M., Bazzani, L.: iedit: Localised text-guided image editing with weak supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7426–7435 (2024)
3. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)* **42**(4), 1–10 (2023)
6. Chen, H.J., Hui, K.M., Wang, S.Y., Tsao, L.W., Shuai, H.H., Cheng, W.H.: Beauty-glow: 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)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **34**, 8780–8794 (2021)
8. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 2969–2977 (2025)
9. Gu, Q., Wang, G., Chiu, M.T., Tai, Y.W., Tang, C.K.: Ladrn: Local adversarial disentangling network for facial makeup and de-makeup. In: *Proceedings of the IEEE/CVF International conference on computer vision*. pp. 10481–10490 (2019)
10. Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards fast, accurate and stable 3d dense face alignment. In: *European Conference on Computer Vision*. pp. 152–168. Springer (2020)
11. 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)
12. Hou, C., Wei, G., Chen, Z.: High-fidelity diffusion-based image editing. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 2184–2192 (2024)
13. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
14. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8362–8371 (2024)
15. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
16. 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)
17. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724* (2019)
18. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18750–18759 (2022)
19. Kim, M., Su, Y., Liu, F., Jain, A., Liu, X.: Keypoint relative position encoding for face recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 244–255 (2024)
20. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)

21. Li, P., Huang, Q., Ding, Y., Li, Z.: Layerdiffusion: Layered controlled image editing with diffusion models. In: SIGGRAPH Asia 2023 Technical Communications, pp. 1–4. ACM (2023)
22. 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)
23. 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)
24. Lu, Q., Yang, X., Taketomi, T.: Beautybank: Encoding facial makeup in latent space. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4183–4193. IEEE (2025)
25. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023)
26. 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. pp. 13305–13314 (2021)
27. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (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. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
30. 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)
31. 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)
32. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* **35**, 36479–36494 (2022)
33. Sun, Y., Yu, L., Xie, H., Li, J., Zhang, Y.: Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24584–24594 (2024)
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., 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)
36. 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)

37. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
38. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
39. 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)
40. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* **34**, 12077–12090 (2021)
41. 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. pp. 1102–1110 (2023)
42. 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)
43. Zhang, S., Zhang, Y., Zheng, Q., Ma, R., Hua, W., Bao, H., Xu, W., Zou, C.: 3d-scenedreamer: Text-driven 3d-consistent scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10170–10180 (2024)
44. 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)
45. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)
46. Zhang, Z., Han, L., Ghosh, A., Metaxas, D.N., Ren, J.: Sine: Single image editing with text-to-image diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6027–6037 (2023)
47. Zhao, Y., Po, L.M., Cheung, K.W., Yu, W.Y., Rehman, Y.A.U.: Sscan: Saliency map-guided colorization with generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology* **31**(8), 3062–3077 (2020)
48. Zhu, J., Liu, S., Li, L., Gong, Y., Wang, H., Cheng, B., Ma, Y., Wu, L., Wu, X., Leng, D., et al.: Flux-makeup: High-fidelity, identity-consistent, and robust makeup transfer via diffusion transformer. arXiv preprint arXiv:2508.05069 (2025)