Multimedia Tools and Applications (2025) 84:43805–43826
https://doi.org/10.1007/s11042-025-20874-3

Gorgeous: Creating narrative-driven makeup ideas via image prompts

Jia Wei Sii[1] · **Chee Seng Chan**[1]

Received: 5 December 2024 / Revised: 6 April 2025 / Accepted: 16 April 2025 /
Published online: 5 May 2025

Abstract
We introduce *Gorgeous*, a diffusion-based generative method that redefines digital makeup application by enabling the generation of creative makeup designs through image prompts. Unlike conventional makeup transfer techniques that primarily replicate existing styles, *Gorgeous* is the first to empower users to incorporate narrative elements into makeup ideation via image-based prompts. This approach allows for the creation of makeup concepts that visually convey user expressions, offering innovative and imaginative makeup ideas for real-world application. To achieve this, *Gorgeous* establishes a foundational learning framework, ensuring that the model first comprehends "what makeup is" before integrating narrative elements. A pseudo-pairing strategy, leveraging a face parsing and content-style disentangling network, effectively addresses challenges associated with unpaired data, enabling the model to be trained on bare-face images. Users can input reference images that represent their conceptual ideas (e.g., fire), from which *Gorgeous* extracts contextual embeddings to guide a novel makeup inpainting algorithm. This process facilitates the generation of creative, narrative-driven makeup designs tailored to specific facial regions. Comprehensive experiments validate the effectiveness of *Gorgeous*, demonstrating its potential to establish a new paradigm in digital makeup artistry and application. Code released at: https://github.com/JiaWeiSii/gorgeous.

Keywords Makeup Generation · Stable Diffusion · Textual Inversion · ControlNet · Inpainting

1 Introduction

Makeup, as both an art form and a storytelling tool [1], has evolved beyond personal expression to become an integral component of various creative industries, including film, theater, cosplay, and fashion. In these domains, makeup serves as a transformative medium, enabling individuals to embody characters and narratives that resonate with audiences. However,

✉ Chee Seng Chan
cs.chan@um.edu.my

Jia Wei Sii
17069558@siswa.um.edu.my

[1] CISiP, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

designing makeup that effectively conveys specific visual stories remains a significant challenge, often hindered by a lack of creative inspiration and time constraints.

Existing digital makeup technologies predominantly focus on makeup transfer [2–5], where makeup styles are replicated from reference faces. While these approaches allow users to experiment with diverse makeups virtually, they are fundamentally limited to reproducing existing makeup designs and cannot support the creative objective of embedding narrative elements into makeup. To address this gap, we introduce *Gorgeous*, a novel generative model that, for the first time, enables users to incorporate narrative-driven elements into makeup design. Unlike conventional methods that rely on textual prompts, which can be difficult for users to articulate in the context of makeup, *Gorgeous* utilizes image prompts as intuitive visual cues, allowing users to effectively communicate their creative intent, as illustrated in Fig. 1.

Gorgeous is built upon a foundational learning and formatting framework, *MaFor*, powered by ControlNet, ensuring the model learns "what makeup is" before integrating narrative elements into its generative process. While ControlNet has been widely used for structure-based conditioning in generative models-such as edge and pose guidance-we extend its applicability to the makeup domain, where a bare face serves as the conditioning input to guide the generation of narrative-driven makeup. One of the key challenges in training generative models for makeup synthesis is the lack of paired data, as there are no direct mappings between bare faces and their narrative-enhanced makeup counterparts. To overcome this, we employ a pseudo-pairing strategy, leveraging a face parsing and content-style disentangling network, that enables *Gorgeous* to learn makeup generation directly from bare faces. Additionally, to allow users to narrate their expressions through image prompts, *Gorgeous* extracts contextual embedding from user-provided image prompts (e.g., fire), which then guides a makeup inpainting algorithm. This algorithm extends the concept of inpainting beyond conventional missing-region reconstruction, instead utilizing it in conjunction with *MaFor* to apply narrative-driven makeup directly onto bare faces. Our contributions are

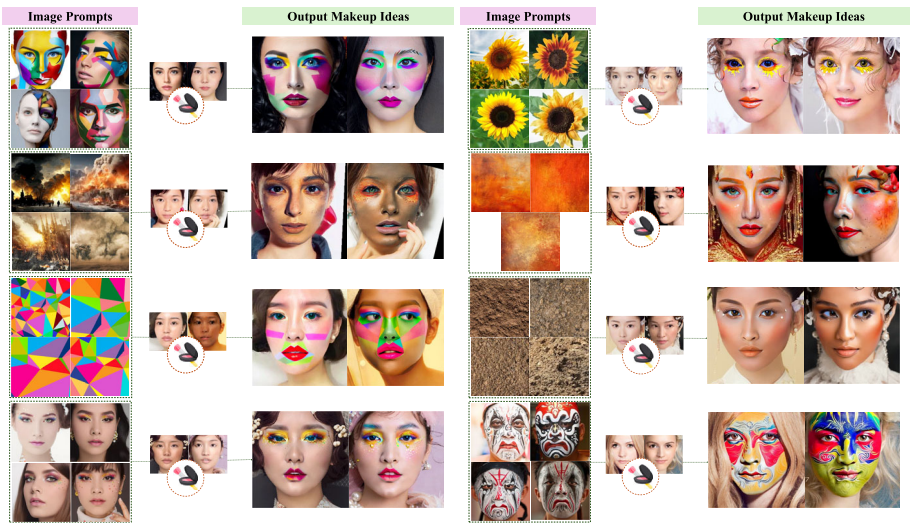

Fig. 1 Provide any image prompt-whether a serene landscape or a dynamic scene-and *Gorgeous* will translate your narrative into creative makeup designs. Perfect for moments when expressing meaning through makeup feels challenging, *Gorgeous* transforms inspiration into artistic visualizations

manifold: **(i)** We propose *Gorgeous*, the first generative model in the makeup industry that enables **narrative-driven makeup synthesis using image prompts**; **(ii)** We develop ***MaFor***, a novel makeup learning and formatting framework that trains the model **from bare faces to fully made-up makeup faces** while addressing the critical issue of unpaired data. This is achieved through a content-style disentanglement technique, which facilitates the generation of pseudo-paired training data; **(iii)** We explored **textual inversion** as a mechanism for learning contextual embeddings from image prompts, ensuring **contextually relevant makeup generation**; **(iv)** We introduce a **unique adaptation of image inpainting** for the makeup domain, integrating it with *MaFor* to push the boundaries of traditional inpainting techniques. Our method successfully generates creative, user-defined makeup directly on facial structures, reflecting artistic ideas provided by the users. Extensive experiments demonstrate that *Gorgeous* generates unique, user-defined makeup styles enriched with narrative elements extracted from image prompts. Our work introduces a new paradigm in digital makeup artistry, bridging creative ideation with generative AI for both virtual and real-world applications.

2 Related works

2.1 Facial makeup

Existing methods such as traditional face-warping makeup techniques [6, 7] and GAN-based [2–4, 8–14] and diffusion-based [15] makeup transfer approaches primarily focus on duplicating makeup from one face to another. These methods are limited in their ability to create personalized and unique makeup designs, as they require an existing makeup look on a source face to initiate the transfer. In contrast, makeup recommendation systems [16–20] generate suggestions based on predefined rules and user characteristics. However, these systems do not allow users to integrate narrative elements into the makeup, limiting the creation of story-driven, imaginative designs tailored to a user's creative expressions or thematic intent. Both makeup transfer and recommendation systems operate within fixed paradigms, restricting innovation in generative makeup artistry.

2.2 Style transfer

Style transfer techniques [21–40] enable the adaptation of aesthetic elements or visual styles from one image to another while preserving content structure. However, similar to makeup transfer methods, these techniques are not directly applicable to makeup generation because they apply global transformations, altering the entire image, which can distort facial structures (e.g., cartoon-like effects). Instead, makeup should be localized and restricted to enhancing specific facial regions while preserving the natural structure of the face.

2.3 Diffusion-based generative methods

Recent diffusion models [41–56] have demonstrated exceptional ability in visual content generation, particularly in text-to-image generation [41–47]. However, relying solely on textual descriptions presents challenges in controlling the generated output, leading to ambiguities and inaccuracies. To address this, two major advancements have been introduced: (1) Text-guided image-to-image translation or editing methods [44, 49, 52, 53, 53, 54, 56–58].

Unlike traditional text-to-image-generation models, which synthesized images from pure noises, text-guided image-to-image translation applies targeted modifications to an existing input image, guided by textual prompts. For example, InstructPix2Pix [49] leverages diffusion models trained on text-conditioned image edits to allow controlled image modifications. Plug-and-Play Diffusion [52] enables users to edit images by conditioning on specific textual attributes without requiring retraining. SDEdit [58] refines images by gradually denoising them while incorporating textual constraints. These approaches improve control over the input image with flexible textual guidance. However, they still inherit limitations from text-only conditioning, as textual descriptions may not always precisely specify desired transformations, especially in fine-grained applications such as artistic or thematic makeup generation; (2) Personalized image-conditioned approaches such as Textual Inversion [50] and DreamBooth [51]. While text-guided image-to-image translation modifies an existing image based on a text prompt, personalized image-conditioned approaches take a different route. These methods learn new visual concepts directly from a set of reference images, enabling fine-grained and identity-preserving generation without requiring an explicit text description of the concept. Although these methods require text at inference time, their reliance on text differs significantly from traditional text-guided generation. Text here is used as a retrieval key, not as a semantic descriptor. The learned token (e.g., "V") does not carry meaningful semantic information* but acts as a placeholder that retrieves the learned visual style when placed in a text prompt. Hence, unlike standard text prompts, users do not need to explicitly describe the makeup concept.

In Gorgeous, we prioritize flexibility and efficiency in user-driven makeup generation. Unlike DreamBooth, which requires fine-tuning the entire model for each new concept (making it computationally expensive), Textual Inversion provides a lighter alternative by learning only a single token that encapsulates the concept. Moreover, unlike IP-Adapter [59], which only conditions on a single reference image, Textual Inversion enables learning from multiple reference images, capturing a broader contextual understanding, enabling Gorgeous to generalize across diverse and narrative-driven makeup designs.

3 Methodology

3.1 Preliminary: Stable diffusion and ControlNet

Latent diffusion models (LDMs) [60] operate on a compressed latent space $\mathcal{Z}$ derived from the original data space $\mathcal{X}$ through a VAE encoder $\mathcal{E}$. The generative process involves learning a denoising model in this latent space. For conditional generation, such as with text guidance, the model incorporates an additional text condition y provided by a text encoder τ_θ. The objective function is typically formulated as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,p,\epsilon\sim\mathcal{N}(0,I),t\sim\mathcal{U}(1,T)}\Big[\|\epsilon - \epsilon_\theta(z_t, p, t)\|_2^2\Big], \tag{1}$$

where $z = \mathcal{E}(x)$ represents the latent variable encoded from the data $x \in \mathcal{X}$, p is the text condition (e.g., a prompt) provided as input (usually encoded by a pretrained text encoder such as CLIP text encoder [61]), ϵ is Gaussian noise added during the forward diffusion process, z_t is the latent variable at time step t obtained through the diffusion process, and $\epsilon_\theta(z_t, p, t)$ is the model's prediction of the noise at time step t conditioned on p. This loss encourages the model to denoise z_t given text condition p and recover the original latent

variable z_0 that aligns with p, enabling controllable text-to-image generation. A well-known example is Stable Diffusion (SD) [60].

ControlNet [47] is an extension of diffusion models designed to guide the generation process using additional input conditions, such as edge maps, segmentation masks, or other spatial/structural information. It incorporates these conditions into the model while preserving the ability to generate diverse outputs. The key idea is to adapt the diffusion process with a trainable control branch. The ControlNet framework modifies the standard diffusion loss as follows:

$$\mathcal{L}_{\text{ControlNet}} = \mathbb{E}_{z,p,c,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(z_t, p, t, \mathcal{C}_\phi(c))\|_2^2\Big], \tag{2}$$

where c is the conditioning input (*e.g.*, edge map, segmentation mask) that guides the generation. $\mathcal{C}_\phi$ is a trainable copy of SD's U-Net [62] alongside the main denoising branch ϵ_θ. The control branch extracts features from c, which are injected into the denoising branch at multiple layers to ensure spatial alignment between the generated output and the condition c. As a result, ControlNet enables the use of not only textual but also spatial conditions to achieve controllable image generation.

3.2 Overview

Technically, our goal is to design a model to generate makeup ideas, where:

(i) complex textual descriptions are unnecessary;
(ii) users can obtain the makeup ideas that are integrated with their desired narrative elements by providing image prompts as inspiration;
(iii) The generated output must align with the essence of makeup, ensuring it remains distinctly makeup-like while being contextually relevant to the provided image prompts;

As shown in Fig. 2, *Gorgeous* utilizes a pretrained Stable Diffusion model and introduces three key modules: (i) Makeup Learning and Formatting *MaFor*, (ii) Context Learning *CL*, and (iii) Makeup Inpainting Pipeline *MaIP*. *MaFor* learns fundamental attributes of "what makeup is" and formatting narrative elements into a makeup-like form. *CL* encodes user-provided image prompts into meaningful embedding, capturing the narrative elements for integration into makeup design in *MaIP*. *MaIP*, which incorporates *MaFor*, applies makeup exclusively to facial area. It ensures the generated output adheres to the makeup form learned in *MaFor*, seamlessly resembling the physical makeup on face while reflecting user's narrative image inputs. Each module and its purpose will be detailed in the following section.

3.3 Makeup learning and formatting *MaFor*

MaFor is the essential piece in the entire *Gorgeous* model. Without it, *Gorgeous* would generate arbitrary outputs (e.g., non-makeup, unrecognizable faces, as shown in Fig. 6 despite containing narrative elements from image prompts. This necessity became evident during our initial experiments using only *CL* and *MaIP* without *MaFor*, where we faced two significant challenges: (i) The model often "replaced" the face with content learned from *CL*, conflicting with the goal of applying makeup on face-where the face itself must be preserved as the foundation; (ii) While *MaIP* without *MaFor* generated content, it frequently did so in a random style. Although contextually relevant, the output lacked a proper makeup format, resembling a rough drawing on the face rather than true makeup. To resolve these issues, we introduced *MaFor* as a foundational module that is detailed next.

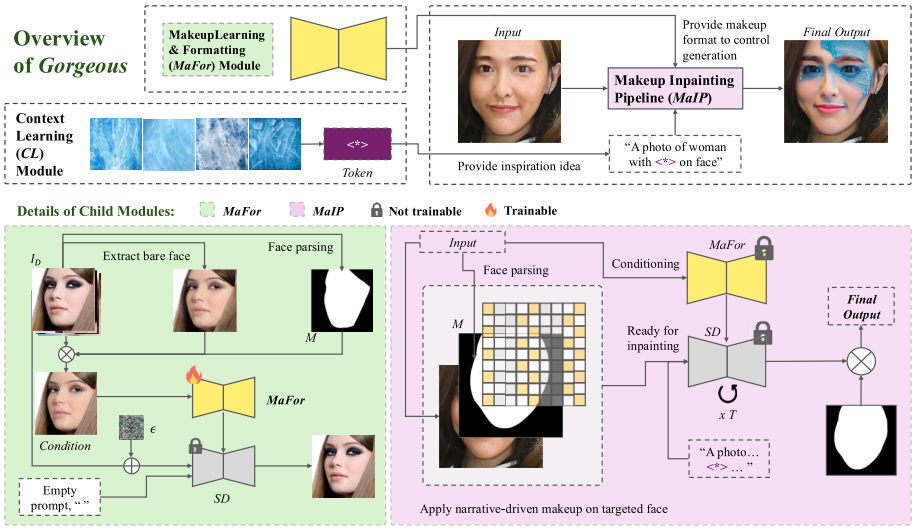

Fig. 2 Overall *Gorgeous* architecture. Given a set of image prompts, these images are first processed to extract key narrative elements, which are embedded into a placeholder token. This token guides *MaIP* to generate makeup ideas for user. However, *MaIP* initially struggles to produce makeup-specific outputs, treating target regions as generic holes due to inpainting limitations (refer to Fig. 6(A)). To address this, *MaFor* ensures the output is makeup-like by: (i) learning "what makeup is" as a pretrained model to be used in *MaIP* during inference, and (ii) transforming narrative tokens into makeup-specific representations during inference. This process refines *MaIP*'s inpainting to apply narrative-inspired makeup ideas to targeted facial regions, retaining both the essence of makeup and the user's narrative image inputs

ControlNet To enable *Gorgeous* to learn "what makeup is", we chose ControlNet [47] as the foundation of *MaFor* for its powerful conditioning capabilities. By using the bare face as the conditioning input, we hypothesized that the model would generate a makeup-applied face that adheres to the facial structure of the input, preserving its natural features. This approach, analogous to ControlNet's use with edge or depth maps, ensures that the output consistently aligns with the input condition. Qualitative experiments confirm this hypothesis, as the generated makeup naturally conforms to the facial structure, effectively mimicking the process of applying makeup directly onto the face.

Challenge with ControlNet-Unpaired Data ControlNet requires paired data-bare face as the conditioning input and a makeup-applied face as the expected output. However, we lack such datasets and only have access to makeup-applied faces. To address this, we use the unsupervised domain transfer method, LADN [63] to generate pseudo paired data. Specifically, LADN performs a "de-makeup" process by transferring a non-makeup style to a makeup-applied face, yielding a pseudo bare face as $I^*_{\text{bare}} = \text{LADN}(I)$, where I is the makeup-applied face.

Domain transfer can unintentionally alter non-facial regions. Hence, we use a face parsing module [64] to segment the facial region and blend I with I_{bare}, restoring bare face, avoiding alterations to non-facial areas:

$$I_{\text{bare}} = I^*_{\text{bare}} \cdot M + I \cdot (1 - M), \tag{3}$$

where M is the segmentation mask of facial region with $M_{ij} = 1$ for facial region and $M_{ij} = 0$ otherwise. A Gaussian blur is applied to M to smooth edges during blending.

Training With the pseudo-paired dataset D_{paired}, we train *MaFor* with ControlNet [47] to condition on the I_{bare} and generate makeup-applied faces. This ensures the model preserves facial identity and adheres to a proper makeup format. Training with makeup images as the target further reinforces this behavior.

To focus the model on the semantic relationship between bare and makeup-applied faces, we use an empty string ("") as the text prompt during training. *MaFor* is optimized using the diffusion loss function:

$$\mathcal{L}_{\text{ControlNet}} = \mathbb{E}_{z,c,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(z_t, \text{“”}, t, \mathcal{C}_\phi(c))\|_2^2\Big]. \tag{4}$$

where z are the latents of makeup-applied faces I, and c is the corresponding bare face I_{bare} obtained through Eq. 3 used as the spatial condition.

3.4 Context learning *CL*

With *MaFor* ensuring outputs adhere to a makeup format, *CL* is introduced to integrate narrative elements into the makeup design via image prompts. We prefer image prompts to complex text descriptions because visual cues are more intuitive for users to convey their ideas.

Here, we use textual inversion [50] to capture and encode the narrative context from user-selected image prompts into a text embedding v. This embedding serves as the inspirational source for *MaIP*, enabling the generation of visually cohesive, narrative driven makeup designs:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z,v,t,\epsilon}\Big[\|\epsilon - \epsilon_\theta(z_t, v, t)\|_2^2\Big], \tag{5}$$

$$v^* = \arg\min_v \mathcal{L}_{\text{LDM}} \tag{6}$$

where v^* is the optimal text embedding. Unlike conventional textual inversion prompts (e.g., "*a photo of a* $\langle * \rangle$"), we employ more directed prompts like "*a photo of a woman with* $\langle * \rangle$ *on face*" to explicitly guide the learning process in interpreting "$\langle * \rangle$" as a makeup style. Note that ϵ_θ is frozen and $\mathcal{C}_\phi$ is not employed here.

We chose textual inversion over alternatives like IP Adapter [59] because it restricts input to a single image, textual inversion allows multiple images, enabling a richer and more accurate context capture. Although a single image with IP Adapter can be controlled using methods like bounding boxes to specify areas of interest, this approach is indirect and less intuitive. Textual inversion provides a more straightforward way to encode the desired narrative elements, aligning better with user preferences, and ensuring a precise representation of their ideas. We would further discuss this with empirical evidences in the Ablation Study in Section 6.

CL is trained independently of *MaFor*, allowing focused training for specific makeup ideas without retraining the entire system. This design ensures efficiency, as *MaFor* requires training only once, while *CL* can adapt to new user inputs with minimal computational cost. For a detailed breakdown of computational costs, please refer to Section 7.

3.5 Makeup inpainting pipeline *MaIP*

To generate the final narrative-driven makeup representation I_{gen} inspired by the user's image prompt, we design a training-free Makeup Inpainting Pipeline *MaIP*. *MaIP* applies the generated makeup idea specifically to the targeted facial regions, guided by a segmentation

mask M, while *MaFor* ensures the output adheres to the makeup format and aligns with the structure of the bare face.

As discussed in the *MaFor* section, using image inpainting alone led to challenges, such as failing to preserve the face as the foundation and generating outputs that lacked a proper makeup format. By integrating *MaFor* with *MaIP*, we ensure that makeup is applied precisely to the face while maintaining its natural structure and proper makeup characteristics.

Users can input prompts such as p ="*a photo of a woman with* $\langle * \rangle$ *on face*", where $\langle * \rangle$ represents the makeup idea derived from *CL*. During the denoising process, *MaIP* applies the narrative-driven makeup, guided by *MaFor*, to produce I_{gen}:

$$\begin{aligned} \epsilon_{\text{uncond}} &= \epsilon_\theta(z_t, \text{“”}, t, \mathcal{C}_\phi(c)), \\ \epsilon_{\text{pred}} &= \epsilon_\theta(z_t, p, t, \mathcal{C}_\phi(c)), \\ z^*_{t-1} &= \epsilon_{\text{uncond}} + g \cdot (\epsilon_{\text{pred}} - \epsilon_{\text{uncond}}), \\ z_{t-1} &= z^*_{t-1} \cdot M^* + c_{t-1} \cdot (1 - M^*), \end{aligned} \tag{7}$$

where g is the guidance scale, p is the text prompt, c_{t-1} is the noised latent of the bare face, z_{t-1} is the denoised latent, both at timestep $t-1$, and M^* is the downsampled segmentation mask. This is to ensure non-face area's latents are replaced by the non-face area of c. At $t = 0$, z_0 is decoded to I_{gen}.

To address quantization errors during decoding, we blend the generated output with the bare face to preserve non-facial details:

$$I_{\text{final}} = I_{\text{gen}} \cdot M + I_{\text{bare}} \cdot (1 - M). \tag{8}$$

This ensures that the makeup is localized to the facial regions while maintaining the integrity of non-facial areas.

4 Experiments

4.1 Datasets

We use two types of datasets: (i) **BeautyFace dataset** [4] with 2,447 unpaired makeup images for training via pseudo-pairing in *MaFor* and 272 images for testing. (ii) **Image prompts** categorized as (1) images with makeup faces (Style 1) for fair comparisons with makeup transfer methods, which rely on faces (Fig. 4), and (2) images without faces (Style 2) to evaluate *Gorgeous*'s generalization beyond facial contexts.

4.2 Evaluation metrics

We employ three key metrics for evaluation: (i) **CSD Similarity** [65], measuring contextual relevance between generated makeup and image prompts; (ii) **DreamSIM** assessing perceptual similarity from a human perception standpoint; Both are computed with cosine similarity. We chose CSD and DreamSIM over CLIP/DINO scores (as used in [51]) because CLIP/DINO are trained to semantically understand images but may overlook low-level stylistic features like color gradients and textual variations, which are essential in makeup evaluation; and (iii) **Fréchet Inception Distance (FID)** [66], quantifying how well generated makeups align with real-world makeup distributions (BeautyFace dataset).

While these metrics are not explicitly designed for makeup evaluation, their combination provides a meaningful evaluation framework for our specific goal: ensuring that the generated makeup aligns with the narrative and conceptual elements in the image prompts. Since our model is designed to generate conceptual, inspirational makeup ideas rather than realistic makeup textures, attributes such as makeup color accuracy, textual realism, or fine detail integration are not the primary evaluation objectives in this phase of our work. Instead, our evaluation prioritizes the effectiveness of concept translation from the prompts to makeup designs, which is crucial for inspiring real world application.

Furthermore, as makeup generation is inherently subjective, we complement these quantitative metrics with a user study (Section 5.2.1) to assess the real-world relevance and user preferences. Given the absence of established domain-specific evaluation metrics for generative makeup, our methodology balances both objective computational metrics and human validation.

4.3 Baselines

We benchmark *Gorgeous* against established methods across three domains: **(1)** Makeup transfer (i.e., EleGANt [2], SSAT [3], BeautyREC [4], and StableMakeup [5]); **(2)** Style transfer (i.e., InST [39], InstantStyle [67], InstantStyle (+MaFor)); **(3)** Image-to-image translation/generation (i.e., I2I SDXL [55], InstructPix2Pix [49], Stable Diffusion Inpainting [60], Inpainting with Textual Inversion [50][1]).

4.4 Implementation details

MaFor (ControlNet [47]) and *CL* (textual inversion) were implemented using SDv2.1[2] [60] with images resized to 512 × 512. MaFor was trained over 15,000 steps with a batch size of 1, gradient accumulation of 4, and a learning rate of 1e-4. For *CL*, each style token was trained for 5,000 steps with a learning rate of 1e-5 and batch size of 1. During inference, guidance scale g ranged from 3 to 20, and inference steps varied from 30 to 100 based on desired makeup intensity.

5 Evaluations

5.1 Qualitative evaluation

In Fig. 3(A), Style 1(a-d) indicate image prompts with faces present, while Style 2(a-d) depict image prompts without faces, representing narrative ideas. In Fig. 3(B), we evaluate *Gorgeous* with makeup transfer methods. **(i) Gorgeous** generates makeups weaved with narrative elements, handling both facial (Style 1) and non-facial (Style 2) prompts. This naturally results in unique designs, as narrative-driven makeups inherently differ from image prompts. While uniqueness is not the goal, it demonstrates *Gorgeous*'s ability to incorporate narrative elements. **(ii)** Makeup transfer methods (EleGANt [2], SSAT [3], and BeautyREC [4]) fail to replicate existing makeups accurately-although this is not our main concern, it is

[1] This means using the token learned through *CL*.

[2] https://huggingface.co/stabilityai/stable-diffusion-2-1

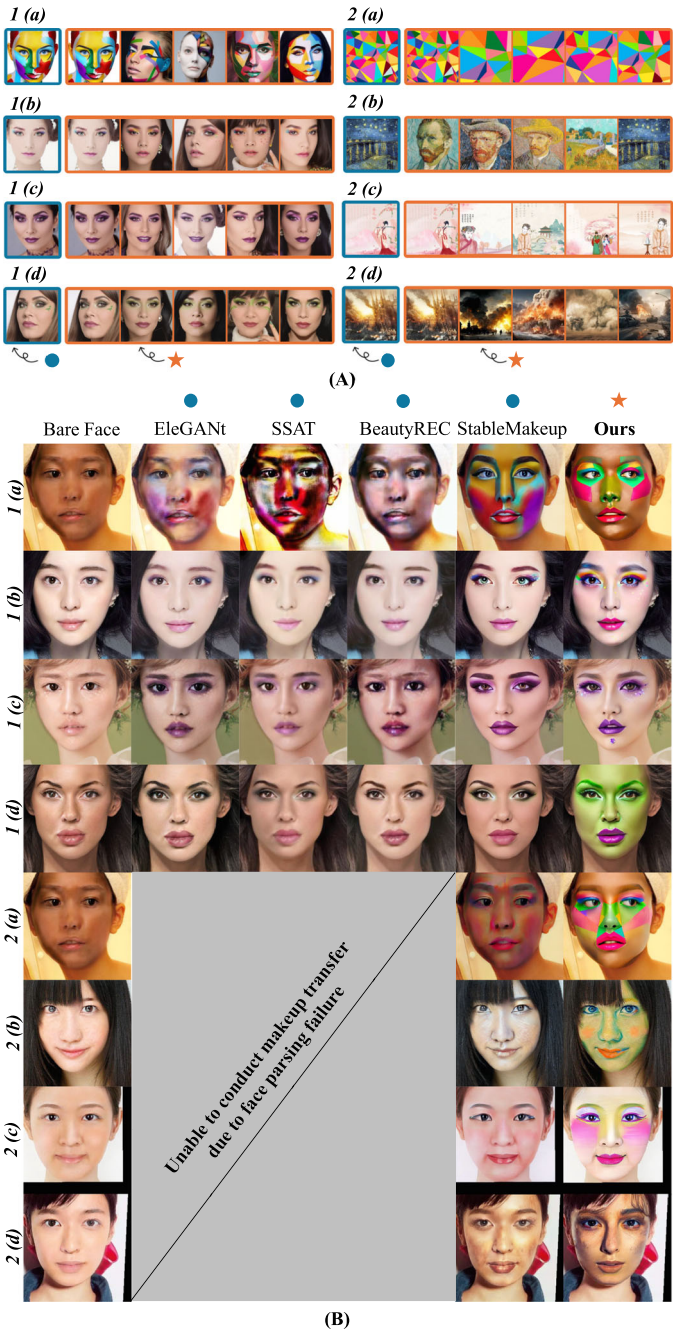

Fig. 3 (*A*) presents images prompts: **(i) Style 1** including images with faces; **(i) Style 2** comprising images without faces. (*B*) displays our qualitative results for both styles, compared with state-of-the-art makeup transfer (i.e., EleGANt, SSAT, BeautyREC) method. *Blue circle* shows image prompts used for makeup transfers; *orange star* for *Gorgeous*

still important to highlight it for fair comparisons. Their failure to reproduce, especially exaggerated makeups (e.g., Style 1(a)) limits the direct comparability of our results. Nonetheless, this comparison is necessary to emphasize our distinct objectives-makeup transfer focuses on replication versus *Gorgeous* focuses on narrative-driven makeup. **(iii)** As explained in Fig. 4, EleGANt, SSAT, BeautyREC rely heavily on face parsing, making them incapable of processing non-facial prompts (Style 2). This causes the need to categorize our comparisons (Style 1 and 2) to evaluate the applicability. **(iv)** StableMakeup improves upon traditional transfer methods by bypassing face parsing and generating output from non-facial prompts. However, it struggles to produce contextually relevant makeup designs, failing to integrate narrative elements.

In Fig. 5, we further evaluate *Gorgeous* against style transfer and text-guided generative methods. **(i)** Common style transfer (i.e., InST, InstantStyle) blend styles globally, do not adapt them into makeup formats. **(ii)** We attempt a hybrid approach, combining *MaFor* with InstantStyle, to format the style into makeup while retaining facial identity. While this hybrid preserves the face structure, it lack *Gorgeous*'s ability to produce narrative-driven makeups. This highlights the importance of *CL* for encoding narrative elements and *MaIP* for ensuring these elements are applied as contextually relevant makeup on the target face. **(iii)** Text-guided diffusion-based models, including SDXL [55] (via SDEdit [58]), InstructPix2Pix [49], image inpainting with Stable Diffusion [60], the inpainting with Textual Inversion [50, 60][3] are used to integrate narrative elements through simple text prompts, aligning with our goal of providing users a straightforward process. However, simple text prompts often struggle to capture complex narrative ideas effectively, making these methods less suitable for intricate makeup designs. **(v)** Additionally, these methods generate makeup on random faces, failing to preserve the identity of the target face. In contrast, *Gorgeous* ensures that narrative-driven makeup designs are contextually relevant, in proper makeup form, and applied directly to the target face. By utilizing image prompts instead of relying solely on text, *Gorgeous* simplifies the user experience while maintaining superior performance.

5.2 Quantitative evaluation

In Table 1, we computed scores for all methods for styles 1 and 2 (a-d), as illustrated in Fig. 3(A). Each style was evaluated using 272 test images of bare faces, respectively. **(i) Gorgeous:** Our method achieves the lowest FID [66] scores (53.29 for Style 1; 89.84 for Style 2), indicating its superior ability to render makeup that aligns with the BeautyFace dataset while incorporating narrative elements. *Gorgeous* also performs comparably well in CSD [65] and DreamSIM [68], indicating its effectiveness in generating contextually relevant makeups in term of style and content alignment (CSD) and perceptual similarity from a human perspective (DreamSIM).

(ii) Makeup Transfer Methods: In Style 1, BeautyREC achieves the best FID (37.82), while StableMakeup has the highest CSD (0.64) and DreamSIM (0.72). However, these methods are limited to replicating existing makeups and cannot incorporate narrative elements. Style 2 evaluations are absent due to their reliance on face parsing, highlighting their inability to process nonfacial prompts.

(iii) Style Transfer Performance: InST achieves high scores (CSD: 0.60 for Style 1, 0.29 for Style 2) but has high FID values (119.35 for Style 1, 202.00 for Style 2), indicating less

[3] This uses the same token learned by *Gorgeous* through *CL*.

Fig. 4 Demonstration of face parsing on facial versus non-facial images in makeup transfer. (Left) showcases successful parsing of facial features in images with faces, enabling makeup transfer. Conversely, (Right) shows the failure of face parsing on non-facial images; the lack of detectable facial features leads to inaccurate parsing maps. This parsing failure prevents the direct application of most makeup transfer techniques

adherence to makeup form. Adding *MaFor* to InstantStyle improves CSD, DreamSIM, and FID compared to only InstantStyle in both styles, but it still lags behind Gorgeous.

(iv) I2I translation: Inpainting + TI scores well in DreamSIM (0.63 for Style 1, 0.46 for Style 2) and CSD (0.34 for Style 2) but struggles with FID (128.37 for Style 1, 250.81 for Style 2), indicating a failure to balance contextual relevance, makeup structure, and adherence to facial identity.

In summary, *Gorgeous* excels across all metrics, by generating narrative-driven makeup designs that align with makeup formats, maintain face identity, and handle both facial and non-facial prompts effectively.

5.2.1 User study

Although the combination of quantitative metrics such as CSD, DreamSIM, and FID provides valuable benchmarks, user preferences for narrative-driven makeup designs remain subjective. To evaluate this, we conducted a user study that included 100 participants of diverse backgrounds. Among them, 82% were female, 18% were male. Participants were categorized into three age groups: 18-30 (73%), 31-40 (25%), and 40-50 (2%). Additionally, 22% of participants claimed that themselves are makeup professionals with 6% among them identifying themselves as cos-players, stage actors, or fashion enthusiasts who frequently use makeup artistically. Whereas, the rest of the participants are general users (78%).

We acknowledge that the distribution of the participants is not perfectly balanced, with a higher proportion of general users compared to makeup professionals. This reflects the primary target audience of digital makeup applications, which is predominantly general users rather than industry professionals.

Each participant reviewed the results from Fig. 3 and Fig. 5, along with the corresponding bare face inputs and image prompts. Participants were asked to vote for the makeup design they found most visually appealing and contextually relevant to the image prompts.

As shown in Table 2, *Gorgeous* consistently received the highest votes in most styles, demonstrating its ability to produce visually appealing and narrative-driven makeup designs. For Style 1, *Gorgeous* led in (a), (b), and (c), while ranking second in (d), slightly behind StableMakeup, which secured 35%. For Style 2, *Gorgeous* outperformed all methods overwhelmingly, receiving more than 70% votes in all styles. These results highlight *Gorgeous*'s ability to create visually appealing narrative-driven makeup designs that resonate with users, particularly in scenarios that require contextual relevance and creativity.

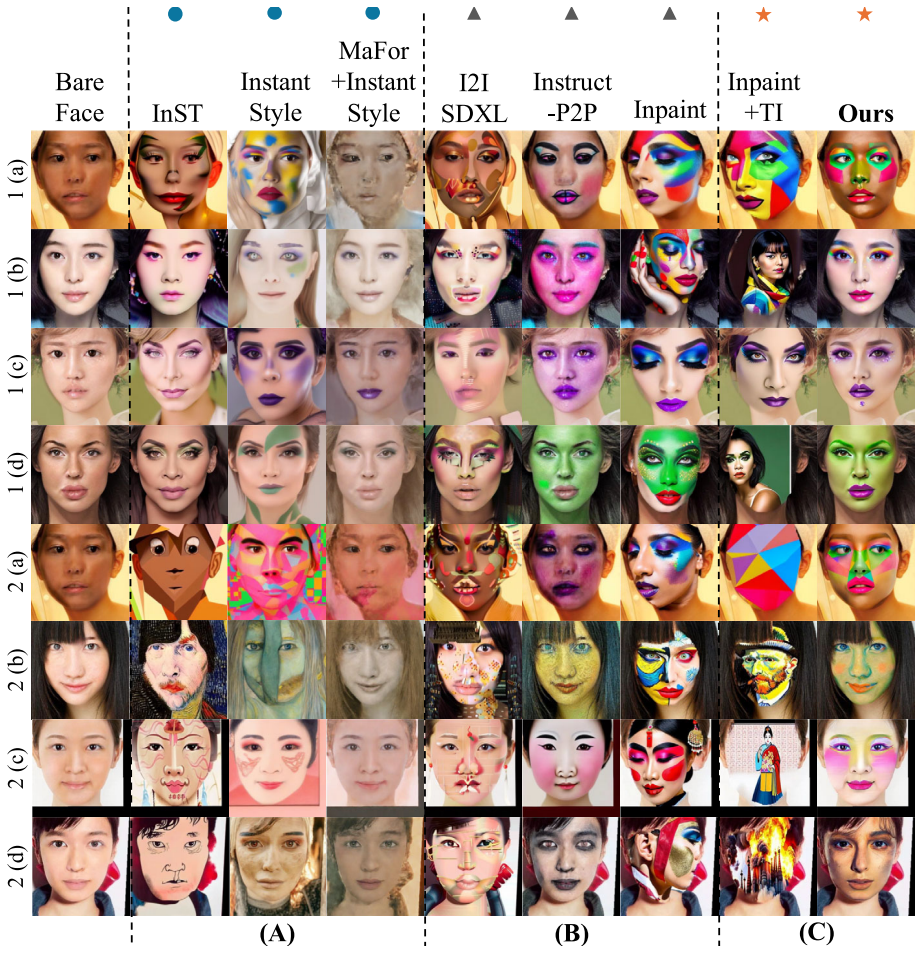

Fig. 5 Other qualitative results compared with state-of-the-art style transfer and text-guided image-to-image generation/editing methods, while the inputs are *Bare Face*. Simple text prompts are used for different styles: [1(a)]-"A photo of a woman with stained makeup on face"; [1(b)]-"A photo of a woman with colorful dotted makeup on face"; [1(c)]-"A photo of a woman with purple makeup on face"; [1(d)]-"A photo of a woman with green makeup on face"; [2(a)]-"A photo of a woman with stained makeup on face"; [2(b)]-"A photo of a woman with Van Gogh makeup on face"; [2(c)]-"A photo of a woman with traditional Chinese makeup on face"; [2(d)]-"A photo of a woman with war makeup on face"

6 Ablation study

6.1 Impact of each component of Gorgeous

To further validate the effectiveness of each component in *Gorgeous*, we conduct an ablation study by systematically removing key modules and evaluating their impact on narrative-driven makeup generation. In Fig. 6, we present the qualitative results of module removal, and Table 3 provides the corresponding quantitative analysis. **(i)** Eq. (8) **in MaIP:** Omitting Eq. (8) results in slight artifacts in non-facial areas, disrupting consistency by altering regions that should remain unchanged. While the impact on CSD Similarity (-0.01) and DreamSIM (-

Table 1 Quantitative evaluation with competitive methods

Methods	Average of Style 1			Average of Style 2		
	CSD ↑	SIM ↑	FID ↓	CSD ↑	SIM ↑	FID ↓
(Makeup transfer)						
EleGANt	0.48	0.54	45.13	–	–	–
SSAT	0.45	0.56	58.92	–	–	–
BeautyREC	0.39	0.49	**37.82**	–	–	–
StableMakeup	**0.64**	**0.72**	64.18	0.13	0.22	**48.26**
(Style transfer)						
InST	0.60	0.67	119.35	0.29	0.29	202.00
InstantStyle	0.20	0.29	279.18	0.09	0.20	309.05
InstantStyle (+MaFor)	0.42	0.51	63.62	0.14	0.20	80.36
(Image-to-image translation)						
I2I SDXL	0.48	0.55	142.27	0.15	0.22	141.75
InstructPix2Pix	0.51	0.54	94.09	0.17	0.16	101.13
Inpainting	0.59	0.57	146.65	0.16	0.23	169.72
Inpainting+TI	0.58	0.63	128.37	**0.34**	**0.46**	250.81
Ours (Gorgeous)	0.61	0.61	53.29	0.21	0.28	89.84

**Makeup transfer scores are absent for Style 2 due to their inability to handle non-facial images. CSD and DreamSIM (SIM) are measured in cosine similarity, with bold values indicating the best scores

0.01) is minor, these metrics primarily assess conceptual alignment and perceptual similarity rather than distribution consistency. In contrast, FID, which evaluates how well the generated image follows the statistical distribution of real-world makeup samples, increases to 90.75 from the full model's 89.84. Since facial regions remain unchanged, the increased FID here reflects discrepancies in non-facial regions rather than the intended makeup transformation. This highlights the importance of Eq. (8) in preserving a stable distribution, ensuring that

Table 2 User study

Methods	Style 1 (%) ↑				Style 2 (%) ↑			
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
EleGANt	0	28	1	3	0	0	0	0
SSAT	0	17	1	3	0	0	0	0
BeautyREC	0	5	2	1	0	0	0	0
StableMakeup	3	2	15	**44**	0	0	0	8
InST	0	0	2	9	2	1	6	0
InstantStyle	0	0	0	0	0	0	0	0
InstantStyle(+MaFor)	0	0	0	0	0	0	0	0
I2I SDXL	0	0	0	1	2	2	2	0
InstructPix2Pix	0	0	0	1	1	12	2	18
Inpaint	5	0	0	0	0	4	4	0
Inpaint+TI	6	5	0	3	0	1	0	2
Ours (Gorgeous)	**86**	**43**	**79**	35	**95**	**80**	**86**	**72**

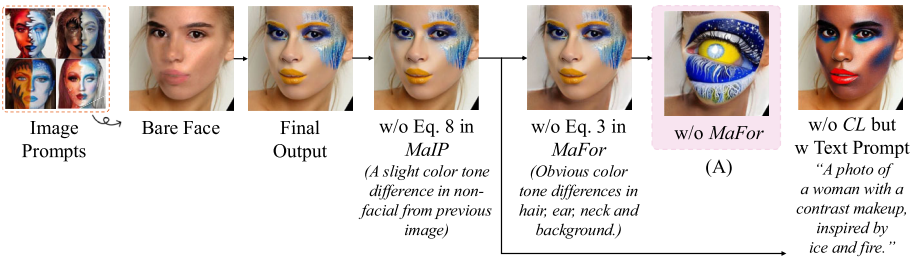

Fig. 6 Ablation Study. (A) shows the challenge addressed in our problem statement: Without MaFor, the model struggles with makeup-specific outputs, treating target regions as generic inpainting tasks

non-facial areas remain consistent with the input bare faceś distribution rather than introducing unintended artifacts. **(ii)** Eq. (3) **in MaFor:** The removal of Eq. (3) leads to a noticeable deviation in non-facial regions. This is reflected in a slight drop in CSD Similarity (-0.01) and DreamSIM (-0.02), though these metrics primarily assess conceptual alignment. More critically, the FID score rises significantly to 111.81, indicating a distribution shift in non-facial areas. Since FID evaluates statistical consistency, this increase suggests that omitting Eq. (3) makes non-facial regions diverge from their expected appearance, reinforcing the necessity of this equation for maintaining coherence to input bare face. **(iii)** *MaFor* **Removal:** Without *MaFor*, the generated images fail to maintain a recognizable makeup structure. Interestingly, CSD Similarity (0.34) and DreamSIM (0.46) are the highest among all settings, suggesting that the model still captures conceptual alignment well. However, FID deteriorates drastically to 250.81, the worst among all cases, showing that while the narrative elements are detected, the absence of MaFor leads to unstructured makeup synthesis. This confirms that MaFor is crucial for ensuring that narrative-driven designs are structured in a makeup form rather than arbitrary visual effects. **(iv)** ***CL*** **Module Removal but with** ***MaFor***: Eliminating *CL* severely limits the model's ability to incorporate narrative elements (e.g., ice or fire) into the makeup design. This is evident from the lowest CSD Similarity (0.16) and DreamSIM (0.23) scores, indicating a significant misalignment between the generated makeup and the intended concept. While the FID score of 169.72 is still relatively high, it remains better than the case without *MaFor*, reaffirming *MaFor*'s critical role in structuring the makeup application, even when narrative elements are not explicitly encoded. These results demonstrate the critical contribution of each module in achieving consistent, contextually relevant and high-quality makeup designs.

Table 3 Quantitative results of the ablation study

Method	CSD Similarity (↑)	DreamSIM (↑)	FID (↓)
Full Model (Gorgeous)	0.21	0.28	**89.84**
w/o Eq. (8) in MaIP	0.20	0.27	90.75
w/o Eq. (3) in MaFor	0.20	0.26	111.81
w/o MaFor	**0.34**	**0.46**	250.81
w/o CL Module	0.16	0.23	169.72

The best results are highlighted in **bold**

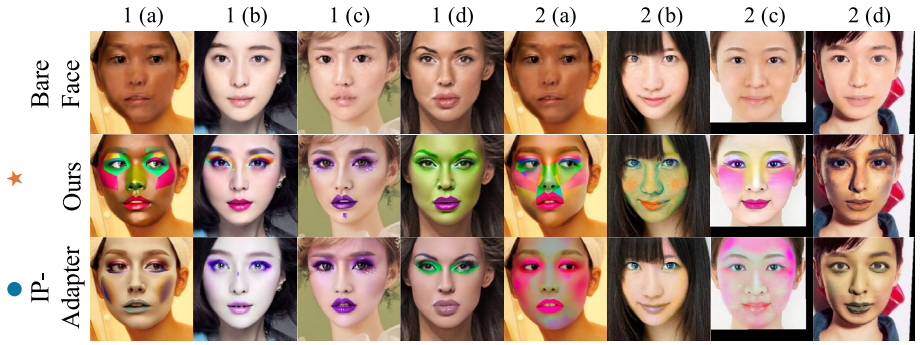

Fig. 7 Comparison of Textual Inversion and IP-Adapter in makeup generation. *Top:* Our method, Gorgeous, integrated with Textual Inverson is able to capture user-provided multi-image concepts. *Bottom:* IP-Adapter fails to capture intended contextual elements which can be presented through shared characteristics among multiple image prompts

6.2 Additional study on textual inversion vs. IP-Adapter in *CL* Module

To further justify our choice of Textual Inversion over IP-Adapter, we conducted an additional empirical study by replacing Textual Inversion with IP-Adapter in the *CL* (Context Learning) Module, while keeping all other settings in Gorgeous unchanged.

As demonstrated in Fig. 7, **IP-Adapter struggles to capture user-defined concepts beyond a single reference image.** This is evident in cases where requiring users to provide multiple example images to define an abstract concept-IP-Adapter fails to synthesize makeup that aligns with the intended theme. In contrast, Textual Inversion successfully learns and applies multi-image-based concepts, ensuring accurate, narrative-driven makeup generation.

Additionally, the quantitative results shown in Table 4 show that Gorgeous with **Textual Inversion achieves superior CSD and DreamSIM scores**, confirming better conceptual relevance. While FID is reported for reference, it primarily assesses **MaFor and MaIP components**, rather than the CL module responsible for capturing the makeup context.

These results highlight the importance of Textual Inversion over IP-Adapter for narrative-driven makeup generation, as it enables multi-image concept learning and better user control over style representation.

7 Computational cost

The computational cost analysis was conducted using an Nvidia RTX 3090 GPU with a batch size of 1 and an accumulative gradient of 4. Inference requires approximately 0.117 seconds

Table 4 Quantitative evaluation of Textual Inversion vs. IP-Adapter

Methods	Average of Style 1			Average of Style 2		
	CSD ↑	SIM ↑	FID ↓	CSD ↑	SIM ↑	FID ↓
IP-Adapter	0.55	0.61	57.53	0.20	0.25	66.16
Ours (Gorgeous)	**0.61**	**0.65**	53.29	**0.21**	**0.28**	89.84

Higher CSD and DreamSIM scores indicate better conceptual alignment with user-provided image prompts. Whereas, FID here is for a reference only as it is not used to assess the capability of CL Module, instead, it is for MaFor and MaIP Module

per step, with 50 iterations leading to a total inference time of around 5 seconds, which is comparable to common ControlNet-based models. Training with *MaFor* takes approximately 0.735 seconds per step, running for 15,000 iterations, resulting in a total training time of about 11,000 seconds (approximately 3 hours). For token-based *CL* training, *MaFor* is not utilized, leading to a slightly faster training speed, with each step taking around 0.540 seconds. The number of iterations varies from 500 to 5,000, depending on the complexity of capturing the shared style among the given images. This results in a total training time ranging from 270 to 2,700 seconds (approximately 4.5 to 45 minutes).

8 Application

8.1 Testing *Gorgeous* on wild images

To validate *Gorgeous*'s adaptability, we tested it on wild images collected from online sources[4][5][6], outside the standard dataset. Figure 8 demonstrates its ability to apply narrative-driven makeup to diverse faces, showcasing its real-world applicability without prior tuning.

8.2 Impact of inference steps

Adjustable inference steps (10-100) and guidance scale (g, 3-20) leads to different makeup intensities and details. Figure 9 demonstrates how more steps enhance definition, from minimalist to detailed.

8.3 Impact of guidance scale

Whereas, Fig. 10 illustrates how g modulates the strength and fidelity of makeup application to align with the given prompts, offering flexibility in generating personalized, narrative-driven makeup designs. To gain insight into user preferences regarding the effect of guidance scale on makeup intensity, we collected feedback from the 100 participants who took part in the User Study earlier 2 to determine the most preferred guidance scale setting. Our findings, summarized in Table 5, indicate that the most preferred guidance scales were 6 and 7.5 for the top row, and 20 for the bottom row. While higher values of g enhance vibrancy and stylization, they may also lead to overly dramatic effects that some users perceive as unnatural. However, in some cases where bombastic makeups are needed, lower values are hard to show the makeup, it needs higher guidance scale to present. This aligns with the concern that extreme stylization may not always be ideal for real-world makeup references.

These insights reinforce the importance of adjustable guidance strength in generative models, ensuring that users can customize the intensity of the generated makeup based on their specific needs and preferences.

[4] https://www.pinterest.com/

[5] https://k.sina.cn/

[6] https://www.facebook.com/

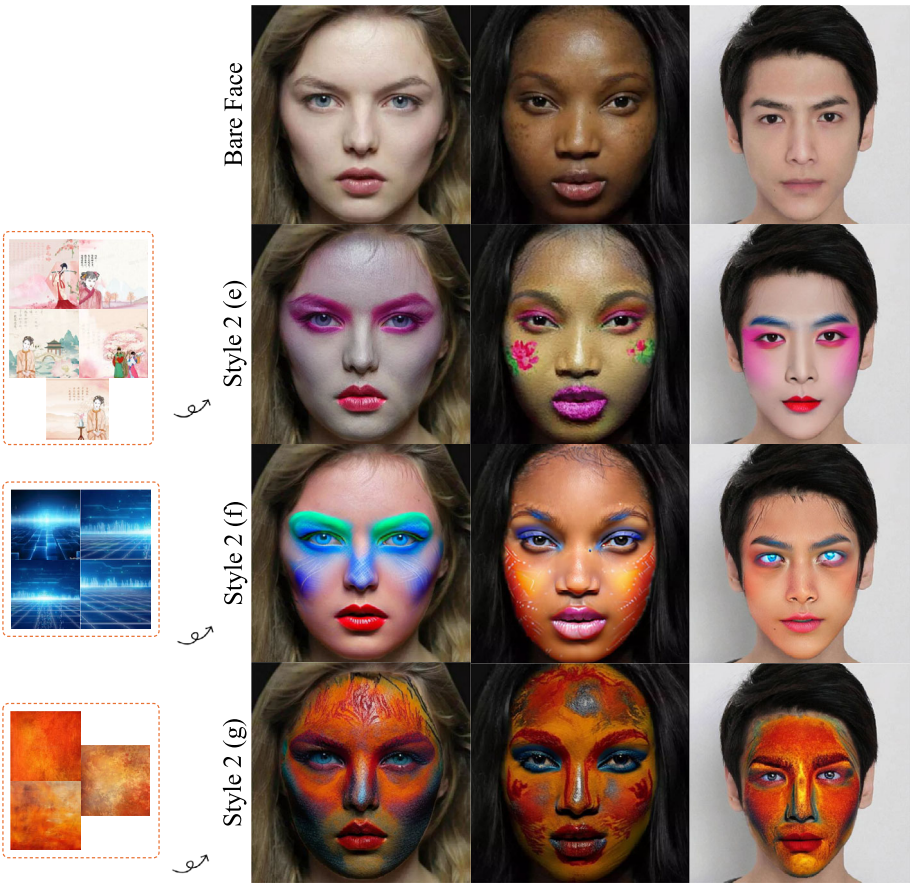

Fig. 8 Application of *Gorgeous* on wild images

Fig. 9 Influence of Inference Steps on Makeup Application. This figure shows the effect of varying inference steps (10 to 100) on narrative-driven makeup generation with *Gorgeous*. Increasing steps enhances definition and vividness, allowing users to control the transformation from subtle to detailed makeup

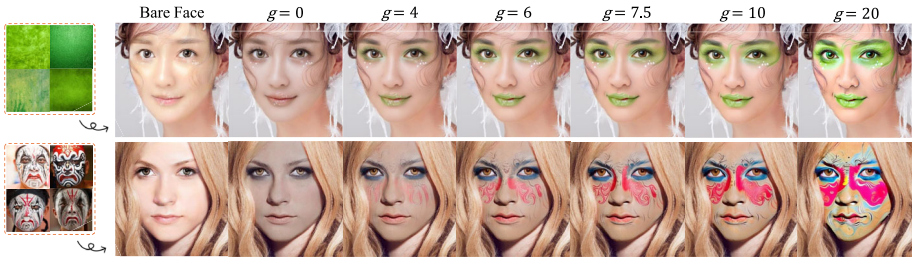

Fig. 10 Impact of guidance scale on Makeup Intensity. It shows the effect of varying guidance scales for two narrative ideas. Adjusting the scale from 0 to 20 alters the intensity, demonstrating *Gorgeous*'s ability to cater personal preferences

9 Limitations

A potential area for future improvement in our work lies in enhancing the textural and color accuracy of the generated makeup. However, in its current phase, our approach is not designed for direct real-world application but rather as a source of inspiration for creative makeup artistry. The primary objective is to generate makeup concepts that align with the narrative elements of image prompts, allowing users to explore and visualize new ideas.

While certain limitations exist in capturing fine-grained texture, precise color blending, or detailed application techniques, these aspects are not critical to evaluating our method at this stage. Instead, our work focuses on conceptual alignment, ensuring that the generated makeup correctly conveys user-intended themes (e.g., fire, ice, galaxy).

In future work, we aim to explore techniques that extend our approach for high-fidelity virtual makeup applications, where attributes like texture detail and fine-tuned shading become more relevant.

10 Conclusion

Gorgeous is the first work that allows narrative-driven makeup generation, empowering users to weave narrative and creative elements into makeup designs using image prompts. Unlike traditional makeup transfer methods, *Gorgeous* introduces a foundational framework, addresses unpaired data challenges, and innovatively integrates image inpainting to generate contextually relevant and visually compelling makeup ideas for real life application. Future work focuses on developing realistic makeup textures and introducing more user-editable features, enhancing usability and bridging the gap between automated generation and personal customization. With its novel approach, *Gorgeous* sets a new standard for creativity

Table 5 User preferences on guidance scale for makeup intensity

Makeup	(Top Row)						(Bottom Row)					
Scale	0	4	6	7.5	10	20	0	4	6	7.5	10	20
Vote	2	29	**31**	20	18	0	12	23	9	7	18	**31**

The top row corresponds to the makeup styles demonstrated in the top row of Fig. 10, while the bottom row corresponds to the makeup styles demonstrated in the bottom row of Fig. 10. Bold values indicate the most preferred guidance scales in each category

in digital makeup, paving the way for advancements in computer vision research and artistic expression.

Acknowledgements We would like to thank all the participants who took part in our surveys for their critical review of the results.

Author Contributions Jia Wei Sii conducted all experiments, code implementation and manuscript writing. Chee Seng Chan contributed to writing the manuscript. The idea of this research was discussed together.

Funding This research is self-funded.

Data Availability BeautyFace dataset can be downloaded from their released code (https://github.com/learningyan/BeautyREC/). For our image prompts dataset, we collected randomly from various online and free platforms such as Pinterest, we will be releasing them upon publication.

Declarations

Competing interests No competing interests.

References

1. Eldridge L (2015) Face Paint: The Story of Makeup. Abrams, New York City
2. Yang C, He W, Xu Y, Gao Y (2022) Elegant: Exquisite and locally editable gan for makeup transfer. In: ECCV. Springer
3. Sun Z, Chen Y, Xiong S (2022) Ssat: A symmetric semantic-aware transformer network for makeup transfer and removal. In: AAAI
4. Yan Q, Guo C, Zhao J, Dai Y, Loy CC, Li C (2023) Beautyrec: Robust, efficient, and component-specific makeup transfer. In: CVPR
5. Zhang Y, Wei L, Zhang Q, Song Y, Liu J, Li H, Tang X, Hu Y, Zhao H (2024) Stable-Makeup: When Real-World Makeup Transfer Meets Diffusion Model
6. Guo D, Sim T (2009) Digital face makeup by example. In: CVPR. IEEE
7. Liu S, Ou X, Qian R, Wang W, Cao X (2016) Makeup like a superstar: Deep localized makeup transfer network. arXiv preprint arXiv:1604.07102
8. Chang H, Lu J, Yu F, Finkelstein A (2018) Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In: CVPR
9. Li T, Qian R, Dong C, Liu S, Yan Q, Zhu W, Lin L (2018) Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: ACMMM
10. Chen H-J, Hui K-M, Wang S-Y, Tsao L-W, Shuai H-H, Cheng W-H (2019) Beautyglow: On-demand makeup transfer framework with reversible generative network. In: CVPR
11. Deng H, Han C, Cai H, Han G, He S (2021) Spatially-invariant style-codes controlled makeup transfer. In: CVPR
12. Jiang W, Liu S, Gao C, Cao J, He R, Feng J, Yan S (2020) Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In: CVPR
13. Lyu Y, Dong J, Peng B, Wang W, Tan T (2021) Sogan: 3d-aware shadow and occlusion robust gan for makeup transfer. In: ACMMM
14. Nguyen T, Tran AT, Hoai M (2021) Lipstick ain't enough: beyond color matching for in-the-wild makeup transfer. In: CVPR
15. Jin Q, Chen X, Jin M, Cheng Y, Shi R, Zheng Y, Zhu Y, Ni B (2024) Toward tiny and high-quality facial makeup with data amplify learning. arXiv preprint arXiv:2403.15033
16. Nguyen TV, Liu L (2017) Smart mirror: Intelligent makeup recommendation and synthesis. In: ACMMM
17. Alashkar T, Jiang S, Wang S, Fu Y (2017) Examples-rules guided deep neural network for makeup recommendation. In: AAAI
18. Alashkar T, Jiang S, Fu Y (2017) Rule-based facial makeup recommendation system. In: FG
19. Perera P, Soysa E, De Silva H, Tavarayan A, Gamage M, Weerasinghe K (2021) Virtual makeover and makeup recommendation based on personal trait analysis. In: ICAC

20. Gulati K, Verma G, Mohania M, Kundu A (2023) Beautifai-personalised occasion-based makeup recommendation. In: ACML
21. Jing Y, Yang Y, Feng Z, Ye J, Yu Y, Song M (2020) Neural style transfer: A review. T-VCG
22. Cai Q, Ma M, Wang C, Li H (2023) Image neural style transfer: A review. Computers and Electrical Engineering
23. Li Y, Wang N, Liu J, Hou X (2017) Demystifying neural style transfer. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI'17
24. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: CVPR
25. Zhang W, Cao C, Chen S, Liu J, Tang X (2013) Style transfer via image component analysis. TMM
26. Liao J, Yao Y, Yuan L, Hua G, Kang SB (2017) Visual attribute transfer through deep image analogy. ACM Transactions on Graphics (TOG). 36:1–15
27. Gu S, Chen C, Liao J, Yuan L (2018) Arbitrary style transfer with deep feature reshuffle. In: CVPR
28. Zhang Z, Liu Y, Han C, Guo T, Yao T, Mei T (2022) Generalized one-shot domain adaptation of generative adversarial networks. NeurIPS
29. Kolkin N, Salavon J, Shakhnarovich G (2019) Style transfer by relaxed optimal transport and self-similarity. In: CVPR
30. Deng Y, Tang F, Dong W, Sun W, Huang F, Xu C (2020) Arbitrary style transfer via multi-adaptation network. In: ACMMM
31. Yao Y, Ren J, Xie X, Liu W, Liu Y-J, Wang J (2019) Attention-aware multi-stroke style transfer. In: CVPR
32. Liu S, Lin T, He D, Li F, Wang M, Li X, Sun Z, Li Q, Ding E (2021) Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: ICCV
33. An J, Huang S, Song Y, Dou D, Liu W, Luo J (2021) Artflow: Unbiased image style transfer via reversible neural flows. In: CVPR
34. Chen H, Wang Z, Zhang H, Zuo Z, Li A, Xing W, Lu D et al (2021) Artistic style transfer with internal-external learning and contrastive learning. In: NeurIPS
35. Zhang Y, Tang F, Dong W, Huang H, Ma C, Lee T-Y, Xu C (2022) Domain enhanced arbitrary image style transfer via contrastive learning. In: SIGGRAPH
36. Wu X, Hu Z, Sheng L, Xu D (2021) Styleformer: Real-time arbitrary style transfer via parametric style composition. In: ICCV
37. Deng Y, Tang F, Dong W, Ma C, Pan X, Wang L, Xu C (2022) Stytr2: Image style transfer with transformers. In: CVPR
38. Cheng B, Liu Z, Peng Y, Lin Y (2023) General image-to-image translation with one-shot image guidance. In: ICCV
39. Zhang Y, Huang N, Tang F, Huang H, Ma C, Dong W, Xu C (2023) Inversion-based style transfer with diffusion models. In: CVPR
40. Huo J, Liu X, Li W, Gao Y, Yin H, Luo J (2022) Cast: Learning both geometric and texture style transfers for effective caricature generation. TIP
41. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125
42. Balaji Y, Nah S, Huang X, Vahdat A, Song J, Kreis K, Aittala M, Aila T, Laine S, Catanzaro B et al (2022) ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324
43. Ding M, Zheng W, Hong W, Tang J (2022) Cogview2: Faster and better text-to-image generation via hierarchical transformers. In: NeurIPS
44. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, Sutskever I, Chen M (2021) Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741
45. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, et al. (2022) Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS
46. Ding M, Yang Z, Hong W, Zheng W, Zhou C, Yin D, Lin J, Zou X, Shao Z, Yang H et al (2021) Cogview: Mastering text-to-image generation via transformers. In: NeurIPS
47. Zhang L, Rao A, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. ICCV
48. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. In: NeurIPS
49. Brooks T, Holynski A, Efros AA (2023) Instructpix2pix: Learning to follow image editing instructions. In: CVPR
50. Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D (2022) An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618

51. Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (2023) Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR
52. Tumanyan N, Geyer M, Bagon S, Dekel T (2023) Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR
53. Kawar B, Zada S, Lang O, Tov O, Chang H-T, Dekel T, Mosseri I, Irani M (2023) Imagic: Text-based real image editing with diffusion models. In: CVPR
54. Kim G, Kwon T, Ye J-C (2022) Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR
55. Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J, Rombach R (2023) Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952
56. Parmar G, Singh KK, Zhang R, Li Y, Lu J, Zhu J-Y (2023) Zero-shot image-to-image translation. In: SIGGRAPH
57. Avrahami O, Lischinski D, Fried O (2022) Blended diffusion for text-driven editing of natural images. In: CVPR
58. Meng C, He Y, Song Y, Song J, Wu J, Zhu J-Y, Ermon S (2021) Sdedit: Guided image synthesis and editing with stochastic differential equations. In: ICLR
59. Ye H, Zhang J, Liu S, Han X, Yang W (2023) Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721
60. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: CVPR
61. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: ICML
62. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: MICCAI
63. Gu Q, Wang G, Chiu MT, Tai Y-W, Tang C-K (2019) Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In: CVPR
64. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV
65. Somepalli G, Gupta A, Gupta K, Palta S, Goldblum M, Geiping J, Shrivastava A, Goldstein T (2024) Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292
66. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS
67. Wang H, Xing P, Huang R, Ai H, Wang Q, Bai X (2024) InstantStyle-Plus: Style Transfer with Content-Preserving in Text-to-Image Generation. https://arxiv.org/abs/2407.00788
68. Fu S, Tamir N, Sundaram S, Chai L, Zhang R, Dekel T, Isola P (2023) Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: NeurIPS