ACFun: Abstract-Concrete Fusion Facial Stylization

Jiapeng Ji, Kun Wei*, Ziqi Zhang, Cheng Deng
School of Electronic Engineering, Xidian University
Xi'an 710071, China
jiapengji777@gmail.com, weikunsk@gmail.com, zqzh9116@gmail.com,
chdeng.xd@gmail.com

Abstract

Owing to advancements in image synthesis techniques, stylization methodologies for large models have garnered remarkable outcomes. However, when it comes to processing facial images, the outcomes frequently fall short of expectations. Facial stylization is predominantly challenged by two significant hurdles. Firstly, obtaining a large dataset of high-quality stylized images is difficult. The scarcity and diversity of artistic styles make it impractical to compile comprehensive datasets for each style. Secondly, while many methods can transfer colors and strokes from style images, these elements alone cannot fully capture a specific style, which encompasses both concrete and abstract visual elements. Additionally, facial stylization often alters the visual features of the face, making it challenging to balance these changes with the need to retain facial information. To address these issues, we propose a novel method called ***ACFun***, which uses only one style image and one facial image for facial stylization. *ACFun* comprises an *Abstract Fusion Module (AFun)* and a *Concrete Fusion Module (CFun)*, which separately learn the abstract and concrete features of the style and face. We also design a *Face and Style Imagery Alignment Loss* to align the style image with the face image in the latent space. Finally, we generate styled facial images from noise directly to complete the facial stylization task. Experiments show that our method outperforms others in facial stylization, producing highly artistic and visually pleasing results.

1 Introduction

Style transfer is a long-standing research topic that aims to generate a new artistic image from an arbitrary input pair of natural images and painting images by combining the content of the natural image and the style of the painting image. With the rapid development of deep learning, optimization-based methods [1–3] and feed-forward based methods [4, 2, 5, 6] continually emerged. These methods have made progress in visual quality and computational efficiency. However, these style transfer techniques can only modify low-level image features such as strokes and colors, and they are unable to adjust high-level semantics such as shape and content. It's like using the style image as a template to paint the content image instead of truly creating a new image based on the style. Style encompasses not only basic visual elements like strokes and colors but also deeper aspects such as structure, layout, and shape. These elements combine to create a unique visual language and form of expression. In the context of facial stylization, the challenges of style transfer are even more significant. Facial recognition relies on distinct and recognizable individual characteristics, which means that style transfer must capture the visual features of the style image and seamlessly integrate them into the target image while preserving facial recognizability.

With the rapid development of diffusion model-based image generation methods, several style transfer methods [7–9] have emerged, but these methods usually require a set of images belonging to the same

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

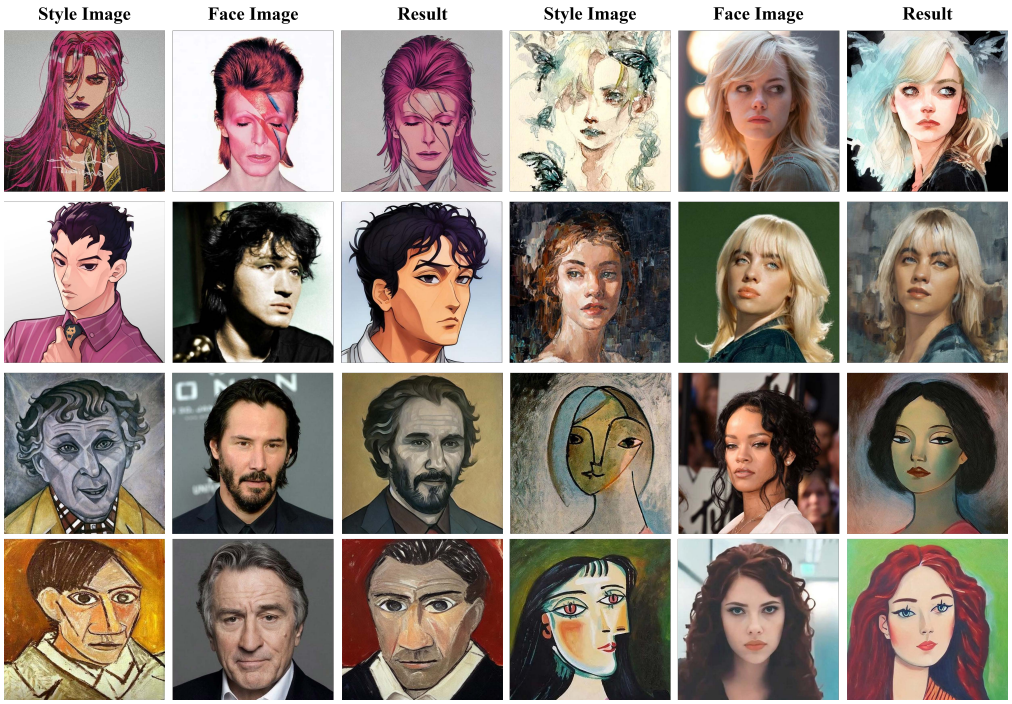

Figure 1: As shown in the figure, the ideal result of facial stylization aims to ensure the faithful restoration of facial information while also significantly deforming the facial image according to the given style image. To achieve high-quality results, it is essential to ensure that the production process not only meets the standards for good generation quality but also encompasses a compelling and evocative artistic atmosphere.

style for training. However, due to the scarcity of high-quality stylized images, it is difficult to obtain them. Moreover, the diversity and complexity of artistic styles make compiling a comprehensive dataset for each style impractical. Subsequently, there are also some Textual-Inversion methods [10, 11] that can use a single image, but this method only fine-tuned the CLIP model as the conditional input, which limited its ability to generate models and often caused it to be heavily affected by style degradation, resulting in simpler results and reduced image quality. Moreover, its style transfer method of adding noise and denoising to the target image reduces the flexibility of the generated results and further limits its final generation performance.

As depicted in Figure 1, the most prominent challenge in facial stylization tasks is that images with artistic styles often display significant differences from real-world facial photos. These differences are evident not only in visual features but also in high-level semantics, such as shape and distribution. In particular, as shown in the lower part of Figure 1, there are notable disparities between abstract style portraits and real portraits, leading to substantial changes in facial features during the stylization process. On the other side of the coin, it is essential for stylized facial images to maintain consistency in facial information so that they remain recognizable to people. In summary, for stylized facial images, balancing the ability to maintain facial images with the drastic semantic changes brought about by stylization is the key to achieving good facial stylization results.

To this end, we propose a facial stylization method that allows deep neural networks to fuse abstract visual elements (deformation, distribution, atmosphere, etc.) from style and facial image with concrete visual elements (color, brushstrokes, lines, etc.) into an imagery visual feature just like artists, to achieve facial image stylization, rather than simply transferring texture and color information. Specifically, we introduced an Abstract-Concrete Fusion Facial Stylization Model (ACFun) to fuse, consisting of an Abstract Fusion Module(AFun) and a Concrete Fusion Module (CFun). The method ACFun uses for facial stylization involves learning abstract and concrete features separately and then combining style images with facial images by aligning them in an imagery feature space formed by the fusion of abstract and concrete features. We first extract abstract features of style and face

through AFun by encoding style images into the CLIP latent space and combining it with the facial description text prompts we use; then, we optimize it using a Textual-Inversion-based method to make it recognizable by CLIP for subsequent generation processes. In addition, by introducing face images during the training process, we implicitly fuse facial information with style abstract information. However, CLIP has limitations in distinguishing between specific style and facial images at the fine-grained level. To address this, we adopt the CFun module, which is a set of trainable parameters embedded in a pre-training generator to enhance its ability to extract concrete features from specific styles and facial images, which greatly improves the quality of the generated results. Finally, the above process is constrained by the Face and Style Imagery Alignment loss we proposed. We demonstrate the effectiveness and excellent facial stylization effect of our method through massive experiments.

2 Related Works

Text-to-Image Generation. The method based on generating adversarial networks [12] has been widely applied in text-to-image generation by training on paired images and text samples. [13] improves condition generation[14] by changing the input conditions from a single class to more complex text descriptions. However, due to the difficulty of training GAN models and the significant progress of large language models [15, 16] and diffusion models, the diffusion models have gradually replaced the position of GAN. The diffusion model[17, 18]can generate realistic images through multiple rounds of iterative refinement. It has been widely used in the image generation task and has achieved significant results [19], attracting the attention of many researchers. Among all these diffusion-based methods, Stable Diffusion [20] is the most typical work. It achieves awe-inspiring results through training on large-scale text image datasets. Afterward, SDXL [21] was introduced with larger model parameters and larger datasets. However, these methods all focus on learning text prompts, meaning that the model can only fit the given reference style through existing knowledge rather than train the generated model to have the ability to generate specific style images. We fine-tune the model to make it more intuitive to learn the target style straightly. Thanks to these powerful pre-trained generative models, we can flexibly apply them to various downstream task applications.

Personalization of Generative Models. How to use a pre-trained text-to-image large model to complete personalized image synthesis tasks has attracted widespread attention [22]. The widely used method in image generation tasks is the inversion-based method, which aims to find a corresponding noise or conditioning embedding to a generated image. Textual Inversion [9] and Hard prompt made easy (PEZ) [23] convert a set of images of an object into corresponding textual representations (such as embedding, token) instead of changing the parameters of the text-to-image model. Dreambooth [7] learns target concepts by fine-tuning the pre-trained model. At the same time, when learning new class concepts, Dreambooth transforms this concept into a special sub-concept of a specific class. Dreambooth achieves the goal of data augmentation by generating original classes, preventing new concept classes from overwriting existing original classes, leading to catastrophic forgetting and affecting the diversity of generated models, leading to overfitting. Lora [8] has also achieved widespread application by transferring the strategy of fine-tuning pre-trained LLM to image generation methods[24], allowing it to save computational and storage costs without the need to train the entire generation network. Although the above methods have achieved good results, they all learn the description of specific concepts through a set of images, which makes it difficult to meet our needs. The method we adopt only requires a pair of images to learn the abstract and concrete features of style images and facial images and ultimately generate a stylized facial image.

Facial Stylization. Facial stylization is currently a widely recognized task in the field of style conversion. Typical style transfer methods such as [25, 26, 4, 6, 27] can effectively transfer the concrete visual features such as texture and color contained in the given style image. However, when facing facial images, especially when transferring real facial images to styles with strong geometric distortions (such as comics, anime, abstract paintings, etc.), the effect is often poor. Although some methods [28, 29] use large-scale paired image data for image-to-image translation training, in practical applications, due to the diversity and scarcity of high-quality style images, collecting such a large number of paired images is not realistic, and some images do not even exist, which makes the above methods difficult to apply in practice. StyleGAN [30] achieves facial stylization through image inversion [31, 29, 32–34], but due to the presence of its latent space, it suffers from problems such as image degradation and limited deformation. Although the method of textual inversion [10, 11] can form certain deformations, the problem of style degradation cannot be ignored. Our method solves the above problem by learning abstract and concrete features separately and aligning images.

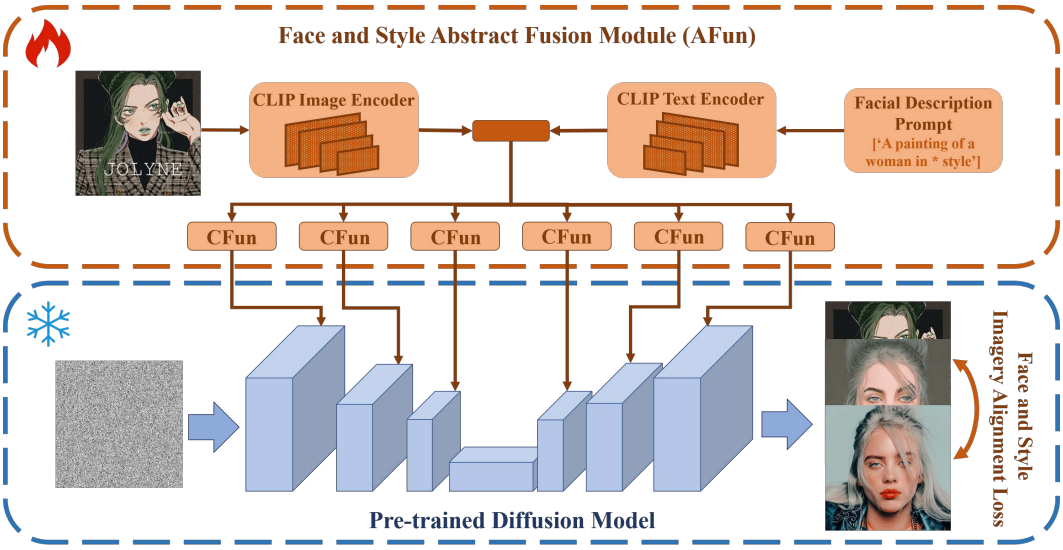

Figure 2: The overall schematic diagram of ACFun consists of two parts: one is the AFun module for extracting abstract features, and the other is the CFun module for extracting concrete features. The entire training process is controlled by our carefully designed Face and Style Imagery Alignment loss. We utilized a facial description text prompt to enhance the fusion of style and facial images.

3 Method

3.1 Overview

In this paper, we propose a novel method named ACFun for image style transfer; the structure and process are shown in Figure 2. Facial style transfer aims to preserve facial identity information while being as close to the target style as possible. Inspired by different artistic styles of expression, we divide visual features into more complex and difficult to describe abstract features, such as shape or atmosphere, and more specific concrete features, such as color or strokes. And ultimately, by combining abstract features and concrete features into imagery features and utilizing them to generate stylized facial images.

Therefore, we first propose an AFun module. By using a CLIP encoder to encode the style image and combine it with a facial description text prompt using the attention mechanism to fully use the information of the reference style image. Then, we optimize it into an embedding $e*$ that can be recognized by the generated model in the subsequent generation process. However, simply using the aforementioned Inversion-based method often leads to style degradation, resulting in the loss of some concrete visual content in the final result, which is often fatal for facial stylization tasks. So, both facial and style images are introduced simultaneously during the training process. To this end, we adopt the CFun module, to improve the flexibility of the network and enable it to have contract generation capabilities, we insert a set of trainable parameters into a pre-trained U-net and adopt the Face and Style Imagery Alignment loss to constrain the above process.

3.2 Face and Style Abstract Fusion Module (AFun)

To fully utilize the pre-trained large-scale model, we hope to learn the style in the target style image as a specific intermediate representation, which is suitable for the pre-trained model. This method uses CLIP as a text encoder for text-to-image generation in stable diffusion. It inputs the text embedding, which is embedded by the CLIP text encoder τ_t as a condition. We use the pseudo-word 'S^*' as a placeholder for the image style and introduce the concept of specific image styles into the pre-trained model through a learnable corresponding embedding e_t^*. However, using only a single pseudo-word S^* as a placeholder, this simple and vague text prompt inevitably leads to us only being able to extract the content of the entire style image, regardless of whether it is an abstract style feature or a specific content feature. This will bring problems due to the content of the style image in our subsequent

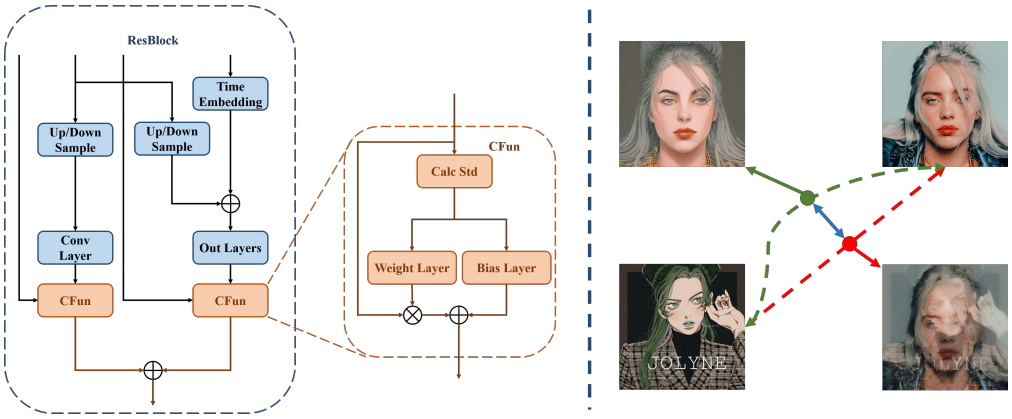

Figure 3: **Left:** The specific insertion method and structure of our CFun. **Right:** Our proposed imagery latent space, where the **red** line represents the VQ visual space, the **green** line represents the imagery latent space, and the **blue** line represents the process of mapping the original image from the VQ space into the imagery latent space through the introduction of abstract features e_a.

facial stylization process, such as style leakage or style coverage. To address this issue, we used facial description text prompts S_f^*. The style and content of the style image can be automatically separated by the description of the facial image, such as "A painting of a woman in S^* style." This is because in the subsequent training process, style images I_s and facial images I_f are introduced, and CLIP naturally implicitly receives information from facial images I_f during the feature update process. We embed the style image I_s using CLIP image encoder τ_i as embedding $e_{i_s}^*$. Then, we fuse it with facial description text prompts embedding $e_{t_f}^*$ to obtain abstract features e_a. This process can be expressed as:

$$\underset{e_{t_f}=e_a}{min}\ E_{I_s,I_f}E_{z_t\sim q(z_t|I_s,I_f)}\|\epsilon-\hat{\epsilon}_\theta(z_t,t,\tau_t(y,S_f^*))\|_2^2, \tag{1}$$

where ϵ stands for an initial noise map $\epsilon\sim N(0,1)$, $\hat{\epsilon}_\theta$ is the denoising network, $t=1\dots T$ stands for the time step for the diffusion model. Since our method only requires a pair of images to complete training, it is difficult to thoroughly learn the style in the background image using a vanilla textual inversion method, and overfitting is prone to occur. To alleviate this problem, we also introduced image information, allowing the model to learn the style and facial abstract visual elements in the image more fully. Benefiting from the CLIP model that has already aligned text with images in its latent space as a multi-modal model, the fusion process of style image embedding $e_{i_s}^*$ with facial description text prompts embedding $e_{t_f}^*$ using an attention mechanism can be represented as:

$$\begin{aligned} Q_n &= W_Q^{(n)}\cdot e_{t_f}^{(n)}, K = W_K^{(n)}\cdot e_{i_s}, V = W_V^{(n)}\cdot e_{i_s}, \\ e_a^{(n+1)} &= Attention(Q_n,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \end{aligned} \tag{2}$$

and set dropout in the attention mechanism to alleviate overfitting problems.

3.3 Face and Style Concrete Fusion Module (CFun)

Even if we have obtained an abstract feature e_a that blends style and facial information, relying solely on this feature to generate images means that we can only fit the target style and facial image using the prior knowledge already present in the pre-trained model. However, in real life, there are numerous style images and facial images, many unseen by pre-trained models, making it challenging to generate the desired results. To solve the above problem, we propose the CFun module for generating concrete visual features. By inserting a set of learnable parameters into U-net, we enhance the flexibility of the network and enable it to learn the visual features of input images. For its insertion position, considering that we do not want to damage the style and information in the facial image excessively. It is too risky to directly insert parameters into the Attention Block for training, as this

may involve overly advanced semantics. Directly changing the Attention Block may result in too drastic advanced semantic changes, ultimately leading to excessively distorted stylized face images. So, we choose to insert the parameters into Resblock, which can make our parameters focus more on specific concrete visual features and indirectly affect the Attention module, thus achieving a balance between the information retention ability of facial images and the high-level semantic changeability brought by stylization. In addition, inspired by AdaIN [4], our insertion module adopts an affine mechanism, which can ensure that the network parameters are not affected during initial training, further improving the stability during training. The specific structure and insertion method are shown on the left of Figure 3. We separate content from the style by calculating mean-standard:

$$CFun(x) = \gamma(\sigma(x))x \oplus \beta(\sigma(x)). \tag{3}$$

And only the style information, i.e., standard $\sigma(x)$, is fused with the network. This is mainly because the abstract features we use are only fused from style images and facial description text prompts at the beginning of training, and no facial image information has been obtained yet. To avoid excessive activation of abstract features and style images in cross-attention during the training process, causing the generated stylized facial images to irreversibly guide to style images, resulting in the final generation result being strongly covered by style.

3.4 Face and Style Imagery Alignment Loss

Finally, we use a Face and Style Imagery Alignment loss to constrain all the above processes. However, as shown on the right of Figure 3, even if encoding using VQ space, the interpolated images represented by simple style images and facial images are meaningless, just a simple superposition of the two images. However, initially, we had already used AFun to fuse style images with facial description text prompts and provided an abstract feature e_a for subsequent training processes. This allowed us to have abstract feature e_a guidance in the subsequent optimization process. After combining the abstract feature with initial noise, the space it projects onto is not simply a VQ latent space composed of simple image features but a latent space composed of semantic features after the combination, we call this latent space, which combines abstraction and concreteness, the imagery space. This also makes our proposed Face and Style Imagery Alignment loss effective. Specifically, by aligning I_s and I_f in the imagery space, we obtain an image I_i that integrates abstract and concrete features of facial image I_f and style image I_s to achieve the goal of facial stylization. We also use the two hyperparameters we establish, β and γ, to control the tendency of the model towards the relationship between facial image content and style during the optimization process. Finally, the pre-trained text-to-image diffusion model DM is used, and the initial noise map $\epsilon \sim N(0, 1)$ and the abstract feature e_a^* obtained in the previous step are given as inputs to generate the image $\hat{I}_i = DM(\epsilon, e_a^*)$, and the diffusion model is fine-tuned using the squared error loss. The specific formula is as follows:

$$\min_{\hat{I}_i} E_{I_i,\epsilon,t,e_a^*}[\beta\omega_t \|DM(\alpha_t I_i + \sigma_t\epsilon, e_a^*) - I_s\|_2^2 + \gamma\omega_t \|DM(\alpha_t I_i + \sigma_t\epsilon, e_a^*) - I_f\|_2^2]. \tag{4}$$

Where $\alpha_t, \sigma_t, \omega_t$ are parameters used to control the noise schedule and sampling quality and are functions of the diffusion process time $t \sim U([0, 1])$. Meanwhile, this is a win-win training strategy. On the one hand, the abstract features e_a of AFun provide a good latent space for the training of the generator. On the other hand, through gradient backward, the abstract features of AFun can also extract concrete information from facial images, providing better quality guidance. Through this method, we can constrain the optimization process of abstract feature fusion for style images and face images in CLIP space, as well as the extraction of concrete features for style images and face images using learnable parameters inserted into the generator.

4 Experiments

In this section, we conduct experiments using images collected from the Internet and provide visual comparisons. We also conducted experiments on both facial stylization and text-to-image generation tasks. Our method has achieved good visual effects in various experiments, demonstrating its effectiveness.

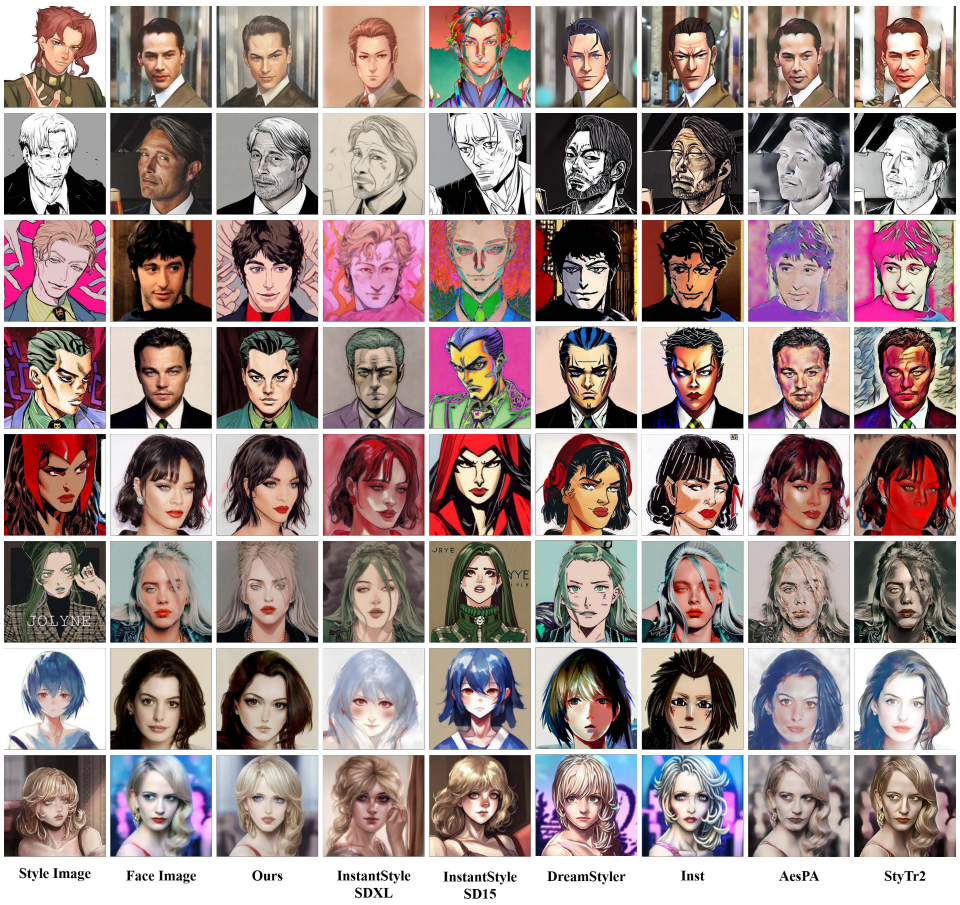

Figure 4: We experimented with different facial and stylistic images and compared our previous high-performance methods. It can be seen that our facial stylization results have a stronger style, while ensuring facial information while harmoniously and naturally integrating with the target style.

4.1 Implementation Details

We trained on a single Nvidia A6000 graphics card, and in the case of a single pair of images, we set the batch size to 1. Each epoch took about 30 seconds, and the overall training process is 5 epochs, which takes only 3 minutes. We set the base learning rate to $1.0e-04$, and the remaining hyperparameters are consistent with Stable Diffusion without changing. Through 40 steps of diffusion, our method can obtain stylized facial images with good results. We set the hyperparameters γ and β to 0.8 and 1.0, respectively, and all subsequent experiments will use this hyperparameter setting method.

4.2 Quality Analysis

We compare our method with the state-of-the-arts image style transfer methods, including InstantStyle [35], DreamStyler [10], Inst [11], AesPA [36], StyTr2 [37]. As shown in Figure 4, the generation results fully demonstrate the excellent performance of our method. Compared with the traditional two-column method on the far right, it can be seen that our method can modify the target image according to the high-level visual elements of the reference style image instead of traditional methods only focusing on low-level visual elements such as color or strokes. Compared to the DreamStyler and Inst inversion-based methods, it can be observed that both methods suffer from serious style degradation, which makes the stylized facial images they generate more sloppy and tend to generate simpler images. InstantStyle pursues fast generation due to its Tuning-Free mode. However, due to its use of ControlNets [38] to control the final generated face, the model only pieces together the

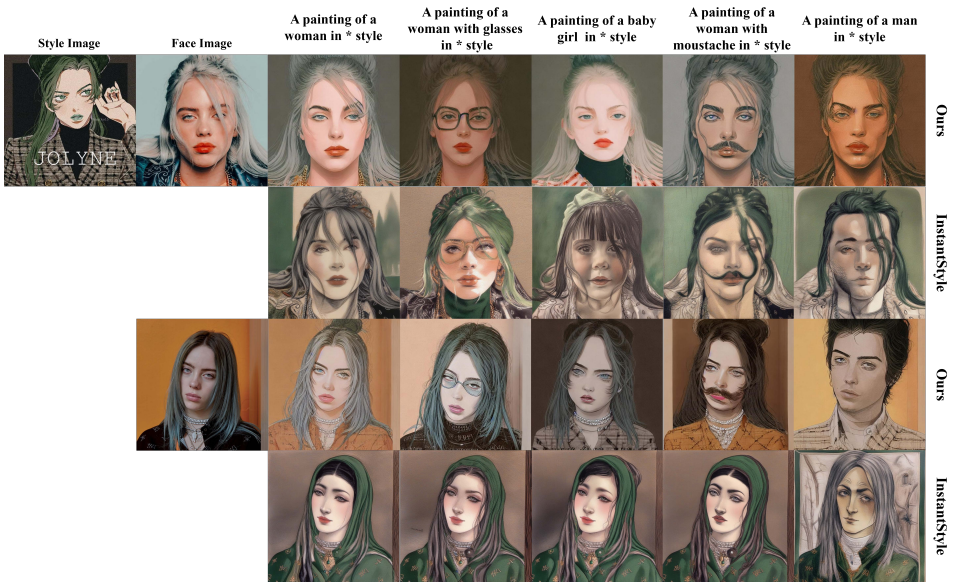

Figure 5: We have made more comparisons with the InstantStyle methods. Our method not only achieves good results in terms of generation quality but also has the ability to generate following text guidance.

elements in the style image according to the guidance by ControlNets rather than truly understanding the face. At the same time, due to the lack of new knowledge introduction, the generated results are not ideal. When using SDXL [21], the generated results also show some competitiveness. However, when using a smaller network like SD1.5, it can be seen that due to the insufficient prior knowledge to support its generation of input images of any style, the results are quite struggling and even generate a lot of noise in some images. On the contrary, although our method only uses SD1.4, a more basic pre-trained generative network, as our backbone model, the injection of new knowledge makes our generation results more effortless. It can be seen that our results faithfully restore facial image information while being close to the reference style image, displaying a highly artistic facial stylization result. We show more quality results and quantitative results in the supplement material.

4.3 Text-to-Image Generation

In addition to focusing on facial stylization generation, our method also has the ability to generate by text guidance since the Stable Diffusion we take as our backbone is originally a text-to-image generation mode. We compare our method with InstantStyle [35] as shown in Figure 5. It can be seen that, our method achieves better visual effects compared to Instantstyle. While ensuring that the style and facial features remain unchanged, we generated corresponding results based on the given text input and displayed good image quality. As Instantstyle is a method of tuning for free, it only relies on changing the attention parameter and subsequently using controllnet to control the final generation result. This also means that the Instantstyle method does not require additional knowledge injection, which makes it lack understanding of the given style and face. On the other hand, the presence of ControlNets can also interfere with the generation results of text participation guidance to some extent. As shown in the bottom line, Instantstyle hardly responds to the first four text inputs. Thanks to the injection of additional knowledge and the high compatibility between directly generating images from noise and text-to-image generation, our method can generate high-quality images that combine style, facial information, and fit text descriptions.

4.4 Ablation Study

We conducted ablation experiments on different modules to demonstrate how we extract the required features and how they work. As shown in Figure 6, we first do not use any CFun modules but only use abstract features from Afun for generation. It can be seen that the generated result is only a blurry

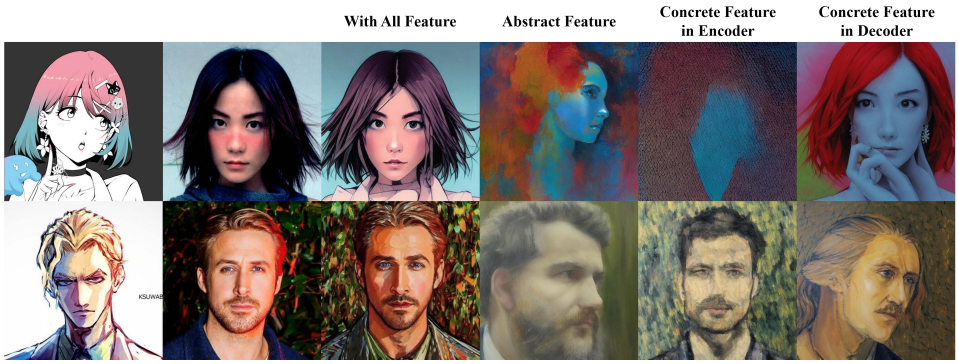

Figure 6: The ablation experiment demonstrated the role of our proposed abstract and concrete features, demonstrating the effectiveness of our proposed separation of learning abstract and concrete features.

image without specific facial information. However, at an abstract level, it can be seen whether it is male or female, and it also provides the colors required for the final generated image. By introducing Cfun into the encoder, it can be seen that the specific features of CFun provide low-level visual information, such as stroke texture. In the decoder, CFun can see that the main concrete features are the structure and posture of the face. The results of the ablation experiment indicate that the feature extraction mechanism is consistent with what we described in the methods section, proving the effectiveness of our method and the rationality of the proposed method. We show more results in our supplement material.

5 Conclusion

We propose a new facial stylization method called ACFun, which can achieve facial image stylization using only one image pair. To adapt to the task of facial stylization, we separate the learning of abstract features and concrete features through AFun and CFun and constrain the above process through a carefully designed Face and Style Imagery Alignment loss. We can transform any facial image into a specific style through short-term training. Numerous experiments have shown that our method exhibits excellent facial stylization and text-guided image generation results compared to state-of-the-art methods. A large number of visual results demonstrate that our method not only has good facial preservation ability but also generates stylized facial images with a strong artistic atmosphere. Meanwhile, the results of ablation research have demonstrated the effectiveness of our proposed method and demonstrated that our proposed AFun module and CFun module indeed learn abstract and concrete features separately. Our method provides excellent convenience for facial stylization generation.

6 Limitation

Even if we use separate learning methods, style is always difficult to describe, which often leads to some low-level or advanced visual features being deeply bound to certain content, resulting in inevitable style leakage problems. A typical example is that stylized facial images will change with the posture in the style image. Furthermore, due to the ambiguity of the style, our method is still unable to accurately extract a few visual elements of the style we want to exist in the style image. Moreover, our method relies on the effectiveness of the pre-trained model. We found that the training data used in the pre-trained model is imbalanced, especially when facing certain specific character images or style images. This can lead to serious semantic binding phenomena, making it difficult to stylize or causing obvious style leakage, which can also impact the performance of our method.

7 Acknowledgements

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), Joint Fund of Ministry of Education of China (8091B022149), National Natural Science Foundation of China (62132016, 62406238), and Natural Science Basic Research Program of Shaanxi (2020JC-23).

References

[1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[2] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

[3] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.

[4] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1501–1510, 2017.

[5] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019.

[6] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6649–6658, 2021.

[7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[10] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 674–681, 2024.

[11] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[13] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[14] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[15] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, August 2024.

[16] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5004–5013, 2024.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[22] Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman Khan, and Fahad Khan. How to continually adapt text-to-image diffusion models for flexible customization? In *Advances in Neural Information Processing Systems*, 2024.

[23] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023.

[24] Gan Sun, Wenqi Liang, Jiahua Dong, Jun Li, Zhengming Ding, and Yang Cong. Create your world: Lifelong text-to-image diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6454–6470, 2024.

[25] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[26] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[27] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021.

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[31] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021.

[32] Min Jin Chong and David Forsyth. Jojogan: One shot face stylization. In *European Conference on Computer Vision*, pages 128–152. Springer, 2022.

[33] Ziqi Zhang, Siduo Pan, Kun Wei, Jiapeng Ji, Xu Yang, and Cheng Deng. Few-shot generative model adaption via optimal kernel modulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[34] Siduo Pan, Ziqi Zhang, Kun Wei, Xu Yang, and Cheng Deng. Few-shot generative model adaptation via style-guided prompt. *IEEE Transactions on Multimedia*, 2024.

[35] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

[36] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22758–22767, 2023.

[37] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11326–11336, 2022.

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

A Quantitative Analysis

We also conducted quantitative experiments based on user study, shown in Table 1, providing 40 pairs of style facial images and their generated results, and collected voting results from 50 volunteers. The voting was conducted from three aspects: style consistency, facial consistency, and overall which image users prefer. The results are shown in the table. It can be seen that our method can achieve good results in style consistency and outstanding performance in facial consistency and overall evaluation. In addition, we also compared our method separately with Instantstyle SDXL, which is shown in Table 2, and it can be seen that users favor our approach.

Table 1: Quantitative Experiments based on user study.

	Ours	InstantStyle SDXL	InstantStyle SD1.5	DreamStyler	Inst
Style Consistency	29.17%	21.67%	**34.17%**	8.33%	6.67%
Face Consistency	**63.33%**	10.83%	6.67%	9.17%	10%
Overall	**54.17%**	10.83%	15.00%	10.83%	9.17%

Table 2: Separately Compared with Instantstyle SDXL

	Ours	InstantStyle SDXL
Style Consistency	**52.27%**	47.73%
Face Consistency	**63.64%**	36.36%
Overall	**71.59%**	28.41%

B Comparative Experimental Analysis

In this section, we conduct a comparative analysis of the prompts within the AFun module discussed in Section 3.2, the two experimental implements within the CFun module detailed in Section 3.3, and the hyperparameters of the Face and Style Imagery Alignment Loss as outlined in Section 3.4.

B.1 *Single pseudo-word Prompt* vs *Facial Description Prompt*

It can be seen in Figure 7 that the facial stylization results using facial description text prompts show better stability and generation quality, which also proves that facial description text prompts improve the accuracy of style and facial learning. Meanwhile, since facial description texts contain more information, we can obtain a better initial abstract feature, which reduces the difficulty of subsequent training. In addition, facial stylization results using only a single pseudo-word will always be affected by artifacts, especially in results with stronger styles. This further demonstrates the effectiveness of our facial description text prompts.

B.2 Hyperparameter Analysis of Imagery Alignment Loss

We also set different γ and β parameters, which are 0.6:1, 0.8:1, and 1:1. It is obviously shown in Figure 7 that as a continues to increase, the style information contained in its facial stylization results will become more dense. At 0.6:1, the style information is almost completely ineffective, and its stylization ability is extremely limited, only staying at color changes or facial shifts. When set to 1:1, dense style information is introduced into the facial image, and even style leakage may occur, such as turning hair into green. This also proves the effectiveness of our proposed Imagery Alignment Loss and the existence of our proposed Imagery Latent Space. We can indeed generate stylized facial images by finding a point in the Imagery Latent Space that lies between the style image and the facial image.

Figure 7: *Single pseudo-word Prompt* vs *Facial Description Prompt* and Hyperparameter Analysis of Imagery Alignment Loss

B.3 *Resblock* vs *Attention Block*

We mentioned our strategy of selecting the insertion parameter position n section 3.3. The experimental results in Figure 8 indicate that inserting training parameters into attention can lead to serious style leakage, resulting in overfitting the given style image. This is because the initial abstract features e_a we provide are obtained by fusing the CLIP encoding e_{i_s} of the style image I_s with the CLIP encoding e_{t_f} of the facial description text prompt t_f. This means that the initial abstract features e_a do not contain facial information. If we use the method of inserting parameters into attention, the style information in the abstract features e_a will be activated before the information in the facial image, leading to serious style leakage and overfitting.

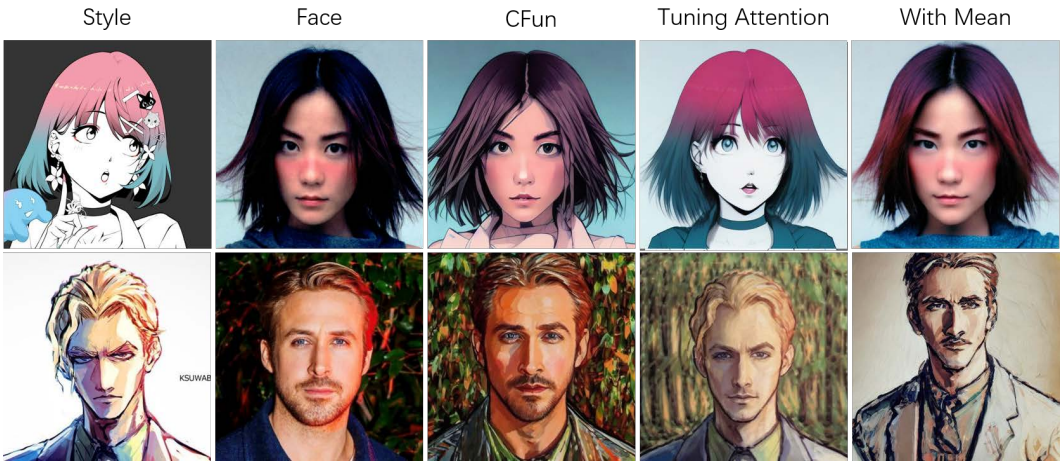

Figure 8: *Resblock* vs *Attention Block* and *Mean* + *Standard* vs *Only Standard*

B.4 *Mean* + *Standard* vs *Only Standard*

The use of Mean becomes even more complex, as shown in the figure. We found that the reverse direction of overfitting in the final result is uncertain, possibly due to the more complex information contained in the Mean and the higher degree of mixing of facial and style information. However, regardless, the excessive content information contained in Mean is not conducive to our facial stylization. By only using Standard, we can avoid this trouble and obtain more stable facial stylization results.

C More Stylization Results

In this section, we conduct an interesting experiment on splicing facial images and present an array of stylized experimental results, all of which demonstrate commendable efficacy.

C.1 Splicing Facial Stylization

We have designed an extremely interesting task to demonstrate the superiority of our proposed method and why we believe that separating abstract and concrete features for learning and finally mixing them into imagery features can effectively complete the task of facial stylization. We use this image to stylize faces by concatenating different parts of different faces into a single face. As a comparison, InstantStyle, due to the presence of ControlNet, directly generates a fragmented stylized facial image based on the cropped image, indicating that it does not understand what a face is. Our images can generate complete and meaningful stylized facial images while ensuring the features contained in each facial feature as much as possible and harmoniously integrating them into one face. This is all due to our approach of abstract, concrete, and imagery features, which makes our model adopt a more knowledge-driven approach rather than a simple feature transfer facial stylization method. This also indicates that even if the given image is fragmented, it does not prevent our method from finding a meaningful feature corresponding to it in the imagery latent space, which proves that our method has a strong abstraction ability.

C.2 More Quality Experimental Results

We conducted experiments on more faces and styles to demonstrate the stability and generalization of our method, and we compared it with InstantStyle-SDXL, which is the best-performing InstantStyle method for comparison. in Figure 10 and Figure 11. Even though we use SD1.4, which has a smaller CLIP and training on a smaller dataset than SDXL, our method appears more adept at facial stylization tasks due to the injection of additional knowledge. It can be seen that the stylized facial images generated by our method have a high artistic level, just like the creation of an artist. Of

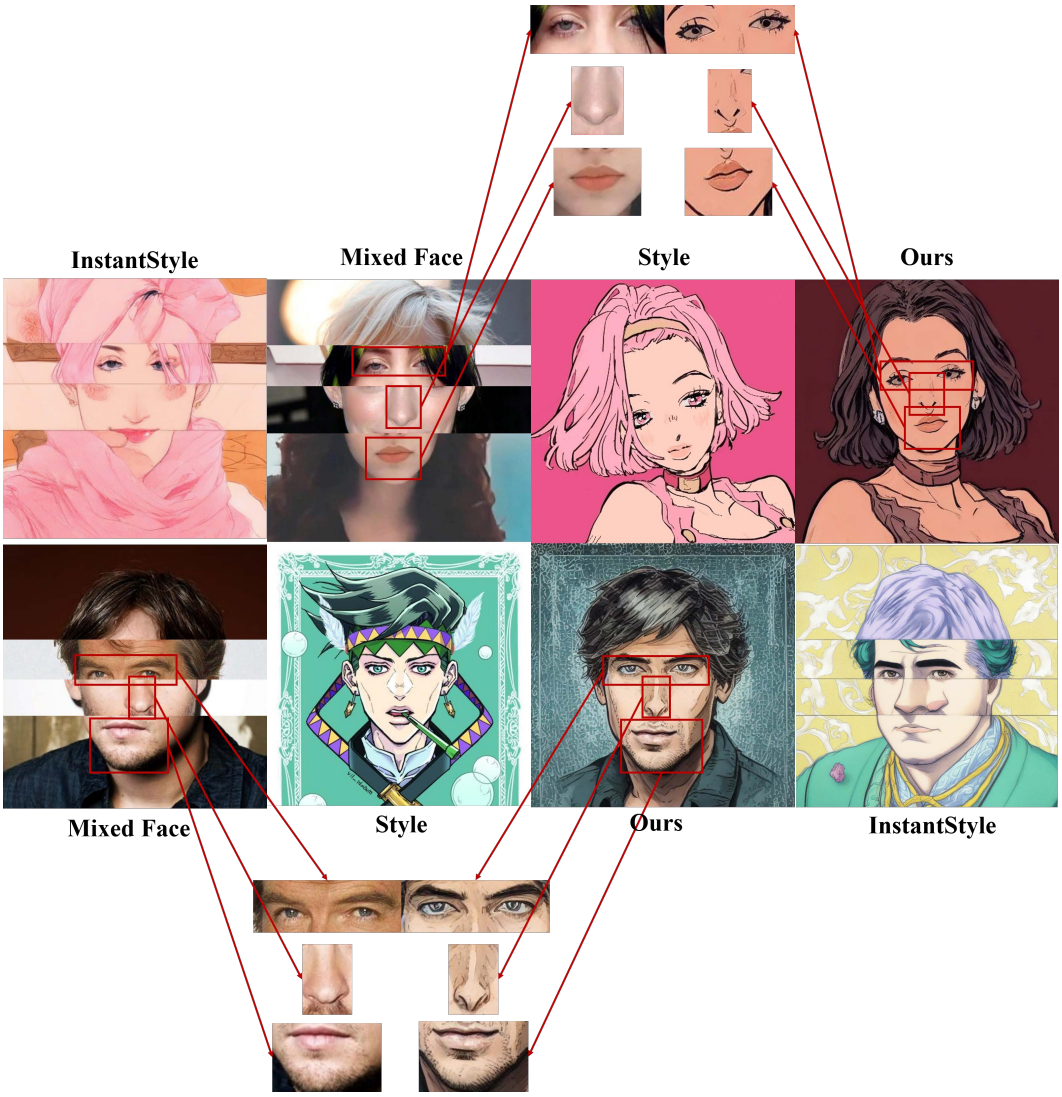

Figure 9: Splicing Facial Stylization

course, as the saying goes:' There are a thousand Hamlets in a thousand people's eyes', and people's understanding of images is subjective, triggering another reflection on me.

D Discussion

Is the Pre-trained Large Model a Natural Brain-like Mechanism? The fundamental idea of artificial intelligence is to simulate the operation of the human brain through artificial neurons. From a neurological perspective, human vision is not simply a peripheral sensory experience but a product of the coordination between the brain and the senses. So, the perceptual process of human "seeing" is not a simple bottom-up information transmission process but rather a top-down information construction process. The processing of facial information and style information by the human brain is unique. This also means that the action of "seeing" is not simply a visual drive but rather a knowledge-driven action.

The human visual mechanism is quite complex, especially when facing faces and artistic-style images. For faces, humans not only receive bioelectric signals from edge sensors but also activate specific brain regions and use more advanced knowledge from the hippocampus to recognize faces jointly. And the same goes for style. As a visual art experience, neuroaesthetics points out that art works

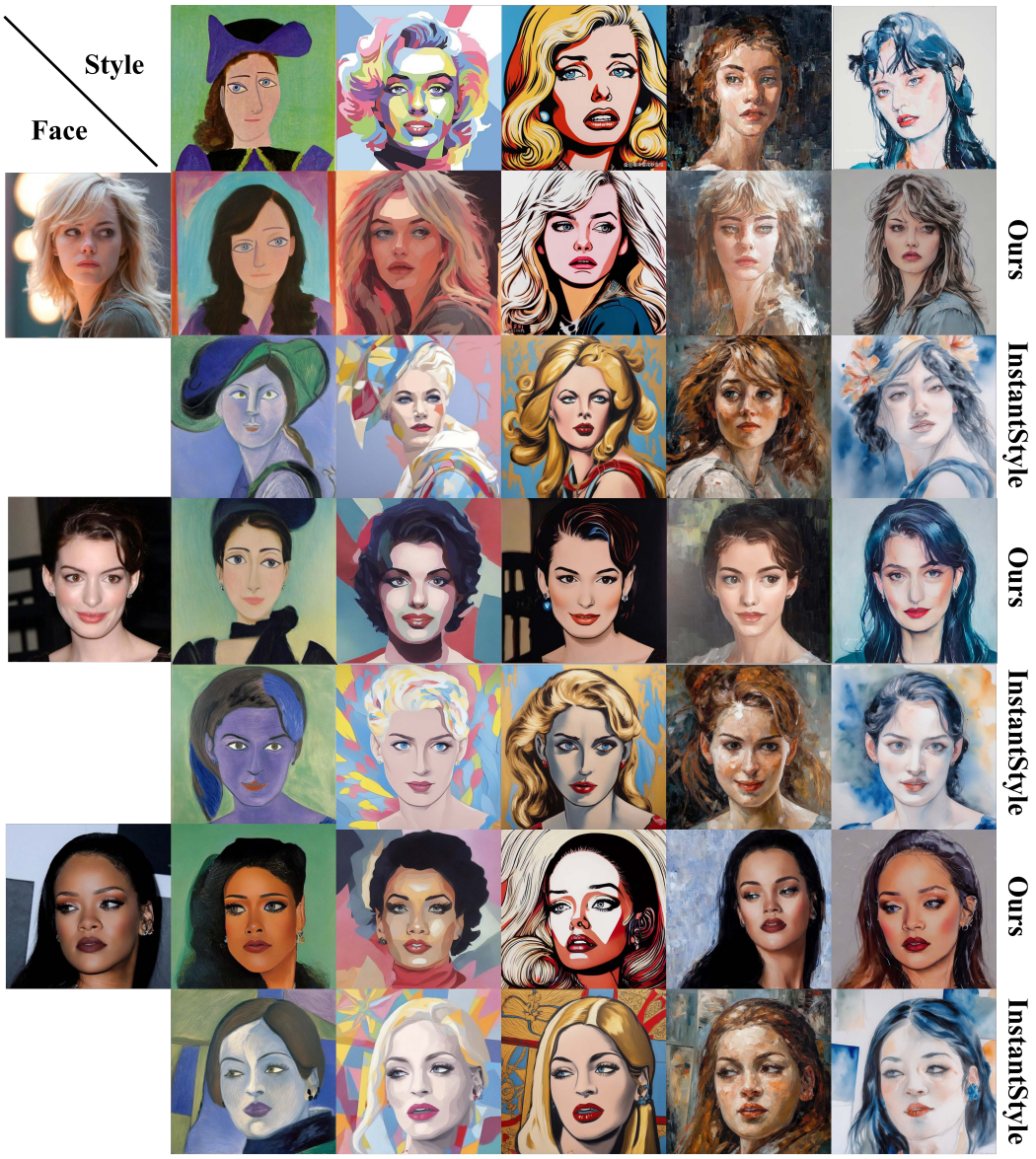

Figure 10: More Quality Experimental Results

will stimulate specific brain regions through visual stimulation, while this process still involves knowledge from the hippocampus. This is why "There are a thousand Hamlets in a thousand peoples eyes". Different people have different understandings and feelings towards the same artistic image, as it involves their past life, experiences, and educational level. Therefore, facial stylization is an extremely complex task which involves both algorithm design and human visual perception. This is why it is difficult to obtain a standard evaluation standard.

And does our method really possess mechanisms and abilities similar to the human brain? We can find a very interesting phenomenon. For the abstract and concrete features we learn, it is like the processing of visual information by the human brain. On the one hand, it is stimulated by edge sensors, and on the other hand, it is knowledge-driven cognition. In addition, concrete features also have vastly different characteristics in encoders and decoders, just like how the human left and right brains focus on different functions and tasks. However, the diffusion model is not a brain-like neural network designed concerning brain structures, which has also sparked our thinking about the relationship between large models and brain-like mechanisms.

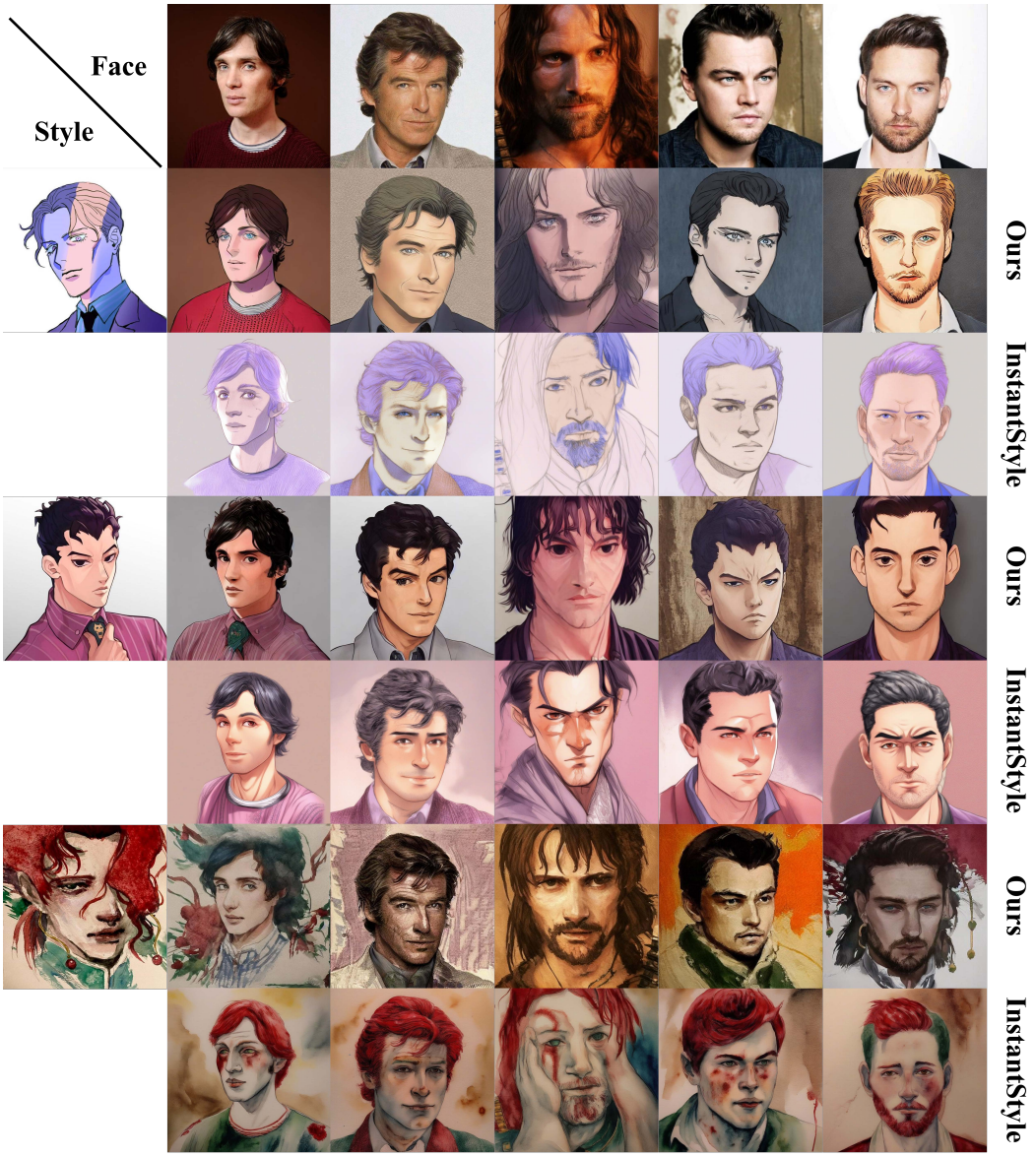

Figure 11: More Quality Experimental Results

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As in the Abstract, our claim is that we present an Abstract-Concrete Fusion Facial Stylization method, which is discussed in Section 3, and make contributions in Section 4

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided a detailed introduction to the implementation details of our experiment in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide open access to the data and code due to proprietary restrictions. However, detailed descriptions of the experimental setup and model architecture are provided in Sections 3 and 4, ensuring the experiments can be understood and replicated by researchers with similar resources.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details of the training and testing settings in Section 4, including data splits, chosen hyperparameters, and the type of optimizer used. This ensures that the experimental results are fully understandable and reproducible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported as previous attribute and affordance benchmarks.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 specifies the type of compute resources used, including the GPU models, memory configurations, and the time required for each experiment, ensuring that the experiments can be accurately reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 3 we cite including Stable Diffusion [20]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.