# GMM Estimator: A Precise Predictor of Style Latent Space

Lingcheng Kong† Junyu Liao† Weihao Nie†

## Abstract

*The generation of high-quality, style-specific images presents persistent challenges for diffusion-based models. While stable diffusion models have achieved impressive results in text-to-image synthesis, enforcing a consistent and nuanced artistic style remains difficult, often requiring costly fine-tuning or extensive prompt engineering. Recent work like Diffusion in Style has improved style transfer by adapting the initial latent tensor distribution, yet it simplifies the latent space by adopting a unimodal Gaussian prior, which limits the expressiveness of generated styles. In this work, we propose methodological advancements to overcome these limitations by capturing the multimodal nature of artistic styles. Our approach leverages advanced statistical techniques to better reflect the complex distribution of style-specific latent tensors. These improvements not only enhance stylistic coherence and visual quality but also maintain computational efficiency. Through this refined modeling of the latent space, we enable diffusion models to more accurately reproduce intricate style features, significantly advancing the fidelity and expressiveness of style-specific image generation. Our findings offer new insights into bridging statistical modeling and artistic generation, paving the way for more precise and efficient diffusion-based style transfer.*

## 1. Introduction

The generation of style-specific images, such as high-resolution Pokemon-style outputs, poses significant challenges for diffusion-based models. While stable diffusion models [23] have demonstrated remarkable capabilities for text-to-image synthesis, enforcing a coherent and nuanced artistic style remains difficult. Traditional approaches often require intensive computational resources and extensive fine-tuning [2, 19], yet still struggle to maintain a consistent style throughout the generated images [18, 28]. Practical evidence from methods like Text-To-Pokemon [18] shows persistent issues with background consistency despite extensive training. Parameter-efficient methods like LoRA [26] offer computational advantages but sacrifice style precision, while prompt engineering approaches [19] and gradient guidance techniques [2] face limitations in style control and generation speed respectively.

A key insight into these challenges is that the style of generated images is tightly coupled with the initial latent tensor, which is typically sampled from a simple Gaussian distribution. The recent work "Diffusion in Style" [3] advances this understanding by modifying the latent tensor distribution based on target style images, improving both fine-tuning efficiency and result quality. This approach demonstrates that adapting the initial latent distribution to better reflect style characteristics can significantly enhance style transfer outcomes without extensive computational resources.

However, the "Diffusion in Style" method introduces its own limitations by computing only the element-wise mean and variance of the VAE image encodings. This simplification assumes a unimodal Gaussian distribution, which fails to capture the multi-modal nature of artistic styles, where distinct sub-styles may coexist within a single artistic domain [7, 21]. These simplifications potentially discard important stylistic attributes that could contribute to a more coherent visual style [12, 16].

To address these challenges, We propose GMM Estimator, a novel method for adapting Stable Diffusion to a target style. The key idea behind Diffusion in Style is to start the denoising process with style-relevant initial latent tensors. We leverages the multi-modal distribution to model the latent encodings of a small set of target style images. This approach allows us to capture the multimodal nature of artistic styles, where multiple substyles may coexist within the same target style. The highlights of GMM estimator are:

(1)The GMM estimator requires only a small amount of images from the target style, typically 50 to 100. This paves the way for numerous practical applications where access to thousands of images in the desired style is not feasible.

(2)By applying our proposed adapted distribution, the model achieves significantly faster convergence compared to the original Diffusion in Style model, allowing for more efficient training and faster adaptation to the target style within the same amount of time.

We evaluate Diffusion in Style quantitatively and quali-

tatively, and compare it to Diffusion in Style and the original stable diffusion. Diffusion in Style consistently outputs better qualitative results than prior arts.

## 2. Related Work

### 2.1. Style Representation through Feature Statistics

Contemporary approaches to computational style analysis often leverage statistical properties of neural network features. This paradigm, introduced in neural style transfer [6], defines style as a statistical distribution over feature spaces, with key formulations including Gram matrices [6], feature correlations [14], and channel-wise statistics [11].

Adaptive instance normalization (AdaIN) [11] enables real-time style transfer by aligning channel-wise statistics of content and style features. Similarly, moment matching techniques [14] demonstrate that controlling first and second-order statistics effectively captures stylistic characteristics. These statistical representations are not only foundational to style transfer but also extend to related tasks like domain adaptation [23] and multi-modal generation [3].

In latent diffusion models, the statistical properties of latent representations critically influence the stylistic characteristics of generated outputs. However, existing approaches often simplify these properties, overlooking complex interdependencies that encode nuanced style features.

### 2.2. Textual Guidance Approaches

The simplest way to control style in text-to-image models is through explicit textual descriptions, such as artist names or aesthetic modifiers. While intuitive, this approach often struggles with precise style control, particularly for structural attributes like "white background" [1]. Even detailed prompts may fail to capture nuanced stylistic elements or produce consistent results across different contexts.

Textual Inversion [5] improves on this by learning new embeddings for specific visual concepts using a small set of example images. However, its reliance on frozen diffusion models limits its ability to fully capture complex or unique styles [24]. Recent work explores compositional prompt engineering [1], combining multiple style descriptors to enhance control, but challenges remain in accurately representing intricate styles.

### 2.3. Model Adaptation Techniques

Direct model adaptation, such as full fine-tuning, achieves high-quality style transfer but at the cost of extensive computational resources and large datasets [2, 19]. For example, Waifu Diffusion and OpenJourney required tens of thousands of training images and hundreds of thousands of iterations, making this approach impractical for many applications.

Parameter-efficient methods like LoRA [10] reduce computational overhead by introducing trainable low-rank matrices to specific layers while keeping the base model frozen. Although efficient, LoRA sometimes compromises stylistic fidelity, particularly for highly distinct styles [26]. Similarly, DreamBooth [24] adapts models using as few as 3-5 examples, but balancing style-specific characteristics with semantic generality remains a challenge.

### 2.4. Inference-Time Control Methods

Gradient guidance techniques steer the diffusion process toward target styles by leveraging auxiliary models such as classifiers or CLIP [17]. These methods adjust the predicted noise during denoising steps to align with desired style characteristics. While effective, they are computationally intensive, requiring additional forward and backward passes through auxiliary models.

Hybrid approaches that combine inference-time guidance with lightweight parameter adaptation offer a balance between efficiency and fidelity, presenting a promising direction for future research in style-controlled image generation.

### 2.5. Noise Distribution Adaptation

An alternative approach involves modifying the noise distribution used in the diffusion process. Standard diffusion models sample latent tensors from a zero-mean, unit-variance Gaussian distribution, which may not align well with specific artistic styles. "Diffusion in Style" [3] adapts the noise distribution by calculating element-wise mean and variance of latent tensors from target style images, significantly improving style fidelity with minimal computational resources.

However, this method ignores interdimensional correlations and assumes a unimodal Gaussian distribution, which fails to capture the multi-modal nature of artistic styles [8, 22]. Many styles encompass distinct sub-styles (e.g., different periods of an artist's work), which require more expressive statistical models.

To address these limitations, advanced techniques such as mixture models and covariance estimation can better represent the complex latent distributions of artistic styles. These methods maintain computational efficiency while significantly improving the fidelity and expressiveness of style-specific image generation.

## 3. Data

### 3.1. Dataset Description

The Few-Shot Pokemon dataset [15] is a curated collection of high-resolution images designed to facilitate the training and evaluation of generative models for style-specific image synthesis. The dataset contains 833 images

of Pokemon characters, each rendered in a consistent artistic style and with a clean white background. These characteristics make it particularly suitable for tasks requiring fine-grained style adaptation, such as diffusion-based models.

The dataset is structured to support few-shot learning scenarios, with a limited number of examples per class. This aligns with the practical need for style adaptation when only a small set of target images is available.

In our work, we will utilize this dataset to validate the effectiveness of our proposed method. Specifically, we adapt Stable Diffusion to the Pokemon style by estimating style-specific latent distributions from a subset of this dataset (e.g. 50 images).

### 3.2. Data Processing

Given that the dataset originally used in "Diffusion in Style" [3] included both images and corresponding textual descriptions, but has since been removed, we opted to use an alternative dataset consisting solely of images. To generate textual descriptions for these images, we employed the BLIP (Bootstrapped Language-Image Pretraining) model [13]. The BLIP model was used to generate captions for each image in the dataset, ensuring alignment with the required format for style transfer tasks.

During this process, we observed that some automatically generated captions were of suboptimal quality, particularly in terms of descriptive accuracy and stylistic relevance. To address this, we manually refined these captions to ensure consistency and improve alignment with the target "in style of Pokémon" format. The resulting dataset adheres to the metadata structure specified in the Hugging Face dataset documentation [4], which supports the integration of image data with corresponding textual annotations.

The processed dataset is accompanied by a `metadata.jsonl` file, wherein each entry specifies the image file name and its corresponding textual description. As an illustrative example, we present an image below alongside its associated textual metadata:

Figure 1. Visual example from the preprocessed dataset.

The textual annotation corresponding to this image is as follows: `"a brown pokemon with a skull on its head and holding a bone in style of pokemon"`

## 4. Methodology

### 4.1. Preliminary: Stable Diffusion Backbone

Our approach is built on the latent diffusion architecture of Stable Diffusion [23], which combines a pretrained VAE, a frozen text encoder (e.g. CLIP), and a U-Net denoiser. The VAE encoder compresses high-resolution images into a compact latent tensor at lower spatial resolution, dramatically reducing compute and memory costs. During training, a noise scheduler progressively corrupts these latents and the U-Net learns to reverse that corruption, conditioned on textual embeddings. Since all denoising happens in this smaller latent space and the VAE remains fixed, the model is both efficient and easily fine-tuned for new tasks.

### 4.2. Modeling the noise distribution with GMM

To overcome the limitations of element-wise mean and variance estimation in the original "Diffusion in Style" framework [3], where style representations are oversimplified, we replace its noise sampled from single–Gaussian distribution $\epsilon \sim \mathcal{N}(\mu_{style}, \Sigma_{style})$ with a richer distribution that faithfully reflects the true and often multimodal structure of the style latents. We achieve this by fitting a diagonal–covariance Gaussian Mixture Model (GMM) over the VAE encoder outputs.

#### 4.2.1 Motivation

In the context of the Pokémon style, there may exist some fine-grained substyles. For example, blue for water-type characters, and red for fire-type ones. Consider the original images in the training dataset. Each image contains one character with consistent style, such as a yellow Pikachu, or a blue Squirtle. These style clusters are challenging to capture with simple approaches (i.e., the unimodal distribution). However, the GMM effectively accommodates these variations, allowing for a more precise representation of the latent space and better capturing the diversity inherent in the dataset.

In the setting of a **Single Gaussian**, we sample $\epsilon$ and $\hat{z}_T$ from

$$\epsilon, \hat{z}_T \sim \mathcal{N}(\mu_{\text{style}}, \Sigma_{\text{style}})$$

The sample tends to fall near the global center of all Pokémon styles, which corresponds to a blurry average of yellow, blue, and green palettes. As a result, the latent representation does not align well with any specific real cluster in the dataset. Consequently, the fine-tuned UNet struggles to generate anything that clearly represents the "Pokémon-style," leading to outputs that lack distinct stylistic features.

In contrast, in the **Multimodal Gaussian** setting, we first sample a component $k$ with probability proportional to how common that style is in the dataset (e.g., $\pi_1 = 0.3$ for blue, $\pi_2 = 0.4$ for yellow, etc.). Then, we sample the noise and the latent tensor $\epsilon$ and $\hat{z}_T$ as follows:

$$\epsilon, \hat{z}_T \sim \mathcal{N}(\mu_k, \Sigma_k)$$

For example, if $k = 1$ (the blue region), the sampled latent tensor falls into the cluster corresponding to the "Water" region. Since the UNet has been fine-tuned on all possible clusters, it is able to denoise the latent tensor into a vivid water-type Pokémon-style image. This approach ensures that the generated image accurately reflects the specific substyle associated with the sampled cluster.

From the visualization of VAE-encoded latents of these images, as shown in Fig. 2, it is evident that the Gaussian Mixture Model (GMM) provides a more effective way to capture the substyles within the dataset. Unlike a unimodal distribution, which assumes homogeneity across the entire dataset, the GMM allows us to model the multimodal nature of the latent space, ensuring that subtle variations across different substyles are better represented.

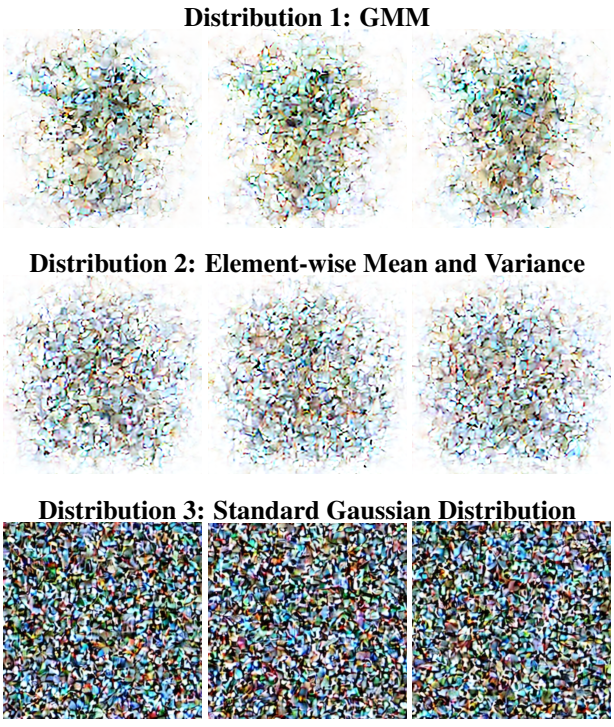


Figure 2. Samples from the different noise distributions. Each row represents a different distribution, with three examples generated using the corresponding method.

#### 4.2.2 Gaussian Mixture Model for Style Latents

To compute the style-adapted multimodal distribution, we first encode the images $i \in I_{\text{style}}$ of the target style with the VAE encoder, getting the latent tensors $\mathcal{E}(i) \in \mathbb{R}^d$. Then, we pick a relatively small number (e.g. 5, determined by cross-validation) for the value of $K$, given a few-shot setting. Based on the selected $K$, we the initialize the mean $\mu_k \in \mathbb{R}^d$ and variance $\sigma_k^2 \in \mathbb{R}^d$, as well as the weights $\pi_k$. Expectation-maximization (EM) is applied with a number of iterations until convergence. Specifically, within each iteration,

- **E-step**

$$r_{ik} = \frac{\pi_k \mathcal{N}\big(\mathcal{E}(i) \mid \mu_k, \text{diag}(\sigma_k^2)\big)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\big(\mathcal{E}(i) \mid \mu_j, \text{diag}(\sigma_j^2)\big)}$$

- **M-step**

$$\pi_k \leftarrow \frac{1}{N}\sum_{i=1}^{N} r_{ik}, \quad \mu_k \leftarrow \frac{\sum_{i=1}^{N} r_{ik}\,\mathcal{E}(i)}{\sum_{i=1}^{N} r_{ik}},$$

$$\sigma_k^2 \leftarrow \frac{\sum_{i=1}^{N} r_{ik}\,(\mathcal{E}(i) - \mu_k)^2}{\sum_{i=1}^{N} r_{ik}}$$

The final obtained multimodal distribution is described as

$$P(\hat{z}_T) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\hat{z}_T \mid \mu_k, \text{diag}(\sigma_k^2)\big).$$

As illustrated in Figure 2, our modified noise distribution better aligns with specific clusters, compared to the unimodal distribution.

### 4.3. Fine-tuning and Inference

In the fine-tuning process, we fine-tune the U-Net on the target style images with our proposed adapted noise distribution. We follow the regular training methods, except that the noise is now sampled from $\epsilon \sim \mathcal{N}\big(\mu_k, \text{diag}(\sigma_k^2)\big)$, where $k$ denotes the cluster selected randomly for each sample. To fine-tune the U-Net, image captions are required. In our setup, we use BLIP [14, 37] to generate these captions. The fine-tuning process follows the typical diffusion training scheme: for each image in a batch, we first encode it into a latent representation $z_0$ using the VAE encoder. We then sample a random diffusion timestep t and apply noise drawn from a Gaussian distribution $\mathcal{N}(\mu_{\text{style}}, \Sigma_{\text{style}})$ to generate a noisy version $z_t$, scaled according to the model's noise schedule $\bar{\alpha}_t$. The U-Net takes $z_t$, the timestep $t$, and the corresponding image caption as input, and predicts the noise component $\hat{\epsilon}$. Training minimizes the mean squared

error between $\hat{\epsilon}$ and the true noise $\epsilon$, updating the U-Net parameters accordingly.

For inference, we first randomly pick a $k$ from $\{1, 2, .., K\}$, then sample the initial latent tensor $\hat{z}_t$ from $\hat{z}_t \sim \mathcal{N}\big(\mu_k, \mathrm{diag}(\sigma_k^2)\big)$. The fine-tuned U-Net then gradually removes noise from the latent tensor. During this process, users can adjust parameters such as the text prompt and the guidance weight to influence the generated image.

# 5. Experiments and results

## 5.1. Convergence Rate

We chose fixed initial learning rate $\alpha = 10^{-4}$ with cosine learning rate decay to fine-tune on the model stable-diffusion-v1-5. In the first 100 iterations, the loss of baseline model decreased from ˜31 to ˜27, while the loss of our new model decreased from ˜31 to ˜22. In the first 400 iterations, the loss of baseline model decreased from ˜31 to ˜10, while the loss of out new model decreased from ˜31 to ˜0.1. This result illustrated that the GMM model gave a more precise distribution of the latent space of the set of style images than the distribution given by mean method.

## 5.2. Evaluation Metrics

We will evaluate our model following established methodologies for text-to-image models [25], employing Pareto curves of CLIP and FID scores across various guidance weights. This approach enables comprehensive assessment of the balance between content alignment and style fidelity.

- **CLIP Score:** The CLIP score will be measured using the ViT-B/32 model [20], quantifying the alignment between textual prompts and the generated images.
- **Normalized FID Score:** For style fidelity evaluation, we will compute the FID score [9] over features of an Inception model [27] trained on the Art-FID dataset [29], using target style images as reference. To enhance interpretability, we will normalize the FID scores against those of the original Stable Diffusion.
- **Few-shot Pokemon Style:** For the few-shot Pokemon style, we will compute the normalized FID using 833 images from the few-shot Pokemon dataset as reference. We plan to compare our method against several baselines, including the classical fine-tuning [18, 28], and LoRA-based fine-tuning [26]. We anticipate that our method will demonstrate superior performance in terms of style-matching while maintaining strong prompt-alignment. We will also conduct a comparative analysis against Text-To-Pokemon [18] approaches, which was fine-tuned for 15k steps on all 833 images of the few-shot Pokemon dataset [15].
- **User Study:** Additionally, we conducted a user study with 36 participants to assess subjective quality factors that may not be fully captured by automated metrics, similar to the approach described in the Diffusion in Style paper [3]. This will provide a holistic assessment of our method's improvements in generating visually pleasing and stylistically consistent images.

This quantitative framework will provide rigorous metrics for assessing our model's performance in both semantic adherence to prompts and faithfulness to target artistic styles.

## 5.3. Results

### 5.3.1 Quantitative Results

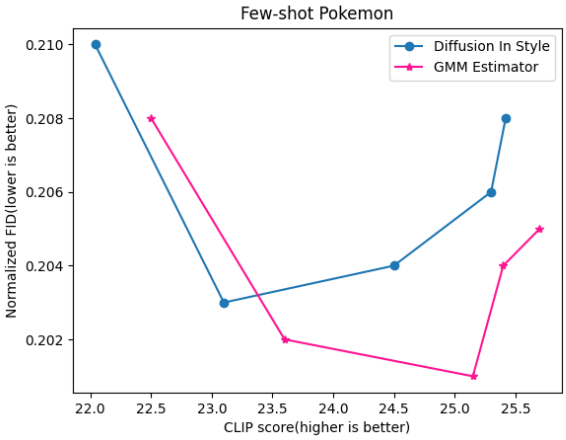


Figure 3. Curves of FID and CLIP scores along a range of guidance weights. Evaluation is performed with a range of guidance weights (6.0, 7.5, 8.0, 10.0, 15.0), leading to a curve for each model. For each point in the figure, all 833 images and text-prompt pair in Few-shot Pokemon dataset are tested to get the mean value.

*What the Metrics Mean:* FID measures how closely the generated images' feature distribution matches that of the real style images. A lower FID indicates that the overall "look and feel" (colors, textures, compositions) of the generated set is more faithful to the few-shot Pokémon style. Meanwhile, CLIP score measures semantic alignment between the text prompt and the generated image, using a pre-trained vision–language model. A higher CLIP score means the model is better at producing images that actually depict what the prompt describes.

Across most of the guidance weights, our GMM curve lies at the bottom-right of the baseline. That is, in most cases, we achieve a lower FID (better style match) at any given CLIP score, and at any given FID achieve an equal or higher CLIP score (better semantic accuracy). This indicates that by replacing the single Gaussian prior with a

GMM, we are capturing Pokémon's real style clusters more accurately across the full range of guidance weights.

#### 5.3.2 Generated images

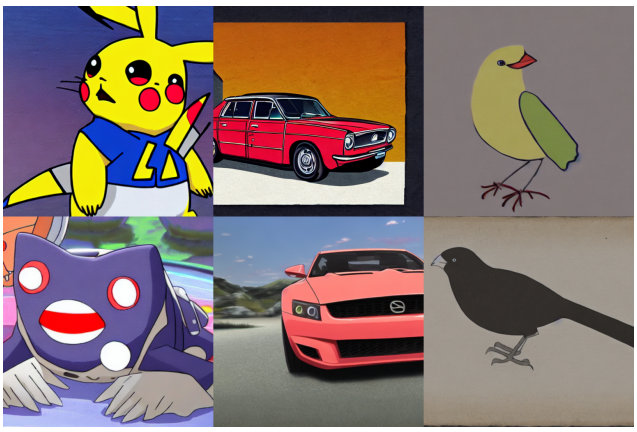

Figure 4. Results from our baseline model. The first line is the results after 100 epochs of training on the Pokemon dataset, while the second line is the results after 5000 epochs. The images demonstrate the model's ability to generate Pokemon-style imagery but also highlight limitations in consistency, background handling, and anatomical refinement.

Figure 5. Results of our proposed multi-modal method against simple mean-and-variance estimation. The first line shows our samples, while the second line contains the reproduction results of unimodal distribution method

All the prompts for the generated images shares the same format (e.g., "*a red car in the style of Pokemon*").

Our initial experiments establish an important baseline using the standard fine-tuning approach provided by the Hugging Face Diffusers framework [28]. We trained a conventional text-to-image diffusion model for 100, 5000 epochs respectively on the Pokemon dataset [15] without implementing our proposed mixture Gaussian estimation or advanced statistical modeling techniques.

As shown in Figure 4, the baseline model generates recognizable Pokemon-style imagery, such as a Pikachu character in a blue outfit (top), a vintage red car depicted in a simplified graphic style (middle), and a stylized bird with flat colors (bottom). In the second line of Figure 4, after additional training epochs, the generated images exhibit slightly improved alignment with the target style compared to those in the first line. These results demonstrate the fundamental capability of diffusion models to learn stylistic elements from a specialized dataset, but also highlight the limitations we aim to address.

Both baseline outputs exhibit several issues that validate our research direction. First, while the model captures some stylistic elements like flat coloring and simplified forms, it struggles with maintaining consistent stylization across different subjects. Second, we observe occasional artifacts and inconsistencies in background handling, particularly in the non-Pokemon subjects. Third, the anatomical proportions and details sometimes appear distorted or unrefined.

It's important to note that these baseline results come from conventional fine-tuning alone, without implementing the element-wise mean and variance estimation proposed in "Diffusion in Style" [3] or our advanced modeling technique. These observed limitations clearly demonstrate the inherent weaknesses of traditional fine-tuning approaches for stable diffusion models. While the baseline model successfully captures some stylistic elements, it struggles with consistency, background handling, and anatomical accuracy—issues that are common in standard fine-tuned models. These results provide a clear reference point that validate our research direction toward more sophisticated statistical modeling techniques.

The effect of adapting the initial latent tensor distribution is illustrated in Figure 5. The top row shows outputs generated with our multi-modal GMM prior, while the bottom row uses the conventional unimodal Gaussian prior. Comparing these to the baseline results in Figure 4, it is clear that tailoring the latent distribution to the target style produces far more faithful Pokémon-style images.

For example, consider the images of bird: even after 5000 epochs of standard fine-tuning, the baseline model still renders it in a generic comic style. In contrast, the model inovlving adapted distribution yields a bird with distinct Pokémon features—an inventive, character-driven silhouette and crisp, expressive eyes—demonstrating how modeling the latent distribution greatly enhances style precision.

Now focus on Figure 5 and compare our proposed method with the unimodal baseline. The top row (GMM prior) and bottom row (unimodal baseline) reveal some differences in consistency:

| Approach/Model | Style (%) | Content (%) |
|---|---|---|
| *GMM (ours)* | $75 \pm 5$ | $54 \pm 3$ |
| Diffusion in Style | $73 \pm 3$ | $56 \pm 4$ |
| Standard Gaussian Distrbution | $22 \pm 3$ | $58 \pm 4$ |

Table 1. **User-study results.** Percentages indicate how likely an image generated by the method is preferred over another. Users were asked to select images that best reproduce the reference style or align with the content described by the reference text.We obtained a total of 450 valid image pair comparisons, from 36 users (excluding rejected) on the campus.

- **Color uniformity:** The results of our GMM method has solid, even color fill with little gradient or noise, while the unimodal baseline contains patches of different colors, resulting a complicated mixture that is not aesthetically pleasing.

- **Outline consistency:** Our method outputs smooth curves, with shapes that matches described character. The baseline, on the other hand, exhibits irregular, broken edges.

Together, these details show that sampling from a multimodal GMM latent prior yields more precise, style-coherent images than a single mean-and-variance Gaussian.

#### 5.3.3 User study

We recruited 36 participants from our social circles, covering various majors, academic years, and genders, with an equal distribution of male and female participants. This diversity in background was intended to minimize subjective biases and provide a more comprehensive evaluation of the generated images.

For each participant, we randomly selected 100 pairs of images and asked them to evaluate two aspects of each pair: the style and content alignment of the generated images. Specifically, for each set of three images, participants were asked to rate how well the images matched the given prompts in terms of style and content. The evaluation for each aspect was binary, with participants choosing between "Very Aligned" and "Not Aligned."

To ensure fairness, participants were provided with the textual prompts but were not informed about which model generated the images. This was done to eliminate any potential bias.

After collecting and analyzing the responses, we derived the conclusions presented in table 1, which shows our user-study results. The results of the user study seem to agree with the reported CLIP/FID scores.

## 6. Future work

In future work, we aim to understand why the original latent distribution $\mathcal{N}(0_d, I_{d\times d})$ limits effective style adaptation. Our current method uses a Gaussian Mixture Model (GMM) to fit the latent space, capturing the multimodal nature of artistic styles and improving style fidelity. However, the number of substyles directly impacts performance and computational cost. We plan to explore how fine-grained the substyle segmentation needs to be to balance efficiency and quality, and whether increasing substyles always improves results or reaches a point of diminishing returns. This could help reduce reliance on fine-tuning while maintaining high-quality style transfer.

Moreover, we noticed that current outputs still suffer from blurry backgrounds and sometimes uneven, over-complex shapes. Given that our current prompts are relatively simple (e.g., "*a red car in the style of Pokemon*"), we can further refine outputs through advanced prompt engineering. For example, we could apply negative prompting by incorporate "no blur," "clean background," or "no texture noise" to explicitly suppress unwanted artifacts. We could also use some multi-attribute templates, such as structure prompts with multiple descriptors (e.g., "a red car with smooth contours, flat cell-shaded colors, single-tone background, in the style of Pokémon") to guide shape and background. These strategies could potentially promise cleaner backgrounds, sharper shapes, and even stronger adherence to the target style, without modifying the diffusion backbone.

## 7. Conclusion

In this work, we present an innovative method for adapting the style of Stable Diffusion. Our method models style-specific latent distributions by applying a Gaussian Mixture Model (GMM) to the VAE encodings of a small set of target style images. Stable Diffusion is subsequently fine-tuned to work seamlessly with these newly modeled style-specific latent tensors, enabling the creation of images in the desired style. Our proposed GMM prior excels at generating a wide variety of objects, even when such objects are not present in the target style images, and achieves superior results compared to existing techniques. Through both qualitative and quantitative evaluations, we demonstrate its effectiveness. This efficient and rapid approach for style adaptation greatly expands the practical use cases of Stable Diffusion in real-world applications.

## References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2022. 2

[2] Cjwbw. Waifu diffusion — replicate repository. https://replicate.com/cjwbw/waifu-diffusion, 2022. Accessed on September 2022. 1, 2

[3] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine S"usstrunk, and Radhakrishna Achanta. Diffusion in style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2251–2259, October 2023. 1, 2, 3, 5, 6

[4] Hugging Face. Hugging face datasets: Load image data with metadata. https://huggingface.co/docs/datasets/v2.4.0/en/image_load#imagefolder-with-metadata, 2023. 3

[5] 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. In *International Conference on Learning Representations*, 2022. 2

[6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 2

[7] Alex Graves, Jacob Menick, and Aäron van den Oord. Associative compression networks for representation learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 1

[8] Alex Graves, Jacob Menick, and Aäron van den Oord. Associative compression networks for representation learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, volume 30, 2017. 5

[10] 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. In *International Conference on Learning Representations*, 2022. 2

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 2

[12] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 1

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapped language-image pretraining. https://github.com/salesforce/BLIP, 2022. 3

[14] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2230–2236, 2017. 2

[15] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Few-shot-pokemon — hugging face dataset. https://huggingface.co/datasets/huggan/few-shot-pokemon, 2022. 2, 5, 6

[16] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2021. 1

[17] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2021. 2

[18] Justin Pinkney. Text-to-pokemon — replicate repository. https://replicate.com/lambdal/text-to-pokemon, 2022. 1, 5

[19] PromptHero. Openjourney v4. https://huggingface.co/prompthero/openjourney-v4, 2023. Accessed on April 2025. 1, 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[21] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 1

[22] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, volume 34, 2021. 2

[23] 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 (CVPR)*, pages 10684–10695, June 2022. 1, 2, 3

[24] 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 *CVPR*, 2023. 2

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 5

[26] Paul Sayak. Pokemon lora — hugging face repository. https://huggingface.co/sayakpaul/sd-model-finetuned-lora-t4, 2023. 1, 2, 5

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[28] Patrick von Platen, Suraj Patil, Anton Lozhkov, Kashif Bertrand, Teven Le Scao, Sylvain Gugger, Lysandre Labrak, Sam Shleifer, Amanpreet McMillan-Major, Younes Belkada, Mitchell Gordon, Nathan Lambert, Leandro Chau, Tiberius Fischer, Sourab Dey, Victor Sanh, Louis Clément, Yoav Saatci, Shamik Shirzad, Thomas Marion, Luke Thompson,

and Daniel Dunbar. Diffusers: Text-to-image generation example. https://github.com/huggingface/diffusers/tree/main/examples/text_to_image, 2023. Hugging Face Diffusers. 1, 5, 6

[29] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 560–576. Springer, 2022. 5