Seamless Dual-Style Image Generation with Diffusion Models

Gabriel Mercier [*1] **Adrien Goldszal** [*1]

Abstract

One little explored frontier of image generation is blending two different styles seamlessly within a single output. We present two methods based on latent diffusion models that outperform simple text-based prompting: noise spatial interpolation and attention weight interpolation between two style prompts. Project code can be found at https://github.com/adriengoldszal/dual-style-image-gen

1. Introduction & Research setting

1.1. State of the art

Latent diffusion models have greatly advanced image synthesis quality, especially through iterative denoising processes (Rombach, 2022; Song et al., 2020). These models, such as Stable Diffusion, enable high-quality image generation and the ability to control the output through methods like classifier-free guidance, which allows for fine-tuned and diverse image generation.

Building upon latent diffusion models, numerous methods have been developed to tackle image editing and style transfer tasks. Some approaches, like the deterministic ODE-based settings, leverage the reversible nature of the process to facilitate easier image editing. In contrast, models such as *Prompt-to-Prompt* (Hertz et al., 2022) retain the stochastic properties of diffusion models, utilizing techniques like freezing and modifying the cross-attention layers to perform style transfer and manipulation.

One notable model in this area is **CycleDiffusion** (Wu & la Torre, 2022), a state-of-the-art stochastic diffusion model for image editing and style transfer. CycleDiffusion uses the noise trajectory from the reverse diffusion process back to the original image to guide denoising under a new conditioning, preserving the structure of an image while allowing for content modification.

[*]Equal contribution [1]Ecole Polytechnique, France. Correspondence to: Gabriel Mercier <gabriel.mercier@polytechnique.edu>, Adrien Goldszal <adrien.goldszal@polytechnique.edu>.

Interpolating between two input images or styles has been explored in previous works (Wang & Golland, 2023), particularly focusing on latent space interpolation. However, smoothly blending styles within a single image is a poorly researched subject and existing methods often face challenges in this task of achieving a seamless fusion of styles while maintaining both visual coherence and distinct stylistic elements within one output.

1.2. Our contribution

Limitations of Textual Guidance. When using a single text prompt specifying the spatial arrangement of two distinct styles, results are very poor. The method fails to produce meaningful results and generates artifacts and inconsistent textures, as described in the following image.1

Figure 1. Example outputs specifying a Van Gogh style on the left and a Minecraft style on the right.
From left to right:
(a) One style dominates (Minecraft)
(b) Irregular style separation
(c) Incoherent style fusion.

Our aim. Leveraging CycleDiffusion's style transfer framework, we aim to achieve a **smooth fusion** of styles, ensuring a continuous and visually coherent transition across different regions of the image. We present two methods for this task: one based on noise spatial interpolation and another using cross-attention weight interpolation, both designed to blend the styles seamlessly. We focus on a transition on the horizontal axis between two styles, on the right and the left, but this method can be adapted to different situations and with more than two styles. This approach opens new possibilities for artistic creation, domain adaptation, and mixed-style rendering.

2. Proposed methodology

2.1. Noise Merging for Dual-Style Diffusion Guidance

During the standard denoising process of DDIM (Song et al., 2020), the noise term is generated at each step using the trained U-Net neural network architecture. By leveraging the classifier-free guidance formulation (Ho & Salimans, 2021), textual guidance can be incorporated into the denoising process as follows:

$$\hat{\epsilon}_\theta(x_t, t, c_i) = \epsilon_\theta(x_t, t, \varnothing) + s\Big(\epsilon_\theta(x_t, t, c_i) - \epsilon_\theta(x_t, t, \varnothing)\Big)$$

where $\epsilon_\theta(x_t, t, \varnothing)$ is the unconditional prediction, $\epsilon_\theta(x_t, t, c_i)$ is the text-conditioned prediction, and s is the guidance weight.

In our approach, we propose generating two separate noise terms $\hat{\epsilon}_\theta(x_t, t, c_1)$ and $\hat{\epsilon}_\theta(x_t, t, c_2)$ at each denoising step—one for each style. These are then merged using a spatial mask $M(x)$:

$$\hat{\epsilon}_\theta(x_t, t, c_1, c_2) = M(x)\cdot\hat{\epsilon}_\theta(x_t, t, c_1) + (1 - M(x))\cdot\hat{\epsilon}_\theta(x_t, t, c_2)$$

Then, we integrate this merged noise into the original denoising equation:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_\theta(x_t, t, c_1, c_2)\right)$$

and the update step for x_{t-1} becomes:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\,\hat{\epsilon}_\theta(x_t, t, c_1, c_2) + \sigma_t\, z$$

where:

$$\sigma_t = \eta\sqrt{\frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t-1}}}\sqrt{1 - \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}, \quad z \sim \mathcal{N}(0, I)$$

By merging the noise terms before denoising, we ensure that each region of the image receives the correct style influence at every step of the diffusion process. This theoretically allows for a transition between styles while maintaining the structure and coherence of the image.

2.2. Cross-attention weight merging for Dual-Style Diffusion Guidance

Stable-diffusion, on which CyleDiffusion is based, leverages cross-attention between prompt tokens and the image to guide the noise prediction in it's U-Net architecture. More specifically, cross-attention helps learn "which parts of the image to modify" by attending more to certain pixels or regions. By changing the cross-attention architecture, we can apply cross-attention on two separate conditionings c_1 and c_2 and interpolate using the same spatial mask $M(x)$ as follows :

$$\text{Attention}_1(Q, K_1, V_1) = \text{softmax}\left(\frac{QK_1^T}{\sqrt{d_k}}\right)V_1$$

$$\text{Attention}_2(Q, K_2, V_2) = \text{softmax}\left(\frac{QK_2^T}{\sqrt{d_k}}\right)V_2$$

Then, these two attention results are merged using the spatial mask $M(x)$, which interpolates between them:

$$\text{Merged Output} = M(x)\cdot\text{Attention}_1 + (1 - M(x))\cdot\text{Attention}_2$$

By doing so, we push certain parts of the image to attend more to prompt tokens of c_1 and others to prompt tokens of c_2, helping create a smooth style fusion on the image between the two prompts.

3. Experimental Setting

We utilized the default architecture provided by the **CycleDiffusion** codebase, selecting **Stable Diffusion v1.4** from the Hugging Face library as our base model. The experiments were conducted with the following hyperparameter settings:

- $\eta = 0.1$
- encoder guidance scale $s_{\text{encoder}} = 1$
- decoder guidance scale $s_{\text{decoder}} = 20$
- number of denoising steps = 100
- skip steps = 10
- upsampling temperature = 1

For each method, we generated **7 images** using samples from the dataset provided by the CycleDiffusion paper and repository.

3.1. Metrics

To thoroughly test and validate our two style fusion methods, we not only qualitatively analyze the resulting images, but also compare **7 metrics** which measure **pixel, feature and semantic clip-related** elements. Some metrics focus on the image reconstruction quality by comparing the original and generated images, others focus on image quality, and finally, some metrics focus on the styles generated and its distribution over the generated images. 3 2 1

Similarity-based metrics	
Metric	**Description and Formula**
L2 Distance	Measures the euclidean distance between the generated image and the original image.
LPIPS (Learned Perceptual Image Patch Similarity)	Uses deep neural network feature maps (e.g., VGG19) to compute perceptual similarity. Unlike L2, LPIPS captures high-level structural and semantic information.

Table 1. Similarity / Reconstruction Metrics

Image quality metrics	
Metric	**Description and Formula**
SSIM (Structural Similarity Index)	Evaluates perceptual similarity by comparing luminance, contrast, and structural details between two images.
PSNR (Peak Signal-to-Noise Ratio)	Quantifies how much noise or distortion has been introduced in the generated image compared to the original.
Smoothness (Total Variation)	Measures the spatial smoothness of an image by computing the total variation.

Table 2. Image quality Metrics

CLIP-Based metrics	
Metric	**Description and Formula**
CLIP Similarity	Measures the semantic alignment between an image and a textual prompt using CLIP embeddings.
Directional CLIP	Measures how well the semantic changes between the original and generated images align with an expected direction in the latent space.

Table 3. CLIP-Based metrics

Remark : To better measure the style transfer capabilities of our two techniques, a **Gram Matrix based metric** was implemented, comparing the distance between the output of reference images for each style through the convolutional layers of VGG19 and the generated images. For each style, 3 images were used. However, this seemed to not be enough and the results were too similar and did not add to our analysis. It would have been interesting to test with a wider, and more diverse variety of reference images to more acurately judge the style transfer.

3.2. Testing setup

We compare our 2 methods with **two baselines** :

- **Pixel by pixel** merge of two images, which are outputs of the diffusion model for the two prompts separately.
- **One Prompt** only specifying the spatial arrangement as described above.

We thoroughly test these methods in different scenarios :

- **Spatial Mask** We test two different spatial masks representing a Linear and a Logistic (with a sharpness of 20) interpolation.

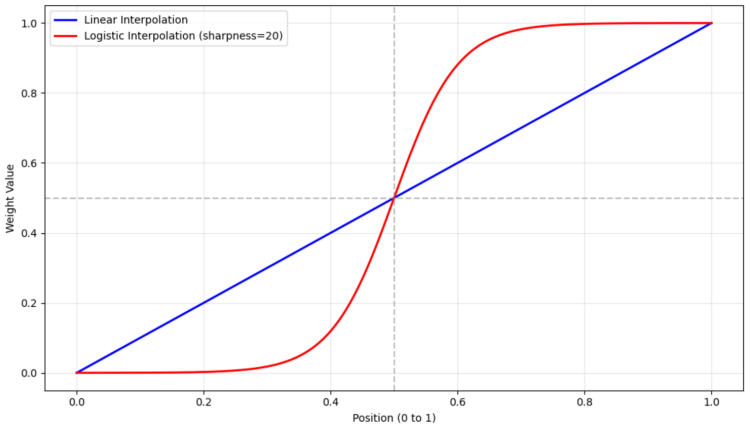

Figure 2. Comparison of linear and logistic spatial masks

- **Styles studied** Two style pairs are tested : Minecraft / Van Gogh and Mosaic / Andy Warhol Pop Art.

Fixed prompts and parameters are defined for even comparison between methods. The prompts are as follows :

- MINECRAFT : A Minecraft-inspired rendering of [prompt], featuring distinct pixelated textures, blocky 3D cube structures, limited color palette with no gradients, sharp right angles and perfect squares, characteristic voxel-based terrain with visible block edges.
- VAN GOGH : A Van Gogh-style painting of [prompt], with bold, swirling brushstrokes, rich textures, and vibrant, expressive colors reminiscent of Starry Night
- MOSAIC : A highly detailed mosaic of [prompt], made of small, colorful tiles with visible grout lines, creating a textured and handcrafted appearance
- ANDY WARHOL POP : A vibrant Andy Warhol-style pop art image of [prompt], featuring bold, contrasting colors, high saturation, thick outlines, and a repeated or silkscreen-like print effect.

4. Qualitative Results

In nearly all our examples, qualitative analysis clearly demonstrates that the epsilon interpolation method produces superior results, with both styles distinctly visible. In contrast, the cross-attention interpolation method sometimes causes one of the styles to dominate in the image. See table 4

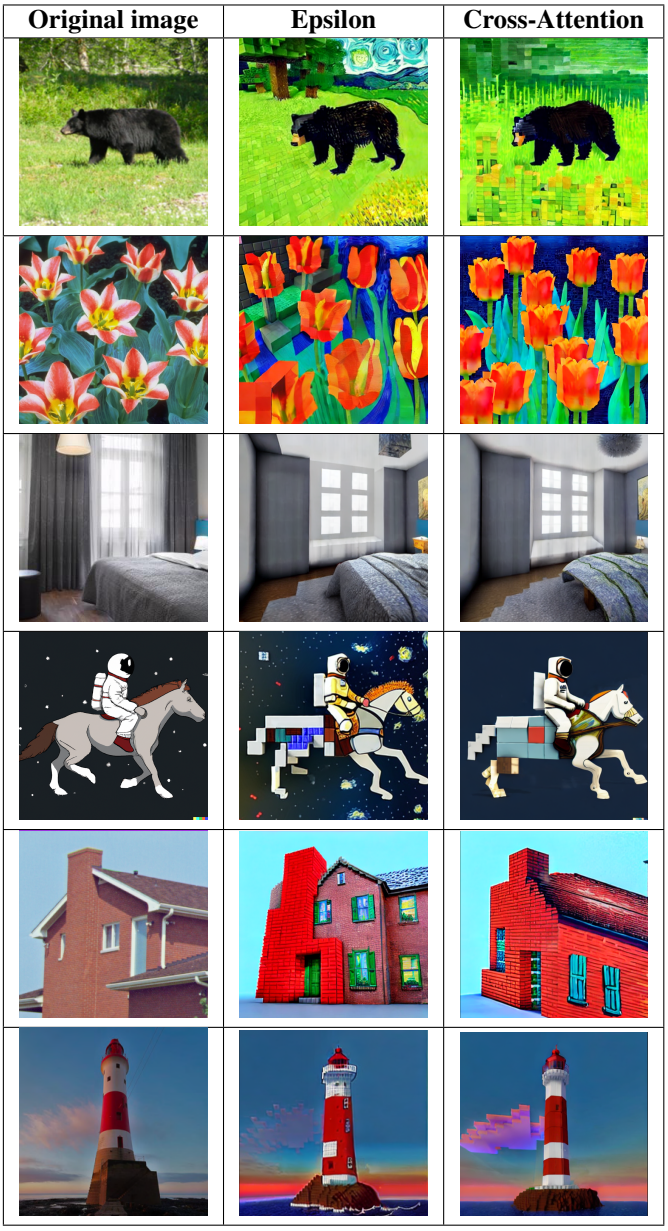

Table 4. Comparison of style fusion results using Epsilon Interpolation and Cross-Attention Interpolation.

5. Quantitative Results

Metric	Linear			2 prompts
	Cross Attn	Epsilon	Pixel	
Similarity				
L2	200±14	214±15	**164±11**	198±12
LPIPS	0,57±0,02	0,59±0,03	**0,54±0,02**	0,58±0,02
Image Quality				
PSNR	13,2±0,6	12,7±0,8	**14,9±0,6**	13,3±0,6
Smooth.	0,48±0,09	1,01±0,29	0,33±0,09	**0,26±0,12**
SSIM	0,48±0,04	**0,44±0,04**	0,53±0,04	0,50±0,04
CLIP				
CLIP	0,32±0,01	0,34±0,02	0,35±0,01	**0,38±0,01**
CLIP r.	0,31±0,01	0,31±0,01	**0,34±0,01**	0,24±0,01
CLIP l.	0,34±0,01	0,34±0,02	0,32±0,01	0,25±0,01
DCLIP	0,10±0,01	0,13±0,02	0,14±0,02	**0,19±0,02**
DCLIP r.	0,09±0,02	0,09±0,01	**0,12±0,01**	0,11±0,01
DCLIP l.	0,11±0,01	**0,13±0,01**	0,13±0,02	0,10±0,01

Table 5. Comparison of different style fusion approaches using various metrics for the linear interpolation. Best results in **bold**, our approach highlighted in yellow.

Metric	Logistic		
	Cross Attn	Epsilon	Pixel
Similarity			
L2	196±15	208±14	**181±11**
LPIPS	0,55±0,03	0,59±0,03	**0,54±0,02**
Image Quality			
PSNR	13,6±0,8	12,9±0,8	**14,1±0,6**
Smooth.	0,41±0,10	0,92±0,25	0,39±0,10
SSIM	0,49±0,05	0,45±0,04	**0,51±0,04**
CLIP			
CLIP	0,32±0,01	0,34±0,02	0,35±0,01
CLIP r.	0,31±0,01	0,31±0,02	**0,34±0,01**
CLIP l.	0,33±0,01	0,34±0,02	0,32±0,01
DCLIP	0,10±0,02	0,12±0,01	**0,14±0,02**
DCLIP r.	0,09±0,02	0,10±0,01	**0,13±0,01**
DCLIP l.	0,11±0,01	**0,13±0,01**	0,12±0,02

Table 6. Comparison of different style fusion approaches using various metrics for logistic interpolation (sharpness 20). Best results in **bold**, our approach highlighted in yellow.

As expected, **pixel interpolation** of images from a state-of-the-art model **outperforms nearly all metrics**. **Cross-attention** method achieves **superior image reconstruction**, as evidenced by higher similarity and image quality metrics. **CLIP measurements** highlight the dominance of the **epsilon method** in style fusion showing both styles have better results.

Remarks on Cross-Attention Interpolation & Results

This discrepancy could be explained by the fact that cross-attention is only one of the many layers in the U-Net. Other components may **smooth out** the interpolation effect, reducing the visibility of both styles unless additional guidance is applied to reinforce stylistic separation.

In fact, strengthening the guidance with a higher guidance scale, and increasing the sharpness often produce better results with the cross-attention. Further testing is certainly required to analyse this behaviour more in depth.

Figure 3. Minecraft / Van Gogh Cross-attention interpolation with sharpness 70 & decoder guidance of 30

In addition, tried doing **multiple passes through the cross-attention** instead of a single one, by passing the resulting attention weights back into the function, with the idea of strenghtening the style transfer. However, this seemed to only accentuate the artifacts on the image and smoothen it out more, with a full two passes rendering the final image completely blurry. A weighted sum of multiple pass outputs was done to mitigate this, but the style transfer didn't appear qualitatively better so the single pass method was kept.

6. Conclusion & Future Work

In this work, we present two novel methods for dual style fusion on images using diffusion models, allowing for smooth interpolation between two distinct styles on a single image. By leveraging the U-Net architecture, we apply epsilon interpolation, which interpolates noise generated from two prompts, and cross-attention interpolation, which operates on cross-attention weights. Our results demonstrate that these methods outperform traditional text-based prompting, providing more refined and accurate style fusion.

Further work should focus on a more in-depth evaluation of these methods, using a larger dataset of images and styles to perform a more comprehensive comparison. The uncertainties in some of our data suggest that additional testing is needed to fully understand the performance across different metrics. It would also be valuable to explore style-transfer quality by revisiting the Gram Matrix technique with a broader variety of images for each style.

Moreover, future research could investigate improved ways to leverage cross-attention layers for style transfer, particularly through multiple passes, to further refine the transfer process and compare the optimal parameters with epsilon interpolation. Finally, exploring direct utilization of better style embeddings, drawing inspiration from recent research (Li et al., 2024), could lead to improved results in style fusion.

References

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv*, 2021.

Li, W., Fang, M., Zou, C., Gong, B., Zheng, R., Wang, M., Chen, J., and Yang, M. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 123–140, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-73390-1_7.

Rombach, R. e. a. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv*, 2020.

Wang, C. J. and Golland, P. Interpolating between images with diffusion models. *arXiv preprint*, 2023.

Wu, C. H. and la Torre, F. D. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance, 2022. URL https://arxiv.org/abs/2210.05559.