Guided by Style: Fine-Grained Modulation in Multi-Style Artistic Transfer

Catherine Zhang
Stanford University
450 Jane Stanford Way
czhang7@stanford.edu

Christina Ba
Stanford University
450 Jane Stanford Way
cba1@stanford.edu

Abstract

We propose a novel diffusion-based framework for artistic multi-style transfer that uniquely combines compositional denoising and classifier-free guidance (CFG) to enable fine-grained control over both content preservation and stylistic blending. Building on a pretrained Stable Diffusion model, our method introduces a principled way to modulate the influence of multiple text-based artistic prompts during the denoising process, allowing users to adjust the relative strength of each style in a controllable and interpretable manner.

We evaluate our approach through both quantitative perceptual similarity metrics (LPIPS) and qualitative evaluations from human- and LLM-based evaluations. We find that a noise strength of 0.47 and a CFG scale of 2.4 offer an effective balance between stylization and content fidelity. Compared to a baseline diffusion method (InstructPix2Pix), our system demonstrates superior style alignment, flexibility, and responsiveness to user-specified guidance, highlighting the expressive potential of compositional diffusion models as interactive tools for human-centered artistic creation.

1. Introduction

Artistic style transfer, the task of rendering an image in the stylistic appearance of another, has seen rapid progress with the advent of deep generative models. While traditional methods focused on transferring a single style, recent research has introduced more expressive frameworks that allow for compositional generation, blending multiple styles and controlling their influence during the image synthesis process. However, effectively controlling style strength, content preservation, and multi-style composition remains a nuanced problem as artistic style is inherently qualitative, subject, and human in interpretation.

In this project, we aim to investigate and extend compositional visual generation using diffusion models, inspired by the framework proposed in *Compositional Visual Generation with Composable Diffusion Models* [11]. Our motivation stems from the observation that most real-world artistic expression is not monolithic; artists often combine multiple influences, and any practical style transfer system should allow users to do the same with fine-grained control.

The input to our algorithm is a content image and a list of text prompts, each representing a desired artistic style (e.g., ”in the style of Monet”, ”in the style of Picasso”). We use a pre-trained text-to-image Stable Diffusion model [2], augmented with a compositional denoising procedure [11], to produce an output image that fuses the visual styles described in the prompts while preserving the content of the original image.

Our system also supports varying the content strength, relative style weights and classifier-free guidance scale to control the extent of content preservation, stylization and prompt adherence. We explore both qualitative and quantitative evaluations of the generated images, using perceptual metrics like LPIPS (Learned Perceptual Image Patch Similarity) for content similarity and human survey data coupled with `GPT-4o` evaluation for style and weight alignment. Using this evaluation schema, we aim to better understand how different compositional and control parameters affect output quality and to identify the tradeoffs between stylization and content fidelity.

This project explores the intersection between progressing, generative technology and the innately human domain of artistic expression. By enabling nuanced control over how visual styles are combined, weighted, and rendered, our system aims to demonstrate new ways of thinking about creativity, where machine learning models do not replace the artist, but instead become a tool for exploration and reinterpretation. Through compositional diffusion and guided experimentation, we demonstrate how technology can participate in, amplify, and evolve the creative process.

2. Related Work

2.1. CNN-Based Neural Style Transfer

The baseline of modern content and style-separated neural style transfer was introduced by the work of *Gatys et al.*[4]: a CNN-based neural style transfer that explicitly separated the two variables using deep features. In their method, a pretrained convolutional network (VGG) serves as a feature extractor: the content of an image is captured by high-level feature maps, while the style is represented by correlations (Gram matrices) of features in multiple layers. It iteratively optimizes a white-noise image to match the content features of the input photo with the Gram-matrix statistics of a painting to recombine content and style in a novel way, producing impressive high-quality painterly outputs. However, a major weakness is its computational cost: the slow optimization (hundreds of iterations) is required for each new image, making it impractical for real-time applications; the method used also often produced style or color distortions. The latter issue was tackled by subsequent papers that used methods like a Markov Random Field to better preserve local color and texture patterns [8].

2.2. Feed-Forward Perceptual Loss Networks

Faster feed-forward networks to directly output stylized images were developed as a response to the inefficiency of initial optimization-based stylization. An especially innovative approach was taken by *Johnson et al.*[7] where such a network was trained with perceptual loss to approximate optimization in one forward pass. This loss metric was the same VGG-based content and style loss deployed by *Gatys et al.*[4], but *Johnson et al.*[7] uses a transform net that is guided by fixed pre-trained VGG features during training and learned to map any input photo to a specific style output without the need for pixel-level supervision. Johnson's model was able to produce stylized results three orders of magnitude faster than the iterative method used in *Gatys et al.*[4]. The same method was adopted to texture in *Ulyanov et al.*[15]. Strengths of feed-forward perceptual loss networks is most obviously speed, while the use of perceptual loss was able to deploy deep feature knowledge without the need for ground-truth stylized images for training. However, these models lack flexibility, requiring training a new network from scratch for additional styles. While we initially tried to expand our project off the model by *Johnson et al.*[7], the initial overhead of training a feed-forward network from scratch on each new style was unproductive, thus pushing us to defer to the stable diffusion network approach we will later cover. These drawbacks were addressed in later papers, which implemented instance norm in place of batch norm that successfully pushed the model to refocus on the style statistics and become more agnostic to content statistics[16].

2.3. Multi-style Interpolation

The idea of multi-style transfer has been explored by *Dumoulin et al.* [3], who transform style embeddings into a vector space using conditional instance normalization. This enables a single deep network to capture a diverse range of artistic styles. By blending the normalization parameters of two styles, the network can generate a novel mixed style, offering users a continuum of artistic effects. Integral to multi-style, this model demonstrated a practical way to achieve many styles in one model and introduced the notion of a style embedding space.

Adaptive multi-style transfer was central to our input design, building on the work of *Huang et al.* [6], which enables style transfer without training on specific styles by applying new style statistics at runtime. This flexibility allows for continuous style interpolation by computing and blending AdaIN outputs from two different style images. *Li et al.* [9] propose another approach to arbitrary style transfer by aligning content and style feature distributions. Their method removes correlations from content features and recolors them using the style's covariance, eliminating the need for style-specific training through an autoencoder with embedded WCT. Both *Huang et al.* [6] and *Li et al.*[9] represent state-of-the-art approaches to style transfer for their broad applicability, speed, and flexibility although lacking in specificity for each individual style.

2.4. Diffusion-Based Style Transfer

Most recently, diffusion models have emerged as a powerful image generation paradigm for applications to style transfer. One application is image-to-image diffusion, as demonstrated by *Meng et al.* [12], who re-render input images in a new style (e.g., "in the style of Van Gogh") by adding noise and then denoising under a different textual condition, an approach that was instrumental to our later development. Greater precision in content and style is achieved by *Hu et al.*[5], a training-free style transfer pipeline that combines visual and textual conditioning to inject content and style separately, enabling more controllable stylization.

Most relevant to our method is the work on Composable Diffusion Models by *Liu et al.* [11], which interprets diffusion models as energy-based models. They demonstrate that multiple diffusion processes can be composed by multiplying their probability densities, allowing style aspects to be layered rather than simply averaged. Diffusion based models are fast, high quality, and flexible, and combined with compositional guidance and proper hyperparameter tuning, representing the frontier of artistic style transfer.

3. Methods

Our project leverages the architecture and denoising capabilities of pretrained diffusion models to perform compositional multi-style transfer on images. Our method involves (1) encoding a source image into the latent space, (2) adding noise with a specified strength, (3) guiding the denoising process using text prompts and classifier-free guidance (CFG), and (4) decoding the latent output to produce a stylized image. We build on the publicly available Stable Diffusion 2.1 model [2] and extend it with compositional guidance as introduced by *Liu et al.* [11]. All model weights are frozen during generation; our contributions focus on algorithmic composition and experimental evaluation.

3.1. Image Encoding and Latent Noise Injection

We first encode the source image $x \in^{512x512x3}$ into a latent representation z_0 using the VAE encoder provided by the Stable Diffusion pipeline. To simulate a forward diffusion step, we perturb z_0 with Gaussian noise scale by a user-inputted **strength** parameter $s \in [0,1]$, which corresponds to a timestep index t. The resulting noisy latent z_t is:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

3.2. Conditional Denoising with Text Prompts

The core of our method relies on a pretrained conditional diffusion model, specifically Stable Diffusion 2.1[2], which performs image generation and editing by iteratively denoising a latent representation. At each timestep t, the model receives a noisy latent z_t and a conditioning embedding $\tau \in^d$, produced by a frozen CLIP text encoder from the input prompt p. During training, the model aims to minimize the denoising score matching loss between the predicted noise and true noise $\epsilon\, \mathcal{N}(0, I)$ given by:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{z_0,\epsilon,t,p}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau)\|^2\right].$$

For the purposes of our project, we pass in the noisy latent of our source image to this model to preserve content rather than pure image generation along with a prompt relating to artistic style. During inference, the denoising model predicts the noise that was originally added to the clean latent, given by $\hat{\epsilon}_\theta = \epsilon_\theta(z_t, t, \tau)$.

3.3. Compositional Denoising with Prompt Guidance

Given z_t, we perform reverse diffusion using the pretrained conditional diffusion model. Following the composition method described by *Liu et al.* [11], we guide the denoising process by aggregating the predicted noise across multiple prompts. At each timestep t, the network predicts the noise $\epsilon_\theta(z_t, t, \tau_i)$ for each style prompt and are composited using either:

- **Uniform averaging** when no weights are provided:

$$\hat{\epsilon}_t = \frac{1}{N} \sum_{i=1}^{N} \epsilon_\theta(z_t, t, \tau_i)$$

- **Weighted composition** when prompt weights w_i are provided:

$$\hat{\epsilon}_t = \frac{1}{\sum_i w_i} \sum_{i=1}^{N} w_i \cdot \epsilon_\theta(z_t, t, \tau_i)$$

In addition, we implemented classifier-free guidance (CFG) which is a widely used technique in conditional diffusion models that improve the alignment of generated samples with conditional signal. CFG leverages the same model to generate both conditional and unconditional predictions, interpolating between them during inference. Given a user-inputted guidance scale γ, we produce a guided prediction using:

$$\hat{\epsilon}_{\text{guided}} = \hat{\epsilon}_{\text{uncond}} + \gamma \cdot (\hat{\epsilon}_{\text{cond}} - \hat{\epsilon}_{\text{uncond}})$$

This process is repeated for all timesteps, producing the final latent z_{stylized}.

3.4. Image Decoding

The final denoised latent image z_{stylized} is decoded back into image space using the pretrained VAE decoder to produce $\hat{x}$, our final stylized image composed by the content of our source image and the influence of multiple style prompts, noise strength, weights, and guidance values.

4. Dataset

4.1. Pre-training Corpus for Backbone Diffusion Model

Our method inherits all pre-training and visual priors from Stable Diffusion v2.1 [2]. This model was trained from scratch on a filtered subset of LAION-5B, a 5.85b image–text pair crawl derived from Common Crawl. Stable Diffusion v2.1 is fine-tuned from stable-diffusion-2 [14] with an additional 55k steps on the same dataset, and then fine-tuned for another 155k extra steps with punsafe$= 0.98$.

For preprocessing in the training of the stable diffusion model, images were center-cropped or padded to square $768x768$ patches, converted to latent space via a frozen auto-encoding VAE, and pixel-values are normalized to $[-1, 1]$. During training the model sees no additional data augmentation besides the stochastic forward-diffusion noise. Text conditioning is supplied by OpenCLIP-ViT/H/14 embeddings of the captions. No hand-crafted features are used and all supervision comes from paired text.

4.2. Evaluation Set

To evaluate our multi-style modifications, we curated a 70-image subset from the MS-COCO 2014 test split [10] and resized to 512x512 pixels. The images were hand selected to ensure diversity in:

- Subject matter: people (19), natural landscapes (18), street/architecture (17), isolated objects (16).
- Structural complexity: simpler scenes of foreground against background, vs. more complex multi-object scenes.

To ensure controlled evaluation, the content images were held constant while systematically varying parameters such as style weight and content strength. For relative style weighting experiments, we tested each configuration on batches of 5 images to assess consistency and visual coherence. See Figure 1 for a visual.

Figure 1: Sample of three images from our COCO2014 subset

5. Experiments and Results

Our experiments were performed on a set of four artists we handpicked for their distinctive and recognizable styles: Vincent van Gogh, Claude Monet, Salvadore Dali, and Pablo Picasso. Examples of each style are shown in Figure 2.

5.1. Tuning Hyperparameters

5.1.1 Noise Strength

We performed a sweep for the the noise strength added to the source image in order to determine optimal value to balance the content versus style tradeoff. To quantify how noise strength correlates with content preservation, we used Learned Perceptual Image Patch Similarity (LPIPS) as our metric. LPIPS measures how perceptually different two images are from a human visual perspective by comparing deep feature maps extracted from a pretrained network (VGG). We chose this metric for measuring content preservation over other measure such as Mean Squared Error(MSE), because it detects content preservation even when style changes through being sensitive to semantic and structural changes instead of smaller changes such as lighting and color shift. In contrast, MSE penalizes all pixel shifts which would blur style influence with content preservation. In particular, we ran a range of noise strength values from 0.1 to 0.9 with fixed prompt styles ”Van Gogh” and ”Picasso” on 70 source images and plotted strength v. computed LPIPS with the original source image, as shown in Figure 3. As expected, we observe that higher noise strength is correlated with less content preservation and the opposite is true of lower noise strength.

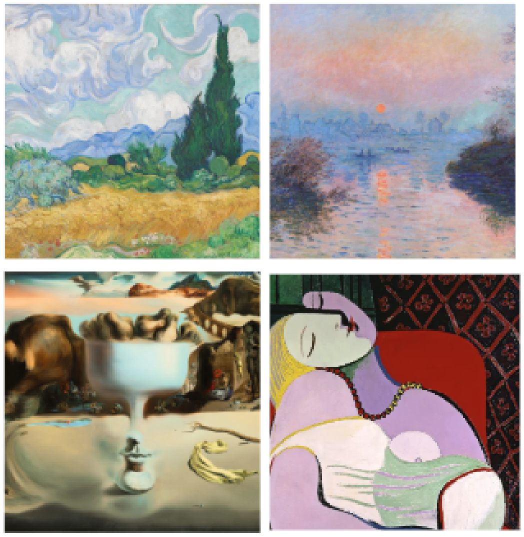


Figure 2: Style samples: Van Gogh, Monet, Dali, Picasso (left to right, top to bottom)

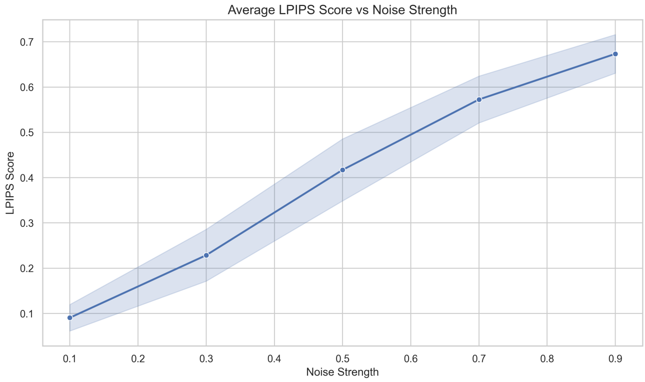


Figure 3: Noise Strength versus Content Preservation

Qualitatively, we relied on two iterations of forced-choice protocol via human evaluation, first on a broader range of strengths and then on a more focused strength range based off initial respondents. Out of 30 respondents, we take the average of each individual respondent’s preferred strength to the following on the same image run with three different artist prompt combinations.

- **Iteration 1**: Given the source image, choose the image that most closely resembles a modification of [insert prompt, e.g. turn this image into the styles of Van Gogh and Picasso] with options of strengths 0.1 to 0.9.

From this initial surveying, we observed that the optimal noise strength was concentrated near 0.5 as seen in Figure 4. To further refine the optimal weight, we conducted a second round of evaluation focused on the narrower strength range of 0.4 to 0.6, which emerged as the most frequently selected interval in the initial iteration.

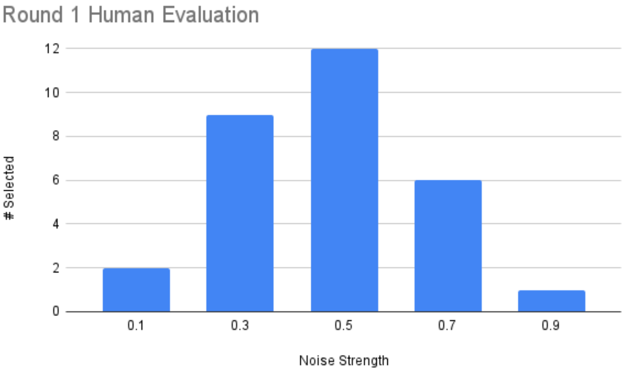


Figure 4: Iteration 1 Human Evaluation Results

- **Iteration 2**: Given the source image, choose the image that most closely resembles a modification of [insert prompt] with options of strengths 0.4 to 0.65.

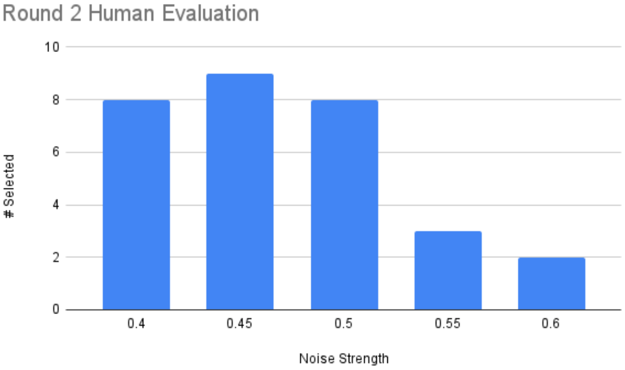


Figure 5: Iteration 2 Human Evaluation Results

As reflected in Figure 5, our qualitative evaluation resulted in an averaged best strength of **0.47**.

5.1.2 Style Weight

To quantify the effectiveness of intentionally weighting prompts on the output image, we ran a varied set of weights on 70 images and used Contrastive Language-Image Pretraining (CLIP) [13] to measure the similarity between outputted images and each style prompt. When comparing the relative similarity between prompts with higher weights and prompts with lower weights, we did not notice any correlation in the computed CLIP similarities as seen in Figure 6. This indicates that CLIP similarity does not accurate reflect human perception of style influence since the outputted images were very visually different in style as seen by Figure 7 but received similar CLIP similarity scores across prompts. Thus, we proceeded with a qualitative analysis of the weights applied to multi-style transfer.

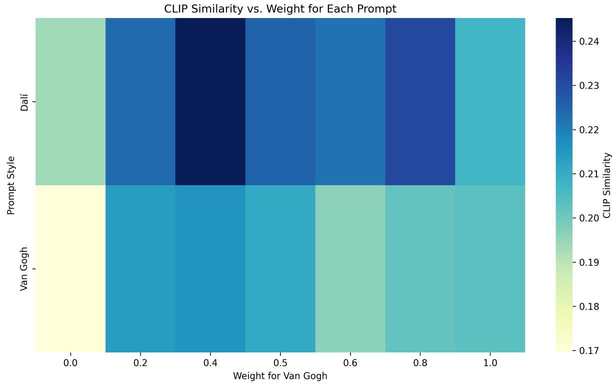


Figure 6: Heatmap: Prompt Weight and CLIP Similarity

To understand how the relative style weights influence perceived stylization, we generated three weight splits—$[0.2, 0.8]$, $[0.5, 0.5]$, and $[0.8, 0.2]$—for every pair drawn from $\{$Monet, Van Gogh, Picasso, Dali$\}$. For each setting we sampled five content images from our COCO2014 subset and rendered them with the optimal noise strength found earlier ($s = 0.47$).

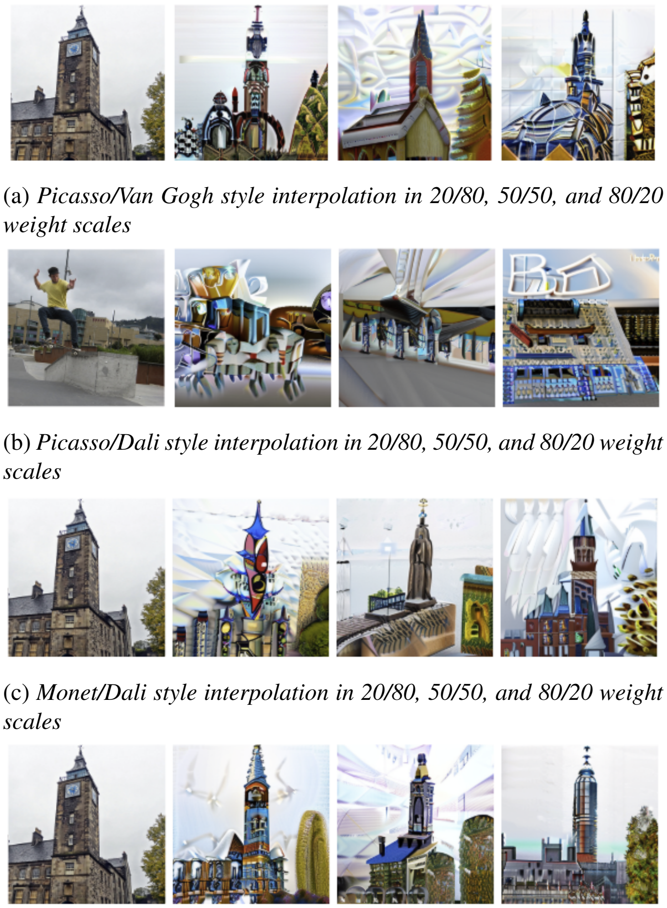
Because manual expert rating was infeasible, we were recommended to adopt a lightweight LLM-assistant evaluation. Each batch (source image + three stylized candidates) was fed to `GPT-4o` with the prompt: *"Which candidate most closely matches a α/β split between Artist A and Artist B?"*. The model returned a rank ordering that we treat as a proxy for human preference.

Observations:

- **Style dominance.** Highly abstract styles (e.g. Dali) tend to overwhelm more representational ones (e.g. Monet). In these cases, an asymmetric $[0.2, 0.8]$ split *in favor of the weaker style* produced images that `GPT-4o` judged closer to an "even-looking" blend than the nominal $[0.5, 0.5]$ setting. As an example, 7c shows strong Dali influence even in the 80/20 Monet/Dali split; in this case a 20% Dali interpolation looks more equally weighted than the 50/50 case.

- **Non-linear mixing.** Pairing two very strong, idiosyncratic styles (e.g. Picasso + Dali) often yielded hybrids that looked like neither parent style as seen in Fig 7b, suggesting destructive interference in the guidance vectors.
- **Best pairings.** The most visually coherent blends in all five test images were Picasso + Van Gogh(7a) and Monet + Van Gogh (7d). These pairs preserved recognizable stylistic components from both artists while maintaining decent content structure.

Figure 7 shows notable samples from our weight testing schema along with their source images on the left.



(a) *Picasso/Van Gogh style interpolation in 20/80, 50/50, and 80/20 weight scales*

(b) *Picasso/Dali style interpolation in 20/80, 50/50, and 80/20 weight scales*

(c) *Monet/Dali style interpolation in 20/80, 50/50, and 80/20 weight scales*

(d) *Monet/Van Gogh style interpolation in 20/80, 50/50, and 80/20 weight scales*

Figure 7: Sample weight experiment results for four style combinations

5.1.3 Classifier Free Guidance

Replicating our quantitative experiment on noise strength, we performed a sweep for guidance scales from 1 (default no guidance) to 12.5 using LPIPS to measure the effect of guidance scale on content preservation on 70 source images with fixed prompts ”Van Gogh” and ”Picasso.” As seen in Figure 8, we see a clear trend between lower guidance scale and higher content preservation. This matches our expectation since more focus on conditioning to the style prompt will decrease content preservation. While the overall trend resembles that of noise strength, we observe greater deviation from the average (measured by standard deviation), suggesting that guidance scale has a less direct impact on content preservation.

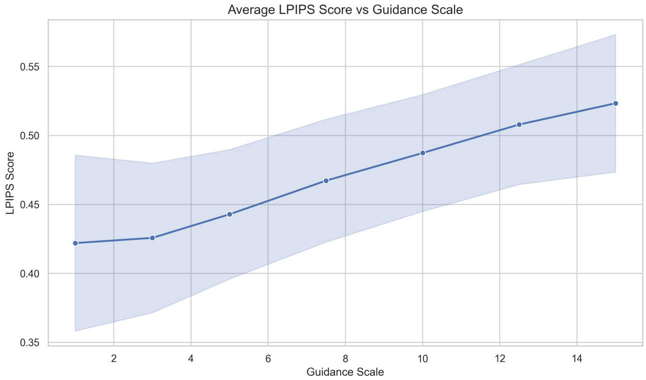


Figure 8: Guidance Scale (CFG) versus Content Preservation

For a qualitative analysis, we see clearly that increased guidance scale increases the amount of ”stylization” of the image, as seen in Figure 9. In line with the quantitative differences observed between noise strength and guidance scale, visual inspection of the output images reveals that guidance scale better preserves content, even while intentionally incorporating more style. Next, we surveyed 30 people to choose the visually optimal guidance scale for a set of images. Our results in Table 1 show the comparison of optimal guidance scale vs. image complexity, demonstrating that more complex images require a lower guidance scale to preserve content compared to simpler images as shown in Figure 10. Survey participants also indicated a threshold of **7.5** for which any guidance scale beyond this point rendered image with large content loss without visual appeal. Given these results, we averaged our best guidance scale to **2.4**.

5.2. Baseline Comparison

As a baseline model for comparison, we used InstructPix2Pix by *Brooks et al.* [1] for their advanced image alteration capabilities. InstructPix2Pix is built on the same Stable Diffusion backbone [2] as our model, so quality gaps can be traced to guidance and fine-tuning strategies rather than pre-training. Crucially, iP2P accepts only a *single* tex-

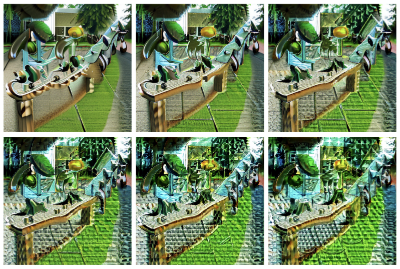


Figure 9: Guidance Scale: 1, 3, 5, 7.5, 10, 12.5 (left to right, top to bottom)

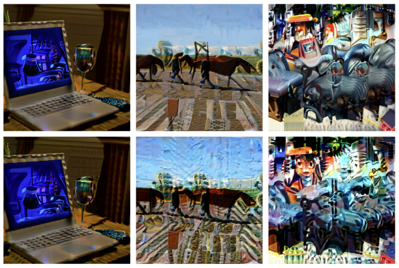


Figure 10: Guidance Effect for Simple, Standard, and Complex Image

Image Type	Optimal Guidance Scale
Simple	3.2
Standard	2.7
Complex	1.3
Average	2.4

Table 1: Optimal guidance scales for different image complexities.

tual instruction and therefore lacks an explicit mechanism to balance or fuse multiple artistic styles. Introducing our compositional guidance module on top of the same backbone isolates the contribution of explicit multi–style control.

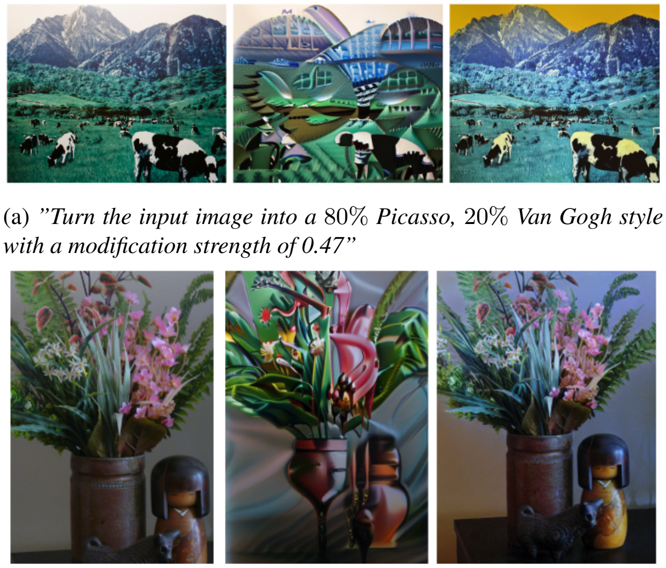
Our experimentation schema included a human-evaluation qualitative analysis of five-image batches in two-option forced-choice protocol run on various artist style combinations of the aforementioned four artists. We ran the same image, style weights, and prompt through our model (with best hyperparameters found earlier of 0.47 strength and 2.4 guidance) and on iP2P.

Thirty participants completed a two-alternative forced-choice survey: given the source content and the text prompt, they selected the output that “better reflects the stated style mixture while preserving content.” Across votes, 63% reported our model as more accurate to the style weights and given prompt.

Qualitative comparison of outputs produced the following observations:

- *Inability of baseline to handle multiple outputs.* iP2P typically reproduced only the style with the higher weight, ignoring the secondary artist; our method retained salient cues from both artists even for asymmetric mixes.
- *Style–content trade-off.* When CFG scale was lowered to improve content fidelity, iP2P outputs became weakly stylized; compositional guidance maintained stronger texture transfer at the same content level.

Sample survey choices are shown in Figure 11. Overall, the experiment confirms that explicit, weight-aware composition yields perceptibly better multi-style transfers than our baseline.



(a) *”Turn the input image into a* 80% *Picasso,* 20% *Van Gogh style with a modification strength of 0.47”*

(b) *”Turn the input image into a* 80% *Monet,* 20% *Dali style with a modification strength of 0.47”*

Figure 11: Sample benchmark comparisons. Left to right: source image, our model’s output, benchmark output

6. Conclusion and Future Works

In this project, we presented a diffusion-based framework for multi-style artistic transfer using compositional denoising, classifier-free guidance, and tunable control over style weights and content preservation. By leveraging Stable Diffusion’s pretrained latent denoising model, we successfully synthesized stylized outputs that reflected combinations that reflected combinations of the artistic styles of Monet, Van Gogh, Picasso, and Dalí. Through a quantitative and qualitative hyperparameter sweep, we found that a noise strength of approximately 0.47 and a classifer-free guidance of 2.4 offered the best tradeoff between stylization and content preservation. Our human and LLM-assisted evaluation highlighted the model's ability to capture stylistic nuance and respect user-specified weightings more effectively for multi-style transfer than our baseline model (InstructPix2Pix).

Interestingly, we observed that style dominance and non-linear mixing effects introduced challenges in balancing strongly abstract styles like Dalí against more structured ones like Monet. In these cases, compensatory weighting helped—but not always linearly—underscoring the inherent complexity of composing learned style priors. Overall, our model offered greater flexibility and control than single-style approaches, especially in capturing multi-style blends with both visual coherence and content awareness.

For future work, we would explore style-specialized finetuning of the underlying diffusion backbone for each artist prompt. Currently, all guidance relies on a general-purpose text-conditioned model; while this allows for flexibility, it lacks style-specific depth. Training or fine-tuning smaller diffusion heads per style on dedicated datasets could improve style fidelity, reduce mode collapse when mixing styles, and allow for richer prompt embeddings beyond CLIP alone. With more compute, we would also investigate adversarial training such as adding a discriminator network to better disentangle style and content features during guidance, enabling more precise multi-style interpolation.
Ultimately, our goal is to move toward a model where artistic intent can be expressed with greater subtlety and control, making machine learning a true collaborator in creative workflows. Yet in doing so, we are reminded that artistic style: its gestural representations, innate creativity, and unique distinction, is an inherently human quality. While models may mimic brushstrokes or color palettes, the spirit behind the art remains something deeply personal and irreducibly human.

7. Contributions

Catherine implemented the model pipeline, which includes image encoding and compositional denoising, along with advanced features such as style weighting and classifier-free guidance. She also ran quantitative tests for hyperparameter tuning. Christina performed qualitative testing of all hyperparameters via human and LLM-evaluation and evaluated our model against our InstructPix2Pix baseline. Both authors contributed to ideation, experimentation, and writing the report.

In our project builds on top of Stable Diffusion 2.1 [2] and uses InstructPix2Pix [1] and CLIP [13] for baseline evaluation.

References

[1] T. Brooks, A. Holynski, and A. A. Efros. Instruct-pix2pix: Learning to follow image editing instructions, 2023. `https://github.com/timothybrooks/instruct-pix2pix`.

[2] CompVis and StabilityAI. Stable diffusion v2.1. `https://huggingface.co/stabilityai/stable-diffusion-2-1`, 2022. Accessed: 2025-06-03.

[3] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations (ICLR)*, 2017.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423. IEEE Computer Society, 2016.

[5] Y. Hu, C. Zhuang, and P. Gao. Diffusest: Unleashing the capability of the diffusion model for style transfer. *arXiv preprint arXiv:2410.15007*, 2024.

[6] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science (LNCS)*, pages 694–711. Springer, 2016.

[8] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2479–2486. IEEE, 2016.

[9] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[11] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022.

[12] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. `https://github.com/openai/CLIP`.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[15] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1349–1357. PMLR, 2016.

[16] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4105–4113. IEEE, 2017.