

MULTISTYLE IMAGE TRANSFER USING DIFFUSION TECHNIQUE

Yashpal Singh

Department of Computer Science & Engineering, MMMUT Gorakhpur

E-Mail: yashcs0117@gmail.com

Abstract

Image style transfer renders the content of one image in the visual style of another. Most methods transfer a single style at a time, yet artists and designers often want to combine several styles, control how strongly each one appears, or apply different styles to different regions of the same image. Diffusion models, which generate images by gradually removing noise, have recently replaced convolutional and adversarial methods as the strongest tools for image synthesis and offer new ways to achieve this control. This study examines multistyle image transfer with diffusion models and proposes a training-free framework, Multi-Style Diffusion Transfer (MSDT), built on a pretrained latent diffusion model. The framework inverts the content image into the diffusion latent space, encodes each style image separately, and combines the styles through three mechanisms: weighted blending of style embeddings, injection of style features into the self-attention layers of the denoising network, and spatial masks that assign styles to regions. A timestep-dependent schedule controls when each style acts, and a multi-condition form of classifier-free guidance sets the strength of each style. The study also analyses how the noise schedule governs the trade-off between content preservation and stylisation. Computations for three widely used schedules show that, at a transfer strength of 0.5, the fraction of the original content signal retained ranges from 0.28 for the linear schedule to 0.70 for the cosine schedule, so the same strength setting produces very different results across models. Finally, the study sets out an evaluation protocol based on content fidelity, style fidelity, style balance and human judgement, and discusses limitations and ethical questions, including the rights of artists whose styles are reproduced.

Keywords: style transfer; diffusion models; multistyle synthesis; latent diffusion; attention injection; classifier-free guidance; image generation

1. Introduction

Style transfer is the task of producing a new image that keeps the structure and objects of a content image while taking on the colours, textures, brushwork and visual character of a style image. The modern field began when Gatys, Ecker and Bethge (2016) showed that the feature correlations of a convolutional neural network capture artistic style and that an image could be optimised to match the content features of one picture and the style statistics of another. Faster feed-forward networks soon followed (Johnson et al., 2016), as did methods that could transfer any style in a single pass by matching feature statistics (Huang & Belongie, 2017).

Two developments have since changed the field. The first is the growing demand for more flexible control. A single style is often not enough: a designer may want an image that is part watercolour and part ink drawing, a film-maker may want to blend the palettes of two artists, and a graphic artist may want the sky in one style and the foreground in another. Multistyle transfer, in which several styles are combined in controlled proportions or regions, addresses these needs. Early convolutional approaches learned a separate set of normalisation parameters for each style and could interpolate between them (Dumoulin et al., 2017),

but they were limited to styles seen during training.

The second development is the rise of diffusion models. These models learn to reverse a gradual noising process and have achieved image quality beyond that of generative adversarial networks (Dhariwal & Nichol, 2021). Latent diffusion models, which run the diffusion process in the compressed latent space of an autoencoder, made high-resolution generation practical and are the basis of widely used text-to-image systems (Rombach et al., 2022). Their attention layers, text and image conditioning, and step-by-step generation offer natural points at which several styles can be introduced and balanced.

This study brings these two developments together. Its objectives are:

1. To explain the principles of diffusion models that are relevant to style transfer, including the forward and reverse processes, inversion and guidance.
2. To review existing diffusion-based style transfer methods and assess their support for multiple styles.
3. To propose a training-free framework for multistyle transfer that offers global blending, regional control and timestep control of styles.
4. To analyse how the noise schedule determines the balance between content preservation and stylisation.
5. To set out an evaluation protocol and discuss the limitations and ethical implications of multistyle diffusion transfer.

2. Background

2.1 From optimisation to feed-forward style transfer

In the optimisation approach, style is represented by Gram matrices of network features. For a feature map F with C channels at layer l , the Gram matrix $G = F F^T$ records which features occur together, and the style loss measures the difference between the Gram matrices of the generated and style images (Gatys et al., 2016). Optimisation produces high-quality results but takes minutes per image. Feed-forward networks trained with perceptual losses reduce this to a single pass (Johnson et al., 2016). Adaptive instance normalisation (AdaIN) made arbitrary style transfer possible by aligning the channel-wise mean and standard deviation of content features with those of style features:

$$\text{AdaIN}(x, y) = \sigma(y) \cdot [(x - \mu(x)) / \sigma(x)] + \mu(y) \quad (1)$$

where x and y are content and style features and μ and σ denote channel-wise mean and standard deviation (Huang & Belongie, 2017). The whitening and colouring transform extended this idea by matching the full covariance of features (Li et al., 2017). These methods remain fast, but they often produce artefacts, struggle with complex semantic content and offer limited control over how styles combine.

2.2 Diffusion models

A denoising diffusion probabilistic model defines a forward process that adds Gaussian noise to an image x_0 over T steps according to a variance schedule $\beta_1, \dots, \beta_T$ (Ho et al., 2020). Writing $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s \leq t} \alpha_s$, the noisy image at step t can be sampled directly as:

$$x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{(1 - \bar{\alpha}_t)} \cdot \epsilon, \quad \epsilon \sim N(0, I) \quad (2)$$

A neural network $\epsilon_\theta(x_t, t, c)$, usually a U-Net, is trained to predict the added noise given the noisy image, the timestep and an optional condition c , by minimising:

$$L = E_{x_0, \epsilon, t} [\| \epsilon - \epsilon_\theta(x_t, t, c) \|^2] \quad (3)$$

Generation starts from pure noise and applies the learned network step by step to remove noise. The deterministic sampler known as DDIM allows sampling in fewer steps and, importantly for style transfer, can be run in reverse to map a real image to the noise that would regenerate it, a process called inversion (Song et al., 2021). The DDIM update from step t to step $t-1$ is:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot [(x_t - \sqrt{(1 - \bar{\alpha}_t)} \epsilon) / \sqrt{\bar{\alpha}_t}] + \sqrt{(1 - \bar{\alpha}_{t-1})} \cdot \epsilon_\theta \quad (4)$$

Latent diffusion applies this process to the latent code $z = E(x)$ of a pretrained variational autoencoder rather than to pixels, which greatly reduces computation; conditions such as text are introduced through cross-attention layers (Rombach et al., 2022).

2.3 Guidance and conditioning

Classifier-free guidance strengthens the effect of a condition by combining conditional and unconditional predictions of the same network (Ho & Salimans, 2022):

$$\hat{\epsilon} = \epsilon_{\theta}(x_t, \emptyset) + s \cdot [\epsilon_{\theta}(x_t, c) - \epsilon_{\theta}(x_t, \emptyset)] \quad (5)$$

where $s > 1$ is the guidance scale. Image prompts can be used as conditions through image encoders such as CLIP (Radford et al., 2021), and lightweight adapters such as IP-Adapter add decoupled cross-attention layers so that a pretrained text-to-image model can be guided by reference images (Ye et al., 2023). These mechanisms are the building blocks of most diffusion-based style transfer methods.

3. Related Work

Diffusion-based style transfer methods fall into four broad families, which differ in whether they require training and in how readily they support multiple styles:

- **Noise-and-denoise editing.** SDEdit adds a controlled amount of noise to the content image and denoises it under a new condition, trading realism against faithfulness through the starting timestep (Meng et al., 2022). It is simple and training-free but offers only coarse control over style.
- **Fine-tuning and embedding learning.** Textual inversion learns a new word embedding that represents a visual concept (Gal et al., 2023), DreamBooth fine-tunes the whole model on a few images (Ruiz et al., 2023), and StyleDrop learns style from a single reference image with adapter tuning (Sohn et al., 2023). Inversion-based style transfer learns a textual representation of an artwork's style (Zhang et al., 2023). Low-rank adaptation reduces the cost of fine-tuning (Hu et al., 2022), and ZipLoRA merges separately trained subject and style adapters (Shah et al., 2024). These methods are accurate but need training for each style, and combining several learned styles can cause interference.
- **Attention manipulation.** StyleAligned shares self-attention between images generated in a batch so that they adopt a common style (Hertz et al., 2024). StyleID injects the keys and values of a style image into the self-attention of the content image during denoising, preserves the content queries and adjusts the initial latent with AdaIN, without any training (Chung et al., 2024). Attention-level methods give fine control and are well suited to combining styles.
- **Image-prompt adapters.** IP-Adapter conditions generation on image embeddings (Ye et al., 2023), and InstantStyle shows that injecting style features only into specific attention blocks separates style from content more cleanly (Wang et al., 2024).

Table 1. Representative diffusion-based style transfer methods and their support for multiple styles

Method	Training required	Main mechanism	Support for multiple styles
SDEdit (Meng et al., 2022)	No	Partial noising and conditional denoising	Indirect, through mixed text prompts
Textual inversion (Gal et al., 2023)	Yes, per concept	Learned word embedding	Several tokens in one prompt; interference possible
InST (Zhang et al., 2023)	Yes, per style	Inversion of artwork into a textual embedding	Not designed for blending
StyleDrop (Sohn et al., 2023)	Yes, adapter tuning	Adapter trained on one style image	Limited
ZipLoRA (Shah et al., 2024)	Yes, per adapter	Merging of subject and style low-rank adapters	Two adapters merged
StyleAligned (Hertz et al., 2024)	No	Shared self-attention across a batch	One reference style per batch
StyleID (Chung et al., 2024)	No	Key-value injection with query preservation	Single style; extensible to several
InstantStyle (Wang et al., 2024)	No	Style features injected into selected blocks	Weighted image prompts possible

The review shows that training-free attention-based methods offer the best starting point for multistyle transfer, because they work with any reference image, need no additional training and expose internal representations that can be weighted and combined. What is missing is a unified framework that supports global blending, regional assignment and temporal control of several styles at once.

4. Proposed Framework: Multi-Style Diffusion Transfer

The proposed framework, Multi-Style Diffusion Transfer (MSDT), operates on a pretrained latent diffusion model and requires no training. Its pipeline is shown in Figure 1.

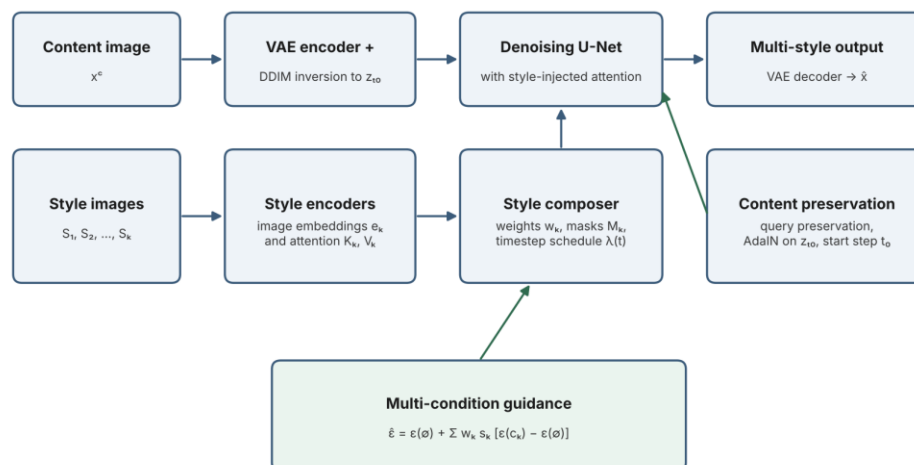


Figure 1. Pipeline of the proposed Multi-Style Diffusion Transfer framework

4.1 Content inversion

The content image x^c is encoded by the autoencoder to z_0^c and mapped by DDIM inversion to a noisy latent

z_0 at a starting step t_0 . The choice of t_0 sets the transfer strength: a small t_0 keeps most of the content structure, while a large t_0 allows stronger stylisation. Inversion, rather than random noising, ensures that the starting latent corresponds to the content image and reduces loss of detail (Song et al., 2021). Where exact reconstruction is important, null-text optimisation can further improve inversion fidelity (Mokady et al., 2023).

4.2 Style encoding

Each of the K style images $S_1, \dots, S_K$ is processed in two ways. First, an image encoder produces an embedding e_k that is used as an image prompt through decoupled cross-attention. Second, each style image is itself inverted, and the keys K_k and values V_k of its self-attention layers are stored at every timestep. These two representations capture complementary aspects of style: embeddings capture global character such as palette and mood, while attention features capture local texture and stroke patterns.

4.3 Style composition

The core of the framework is the style composer, which combines the K styles through three mechanisms:

1. **Global blending of embeddings.** A combined image prompt is formed as a weighted sum of style embeddings, $e = \sum_k w_k e_k$, with $\sum_k w_k = 1$. The weights w_k let the user set the share of each style.
2. **Attention injection with weights.** In selected self-attention layers of the U-Net, the content query Q^c attends to the keys and values of all styles. The attention output is computed as a weighted mixture: $A = \sum_k w_k \cdot \text{softmax}(Q^c K_k^T / \sqrt{d}) V_k$, where d is the key dimension. Following the observation that preserving the content query keeps spatial structure (Chung et al., 2024), the queries are taken from the content latent and only keys and values come from the styles.
3. **Regional masks.** For regional multistyle transfer, a set of spatial masks M_k that sum to one at every location assigns each style to part of the image. The attention output becomes $A = \sum_k M_k \odot \text{softmax}(Q^c K_k^T / \sqrt{d}) V_k$, so that each region draws its texture from its assigned style while boundaries remain smooth if the masks are softened.

A timestep schedule $\lambda_k(t)$ further controls when each style acts. Early, high-noise steps determine global layout and colour, while later steps determine fine texture (Hertz et al., 2023). Assigning one style a high weight in early steps and another in late steps therefore produces an image with the colour composition of the first style and the surface texture of the second, a form of control that single-style methods do not offer.

4.4 Content preservation

Three measures protect the content structure. The starting step t_0 limits how much of the content signal is destroyed; the content queries are preserved in attention injection; and the mean and standard deviation of the initial latent are aligned with the weighted statistics of the style latents using AdaIN, which transfers colour without disturbing layout. Where stronger structural control is needed, an edge or depth map of the content image can be supplied through ControlNet (Zhang et al., 2023b).

4.5 Multi-condition guidance

Classifier-free guidance is extended to several style conditions $c_1, \dots, c_K$:

$$\hat{e} = \varepsilon_\theta(z_t, \emptyset) + \sum_k w_k s_k \cdot [\varepsilon_\theta(z_t, c_k) - \varepsilon_\theta(z_t, \emptyset)] \quad (6)$$

where s_k is a guidance scale for style k . Separate scales allow a weak style to be strengthened without increasing the overall guidance, which otherwise tends to over-saturate colours. Because each additional condition requires an extra network evaluation, the cost of sampling grows linearly with K ; for large K , the embedding-blending route described above can be used instead, at a single extra evaluation.

4.6 Procedure

The complete procedure can be summarised as follows:

1. Encode and invert the content image to z_0 ; encode and invert each style image, storing its attention

keys and values.

2. Set style weights w_k , optional masks M_k , timestep schedules $\lambda_k(t)$ and guidance scales s_k .
3. Align the statistics of z_{t_0} with the weighted style statistics using AdaIN.
4. For $t = t_0$ down to 1, run the U-Net with style-injected attention, combine predictions by equation (6) and update the latent by equation (4).
5. Decode the final latent with the autoencoder decoder to obtain the multistyle image.

4.7 Implementation considerations

The framework can be implemented on any latent diffusion model that exposes its attention layers, such as the openly available Stable Diffusion family built on the latent diffusion architecture (Rombach et al., 2022). Several practical choices influence the results:

- **Number of sampling steps.** DDIM sampling with 30 to 50 steps usually offers a good balance between quality and speed; inversion should use the same number of steps as generation so that the inverted latent matches the sampling trajectory.
- **Choice of attention layers.** Injecting style features into the decoder (up-sampling) layers of the U-Net affects texture and colour more than layout, while injection into the bottleneck and encoder layers also alters structure. Restricting injection to selected blocks reduces content leakage from the style images.
- **Weight normalisation.** Style weights should sum to one so that the total amount of style remains constant as the mix changes; otherwise adding a style also increases overall stylisation.
- **Mask preparation.** Masks can be drawn by the user or produced by a segmentation model. Blurring their edges by a few latent pixels avoids visible seams between regions in different styles.
- **Memory.** Storing the keys and values of K style images at every timestep increases memory use roughly in proportion to K . Storing them only for the injected layers, or recomputing them on the fly, keeps memory manageable.

4.8 Computational complexity

Let N be the number of sampling steps and C the cost of one forward pass of the U-Net. Single-style attention injection requires roughly N passes for inversion of the content image, N passes for inversion of the style image and N passes for generation, giving a cost of about $3NC$. In the proposed framework, each additional style adds one inversion of N passes, and multi-condition guidance adds one conditional evaluation per style at each generation step. The total cost is therefore approximately:

$$\text{Cost} \approx N \cdot C \cdot [1 + K + (1 + K)] = N \cdot C \cdot (2 + 2K) \quad (7)$$

for K styles with separate guidance, or about $N \cdot C \cdot (3 + K)$ if the styles share a single blended guidance term. Style inversions can be computed once and reused for many content images, so in practice the incremental cost of transferring a fixed set of styles to a new image is only the content inversion and the generation steps.

5. Noise Schedule and the Content–Style Trade-off

The starting step t_0 is the main control of transfer strength, and many tools expose it to users as a single ‘strength’ value between 0 and 1. Equation (2) shows that, after noising to step t_0 , the content contributes to the latent with coefficient $\sqrt{\bar{a}_{t_0}}$ and noise with coefficient $\sqrt{(1 - \bar{a}_{t_0})}$. The fraction of content that survives therefore depends not only on the strength value but also on the noise schedule of the model. To quantify this, the coefficients were computed for three widely used schedules with $T = 1000$ steps: the linear schedule of the original DDPM, with β rising from 0.0001 to 0.02 (Ho et al., 2020); the scaled-linear schedule used in latent diffusion, with β rising from 0.00085 to 0.012 on a square-root scale (Rombach et al., 2022); and the cosine schedule (Nichol & Dhariwal, 2021). Table 2 reports the results.

Table 2. Content signal coefficient $\sqrt{\bar{a}_t}$ at different transfer strengths for three noise schedules (T = 1000)

Transfer strength t_0/T	Linear (DDPM)	Scaled-linear (latent diffusion)	Cosine
0.1	0.947	0.946	0.986
0.2	0.812	0.869	0.948
0.3	0.630	0.770	0.887
0.4	0.442	0.653	0.805
0.5	0.280	0.527	0.703
0.6	0.161	0.402	0.584
0.7	0.083	0.288	0.451
0.8	0.039	0.193	0.307
0.9	0.017	0.120	0.155

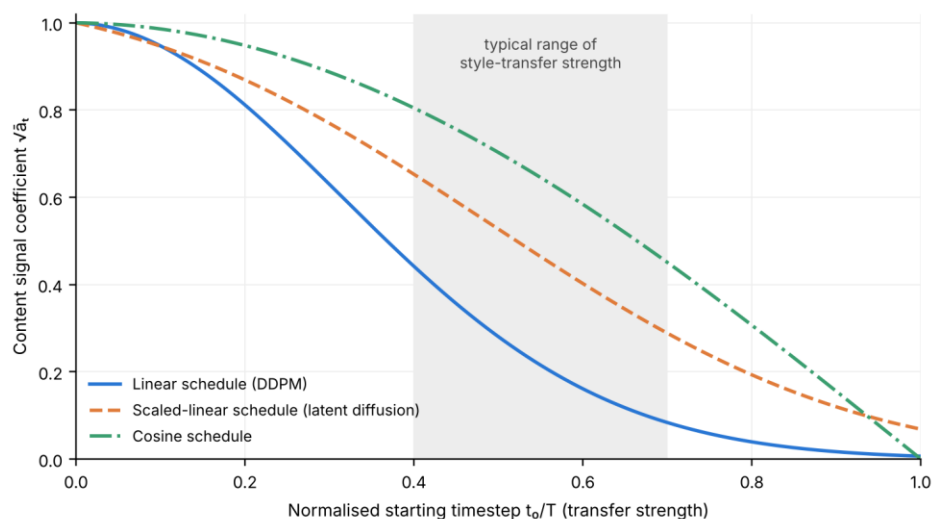


Figure 2. Content signal retained at different transfer strengths for three noise schedules

The results show that the same strength value has very different effects under different schedules. At a strength of 0.5, the linear schedule retains only 28 per cent of the content amplitude, the scaled-linear schedule 53 per cent and the cosine schedule 70 per cent. At a strength of 0.7, the linear schedule has almost erased the content (8 per cent), while the cosine schedule still retains 45 per cent. For style transfer, where the goal is to change appearance while keeping structure, these differences matter a great deal. A strength that produces a pleasing balance in a model with a cosine schedule may destroy the content in a model with a linear schedule.

Two practical conclusions follow. First, transfer strength should be specified in terms of the retained signal $\sqrt{\bar{a}_{t_0}}$, or equivalently the signal-to-noise ratio $\bar{a}_{t_0} / (1 - \bar{a}_{t_0})$, rather than as a fraction of steps, so that settings transfer between models. Second, in multistyle transfer the schedule interacts with the timestep schedule $\lambda_k(t)$: a style assigned to early steps has more influence when t_0 is large, because more of the generation takes place at high noise. A setting in the region where the retained coefficient lies roughly between 0.3 and 0.6 offers a reasonable starting balance between stylisation and structure, and the corresponding strength value can be read from Table 2 for each schedule.

6. Evaluation Protocol

Multistyle transfer requires evaluation along more dimensions than single-style transfer. The study proposes the following protocol:

6.1 Data

- **Content images:** a diverse set of natural photographs, for example drawn from MS-COCO, covering people, animals, landscapes and urban scenes.
- **Style images:** paintings and graphic artworks from a range of movements and media, for example drawn from WikiArt, grouped into pairs and triples of contrasting styles.
- **Masks:** semantic segmentation masks of content images for regional transfer.

6.2 Metrics

- **Content fidelity:** structural similarity and LPIPS perceptual distance between the content image and the output (Zhang et al., 2018), and similarity of edge or depth maps.
- **Style fidelity:** Gram-matrix distance and CLIP image similarity between the output and each style image (Radford et al., 2021), and FID between stylised outputs and the style collection (Heusel et al., 2017).
- **Style balance:** for target weights w_k , the measured share of each style, estimated from normalised CLIP similarities, should match the requested weights; the mean absolute difference between requested and measured shares provides a balance error.
- **Regional accuracy:** for masked transfer, style similarity computed separately inside each mask, and the sharpness of transitions at mask boundaries.
- **Efficiency:** time and memory per image as the number of styles K increases.
- **Human judgement:** paired comparisons in which participants judge content preservation, style resemblance, balance and overall quality.

6.3 Baselines and ablations

Baselines should include a classical arbitrary style transfer method such as AdaIN, single-style diffusion methods applied with mixed prompts, and attention-injection methods applied to one style at a time. Ablation studies should remove each component of the framework in turn: embedding blending, attention injection, AdaIN initialisation, query preservation and the timestep schedule, to show its contribution.

7. Discussion

7.1 Strengths of the diffusion approach

Diffusion models bring three advantages to multistyle transfer. First, their step-by-step generation provides a natural axis of control: different styles can act at different noise levels, separating colour and composition from texture. Second, their attention layers expose representations that can be mixed with simple weights, so that blending and regional assignment require no new training. Third, their strong generative prior allows them to fill in plausible detail where styles conflict, producing more coherent results than feature-statistics methods, which tend to average textures into blur.

7.2 Limitations

- **Computational cost.** Inversion and dozens of denoising steps make diffusion transfer much slower than feed-forward methods, and each additional style adds attention computation and possibly guidance evaluations.
- **Content leakage and style leakage.** Styles may carry semantic content, such as objects in a painting, into the output, while strong content can suppress weak styles. Choosing the attention layers for injection reduces but does not remove this problem (Wang et al., 2024).
- **Imperfect inversion.** DDIM inversion is not exact, especially with guidance, so fine details of the content image can be lost.

- **Style interference.** When styles have conflicting textures, simple weighting can produce muddy results; non-linear combination or regional separation may be needed.
- **Evaluation difficulty.** Style is partly subjective, and automatic metrics capture only some aspects of it, so human evaluation remains essential.

7.3 Ethical considerations

The ability to reproduce and blend the styles of named artists raises questions of consent, attribution and economic harm. Diffusion models are trained on large image collections that may include copyrighted works, and style transfer can produce images in an artist's recognisable manner without their permission. Responsible use requires respecting artists' wishes, avoiding the imitation of living artists without consent, labelling generated images, and preferring style references that are in the public domain or licensed for such use. Multistyle transfer may reduce direct imitation of any one artist by producing new combinations, but it does not remove the underlying questions about the use of creative work.

8. Applications

Multistyle diffusion transfer has applications in several fields:

- **Digital art and graphic design:** exploring new visual styles by blending influences, and producing consistent artwork for books and games.
- **Film, animation and advertising:** creating mood boards and concept art, and adapting a single scene to different visual treatments.
- **Design and fashion:** applying combinations of patterns and textures to product mock-ups and textiles.
- **Education and cultural heritage:** demonstrating the characteristics of artistic movements by showing how a scene would look in their styles, and visualising the blending of traditional art forms.
- **Data augmentation:** generating stylistically varied training images to make computer vision models more robust to changes in appearance.

9. Conclusion

This study examined multistyle image transfer with diffusion models and proposed a training-free framework, Multi-Style Diffusion Transfer, for combining several styles in a single image. Building on the principles of latent diffusion, DDIM inversion and classifier-free guidance, the framework combines styles through weighted image embeddings, attention injection with preserved content queries, spatial masks for regional control and timestep schedules that separate colour and composition from texture. A multi-condition form of guidance sets the strength of each style. The analysis of noise schedules showed that the same transfer strength preserves very different amounts of content in different models, from 28 to 70 per cent of the content amplitude at a strength of 0.5, which argues for specifying strength in terms of the retained signal rather than the fraction of steps. The proposed evaluation protocol adds style balance and regional accuracy to the usual measures of content and style fidelity.

References

1. Chung, J., Hyun, S., & Heo, J.-P. (2024). Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
2. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34* (pp. 8780–8794).
3. Dumoulin, V., Shlens, J., & Kudlur, M. (2017). A learned representation for artistic style. In

- International Conference on Learning Representations.*
4. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations.*
 5. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423).
 6. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations.*
 7. Hertz, A., Voynov, A., Fruchter, S., & Cohen-Or, D. (2024). Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*
 8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30.*
 9. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33* (pp. 6840–6851).
 10. Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598.*
 11. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations.*
 12. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1501–1510).
 13. Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision* (pp. 694–711). Springer.
 14. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017). Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems 30.*
 15. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., & Ermon, S. (2022). SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations.*
 16. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6038–6047).
 17. Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8162–8171). PMLR.
 18. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.
 19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

20. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22500–22510).
21. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., & Jampani, V. (2024). ZipLoRA: Any subject in any style by effectively merging LoRAs. In *European Conference on Computer Vision*. Springer.
22. Sohn, K., Ruiz, N., Lee, K., Chin, D. C., Blok, I., Chang, H., et al. (2023). StyleDrop: Text-to-image generation in any style. In *Advances in Neural Information Processing Systems 36*.
23. Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *International Conference on Learning Representations*.
24. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., & Chen, A. (2024). InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*.
25. Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.
26. Zhang, L., Rao, A., & Agrawala, M. (2023b). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836–3847).
27. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 586–595).
28. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., & Xu, C. (2023). Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10146–10156).