

# DIFF-NST: Diffusion Interleaving For deFormable Neural Style Transfer

Dan Ruta  
University of Surrey

Gemma Canet Tarrés  
University of Surrey

Andrew Gilbert  
University of Surrey

Eli Shechtman  
Adobe

Nicholas Kolkin  
Adobe

John Collomosse  
University of Surrey, Adobe

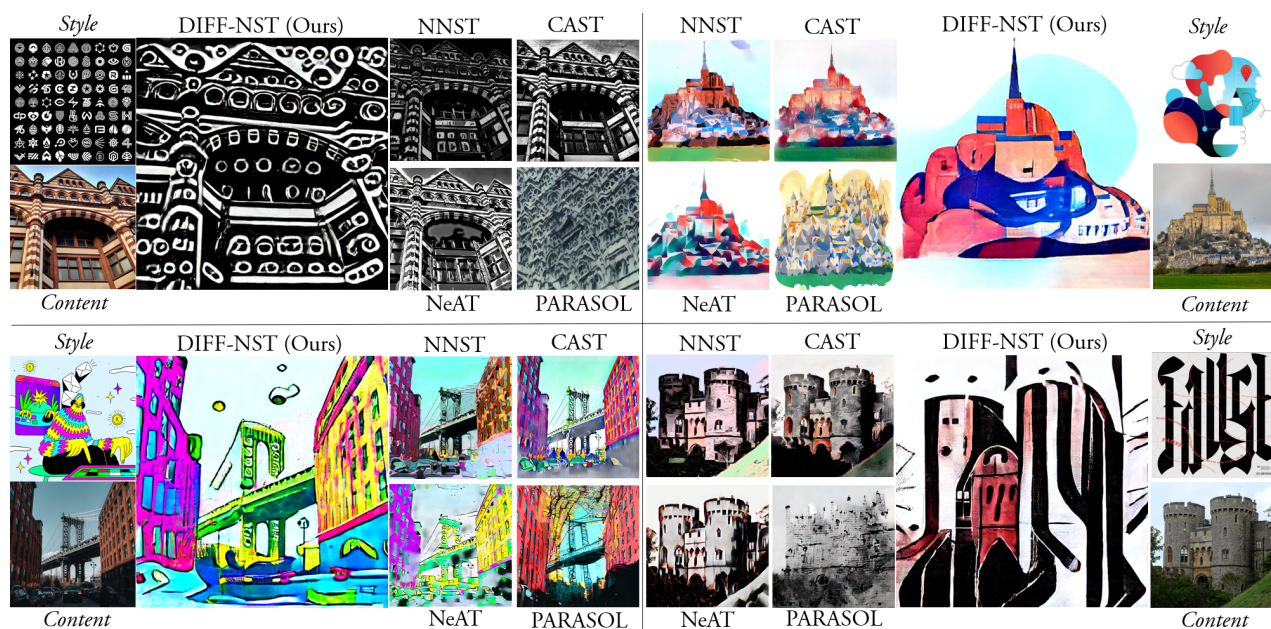


Figure 1. Deformable style transfer using DIFF-NST, compared to baselines: NNST [13], CAST [41], NeAT [25], and PARASOL [31]. Our DIFF-NST method performs style transfer with much stronger style-based form alteration - matching the shapes and structures to those in the style image, not just the colors and textures. More in Fig 9. Zoom for details.

## Abstract

Neural Style Transfer (NST) is the field of study applying neural techniques to modify the artistic appearance of a content image to match the style of a reference style image. Traditionally, NST methods have focused on texture-based image edits, affecting mostly low level information and keeping most image structures the same. However, style-based deformation of the content is desirable for some styles, especially in cases where the style is abstract or the primary concept of the style is in its deformed rendition of some content. With the recent introduction of diffusion models, such as Stable Diffusion, we can access far more powerful image generation techniques, enabling new possibilities. In our work, we propose using this new class of models to perform style transfer while enabling deformable style transfer, an elusive capability in previous models. We show how leveraging the priors of these models can expose new artistic controls at inference time, and we document our findings in exploring this new direction for the field of style transfer.

# 1 Introduction

Neural Style Transfer (NST) aims at re-rendering the content of one image with the distinctive visual appearance of a second style image, typically an artwork. Most prior work has focused on low level style, represented as colors and textures. However, artistic style covers a broader gamut of visual properties, including purposeful geometric alterations to the depicted content, often called *form* [37].

We introduce a novel NST approach that considers not only low level color and texture changes but also higher level style-based geometric alterations to the depicted content. We aim to maintain the object structure to resemble the original content image and remain identifiable as such. But with style-based deformations of the content reflecting the artist’s original intent as they depicted their original subject matter in the exemplar artwork image. Such content deformations have been more challenging to achieve, given a need for a higher level spatial semantic understanding of subject and/or scene information [11].

Learning priors regarding the interplay of artistic style, semantics, and intentional deviations from photo-realistic geometry is non-trivial and not generally a part of NST pipelines. However, recent diffusion-based image generation literature has made impressive progress in modeling various visual concepts [19, 20, 3], accurately modeling how objects fit into the world around them.

We leverage these extensively learned priors in our work, adapting them to NST. We adapt them in our DIFF-NST model to function without text prompts in an exemplar-based setting, similar to more traditional NST. Text-less exemplar-based is desirable for some stylistic edits, as textual prompts would require extensive descriptions of the style, which may be difficult or impossible to articulate fully. We build the first NST model to make significant high level edits to content images. We compare our work to several baselines and show state-of-the-art user preference in user studies.

# 2 Related Work

The seminal work of Gatys’ Neural Style Transfer (NST) [7] enabled neural techniques for transferring the artistic style appearance of a reference artwork to an unstylized depiction of some content - typically a photograph. Follow-up works created feed-forward, optimization free approaches to achieve this [9, 15]. Other techniques for NST emerged, such as optimal transport [12], hypernetworks [24], and Neural Neighbours [13]. Attention based techniques later emerged [17, 16], with further follow-up improvements to contrastive losses [4, 41], and scaling to high resolution with improvements to robustness and detail propagation [25]. Deformation in style transfer has been explored in previous work [11], based on detecting shared keypoints between the style and content, thereby limited by a shared depicted subject. Regarding fine-grained representation space for artistic style, ALADIN [22] introduced the first solution to this training over their fine-grained BAM-FG dataset. This was later evolved into ALADIN-ViT [23] using a Vision Transformer [5] for stronger expressivity, and later as ALADIN-NST [26], with stronger disentanglement between content and style by changing BAM-FG [22] for a fully disentangled, synthetic dataset.

Within the generative image domain, sizeable text-to-image diffusion models such as Dall-e 2 [19], Parti [36], Imagen [27], and e-Diffi [3] have recently made significant advances in image generation fidelity and control, enabling free-form text prompts as an input control vector for guiding image synthesis, with unprecedented quality. These models are trained on large datasets and require prohibitive amounts of computation. Latent Diffusion Models [20] introduced the concept of applying the diffusion process to a smaller, latent representation of images rather than operating in pixel space like the previous works. This dramatically reduces the compute requirements for training and, more importantly, inference. Stability AI [1] democratized comprehensive open access to such models by open sourcing weights for an LDM trained on a subset of the LAION [28] dataset.

Much follow-up research has been enabled and built on these pre-trained weights, known as the Stable Diffusion model. Due to the still prohibitive training costs, several works have studied the personalization of existing pre-trained model weights for new concepts, such as Dreambooth [21], Textual Inversion [6], and Custom Diffusion [14]. Other works have studied enabling new ways to control these models for tasks such as subject-oriented editing [18, 35, 10]. Or focusing on more general image editing based on text-based prompt changes [33, 8, 34]. However, most of these techniques aim at semantic changes or require text-based prompt changes. Text-less exemplar-based stylistic edits have not commonly been explicitly explored with diffusion models. Recently,

PARASOL [31] has used an ALADIN-ViT style embedding to perform style-based image generation, with some capabilities of maintaining content structure.

### 3 Method

To push beyond the traditional boundaries of texture-only style transfer, we wish to leverage the significant learned model priors such as Stable Diffusion [20], having been trained on large amounts of data, with typically inaccessible amounts of compute. In our approach, as shown in Figure 2 we freeze the pre-trained weights and train several modules of fully connected layers in each UNet self-attention block. We interleave pre-extracted content noise used for shapes and composition and the style attention values from the style image. These are used across reverse diffusion timesteps, generating a final stylized image using content and style information extracted from the interleaved data.

#### 3.1 Preliminary analysis of style information in attention space

Prior work [34] has shown that early diffusion timesteps affect an image’s global structural and compositional information, whereas later timesteps affect local fine details. Inspired by this, we set out to determine which timesteps of the diffusion process control style and which control content.

Given a lack of research around exemplar-based Neural Style Transfer with diffusion models, we use a prompt-based model, prompt-to-prompt [8], to carry out this visualization. We use ChatGPT [2] to generate 20 content prompts, and we further define 10 style modifier prompts. With the prompt-to-prompt pipeline (operating over the Stable Diffusion LDM weights), we use the content prompts to generate reference content images, and we combine each content prompt with each style modifier prompt to re-generate the content images with the different explicitly defined styles still using prompt-to-prompt [8]. At the end of the process, we have 20 reference content example images and 200 "stylized" images. During the generation process, we extract attention values for analysis. We average the differences between the content example images’ attention values and each of their 10 stylized variants’, at each timestep. Fig 1 in the supplementary materials visualizes the average differences between these attention values at the diffusion timesteps. The red indicates a larger difference between the original content image and its stylized versions. Given that the structural and compositional information of the example content and their "stylized" counterparts is similar, we can infer that the stylistic differences relate to the higher attention discrepancies found at the later timesteps. This preliminary exploratory experiment clarifies the different effects of diffusion timesteps across the LDM generation process.

An additional preliminary experiment using these prompt-to-prompt images is an analysis of where the style information is captured in the LDM activations. We explicitly focus on the attention mechanism, where  $\mathcal{Q}$ ,  $\mathcal{K}$ , and  $\mathcal{V}$  values are used in the attention process [32]. We generate a base non-stylized image with the content prompt and then stylized variants with style modifier prompts. We extract attention values from the content-only prompt generation and replace the attention values of the stylized generation with those from the content-only generation. Doing so re-generates the original, non-stylized image. However, in our analysis, we observe that interpolating between the  $\mathcal{V}$  self-attention values of the content/style-modified generations (while using only the original content values for the rest) can provide control over the stylization strength. From this experiment, we can infer that most, if not all, style information is captured from just the  $\mathcal{V}$  self-attention values in the LDM. We visualize examples of this style interpolation in the supplementary materials.

#### 3.2 DIFF-NST real image inversion

Our work aims to perform style transfer of existing real user-provided images. As such, the re-styled synthesized image must stay faithful to the provided content image in terms of overall composition and structure. This means we must *edit* the image rather than *re-generate* a semantically similar approximation. We invert the content image through the LDM, similar to previous works such as prompt-to-prompt [8] and diffusion disentanglement [34]. This inversion process extracts the predicted noise at each timestep, as predicted by the UNet modules. To reconstruct the same image using an LDM, this content noise can be injected into the reverse diffusion process, replacing the LDM noise predictions at multiple timesteps. The more timesteps the noises are applied to, the better the reconstruction fidelity, with less freedom of input from the LDM. As shown in the diffusion

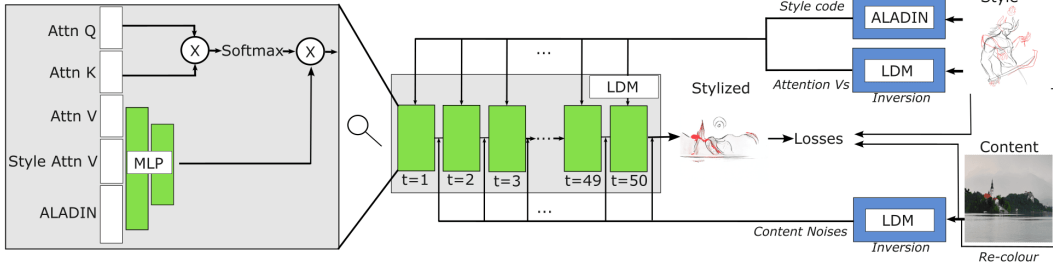


Figure 2: High level visualization of our diffusion-based neural style transfer process. (left) Trainable MLP in the self-attention blocks of the LDM Unet modules. (right) Attention values and ALADIN style codes are extracted from the style image. The content image is re-colored by the style image, after which the LDM extracts content noises from it. These are interleaved into the reverse diffusion process at multiple time steps to generate a stylized version for the loss objective. Green modules are trainable, and blue modules are frozen.

disentanglement work [34], applying changes to the diffusion values from an earlier timestep allows more significant change in image structure.

Similar to these previous works, we use 50 time steps for the forward (inversion) and reverse (re-generation) diffusion processes. However, unlike these previous works, we interleave this noise starting from an earlier time, step 5, rather than 16, to improve reconstruction quality. We apply noise until step 45 instead of 50 to allow the model to self-correct some artifacts. Also, unlike prior work, we do not set the LDM predicted noises to zero for timesteps where pre-extracted content noises are not injected into the diffusion process. We aim to allow the model to generate new details to leverage its learned priors.

A notable trait of image-to-image and image-inversion with diffusion models is that color information is not disentangled from overall image structure across timesteps, as it is with feature activation across layers of a VGG model, for example. Thus, color information must be explicitly handled before inversion. Similar to previous works [38, 25], we pre-adjust the color of the content image through mean and covariance matching. We do this dynamically during training before inversion.

A final consideration is that we aim to perform prompt-less execution of LDMs, given our use of exemplar images for both content and style. As such, we only need to use the model’s unconditional capabilities. Latent Diffusion Models execute two iterations of their model: one with no prompt conditioning and one with prompt conditioning. The output of both branches is joined at every time step via the classifier free guidance (CFG). This exposes prompt control via this adjustable strength. Given that we aim not to use any text prompts anywhere in the process, we, therefore, altogether disable the prompt-conditioned branch of the model execution and use only the un-conditional branch for both inversion and reverse diffusion. The process would function the same if the text prompt were fixed to a generic prompt throughout or if CFG was zero, but this approach saves on compute.

### 3.3 Attention manipulation

We train a set of MLPs across each self-attention module in the LDM UNet blocks. We do not wish to re-train or fine-tune the LDM weights due to large compute/financial requirements. Instead, we train several smaller modules to *hijack* part of the LDM process, similar to how content noises are injected into the diffusion process. We directly target the attention process’s  $\mathcal{V}$  values, generating brand new values for the remaining process to use. We chose the  $\mathcal{V}$  values following our initial exploratory experiments with existing text-prompt-based diffusion image editing techniques such as prompt-to-prompt, where we observed that interpolation between  $\mathcal{V}$  values only is enough to induce stylistic changes between content prompts and style-modified prompts.

Before our reverse diffusion process, similar to the real content image inversion to collect the noise predictions for reconstruction, we additionally invert and fully reconstruct the real style image through the LDM. This time, instead of collecting the predicted noises, we collect the predicted attention  $\mathcal{V}$  values at every location and timestep and interleave them into the reverse diffusion process. Here, the MLPs generate the new  $\mathcal{V}$  values based on an input consisting of the current  $\mathcal{V}$  values, the corresponding  $\mathcal{V}$  values at the same location and timestep of the style image, and the ALADIN style code of the style image, which we also pre-extract. We use both the style attention values and ALADIN, as this provides both global and local style information. Using only the attention values



Figure 3: Visualization of style code ablation. The more disentangled ALADIN-NST [26] embedding carries over less semantic information from the style images.

induces a similar style transfer. Anecdotally, however, using both sources of style information leads to a higher overall perceived quality of style transfer. We use the more recent ALADIN-NST [26] variant of ALADIN, as it is more disentangled, capturing less content information. This helps to avoid semantic content creeping into the stylized image from the style image, as shown in Fig 3.

A final consideration is that we only apply this attention manipulation process to the UNet decoder/upscaling layers, as per ControlNet [39]. Similar to their findings, we notice no perceivable differences in the output quality, but the VRAM consumption and compute costs are lower.

### 3.4 Training process

Diffusion models are typically trained one random timestep at a time, given the nature of focusing the training on noise predictions at individual timesteps. In our case, however, such timestep-localized deltas are not as easy to isolate. We can only guide our model during training based on the final de-noised output image. Moreover, well known existing style losses have been designed to operate in pixel space. They are, therefore, not directly applicable to latent space - though this may be an area of potential future study.

Therefore, we build our training process around unrolling the entire diffusion process, from starting to ending timesteps. We then decode the latent values into pixel space, where we can finally apply standard NST losses amongst the stylized and real style images from our style dataset. We opt to keep these style learning losses similar to previous works to reduce variables and uncertainty from our work. We follow a similar training objective to recent works such as NeAT [25], ContraAST [4], and CAST [41] - described in detail in Sec 3.5. We can report some negative results in using the LDM UNet as a noised feature extractor for computing a VGG-like style loss to avoid the unrolling process - the features extracted by the UNet did not accurately model the image style features.

### 3.5 Training objective

We train our model using well explored training objectives from traditional NST methods to focus solely on the model technique - we most similarly follow training objectives resembling those of NeAT [25], ContraAST [4], and CAST [41]. Between style and stylized images, we use a VGG [30] style loss (Eq. 1), identity loss (Eq. 4), contrastive loss (Eq. 7), sobel-guided patch discriminator (Eq. 9), domain-level discriminator (Eq. 2), and ALADIN loss (Eq. 6). Between the stylized and content images, we use a perceptual loss (Eq. 3), contrastive loss (Eq. 8), and identity loss (Eq. 5). We use Sobel guidance for the patch discriminator, as per NeAT.

Equation 1 shows the VGG style loss, with  $\mu$  and  $\sigma$  representing the mean and standard deviation of extracted feature maps,  $I_s$  represents style image from the style dataset  $S$ ,  $I_c$  represents a content image from the content dataset  $C$  after the color adjustments, and  $I_{sc}$  represents the stylized image.

$$\mathcal{L}_s := \lambda_{\text{vgg}} \left( \sum_{i=1}^L \|\mu(\phi_i(I_{sc})) - \mu(\phi_i(I_s))\|_2 + \|\sigma(\phi_i(I_{sc})) - \sigma(\phi_i(I_s))\|_2 \right) \quad (1)$$

Eq 2 represents the domain-level adversarial loss, as per ContraAST [4], learning to discriminate between generated stylized images and real artworks. Here, a discriminator  $\mathcal{D}$  operates over the stylized image, following our model  $M$  modules. Eq 3 details standard perceptual loss, where  $\phi_i$  represents the pre-trained VGG-19 layer index.

$$\mathcal{L}_{adv} := \lambda_{\text{adv}} \left( \mathbb{E}_{I_s \sim \mathcal{S}} [\log(\mathcal{D}(I_s))] + \mathbb{E}_{I_c \sim \mathcal{C}, I_s \sim \mathcal{S}} [\log(1 - \mathcal{D}(M(I_s, I_c)))] \right) \quad (2)$$

$$\mathcal{L}_{\text{percep}} := \lambda_{\text{percep}} (\|\phi_{\text{conv4}_2}(I_{sc}) - \phi_{\text{conv4}_2}(I_c)\|_2) \quad (3)$$

Eqs 4 and 5 show MSE identity losses between the reconstructed images and the style or content images, respectively. Eq 6 shows the ALADIN loss, with  $\mathcal{A}$  representing the ALADIN model.

$$\mathcal{L}_{\text{id}_s} := \lambda_{\text{identity}} (\|I_{ss} - I_s\|_2) \quad (4)$$

$$\mathcal{L}_{\text{id}_c} := \lambda_{\text{identity}} (\|I_{cc} - I_c\|_2) \quad (5)$$

$$\mathcal{L}_{\text{aladin}} := \lambda_{\text{aladin}} (\|\mathcal{A}(I_{sc}) - \mathcal{A}(I_s)\|_2) \quad (6)$$

Eqs 7 and 8 show contrastive losses as detailed in Sec 4.1, similar to [4] and [41], where  $l_s$  and  $l_c$  are extracted style/content embeddings respectively, using a projection head, and  $\tau$  is the temperature hyper-parameter. The contrastive losses are applied over the averaged attention values per timestep.

$$\mathcal{L}_{s\_contra} := \lambda_c \left( -\log \left( \frac{\exp(l_s(s_i c_j)^T l_s(s_i c_x) / \tau)}{\exp(l_s(s_i c_j)^T l_s(s_i c_x) / \tau) + \sum \exp(l_s(s_i c_j)^T l_s(s_m c_n) / \tau)} \right) \right) \quad (7)$$

$$\mathcal{L}_{c\_contra} := \lambda_c \left( -\log \left( \frac{\exp(l_c(s_i c_j)^T l_c(s_y c_j) / \tau)}{\exp(l_c(s_i c_j)^T l_c(s_y c_j) / \tau) + \sum \exp(l_c(s_i c_j)^T l_c(s_m c_n) / \tau)} \right) \right) \quad (8)$$

The  $\mathcal{L}_p$  term defined in Eq 9 is our patch discriminator  $D_{\text{patch}}$  loss, guided by Sobel Maps ( $SM$ ).

$$\mathcal{L}_p = \lambda_{\text{patch}} \left( \mathbb{E}_{I_s \sim \mathcal{S}} [-\log(D_{\text{patch}}(\text{crop}(I_{sc}, SM_{sc}), \text{crops}(I_s, SM_s)))] \right) \quad (9)$$

Our final combined loss objective is shown in 10 where each term is weighted by their respective  $\lambda$  term. The loss weights are as follows:  $\lambda_{\text{vgg}} = 0.5$ ,  $\lambda_{\text{adv}} = 5$ ,  $\lambda_{\text{percep}} = 6$ ,  $\lambda_{\text{identity}} = 100$ ,  $\lambda_{\text{aladin}} = 10$ ,  $\lambda_c = 1$ ,  $\lambda_{\text{patch}} = 10$ ,  $\lambda_1 = 0.25$ ,  $\lambda_2 = 0.75$ .

$$\mathcal{L}_{\text{final}} := \mathcal{L}_s + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{percep}} + \mathcal{L}_{\text{id}_s} + \mathcal{L}_{\text{id}_c} + \mathcal{L}_{\text{aladin}} + \mathcal{L}_{s\_contra} + \mathcal{L}_{c\_contra} + \lambda_1 \mathcal{L}_{p\_simple} + \lambda_2 \mathcal{L}_{p\_complex} \quad (10)$$

## 4 Experiments and Evaluation

Neural style transfer using diffusion models is a nascent sub-field of research. As such, very few works study this new direction, much less via prompt-less techniques. Despite not being a strictly NST model, PARASOL [31] is currently the only suitable method we can baseline against. We additionally compare against three recent "traditional" NST techniques, NNST [13], NeAT [25], and CAST [41]. These techniques have focused on texture-based style transfer, and as such, their stylized outputs contain a much better match between the style and stylized images' textures. This is reflected in metrics such as SIFID [29], used in NST literature so far that precisely measure such correlations.

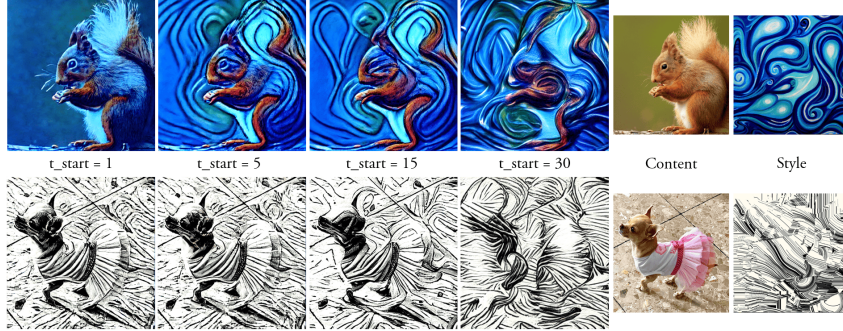


Figure 4: Controlling the style-based content deformation of the stylized image at inference time by varying the starting timestep to apply pre-extracted content noises from the content image inversion.

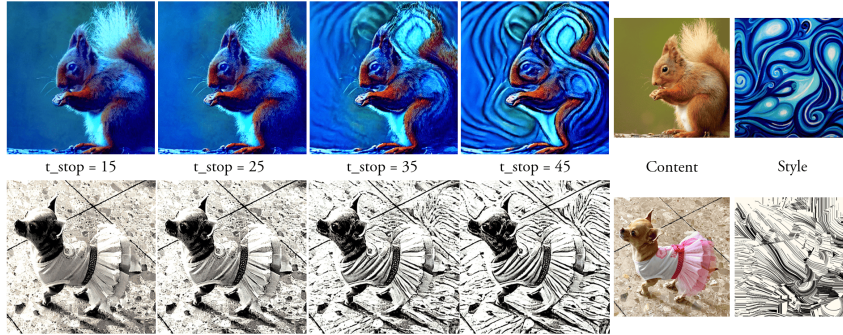


Figure 5: Controlling the stylization strength by varying the stopping timestep at which to apply attention modifications. This inference-time control vector affects the content deformity less than varying noise injection timesteps.

The unrolled approach of training diffusion models does incur a high computation cost. Our technique can train over an LDM at 512px resolution on a GPU with 48GB VRAM at batch size 1. We use gradient accumulation 8 to raise the effective batch size to 8. Inference at 512px fits on 24GB VRAM. We train our model for 3 weeks on a single A100. Like NeAT [25], we use the BBST-4M dataset they introduce, due to its great variety of style data, covering not just fine-art imagery as more commonly found in other datasets. Due to our method and NeAT having been trained using BBST-4M, we aim to use a test set with no overlap with training data. We use the test set from ALADIN-NST [26], which was collected as a test set not overlapping with previous datasets such as BBST-4M. The test set contains 100 content and 400 style images, resulting in 40,000 stylized images. We collect quantitative metrics in Table 1, measuring SIFID [29] and Chamfer for style and color consistency with the style image respectively, and LPIPS [40] for structure consistency with the content. Due to long-running generation times for our method and those of multiple baselines, we randomly sub-sample and use 5,000 images.

Table 1: Quantitative metrics. Lower is better. ↓

Model	LPIPS ↓	SIFID ↓	Chamfer ↓
NeAT [25]	0.624	0.880	24.970
CAST [41]	0.632	1.520	43.864
NNST [13]	0.633	2.007	53.328
PARASOL [31]	0.716	3.297	105.371
DIFF-NST (Ours)	0.656	2.026	45.777

Table 2: User studies for our model, for individual ratings (out of 5), and 5-way preferences (%). Higher is better. ↑

Model	Content Rating ↑	Style Rating ↑	Content Preference ↑	Style Preference ↑
NeAT [25]	3.271	2.952	32.222	26.000
CAST [41]	3.031	2.863	16.756	16.133
NNST [13]	2.937	2.712	21.200	17.778
PARASOL [31]	2.301	2.257	12.400	9.556
DIFF-NST (Ours)	2.751	2.973	17.422	30.533

We present a qualitative random sample of stylizations in Fig 9 and the supplementary materials. We visualize stylizations using our method, the closest technically related work PARASOL [31], and some traditional NST techniques.

The most impactful ablation to report on is experimenting with the style embedding used alongside the style attention values. We show some comparative examples in Fig 3, having tested the regular ALADIN-ViT style embedding and the more disentangled ALADIN-NST variant. The ViT variant introduces some content features from the style image into the stylized image when these features have strong activations - most commonly occurring with faces. Though rare, we mitigate this issue using a fully disentangled style embedding, ALADIN-NST.

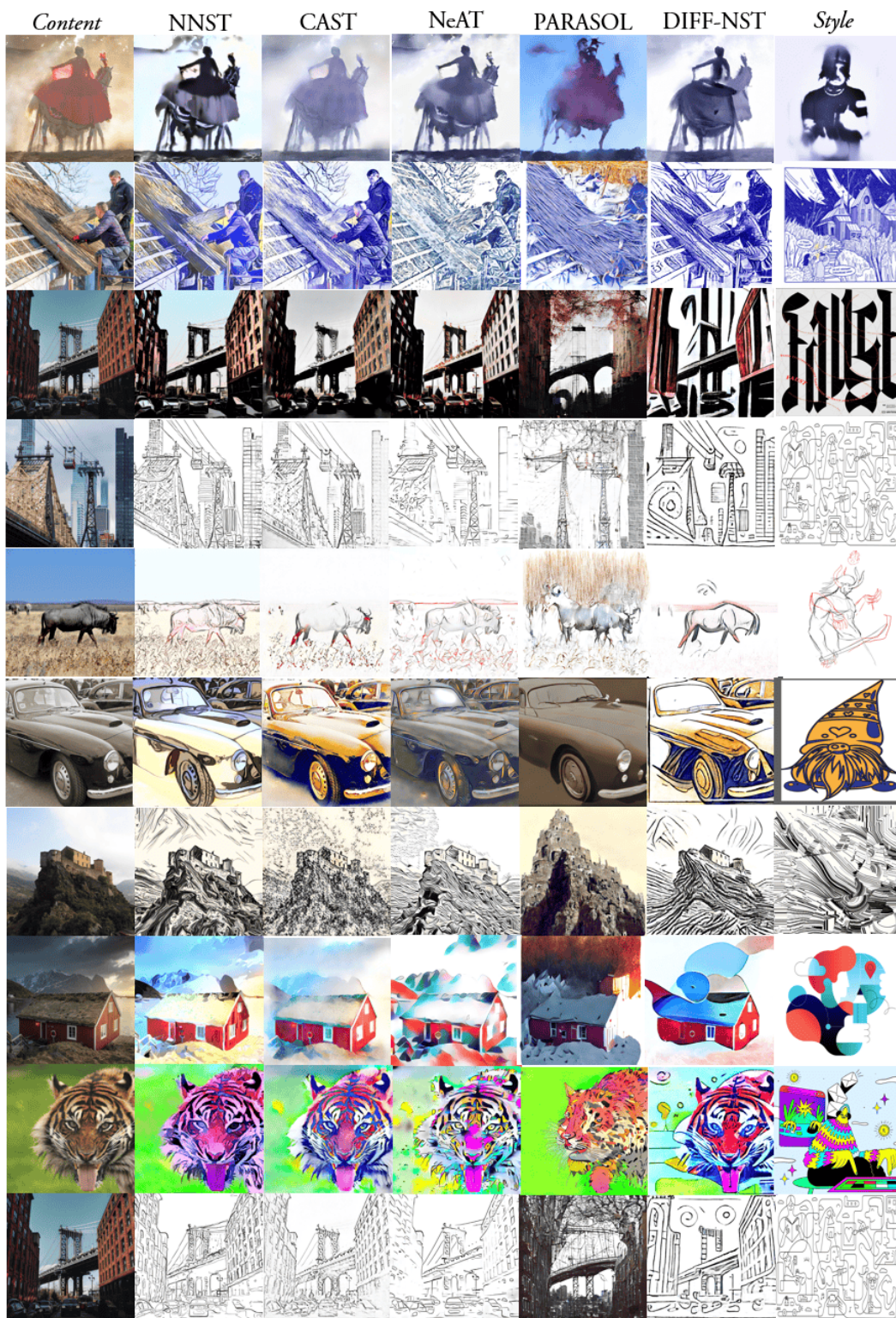


Figure 6: Deformable style transfer, comparing to NNST [13], CAST [41], NeAT [25], and PARASOL [31]. All our figures are generated using images from the ALADIN-NST test set, which were not seen during training. More in the supplementary materials.



## 4.1 User studies

We undertake a pair of user studies to gauge real life human preference amongst our method and the baselines. First, we carry out an individual rating exercise, measuring the content fidelity between the content image and the stylized image, and separately measuring the style consistence compared to the style image. Second, we carry out a 5-way comparison, where we ask workers to select their best preference from randomly shuffled samples. We bin the ratings in the individual exercise to five levels, and we explicitly instruct what each rating level should represent. We include the definitions in the supplementary material. We randomly sub-sample 750 stylized samples from the test set and compare our method against each baseline on Amazon Mechanical Turk (AMT). We collect and average our responses over 5 different workers for each comparison, and show our results in Table 2.

The results indicate that workers are scoring our DIFF-NST method low on the content information, in both the ratings and preference studies. This is a positive result, as it highlights our technique’s more substantial content deformation. The only model which scored lower is PARASOL. However, as seen in our visual comparison figures, PARASOL tends to make significant conceptual changes to the depicted content. It is not so much a technique for style transfer as it is for style-inspired re-generation of similar semantic content. The results for our style-focused experiments indicate that workers prefer our method to baselines in both individual ratings and 5-way preference studies, which signifies a successful transfer of style while still deforming the content.

## 4.2 Inference controls

One key strength of our diffusion-based NST method is control over the structural deformity in the represented content concerning the style image. The reference content information is injected into the diffusion process by applying noises at each time step, pre-extracted from the content image inversion. With diffusion models, the early time steps strongly affect the significant structural components of the image, whereas the later timesteps affect lower level textural information. Therefore, by varying the starting timestep at which these pre-extracted content noises are applied, we can adjust, at inference time, how much the style should deform the content structure. This effect is difficult to evaluate quantitatively, but we show two examples in Fig 4.

An alternative vector of inference-time control is varying the diffusion timesteps in which our method’s attention replacement happens. By stopping at earlier timesteps, less style information is injected into the diffusion process, reducing the stylization strength. Unlike reducing content noise injection, this approach maintains the content structure better and more directly targets the style properties instead of structure. We show examples of this second approach in Fig 5, using the same example images as in Fig 4 for clarity.

## 5 Limitations and Conclusions

One limiting factor of our approach is that textures are not matched to the style image with as much detail and fidelity as traditional NST approaches. This can, however, be alleviated by introducing a conventional NST approach into the pipeline as a post-processing step.

Though rare, due to the one-to-one mapping between the content and style attention values, some structure from some style images sometimes creeps into the stylized image. We can report negative results experimenting with Neural Neighbours [13] in attention space, which resolved this issue, but only at the cost of worse overall stylization quality. This is an area of potential future improvement.

One of the principal challenges with our method has been computation due to the unrolled nature of the reverse diffusion process during training. Future work can explore the adaptation of the style training objective to the latent space instead of pixel space, enabling non-unrolled training.

## 6 Broader Impact

Neural techniques for artistic image editing and generation offer new tools and capabilities for skilled artists to take their work further than before. However, this does make the field easier to enter as a novice. As such, existing novice-level artists may find more competition in this space, reducing work opportunities. As digital art emerged, it offered new capabilities to artists with new tools at the detriment of some artists using physical mediums. Neural techniques can similarly open up new genres of art while reducing some opportunities for some existing digital artists.

## References

- [1] Stability ai. <https://stability.ai/>. (Accessed on 05/16/2023). (Cited on page 2)
- [2] Chatgpt. <https://chat.openai.com/>. (Accessed on 05/16/2023). (Cited on page 3)
- [3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. (Cited on page 2)
- [4] Haibo Chen, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=hm0i-cunzGW>. (Cited on page 2, 5, 6)
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL <https://arxiv.org/abs/2010.11929>. (Cited on page 2)
- [6] 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, 2022. (Cited on page 2)
- [7] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. CVPR*, pages 2414–2423, 2016. (Cited on page 2)
- [8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. (Cited on page 2, 3)
- [9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. ICCV*, 2017. (Cited on page 2)
- [10] Kangyeol Kim, Sunghyun Park, Junsoo Lee, and Jaegul Choo. Reference-based image composition with sketch via structure-aware diffusion model, 2023. (Cited on page 2)
- [11] Sunnie S. Y. Kim, Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Deformable style transfer, 2020. (Cited on page 2)
- [12] Nicholas Kolkin, Jason Salavon, and Greg Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity, 2019. (Cited on page 2)
- [13] Nicholas Kolkin, Michal Kucera, Sylvain Paris, Daniel Sykora, Eli Shechtman, and Greg Shakhnarovich. Neural neighbor style transfer, 2022. URL <https://arxiv.org/abs/2203.13215>. (Cited on page 1, 2, 6, 7, 8, 9)
- [14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2022. (Cited on page 2)
- [15] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *CoRR*, abs/1705.08086, 2017. URL <http://arxiv.org/abs/1705.08086>. (Cited on page 2)
- [16] Xuan Luo, Zhen Han, Lingfang Yang, and Lingling Zhang. Consistent style transfer. *CoRR*, abs/2201.02233, 2022. URL <https://arxiv.org/abs/2201.02233>. (Cited on page 2)
- [17] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. *CoRR*, abs/1812.02342, 2018. URL <http://arxiv.org/abs/1812.02342>. (Cited on page 2)
- [18] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-guided diffusion with inside-outside attention, 2023. (Cited on page 2)
- [19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. (Cited on page 2)
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. (Cited on page 2, 3)
- [21] 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, 2023. (Cited on page 2)

- [22] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: All layer adaptive instance normalization for fine-grained style similarity. *arXiv preprint arXiv:2103.09776*, 2021. (Cited on page 2)
- [23] Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, Jo Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, Zhe Lin, and John Collomosse. Stylelabel: Artistic style tagging and captioning, 2022. URL <https://arxiv.org/abs/2203.05321>. (Cited on page 2)
- [24] Dan Ruta, Andrew Gilbert, Saeid Motiian, Baldo Faieta, Zhe Lin, and John Collomosse. Hypernst: Hyper-networks for neural style transfer, 2022. URL <https://arxiv.org/abs/2208.04807>. (Cited on page 2)
- [25] Dan Ruta, Andrew Gilbert, John Collomosse, Eli Shechtman, and Nicholas Kolkin. Neat: Neural artistic tracing for beautiful style transfer, 2023. URL <https://arxiv.org/abs/2304.05139>. (Cited on page 1, 2, 4, 5, 6, 7, 8, 12)
- [26] Dan Ruta, Gemma Canet Tarres, Alex Black, Andrew Gilbert, and John Collomosse. Aladin-nst: Self-supervised disentangled representation learning of artistic style through neural style transfer, 2023. (Cited on page 2, 5, 7)
- [27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. (Cited on page 2)
- [28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. (Cited on page 2)
- [29] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. *CoRR*, abs/1905.01164, 2019. URL <http://arxiv.org/abs/1905.01164>. (Cited on page 6, 7)
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. (Cited on page 5)
- [31] Gemma Canet Tarrés, Dan Ruta, Tu Bui, and John Collomosse. Parasol: Parametric style control for diffusion image synthesis, 2023. URL <https://arxiv.org/abs/2303.06464>. (Cited on page 1, 3, 6, 7, 8)
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL <http://arxiv.org/abs/1706.03762>. (Cited on page 3)
- [33] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models’ latent space, with applications to CycleDiffusion and guidance. In *ArXiv*, 2022. (Cited on page 2)
- [34] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models, 2022. (Cited on page 2, 3, 4)
- [35] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models, 2022. (Cited on page 2)
- [36] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. (Cited on page 2)
- [37] Nick Zangwill. *The Metaphysics of Beauty*. Cornell University Press, 2001. ISBN 9780801438202. URL <http://www.jstor.org/stable/10.7591/j.ctv1nhmzk>. (Cited on page 2)
- [38] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. Arf: Artistic radiance fields, 2022. URL <https://arxiv.org/abs/2206.06360>. (Cited on page 4)
- [39] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. (Cited on page 5)

- [40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. (Cited on page 7)
- [41] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*, 2022. (Cited on page 1, 2, 5, 6, 7, 8)

## A Prompt-to-prompt Analysis

The base content captions partially generated using ChatGPT for the prompt-to-prompt analysis experiments are:

1. A squirrel eating a burger
2. A hamster on a skateboard
3. A toy next to a flower
4. A car driving down the road
5. A giraffe in a chair
6. A bear wearing sunglasses
7. An octopus in a space suit
8. A hedgehog getting a haircut
9. A sloth running a marathon
10. A cat posing like napoleon
11. A dog with a beard, smoking a cigar
12. A bee flying underwater next to fish
13. A fish with a hat, playing a guitar
14. A bird with a bowtie, playing a saxophone
15. A turtle with a top hat, playing a piano
16. A frog with a cowboy hat, playing a banjo
17. A mouse with a sombrero, playing a trumpet
18. A snake with a beret, playing a violin
19. A rabbit with a fedora, playing a cello
20. A squirrel with a baseball cap, playing a drum

The style modifiers are:

1. A van gogh painting of
2. A graphite sketch of
3. A neon colourful pastel of
4. A minimal flat vector art illustration of
5. A watercolour painting of
6. A psychedelic inverted painting of
7. A pop-art comic book panel of
8. A neoclassical painting of
9. A cubist abstract painting of
10. A surreal dark horror painting of

We visualize results from the preliminary prompt-to-prompt analysis experiments, in Fig 8. The figure shows the first content prompt for the base content, with the subsequent rows interpolating towards style-modified prompts using style prompt modifiers 1, 4, 2, and 8. Although not directly relevant to our study, it was also interesting to note that the stylization strength could be pushed beyond the default strength by pushing the interpolation into over-drive, similar to the technique presented in NeAT [25].

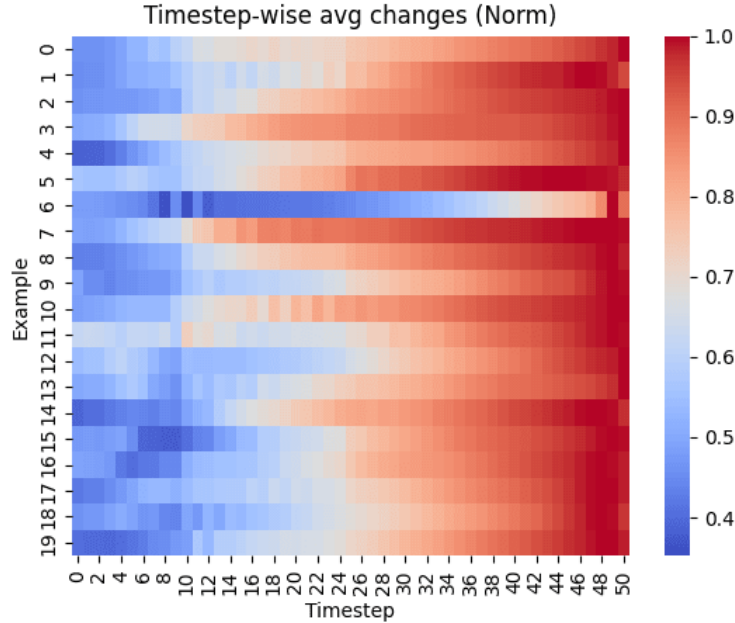


Figure 7: The averaged normalized difference in attention values between reference content and 10 stylized content images in the prompt-based prompt-to-prompt model. The higher difference values (in red) in the later timesteps visualize the effect that earlier timesteps affect coarser structural details, whereas later timesteps affect lower level textural details.

## B Additional details on user studies

We carried out two user studies: an individual rating exercise with defined rating levels, and a 5-way preference comparative exercise. For each, we executed the experiments once for the content, and once for the style.

Our content-focused rating exercise asks the following question: "A photo has been re-generated with a different style. Please rate the structure details of the new image, 1 to 5 as follows:", where we next define the expected judgement criteria for each rating level as follows:

1. The structure is different
2. The structure slightly resembles the photo
3. The structure mostly resembles the photo
4. The structure is the same
5. The structure is the same, including small details

Our style focused rating exercise asks the following question: "A photo has been transformed into the style of the artwork. Please rate the quality of the style, 1 to 5 as follows:", where the rating definitions are:

1. The style is not recognisable
2. The style is recognisable
3. The colours match
4. The textures match
5. The shapes match

The 5-way comparative study presents the following question for the content-focused experiment: "A photo has been re-generated with a different style in 5 ways. Please select the highest quality reconstruction of the photo's structure details", and the following for the style-focused experiment: "A photo has been re-generated with a different style in 5 ways. Please select the most similar artistic style to the artwork"

The workers were fairly compensated. We used 5 different workers for each stylized image, for each question.

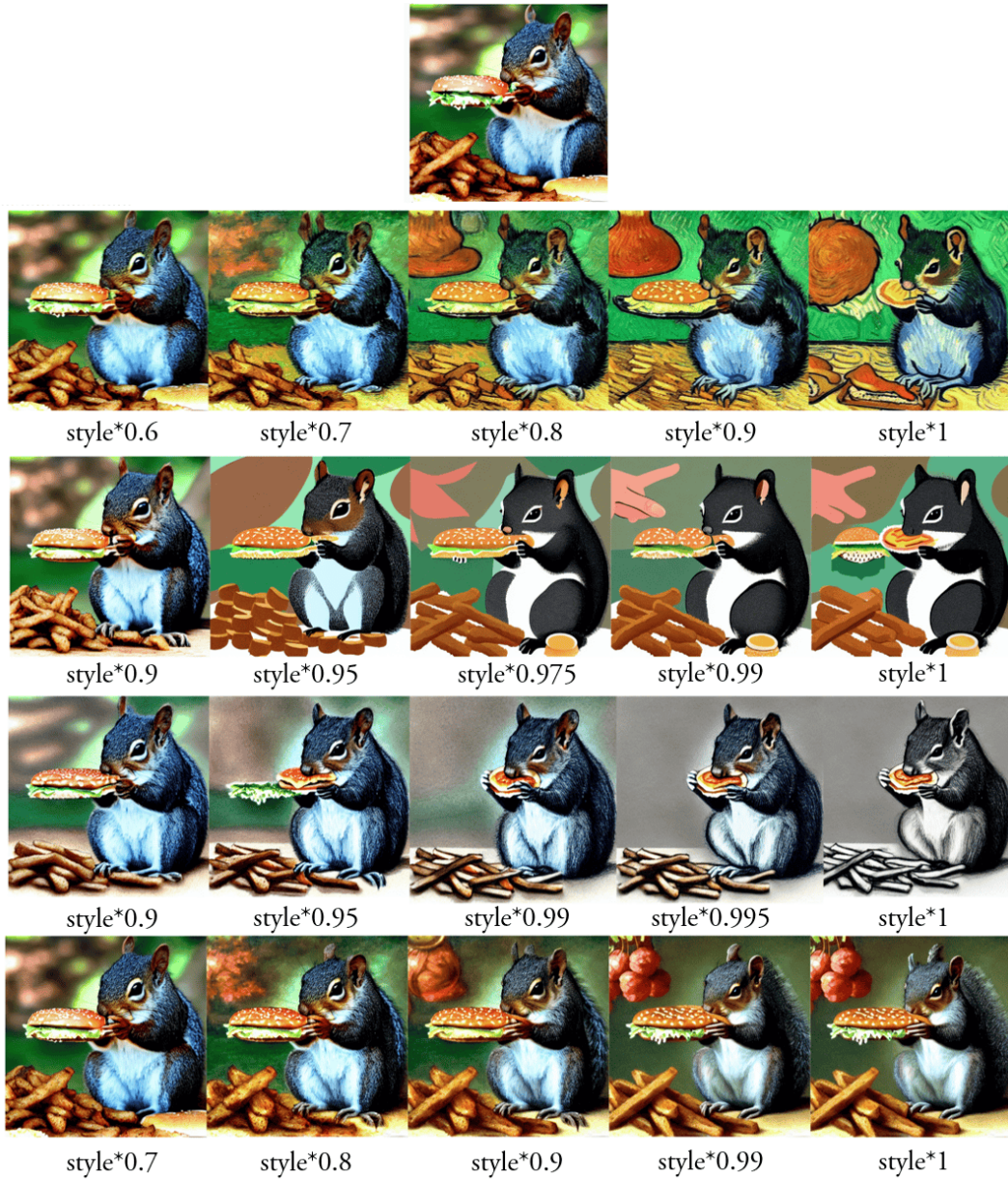


Figure 8: Visualization of stylization interpolation using prompt-to-prompt, changing only the attention V values. The stylization strength displayed represents the interpolation strength between the content and style attention values.

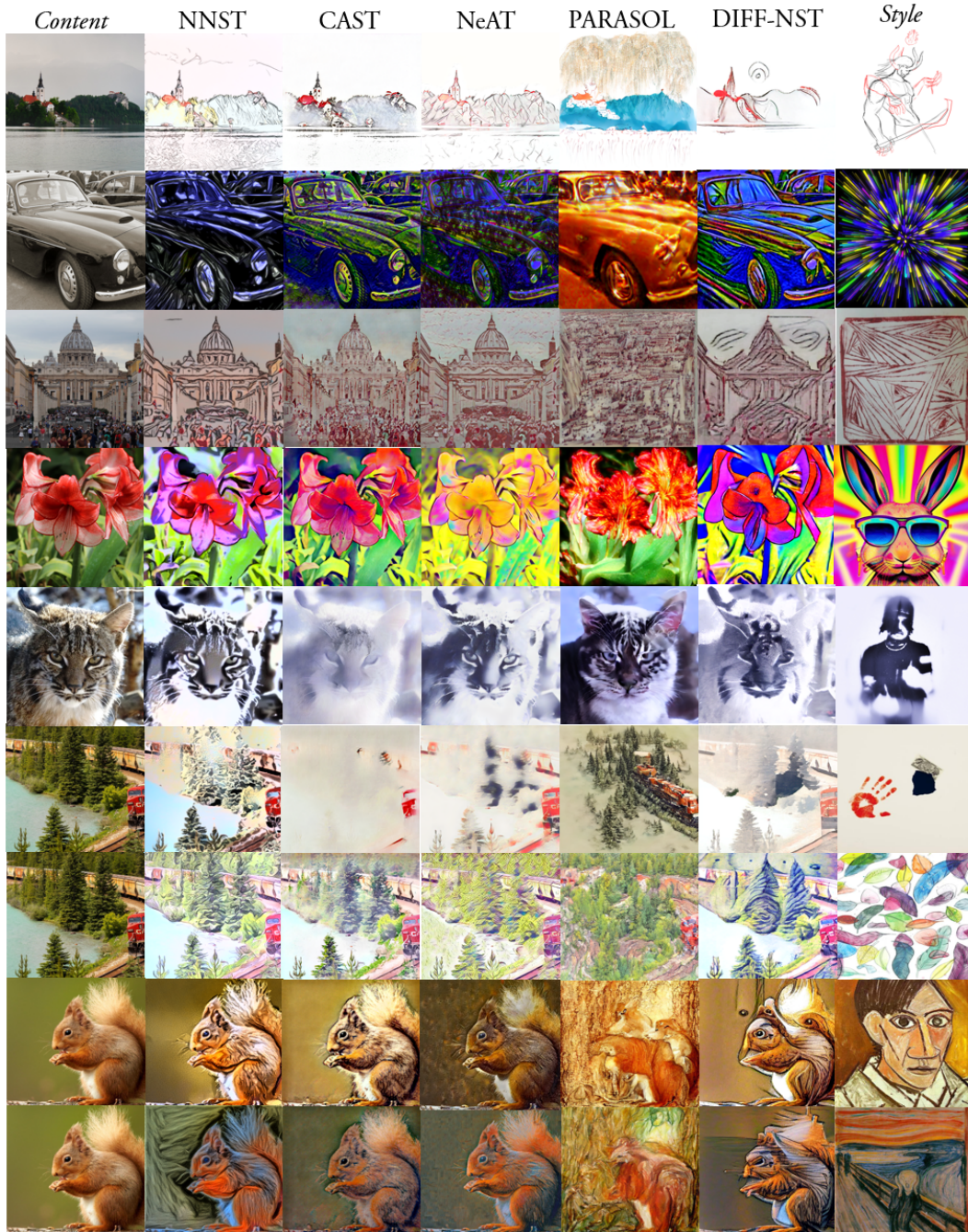


Figure 9: Additional deformable style transfer comparisons