# Opt-In Art: Learning Art Styles Only from Few Examples

**Hui Ren**[1,2,*,†] **Joanna Materzyńska**[*] **Rohit Gandikota**[3] **Giannis Daras**[2]
**David Bau**[3] **Antonio Torralba**[2]

[1]UIUC [2]MIT [3]Northeastern University

{huiren2@illinois.edu, jomat@mit.edu, gandikota.ro@northeastern.edu,
gdaras@mit.edu, d.bau@northeastern.edu, torralba@mit.edu}

Project Webpage

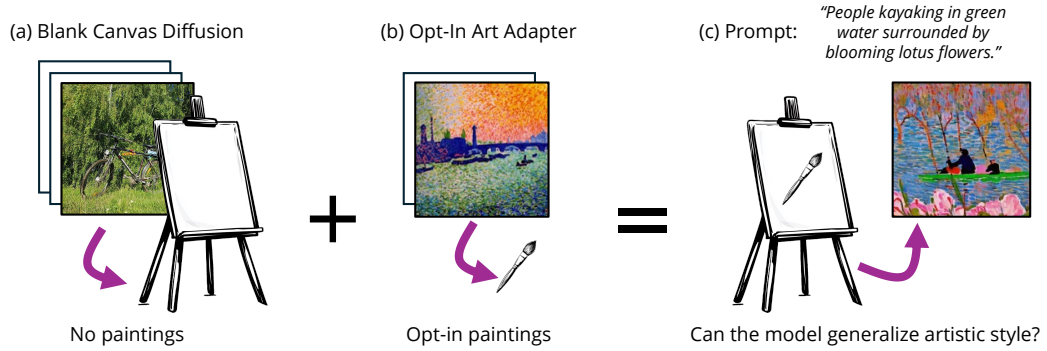



Figure 1: (a) We introduce Blank Canvas Diffusion, a carefully curated text-to-image model trained only on photographs, serving as the pretraining foundation for our model. Our study explores whether a model with no prior exposure to paintings can learn artistic styles using (b) a LoRA Art Adapter trained on a small opt-in sample of an artist's work. (c) We find that it is possible to adapt a model that is trained without paintings to generalize an artistic style, given only few examples.

## Abstract

We explore whether pre-training on datasets with paintings is necessary for a model to learn an artistic style with only a few examples. To investigate this, we train a text-to-image model exclusively on photographs, without access to any painting-related content. We show that it is possible to adapt a model that is trained without paintings to an artistic style, given only few examples. User studies and automatic evaluations confirm that our model (post-adaptation) performs on par with state-of-the-art models trained on massive datasets that contain artistic content like paintings, drawings or illustrations. Finally, using data attribution techniques, we analyze how both artistic and non-artistic datasets contribute to generating artistic-style images. Surprisingly, our findings suggest that high-quality artistic outputs can be achieved without prior exposure to artistic data, indicating that artistic style generation can occur in a controlled, opt-in manner using only a limited, carefully selected set of training examples.

[*]Equal contribution.
[†]Work done while visiting MIT CSAIL.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

# 1 Introduction

In this work, we aim to disentangle a generative model's ability to create artistic imagery from its reliance on prior exposure to human-created paintings, drawings, or illustrations. The recent success of generative models, particularly denoising diffusion models (Ho et al., 2020), introduces yet another tool for artistic expression. Unlike traditional artistic tools such as cameras or animation software, which are rarely questioned in terms of their creative agency, generative modeling challenges conventional notions of authorship and artistic ownership. The ability of these models to generate images in the style of specific artists (Gandikota et al., 2023a) raises pressing questions: Who holds the rights to AI-generated images—the original artist, the model's developer, or the user providing the prompt (Epstein et al., 2023; Whiddington, 2024)? Furthermore, how much of the artistic output is a genuine act of creation versus a statistical mimicry of preexisting styles (Epstein et al., 2023; Somepalli et al., 2023, 2024)? Unlike traditional artistic tools, generative models learn from large datasets of human-created art, enabling them to replicate styles with high fidelity (Simonite, 2022). This blurs the line between inspiration and imitation, raising concerns about creative displacement and style homogenization (Epstein et al., 2023), and highlights the need to examine their dependence on existing artworks.

Existing models are trained on massive datasets like LAION-5B (Schuhmann et al., 2022), but it remains unclear whether their artistic ability stems from direct exposure to art or general visual learning. To probe this, we train **Blank Canvas Diffusion** exclusively on photographs, using the **Blank Canvas Dataset**, a rigorously filtered collection free of paintings, drawings, and illustrations. We then apply the **Art Style Adapter**, a fine-tuning framework that enables the model to learn artistic styles from just a few examples in a controlled, **Opt-In Art** setup. We evaluate via painting similarity, crowd-sourced style judgments, and data attribution. Despite no prior exposure to art, our model effectively replicates artistic styles, performing comparably to models trained on art-rich datasets.

These findings suggest that generative models need not rely on large-scale artistic datasets to learn an artistic style. Even without prior exposure to paintings or illustrations, a model can learn to generate art through minimal fine-tuning on a few examples. This offers a pathway to creative AI systems that respect artistic consent. However, our results also complicate the opt-in model debate: if only a few reference images suffice for style replication, restricting training data may be insufficient to protect artistic styles. This raises concerns about the limits of current copyright protections (Lu et al., 2024), as artists may still have little control over the use of their style. Future regulatory discussions may need to go beyond data restrictions to address attribution, consent, and fair use.

# 2 Related work

| | #sample | paintings | stamp | sculptures | digital art | logo | artwork | sketch | advertisement | drawing | illustration | installation art | mosaic art | tapestry | baroque art | art noveau | pop art | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SA-1B | 10,000 | 36 | 71 | 120 | 14 | 36 | 0 | 0 | 2 | 8 | 4 | 12 | 1 | 3 | 6 | 1 | 2 | 315 ( 3.15%) |
| Blank Canvas Dataset | 10,000 | 0 | 0 | 38 | 1 | 12 | 0 | 0 | 2 | 3 | 0 | 12 | 2 | 1 | 0 | 0 | 0 | 71 ( 0.71%) |

Table 1: Statistics of artistic images found during manual inspection of the SA-1B and Blank Canvas datasets before and after the art filter.

Diffusion models memorize more training data compared to GANs and VAEs, raising ethical concerns (Somepalli et al., 2023, 2024; Carlini et al., 2023). Several mitigation strategies have been proposed: concept erasure methods (Gandikota et al., 2023a,b; Kumari et al., 2023a), opt-out initiatives (Spawning AI Team, 2023), watermarking approaches (Zhao et al., 2023; Min et al., 2024), and training with corrupted data to avoid exact replication (Somepalli et al., 2024). Perhaps the most straightforward approach is curating copyright-free training datasets. Gokaslan et al.(Gokaslan et al., 2024) train a model on Creative Commons images, but still include artistic content. We extend this by training with minimal exposure to paintings, drawings, and artistic media, creating a diffusion model largely devoid of non-photographic art (Fig. 2). We investigate how few artistic images are required post-training to learn specific styles. To adapt our art-agnostic model, we

Figure 2: Examples of images included and excluded from the Blank Canvas dataset. The dataset is curated to remove paintings as well as artistic categories related to paintings, such as drawings and fine art. Examples of images that are close to the removal threshold are shown in b) and c).

use our Art Style Adapter based on LoRA fine-tuning (Hu et al., 2021) and Textual Inversion (Gal et al., 2022). Prior work shows methods like Textual Inversion and Dreambooth (Ruiz et al., 2023) can personalize diffusion models using few user images. Our novel finding is that these adaptation methods work even when the base model has never been trained on the target image type.

## 3 Blank Canvas Diffusion

**Blank Canvas dataset.** Most diffusion models are trained on large-scale datasets that include paintings, drawings, and digital art. Our goal is to create a large-scale dataset composed exclusively of photographs. To this end, we curate the **Blank Canvas dataset** from the SAM-LLaVA-Captions10M dataset (Chen et al., 2023), derived from SA-1B (Kirillov et al., 2023), a dataset of natural, camera-captured images with captions generated by a vision-language model (LLaVA).

Despite its photographic intent, the dataset contains incidental artwork e.g., photos of tapestries, logos, or sculptures. We therefore design a two-stage filtering pipeline to minimize art content. We exclude images via caption-based keyword filtering, then we compute CLIP-based cosine similarity scores to art-related terms and remove high-scoring images. Thresholds are determined by manual inspection. After filtering, the dataset retains 9.11M image-text pairs (4.7% removed via text, 16.7% via image filtering). Manual inspection confirms a reduction in art content: in SA-1B, 315 of 10k images originally contained artwork, versus 72 post-filtering. Similar evaluation on COCO shows a drop from 1.06% to 0.12%. We also estimate that 32.9% of images in LAION-Aesthetic v2 5+ (used to train Stable Diffusion 1.4) contain artwork, suggesting over 191M art images in its training set.

Our **Blank Canvas Diffusion** model uses a latent diffusion architecture (Rombach et al., 2022) with three modules: a VAE, a U-Net, and a Text Encoder. To ensure no exposure to artistic content, we train both the VAE and U-Net from scratch on our filtered dataset. Unlike prior models that use CLIP (Radford et al., 2021), which may encode visual artistic concepts, we use a BERT-based language, only encoder (Devlin et al., 2019). While BERT may contain textual knowledge of art, it has no visual grounding, ensuring that the model is free from any pixel-level artistic priors.

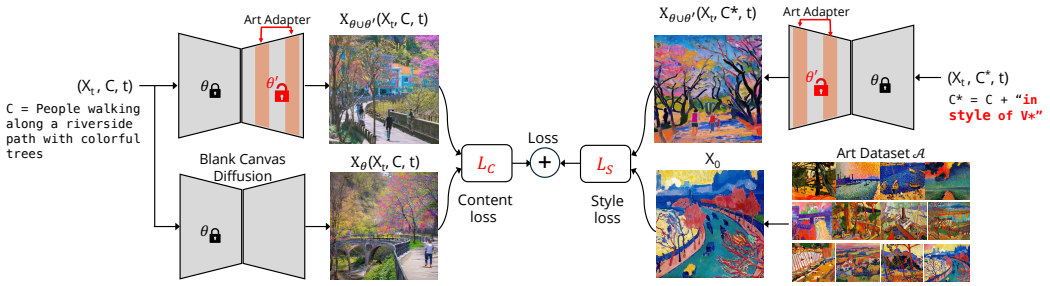



Figure 3: The generated image should match the style of a small exemplar dataset when prompted with a caption $C^*$, which includes a style prefix $V^*$. For example, if $C^*$ = *People walking along a riverside path with colorful trees in the style of* $V^*$, the image should reflect both the scene (content) and the specified artistic style. Content loss ensures that the visual elements of the prompt $C$ = *People walking along a riverside path with colorful trees* are accurate, while style loss maintains the style associated with $V^*$.

## 4 Opt In: Artistic Style Adapter

To opt-in to an art style, we train a LoRA Art-Style Adapter by collecting a few examples of paintings in that style $X_0 \in \mathcal{A}$ and caption the content of the artwork. Captioning can be done automatically or manually. To connect the newly learned style information with specific tokens in the prompt, we append "in the style of V* art" to the content prompt. We use $C^*$ for the final prompt (after our addition). To enable the model to learn this new artistic style, we fine-tune the U-Net module using LoRA (Hu et al., 2021). For a given target artistic image, we define the following loss:

$$\mathcal{L}_{\text{S}}(\theta') = \|\epsilon_{\theta\cup\theta'}(X_t, C^*, t) - \epsilon\|^2, \tag{1}$$

where $\epsilon_{\theta\cup\theta'}$ is the U-Net module with the LoRA updating weights, $t$ is the denoising time step, $X_t$ is the input image at time $t$, and $\epsilon$ is target noise. We refer to this loss as style loss, as it helps the model

implicitly learn the artistic style and link it to the style modification in the prompt. The content loss is defined as follows:

$$\mathcal{L}_{\mathrm{C}}(\theta') = \|\epsilon_{\theta \cup \theta'}(X_t, C, t) - \epsilon_{\theta}(X_t, C, t)\|\|^2 . \tag{2}$$

This loss helps maintain the prompt's content even when the style identifier is omitted from the text. Our final loss is $\mathcal{L} = \mathcal{L}_{\mathrm{S}} + w \cdot \mathcal{L}_{\mathrm{C}}$, where $w$ is the hyper-parameter for the content loss. We combine style and content losses to prevent the model from overfitting to artistic features, allowing it to learn style as a distinct component separate from content. This approach encourages the model to capture the underlying style patterns without embedding them too deeply into the content, enabling it to generate natural images when no specific style is specified. By disentangling style from content in this way, the model learns to apply styles more flexibly while preserving the core content. At inference, we control style by adjusting when art information is introduced. Injecting style earlier makes the image more stylized, later injection preserves natural details with subtle artistic elements.

## 5 Experiments

**Blank Canvas Diffusion** Our Blank Canvas Diffusion follows Stable Diffusion v1.4 architecture (Rombach et al., 2022). We train VAE from scratch, then U-Net on Blank Canvas dataset with frozen VAE, using pre-trained BERT (Devlin et al., 2019). We compare against CommonCanvas-SC (Gokaslan et al., 2024) (30M CC images) and Stable Diffusion v1-4 (600M images), in terms of data, compute time, image quality (FID score) and text and image alignment (CLIP score). Compared to existing models, Blank Canvas Diffusion is trained with significantly fewer resources. It uses only 9 mln image-text pairs and 11,432 A100 GPU hours. In contrast, CommonCanvas-SC is trained on 30 mln images with 73,800 A100 hours, and Stable Diffusion v1.4 (SD1-4) is trained on 600M images using 150,000 A100 hours. Despite this large gap in data and compute, our model achieves a CLIP score of 0.26 and FID of 12.12 on the Blank Canvas test set, competitive with CommonCanvas-SC (0.27 / 13.66) and SD1-4 (0.28 / 17.74). On the COCO-2017 test set, where our model faces a domain mismatch, it obtains 0.23 CLIP and 23.60 FID, while CommonCanvas-SC and SD1-4 perform better (0.27 / 8.23 and 0.27 / 12.54, respectively). These results underscore the competitiveness of our model despite operating at a fraction of the data and compute budget.

**Art Style Adaptation** We adapt styles using our Art Adapter with content loss weight $w = 50$. From WikiArt, we select 17 artists with distinct styles and manually curate 9–50 paintings per artist (avg. 21.88), ensuring stylistic consistency in color, brushwork, and content. We evaluate style similarity using the Contrastive Style Descriptor (CSD) (Somepalli et al., 2024). For each generated image, we compute its mean CSD score against the corresponding artist set. Content preservation is assessed via cosine similarity in ViT-based content features ($\mathrm{ViT_c}$), and CLIP score measures text-image alignment.

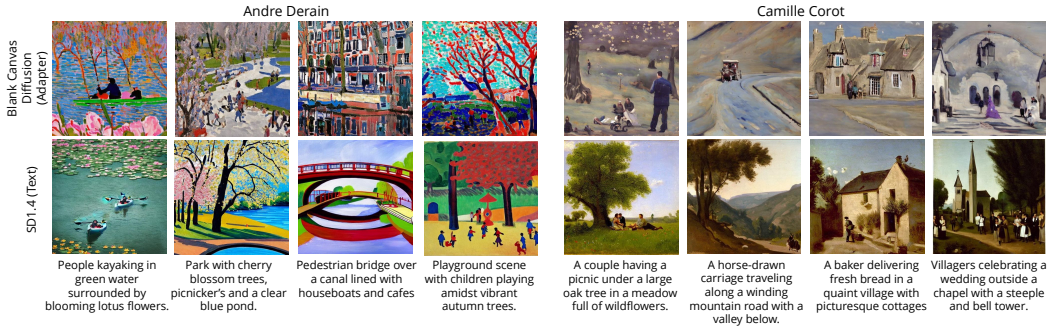



Figure 4: Comparison of Blank Canvas Diffusion art generation (top row) with Stable Diffusion 1.4 generated images (bottom row).

We sample 500 prompts and images from the LAION Pop dataset (Schuhmann and Bevan, 2023), evaluating across 17 artist style sets with results averaged per style. To validate performance, we conduct a user study via Amazon Mechanical Turk, comparing outputs from our Art Adapter and baseline methods on two tasks: Image Stylization and Artistic Style Generation. Participants are shown three reference images from a given artist and asked to select which of two outputs best matches

the style. We also include comparisons with real paintings. To mitigate bias toward photographic realism (Wang et al., 2023b), we include reliability checks, yielding 2,242 responses from 42 users after filtering.

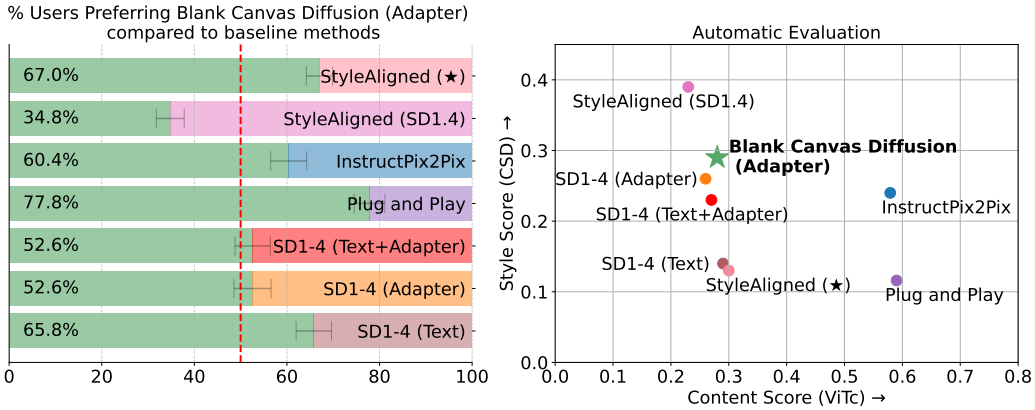



Figure 5: (Left) Results of the Perceptual User Study; Blank Canvas Diffusion with Adapter method (green bar) is preferred over image editing baselines, is on par with Adapter using an SD1.4 backbone and is favored less than StyleAligned (SD1.4). The margin of preference is narrow between the baselines. (Right) Quantitative evaluation; Blank Canvas Diffusion with the Art Adapter achieves a good trade-off between the style and content.

**Image stylization.** We evaluate our method on image stylization using the LAION Pop dataset, aiming to transfer artistic style while preserving image content. Baselines include: **SD1.4 (Adapter)** using a learned style adapter and new token, **SD1.4 (Text)** using only the artist's name, and **SD1.4 (Adapter + Text)** combining both. For all SD1.4 variants, LoRAs are applied across all blocks for optimal performance. Stylization is performed via DDIM inversion to step 800, followed by denoising with a modified prompt and adapter. We also compare to Plug and Play (Tumanyan et al., 2023), InstructPix2Pix (Brooks et al., 2023), and StyleAligned (Hertz et al., 2024), as well as CycleGAN (Zhu et al., 2017) for Monet and Van Gogh. Qualitative results for imitating Van Gogh's style are shown in SM.E. Our method matches or outperforms all baselines in both human and automatic evaluations. In user studies (Fig. 5), participants consistently preferred our outputs over Plug and Play, InstructPix2Pix, and StyleAligned on Blank Canvas Diffusion. While StyleAligned on SD1.4 showed slightly higher preference, it relies on extensive art-pretraining, unlike our model. Automatic metrics support these findings: our method balances style (CSD) and content ($\text{ViT}_c$) fidelity, whereas baselines often favor one at the expense of the other (Tab. 5 (Right)).

**Artistic Style Generation** We address the task of Artistic Style Generation, focusing on creating images in a specific artistic style. Stable Diffusion, known for its ability to replicate styles by simply prompting with artist names, serves as a baseline due to its extensive training on artworks Heikkilä (2022). Additionally, we compare with transferring a style into generated images with the StyleAligned method on both Stable Diffusion and Blank Canvas Diffusion backbones. Qualitative examples presented in Fig. 4. Blank Canvas Diffusion (Adapter) outperforms SD1-4 (Text) in style (CSD 0.34 vs. 0.22) but slightly lags in content preservation (0.21 vs. 0.26). StyleAligned scores 0.47 on Stable Diffusion but drops to 0.22 on Blank Canvas Diffusion, highlighting the need for an artistic backbone. Our perceptual study confirms these trends: 76.2% of participants preferred our method over SD1-4 (Text) for style. Against StyleAligned with SD1-4, our model was chosen 31.5% of the time, reinforcing SD1-4's strong artistic capacity from extensive pretraining. When asked to identify images closest to real artworks, participants selected our method 17.5% of the time and SD1.4 (Text) 11.1%, suggesting our approach better mimics authentic artistic styles.

**Data Attribution** We find that our Art Adapter generalizes well from a small art-style training set, generating novel images consistent with the target style. To understand which training images influenced these generations—and to check whether art content may have inadvertently remained in the training set—we apply the data attribution method from Wang et al. (2023a). Results are shown in Fig. 6, where we retrieve the top five attributed images from both the Blank Canvas Dataset and the Art-Style examples for each output. While stylistic features often dominate, real-world influences from the Blank Canvas Dataset remain significant. In the Picasso-style example, the output reflects

Figure 6: Data attribution experiments on stylized images reveals that while the generated images reflect the distinct artistic styles of each artist, the training images that contributed the most came from both the Blank Canvas dataset and the Art-Style examples.

cubist elements, and attribution reveals all top five matches from the Art-Style dataset. In contrast, for Matisse and Lichtenstein styles, the top attributed images come from the Blank Canvas Dataset. The Matisse-style image shows vivid colors and organic shapes, yet attribution reveals underlying real-world scenes. Similarly, in the Lichtenstein example, the bold comic style overlays content clearly traced back to non-artistic pretraining images.

## 6 Discussion and Limitations

Blank Canvas Diffusion explores whether generative models can learn to imitate artistic style with minimal exposure to traditional artworks. Using a simple Art Style Adapter, we show that models trained on largely art-free data can still generate stylistically rich images. Attribution analysis suggests that patterns in natural imagery may already carry the foundations of visual style, offering insight into how generative models internalize and repurpose structure in the world around them. While we took care to filter explicit artworks, we acknowledge that photography itself is not free of aesthetic intent. Choices in composition, lighting, and framing can carry stylistic influence (Hertzmann, 2022), and some images in our dataset may reflect these qualities. Although not central to our findings, this nuance highlights the difficulty of fully disentangling artistic influence from visual data. Our results suggest that generative models may not require direct exposure to traditional art to develop artistic capabilities, complicating assumptions about what kinds of data drive creativity and raising broader questions about influence, authorship, and control in generative systems.

## 7 Acknowledgments

We thank Alan Kenny for his thoughtful discussions and for graciously allowing the use of his artwork in this research. We also appreciate Yael Vinker and Tamar Rott Shaham for their valuable feedback on our paper. JM is grateful for support from the ONR MURI grant (#033697-00007), IBM (#027397-00163), and Hyundai Motor Company R&D (#034197-00002). RH and DB are supported by Open Philanthropy and NSF grant #2403304.

## References

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 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 (CVPR)*, pages 8795–8805, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

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.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models, 2023a.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. *arXiv preprint arXiv:2308.14761*, 2023b.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Commoncanvas: Open diffusion models trained on creative-commons images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8250–8260, 2024.

Melissa Heikkilä. This artist is dominating ai-generated art. and he's not happy about it. *MIT Technology Review*, 125(6):9–10, 2022.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.

Aaron Hertzmann. The choices hidden in photography. *Journal of Vision*, 22(11):10–10, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023a.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023b.

Yiwei Lu, Matthew YR Yang, Zuoqiu Liu, Gautam Kamath, and Yaoliang Yu. Disguised copyright infringement of latent diffusion models. *arXiv preprint arXiv:2404.06737*, 2024.

Rui Min, Sen Li, Hongyang Chen, and Minhao Cheng. A watermark-conditioned diffusion model for ip protection. *arXiv preprint arXiv:2403.10893*, 2024.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8693–8702, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

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.

Christoph Schuhmann and Peter Bevan. Laion pop: 600,000 high-resolution images with detailed descriptions, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.

Tom Simonite. This artist is dominating ai-generated art. and he's not happy about it. *MIT Technology Review*, 2022.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Spawning AI Team. Spawning ai, 2023.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Sheng-Yu Wang, Alexei A Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7192–7203, 2023a.

Xi Wang, Zoya Bylinskii, Aaron Hertzmann, and Robert Pepperell. A computational approach to studying aesthetic judgments of ambiguous artworks. *Psychology of Aesthetics, Creativity, and the Arts*, 2023b.

Richard Whiddington. Artists land a win in class action lawsuit against a.i. companies. *Artnet News*, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

# Supplementary Material

## A Diffusion Model Background

Diffusion models Ho et al. (2020) represent a class of generative models capable of producing high-quality images by modeling data distributions through successive denoising steps. In diffusion modeling we define a forward process that incrementally introduces noise to the data distribution, transforming it into a Gaussian distribution over time. At any given time step, the relationship between the image and the noise can be expressed as:

$$X_t = \sqrt{1 - \beta_t} \cdot X_0 + \beta_t \cdot \epsilon, \tag{3}$$

where $X_t$ represents the image at time step $t$, $X_0$ is the original image, $\epsilon$ denotes a Gaussian Random Variable with zero mean and unit variance, and $\beta_t$ is the standard deviation of the added noise. Diffusion models optimize for the score function, $\nabla \log p_t(\cdot)$, needed to reverse the corruption process. The score can be learned with supervised learning using the Denoising Score Matching objective defined as:

$$\min_{\theta} \mathbb{E}\left[\|\epsilon_\theta(X_t, C, t) - \epsilon\|^2\right], \tag{4}$$

where $\epsilon_\theta$ is the model and C is the condition, in our case, the text prompt. For computational reasons, the corruption process often occurs in low dimensional latent space Rombach et al. (2022).

## B Artwork Filtering Methodology

Our artwork filtering process operates on both image and caption levels to ensure comprehensive coverage. For image-level filtering, we define a set of concepts to be excluded:

> painting, art, artwork, drawing, sketch, illustration, sculpture, stamp, advertisement, logo, installation art, printmaking art, digital art, conceptual art, mosaic art, tapestry, abstract art, realism art, surrealism art, impressionism art, expressionism art, cubism art, minimalism art, baroque art, rococo art, pop art, art nouveau, art deco, futurism art, dadaism art

Fig. 7 presents a histogram of CLIP scores for images associated with the word "painting". This distribution is derived from a subset of the SA-1B dataset, comprising 11,186 images (0.1% of the complete SA-1B dataset).

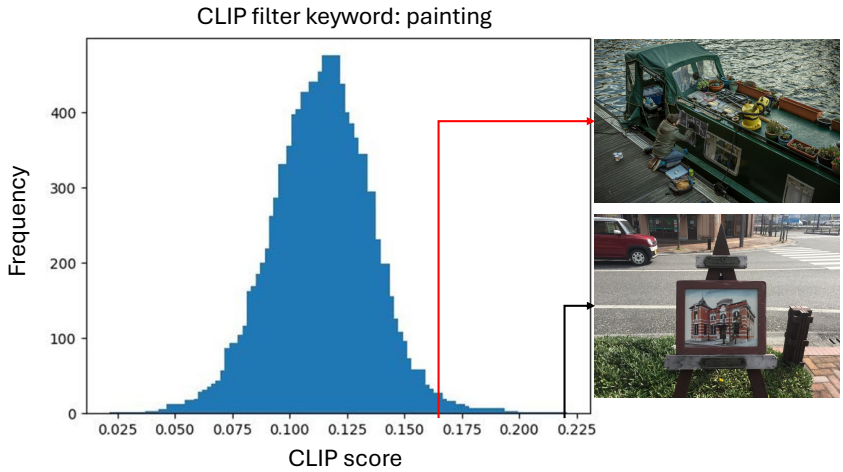



Figure 7: Histogram of the CLIP score of images with the word "painting". The distribution shown is from a subset of the SA-1B dataset. The red line represents the filtering threshold (17) we selected. Our strict threshold aims filters out all the art, even incidental art like a picture of a man painting.

For caption-level filtering, we exclude the following terms:

painting, paintings, art, artwork, drawings, sketch, sketches, illustration, illustrations, sculpture, sculptures, stamp, stamps, advertisement, advertisements, logo, logos, installation, printmaking, digital art, conceptual art, mosaic, tapestry, abstract, realism, surrealism, impressionism, expressionism, cubism, minimalism, baroque, rococo, pop art, art nouveau, art deco, futurism, dadaism

## C Manual Filtering

To inspect the quality of our dataset, one of the authors manually inspected a random sample of 10k images. We did so by writing a simple javascript script that displayed individual images on a page, by clicking the category a result was written in a text file (see Fig. 8 –9).

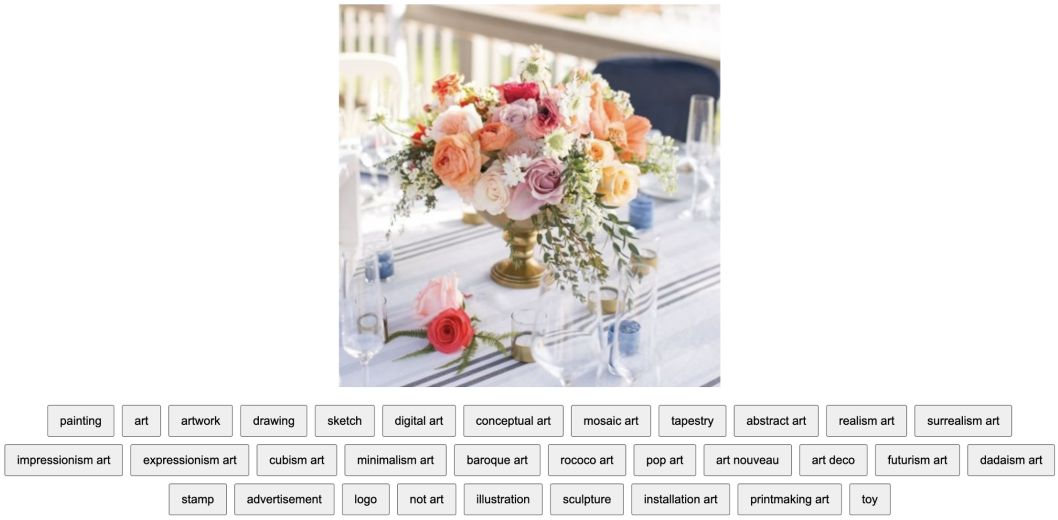



Figure 8: Manual inspection tool

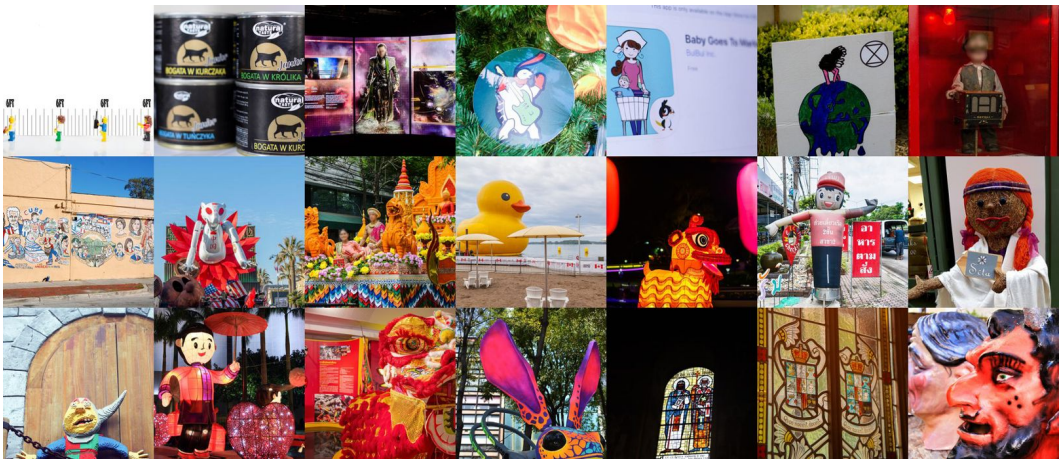


Figure 9: In a random sample of 10000 images (total: 71), we found categories; advertisement (2), digital art (1), drawing (3), installation art (12), logo (12), mosaic art (2), printmaking art (1), and sculpture (38), from left to right, excluding sculpture and logo.

## D Qualitative Results of Blank Canvas Diffusion

We demonstrate qualitative results of the Blank Canvas Diffusion in Fig.10, for comparison, we also include images generated by StableDiffusion 1.4 and CommonCanvas-SC. Our model, despite significantly smaller training set size generates high-quality images faithful to the text prompt.

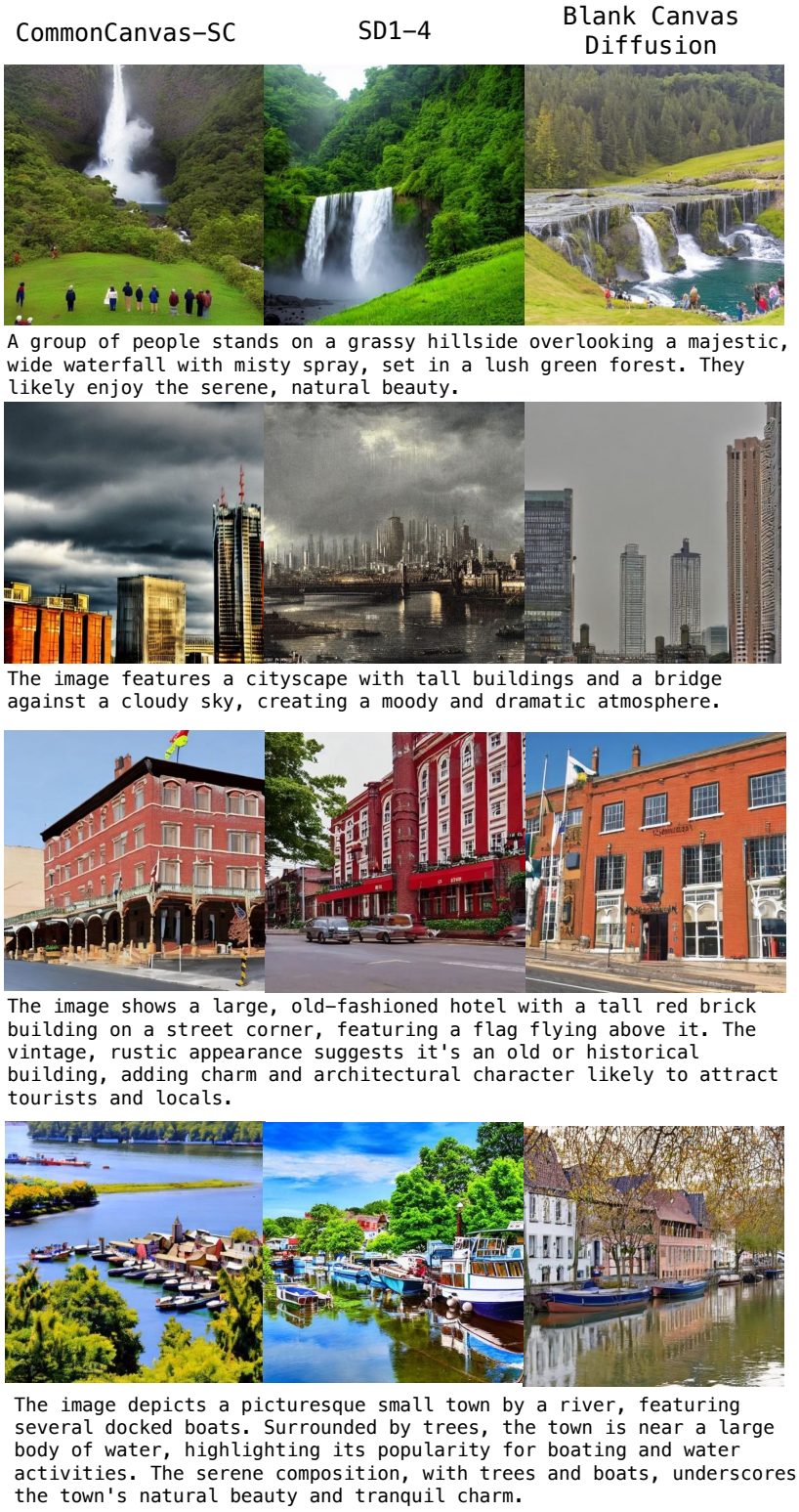



Figure 10: Qualitative comparison of images generated with Blank Canvas Diffusion, Stable Diffusion 1.4 and CommonCanvas-SC model.

## E Qualitative Results of Image Stylization in style of Van Gogh

Figure 11: Comparion of our method and other image stylization baselines for the artist Van Gogh. All captions contain a suffix “in the style of Vincent van Gogh", Our model and SD1.4 + Art Adaptor are prompted with suffix “in the style of V* art".

## F Implementation Details

As detailed in Sec. 6.1, we train both the VAE and the U-Net from scratch, while utilizing a pre-trained BERT text encoder. We train VAE with the mixture of Blank Canvas dataset (SAM) and (filtered) MS-COCO to improve VAE reconstruction quality. While SAM prevents artistic bias, MS-COCO adds non-artistic diverse images. We used 115,294 MS-COCO samples and 104,145 SAM samples. We train the VAE with a batch size of 24, gradient accumulation of 2, and a 2e-4 learning rate for 15 epochs, taking 16 hours. We first train the U-Net under 256 resolution on 7 H100 GPUs, with each GPU using a batch size of 300 and mixed precision of FP16. We apply gradient accumulation of 8 and use a learning rate of 1e-4 with the AdamW optimizer on a 7 H100 GPUs by 41400 steps. We fine-tune the model at a 512x512 resolution for a total of 156,700 steps, with learning rate of 5e-5 and batch size of 90, and apply 10% dropping rate with classifier-free guidance sampling (Ho and Salimans, 2022). The architecture design choices are well motivated to prevent art knowledge leakage from CLIP embedding, otherwise our architecture follows an established baseline SD1.4.

**LoRA Implementation** We found that incorporating low-rank Adapters into the attention, linear, and convolution layers of the UNet’s up block reduces overfitting and improves generation quality, as opposed to introducing LoRA across all UNet blocks. Motivated by the observation that early layers handle global image aspects, which are less style-dependent, we found that injecting LoRA layers only in the UNet’s up block reduces overfitting (Fig. 12). Quantitative evaluation shows this approach achieves a higher style score across 17 artists (0.29 vs. 0.26) while preserving a comparable content score (0.22 vs. 0.23). The learning rate was set to 2e-4 using the AdamW optimizer, and we trained for 1,000 steps with a batch size of 5 and the DDIM noise scheduler. For data augmentation, we resize images with a random scale of 0.9 to 1 and randomly crop with an aspect ratio of 3/4 to 4/3. In the experiments, we use ‘sks’ as the V* token, which serves as a random new token for learning a new art style concept.

Additionally, we conducted an analysis to determine the effect of LoRA rank on the art adapter’s performance. Tab. 2 presents the results of our model with LoRA ranks 1 and 64. Our findings indicate that LoRA rank does not significantly impact model performance. This experiment is done on the image stylization task, the scores are average across 17 artists, with 1.0 LoRA scale.

In our quantitative evaluation, we use a 500-sample subset of the LAION Pop dataset, randomly sampled while excluding images with keywords listed in B. For captions longer than a baseline’s content length, we use only the first sentence.

**Content Loss Strength** We investigated the influence of the content loss weight ($w$) in the art adapter across different models Tab. 3.

The content loss substantially enhances learning performance, with CSD increasing from 0.14 to 0.29 when $w$ is set to 50. This demonstrates that the content loss effectively aids the model in

Figure 12: Comparison of LoRA applied to all layers vs. only the up block of the UNet. Limiting LoRA to the up block reduces overfitting. Adapters train on a 10 images sample of Camille Pissaro's artwork.

| LoRA Rank | CSD↑ | LPIPS↓ | ViTc↑ | CLIPc↑ |
|---|---|---|---|---|
| 1 | 0.29 | 0.62 | 0.28 | 0.22 |
| 64 | 0.21 | 0.59 | 0.32 | 0.25 |

Table 2: Rank analysis of LoRA on style transfer task. We find that a higher rank of LoRA does not improve the model learning performance.

distinguishing between art images and natural images. The effect remains robust across different weight values, with performance remaining nearly constant when $w$ is set to 20 or 100 (up to 0.02 difference in CSD).

| Content Loss scale | CSD↑ | LPIPS↓ | ViTc↑ | CLIPc↑ |
|---|---|---|---|---|
| 0 | 0.14 | 0.57 | 0.33 | 0.25 |
| 20 | 0.29 | 0.62 | 0.28 | 0.22 |
| 50 | 0.29 | 0.62 | 0.28 | 0.22 |
| 100 | 0.27 | 0.61 | 0.28 | 0.23 |

Table 3: Analysis of prior preservation loss weight ($w$) on our model. Experiments are conducted on style transfer, with noise added at the 800th time step. The scores are averages across 17 artists on image stylization task, with 1.0 LoRA scale.

## G Art-Agnostic Model Verification

To verify the art-agnostic nature of our model, we conducted a textual inversion experiment as suggested by Pham et al. (2023). In the experiment we use the same Art Dataset as for training the Art Adapter for Vincent Van Gogh style. Fig. 13 illustrates that our model fails to produce the target style using textual inversion, further confirming its lack of prior artistic knowledge.

Figure 13: Through textual inversion using paintings by van Gogh, we found that, unlike SD1-4, our model cannot generate images in the corresponding style. This indicates that our model cannot be inverted to generate artwork through prompt space searching, demonstrating it has no prior knowledge of art.

## H Model Editing and Controlling Ability

Despite being trained on a significantly smaller and less diverse dataset limited to natural images, our art-agnostic model demonstrates comparable editing and control capabilities to competitive models. This is evident in both single-image editing and customization experiments.

In Fig. 14, we qualitatively illustrate the single-image editing process using the Plug-and-Play method (Tumanyan et al., 2023) applied to our model. We provide editing examples on both real and generated images, demonstrating the model's ability to replace a pyramid with a large mountain, both with and without the artistic adapter (weight 1.5) of van Gogh.

Furthermore, we demonstrate our model's customization abilities using the Dreambooth technique (Ruiz et al., 2023). We learned the concept of a barn using 7 training images from the Custom-

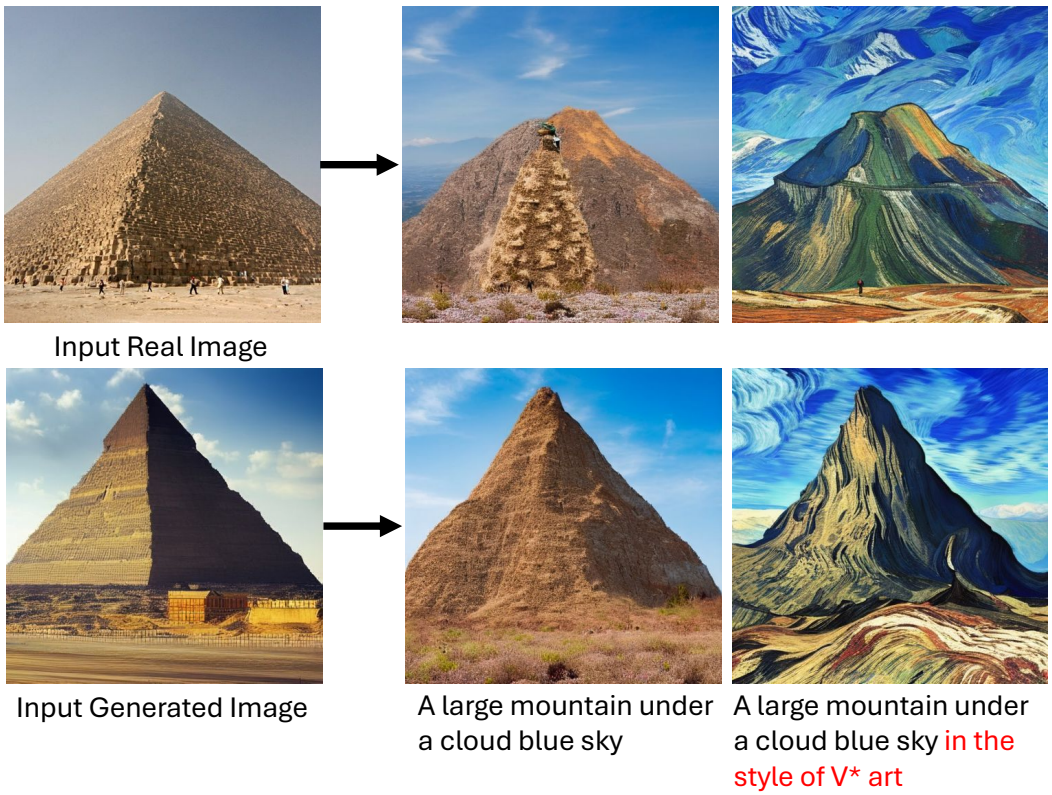



Figure 14: Plug-and-Play editing on our model. We provide both editing on real and generated image examples. We replace a pyramid to a large mountain both without and with the artistic adaptor of van Gogh.

Concept101 dataset (Kumari et al., 2023b). The model was trained to generate the barn in various contexts, utilizing 200 prior samples from Stable Diffusion v1-4, with a prior preservation loss of 1.0, a learning rate of 5e-6, and 250 training steps on 2 GPUs.

Figure 15: Dreambooth editing on our model. We send 7 barn example images to the model and ask it to generate the barn in various contexts.

## I Effect of Applying the Adapter at various Time Steps

We analyzed the effect of the adapter time step on art generation results. Fig. 16 shows the art generation outcomes with different adapter time steps. Intuitively, the model generates more style information when the adapter starts earlier (left) and more content information when the adapter starts later (right).

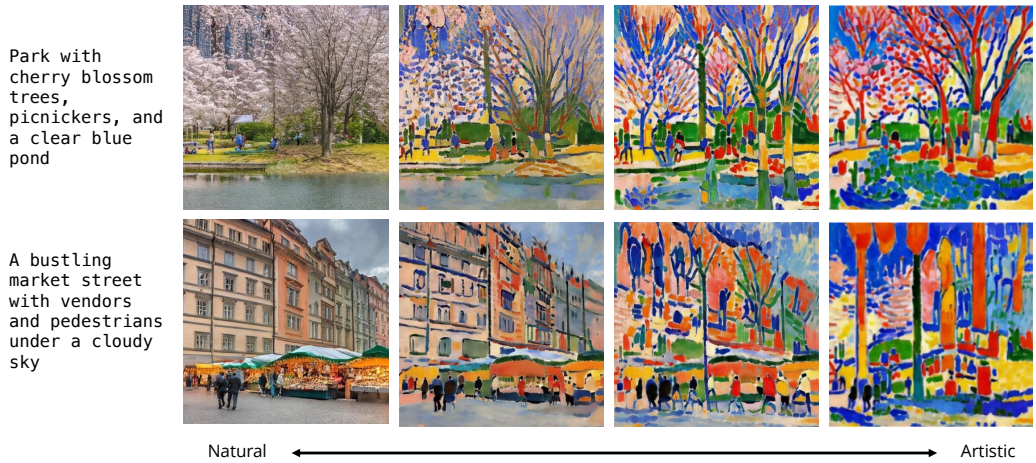



Figure 16: Art generation results using Art Adapter at different timesteps. From left to right: no adapter (column 1), adapter introduced at timestep 800 (column 2), 600 (column 3), and 0 (column 4). This demonstrates how earlier adapter introduction increases artistic influence in the image.

## J User Study

The user study was conducted on AWS with 42 participants. We disregarded the workers that took less than 5 seconds to complete the assignment. To ensure reliability, we included a validation task requiring participants to distinguish a non-artistic image from a painting, ensuring they focused on stylistic qualities rather than content or quality of the images. Only those who passed this task were included in the analysis. This process reduced the number of workers from an initial pool of 88 to 42. Additionally, we selected expert workers and randomized both the image order and methods to minimize potential bias. The user study interface is shown in Fig. 17.

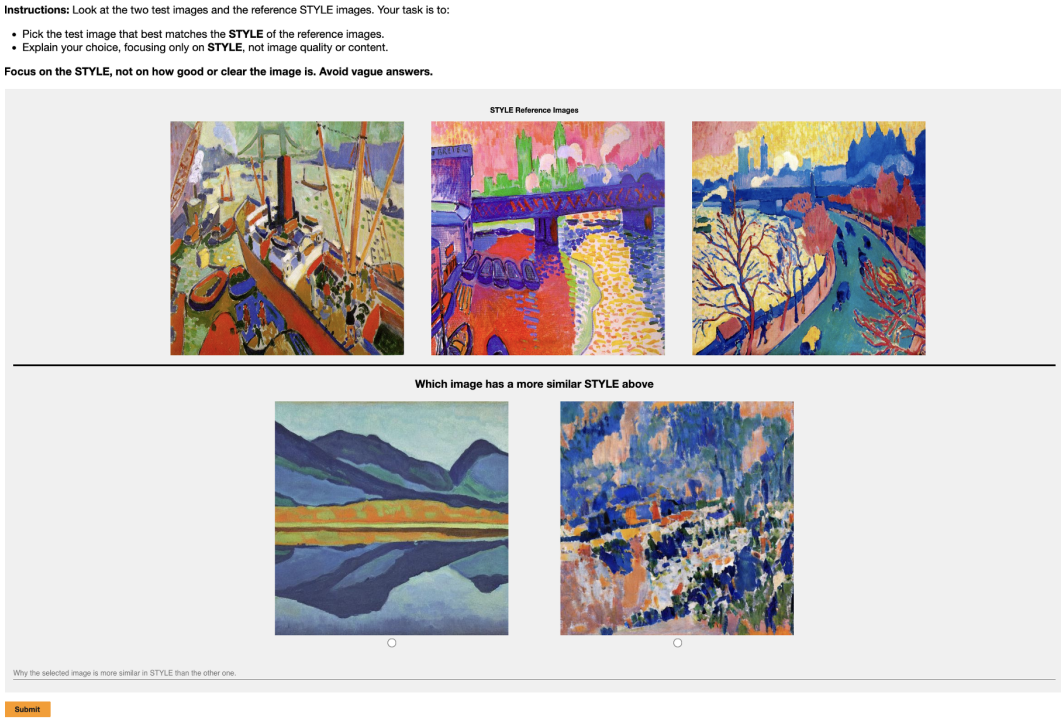
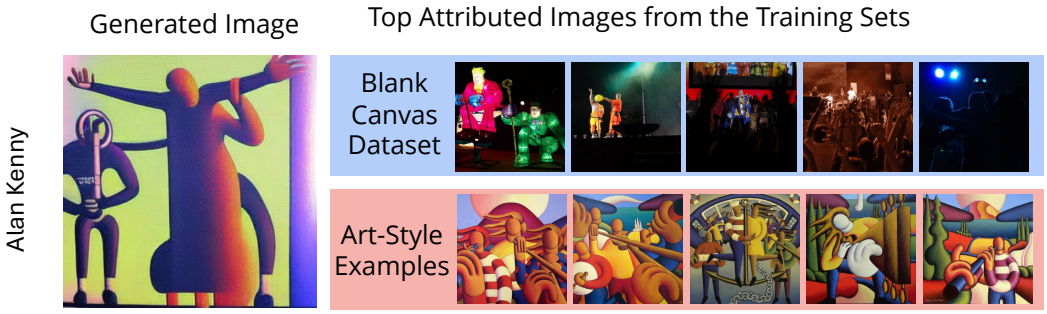



Figure 17: The interface for the user study we conducted, the participants were shown examples of an art style and were asked to select an image that matched it better.

## K Introducing the Art Adapter to the Artist

To explore the artistic community's reaction to AI-generated art, we conduct an interview with the renowned artist Alan Kenny. Upon obtaining Alan's permission, we train an Art Adapter on 11 artworks showing his distinctive style. We describe our work and present Kenny with the generated images imitating his style. In the interview, the artist expresses a blend of astonishment and familiarity when observing the AI-generated art, remarking, "I didn't expect [this quality] if you were using a base model of blank canvas... you probably achieved more than I would have expected for a base model with no information." He acknowledges that the AI has captured aspects of his distinct style to the extent that, "if you were to post some of these images online, I would get people texting me, 'I see your images.' They would spot it, and I spot it." Despite noting that "compositionally, it is weak" and contrasting this with his own "well thought and meticulous" compositions, he recognizes that "there are some very positive things" in the AI's work. The artist describes the experience as "terrifying and a bit exciting at the same time," specifically pointing out how the AI imitates his signature "gradation of the landscape" and "gradation of the shapes." Though he felt his style is largely captured, he admits, "there is kind of originality to them... I see me in them, yes, very strongly... but there is an originality to some of the images.".

In Fig. 18, 19 we present qualitative examples of images generated in the style of Alan Kenny along with the results of the data attribution technique. These examples reveal how natural images inspire features in the generated art (e.g., a stage with musicians) while preserving the characteristics of the artistic style like the use of colors, smooth boundaries and geometric shapes.




Guitarist adjusting strings on stage before a performance.

Figure 18: Generated artwork in the style of Alan Kenny (created and displayed with the artist's permission) showcases the top-5 influential images from the Blank Canvas and Art Datasets.

## L Data Attribution

We present additional results of applying data attribution to the generated art images in Fig. 20. These results illustrate how specific visual elements from the training data, including both natural and small art images, influence the generated outputs. Despite the Art Adapters being trained on a limited set of art images, and the base text-to-image model itself having minimal exposure to graphic art, the attribution analysis points to similarities in the natural images that may enable the model to effectively generalize from few examples.

The attribution method [47] is an established approach, with the similarity metric chosen to closely approximate exact attribution. Although the method is an approximation, no current data attribution method can pin-point training examples more precisely at the scale of training data we need to analyze. We acknowledge this gap and will add the discussion to the main paper.

## M Different Baselines

While many methods transfer style from a reference image to another, direct comparisons are often infeasible due to differences in model architecture and dependencies. For instance, StyleDrop Sohn et al. (2023) is designed specifically for the Muse architecture, making it difficult to separate the

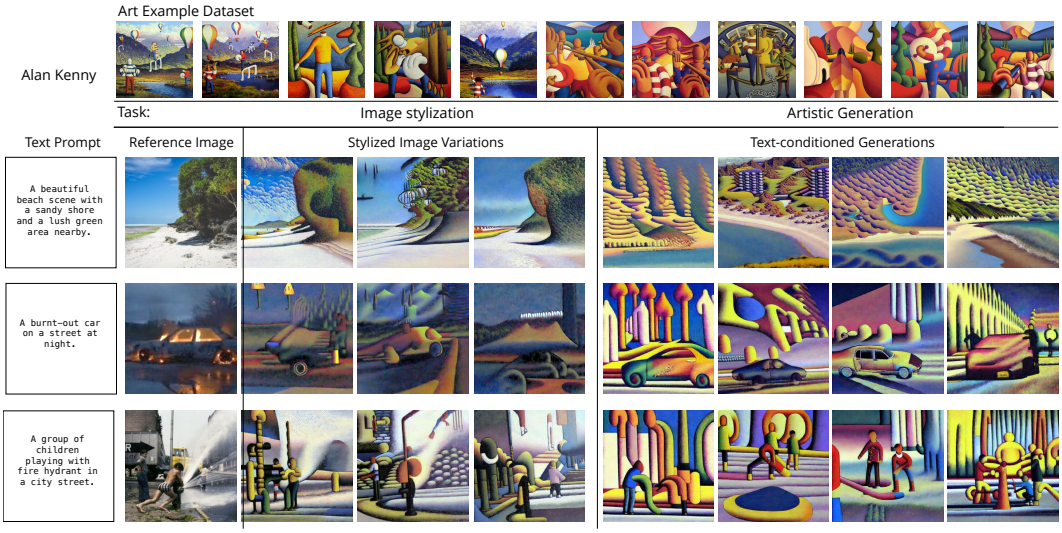



Figure 19: Additional qualitative experiments of the art imitation of the interviewed artist Alan Kenny.

contribution of the adaptation method from the pretrained model's inherent stylization capabilities. Similarly, Visual Style Prompting Jeong et al. (2024) and InstantStyle Wang et al. (2024) are designed primarily for Stable Diffusion XL. Computational constraints prevent us from training a comparable model with our Blank Canvas data, but we encourage others to explore similar experiments.

DeadDiff Qi et al. (2024), while offering advantages to text-to-image adapters, relies on a paired dataset where the reference image and ground truth share style or semantics, which differs significantly from our approach. Our primary goal is to demonstrate that effective style transfer is achievable with a few examples, rather than competing with methods that leverage extensive pretrained knowledge of graphic art.

To disentangle the adaptation method from the pretrained model's capabilities, we applied another baseline, StyleID Chung et al. (2024), to both our Blank Canvas Diffusion model and SD1.4. Similar to StyleAligned, both training-free adaptation methods performed better on SD1.4, leveraging its broad artistic knowledge, but struggled on Blank Canvas Diffusion, highlighting their reliance on pretrained models rich in artistic priors. In contrast, our Art Adapter bridges this gap effectively, demonstrating that focused adaptations within the Blank Canvas framework can achieve compelling results without relying on inherited artistic biases.

While this comparison is not entirely equivalent—StyleAligned and StyleID use a single reference image, whereas our Art Adapter employs multiple style references (in this experiment, we compare five artists: Derain, Miró, Klimt, Picasso, and Lichtenstein, with an average training set of 15)—we were unable to adapt these methods to support multiple references, as doing so falls outside the scope of this work.

It is important to emphasize that our goal is not to compete with models and methods trained on significantly larger graphic art datasets, as such comparisons would be inherently unfair. Instead, our work focuses on a key question: how much graphic art data is truly needed to effectively replicate an artistic style? Our analysis demonstrates that an artistic style can be successfully learned from just a few examples.

## N Additional Qualitative Results

Additional results of art generation (art generation and image stylization) and training images in Figures 21 –37. We show our model's ability to replicate diverse artistic styles: Impressionism (Monet, van Gogh, Corot), Art Nouveau (Klimt), Fauvism (Derain), Abstract Expressionism (Matisse, Pollock, Richter), Abstract Art (Kandinsky), Cubism (Picasso, Gleizes), Pop Art (Lichtenstein,

Figure 20: Additional qualitative experiments showing diverse art generations and top five attributed images from both the Blank Canvas dataset and and Art-Style example dataset.

| Text-To-Image Model | Adaptation Method | CSD_mean↑ | ViTc↑ | CLIPc↑ |
|---|---|---|---|---|
| Blank Canvas Diffusion | Art-Adapter | **0.35** | 0.27 | **0.23** |
| | StyleAligned | 0.12 | 0.31 | 0.22 |
| | StyleID | 0.11 | 0.63 | 0.22 |
| SD1.4 | Art-Adapter | 0.21 | 0.27 | **0.26** |
| | StyleAligned | **0.43** | 0.23 | 0.21 |
| | StyleID | 0.29 | 0.40 | 0.23 |

Table 4: Comparing different art adaptation methods across our Blank Canvas Diffusion model and Stable Diffusion 1.4. Training-free adaptation methods, StyleAligned and StyleID, perform better on SD1.4, benefiting from the model's broad artistic knowledge, but struggle on Blank Canvas Diffusion, showing their reliance on pretrained models rich in artistic priors. In contrast, our Art Adapter effectively bridges this gap, proving that focused adaptations within the Blank Canvas framework can deliver compelling results without depending on inherited artistic biases.

Warhol), Ukiyo-e (Hokusai), Expressionism (Escher), and Postmodern and Geometric Abstraction (Miró, Battiss). The captions and reference images are sampled from the LAION Pop dataset.

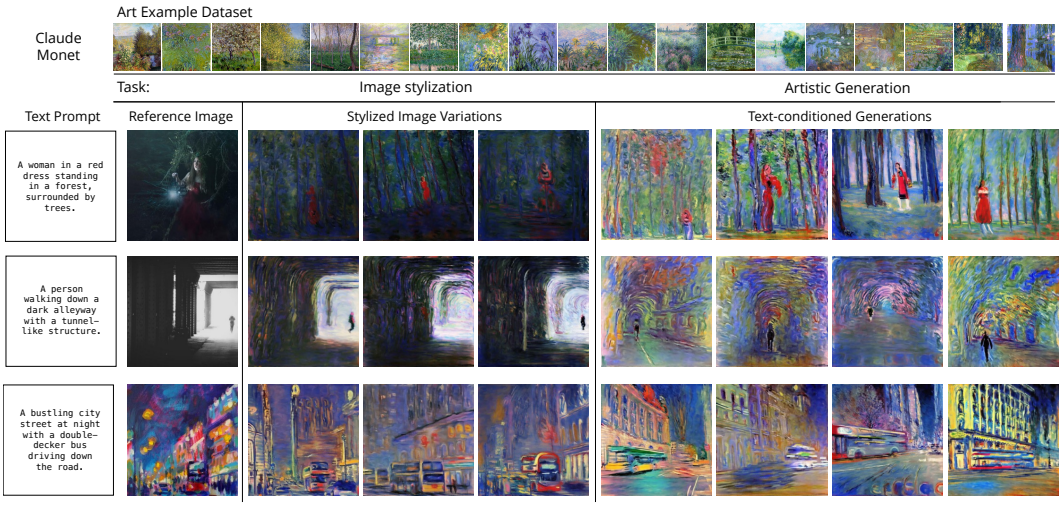



Figure 21: Additional qualitative experiments.

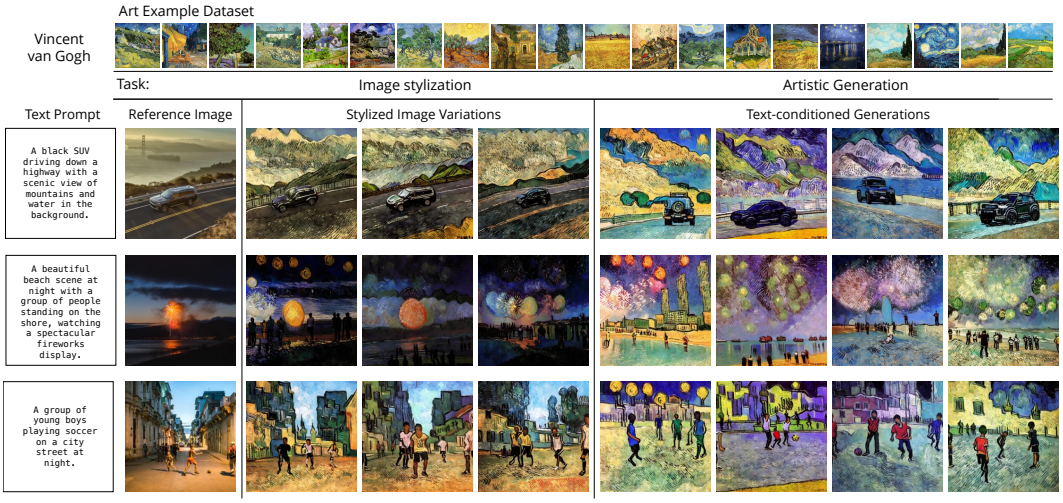



Figure 22: Additional qualitative experiments.

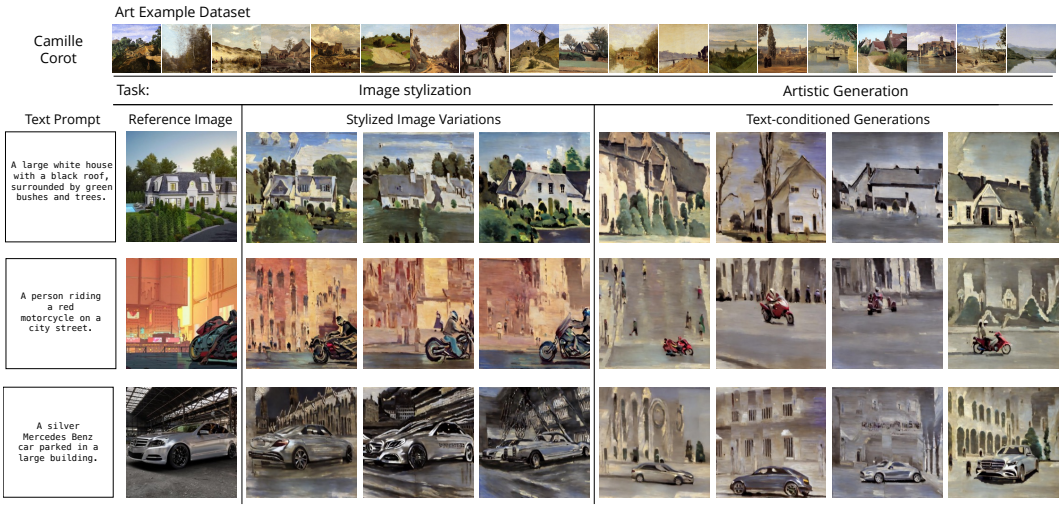



Figure 23: Additional qualitative experiments.

Figure 24: Additional qualitative experiments.

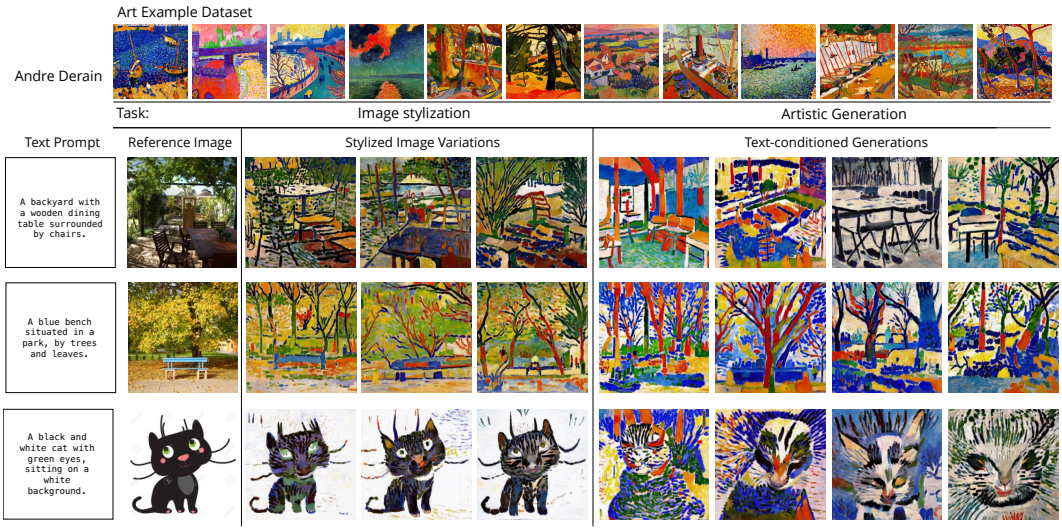



Figure 25: Additional qualitative experiments.

Figure 26: Additional qualitative experiments.

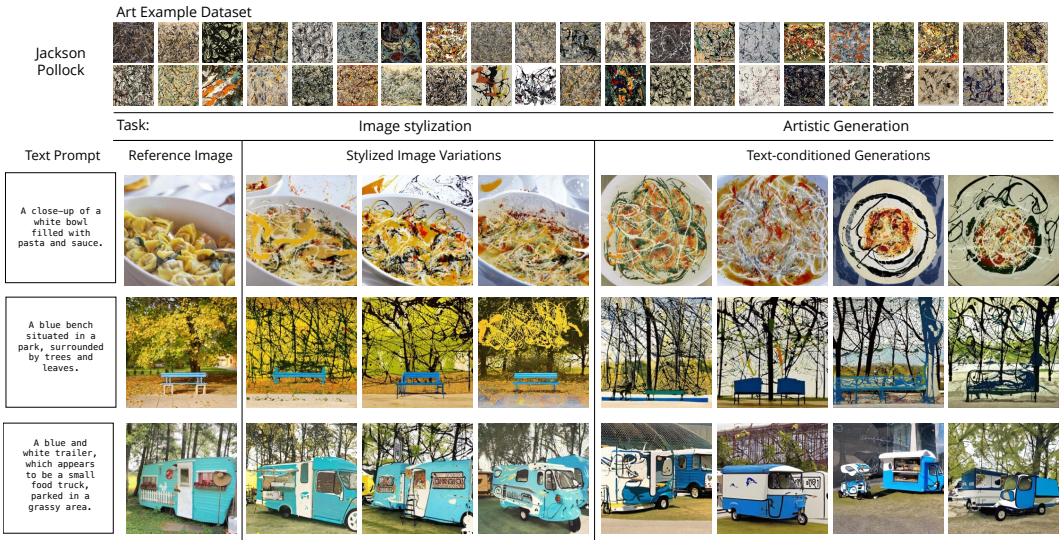



Figure 27: Additional qualitative experiments.

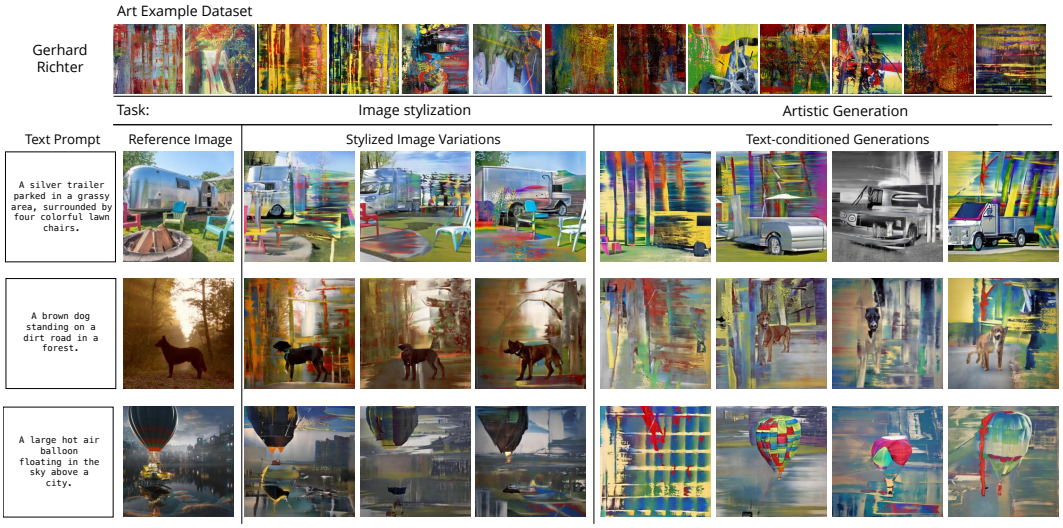



Figure 28: Additional qualitative experiments.

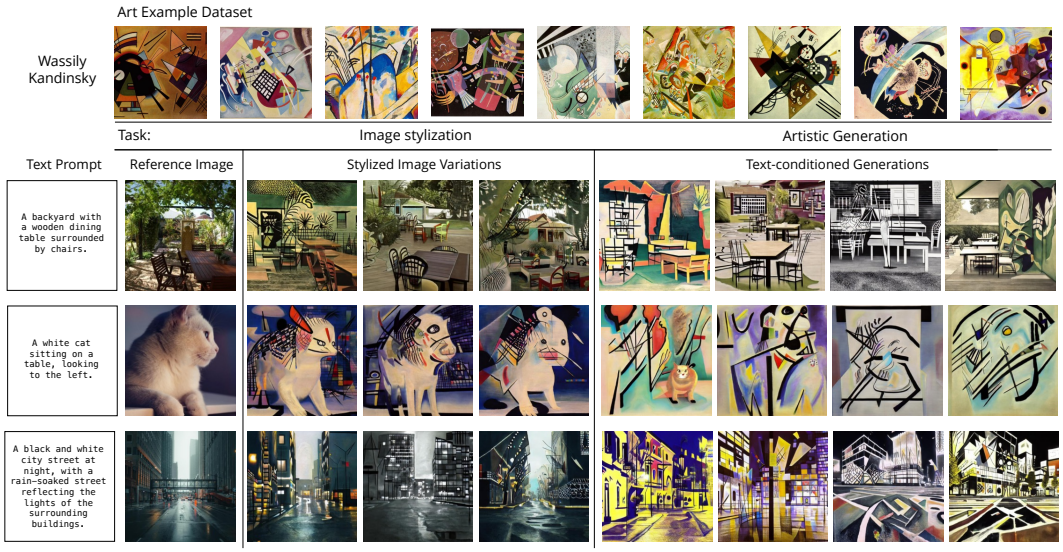



Figure 29: Additional qualitative experiments.

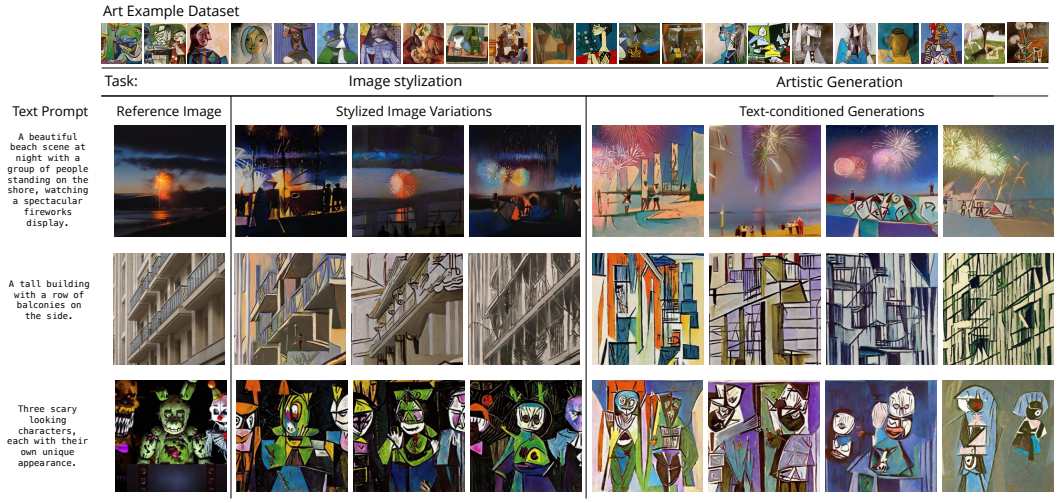



Figure 30: Additional qualitative experiments.

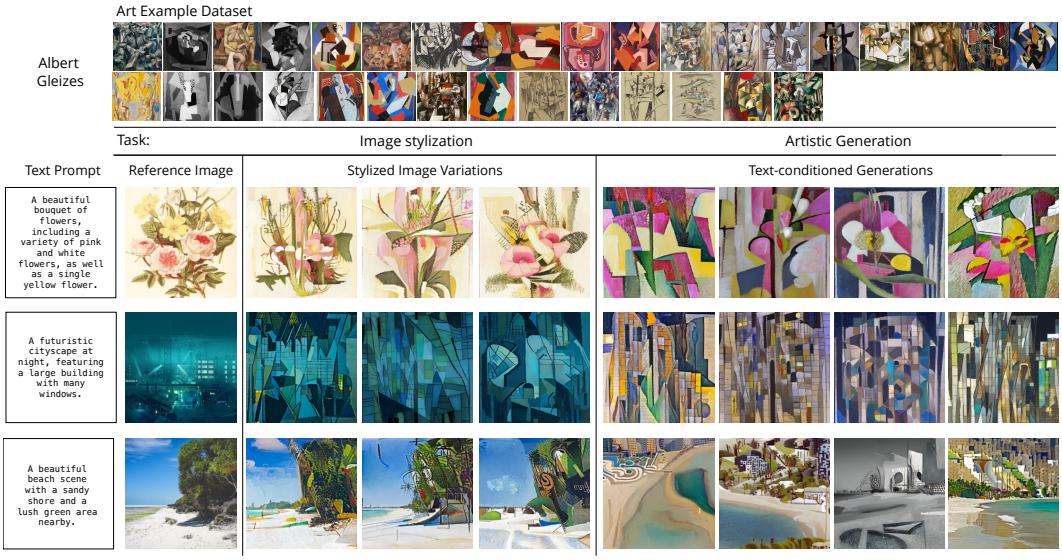



Figure 31: Additional qualitative experiments.

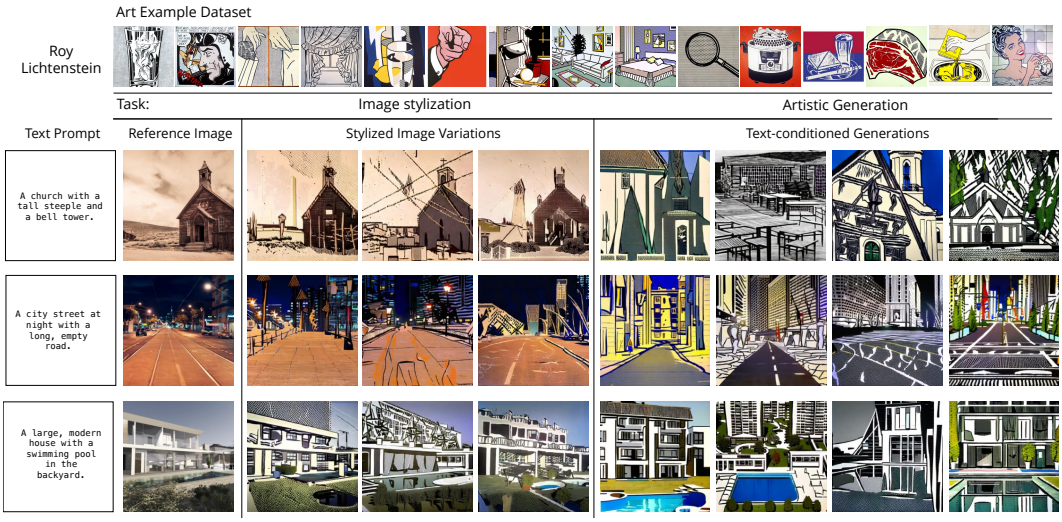



Figure 32: Additional qualitative experiments.

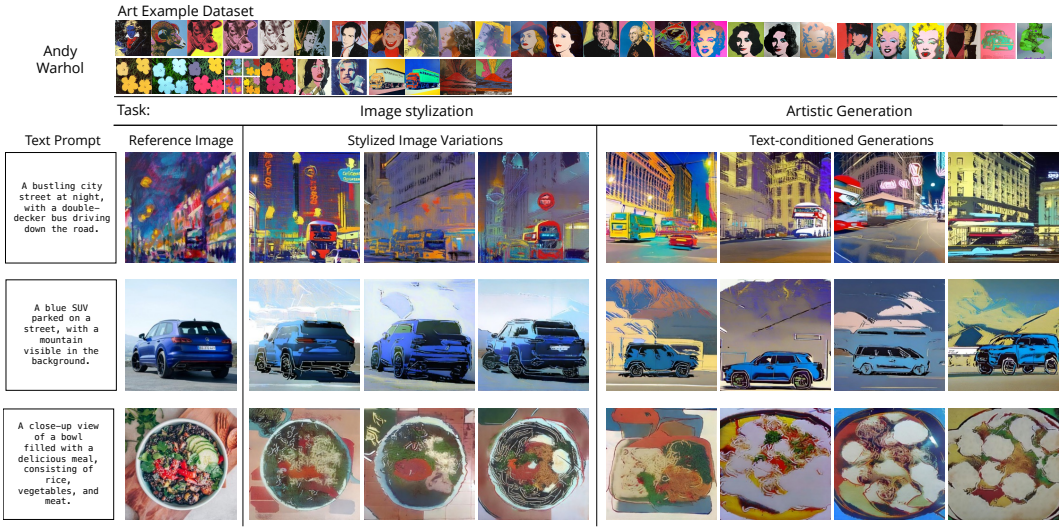



Figure 33: Additional qualitative experiments.

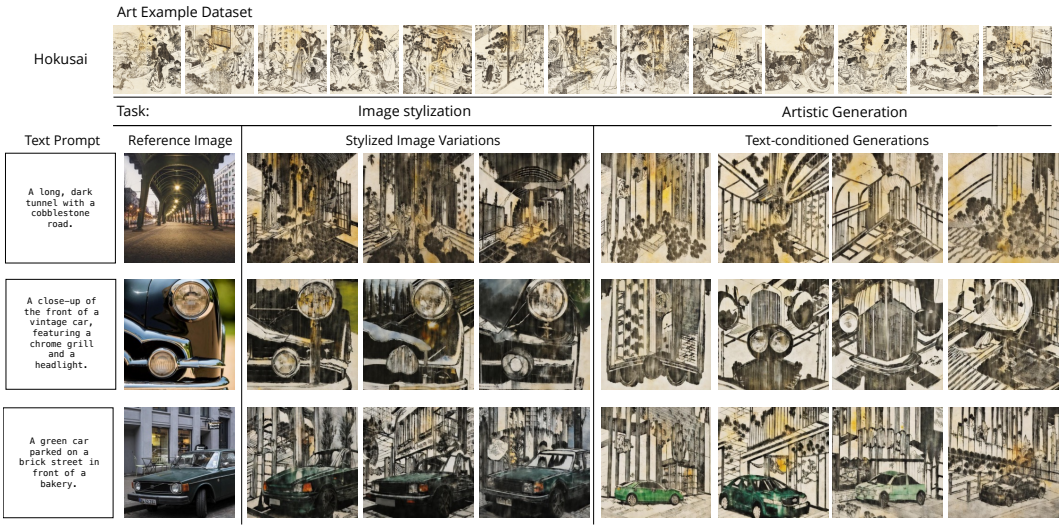



Figure 34: Additional qualitative experiments.

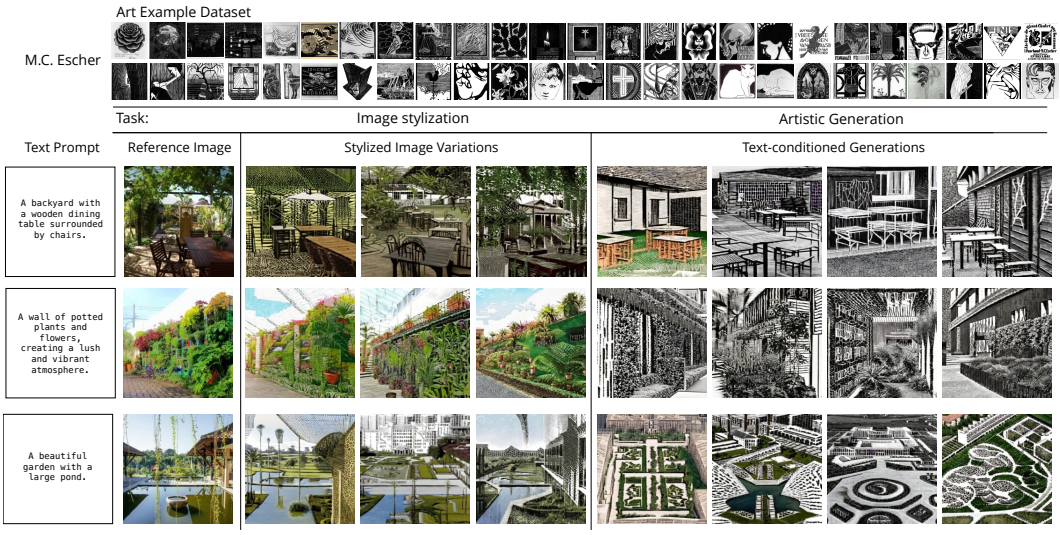



Figure 35: Additional qualitative experiments.

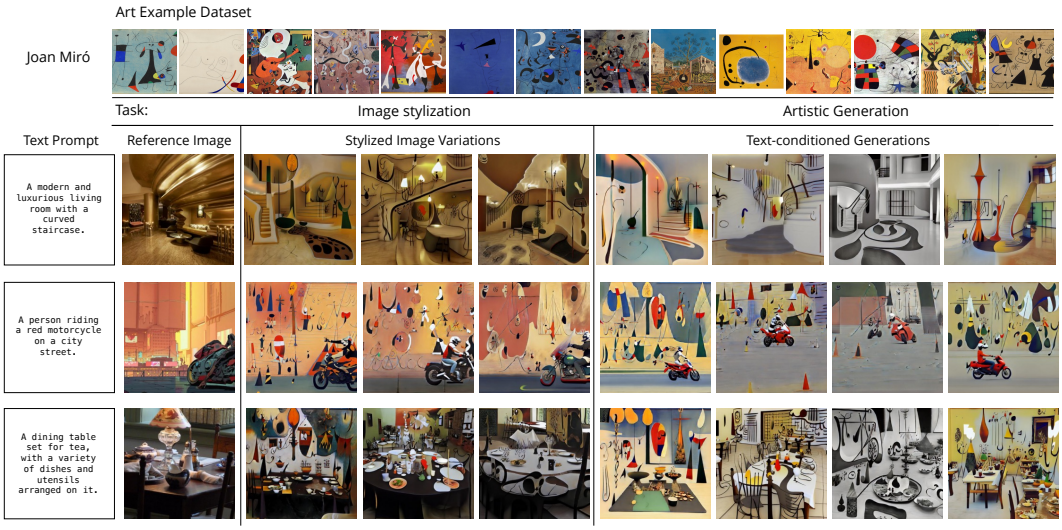



Figure 36: Additional qualitative experiments.

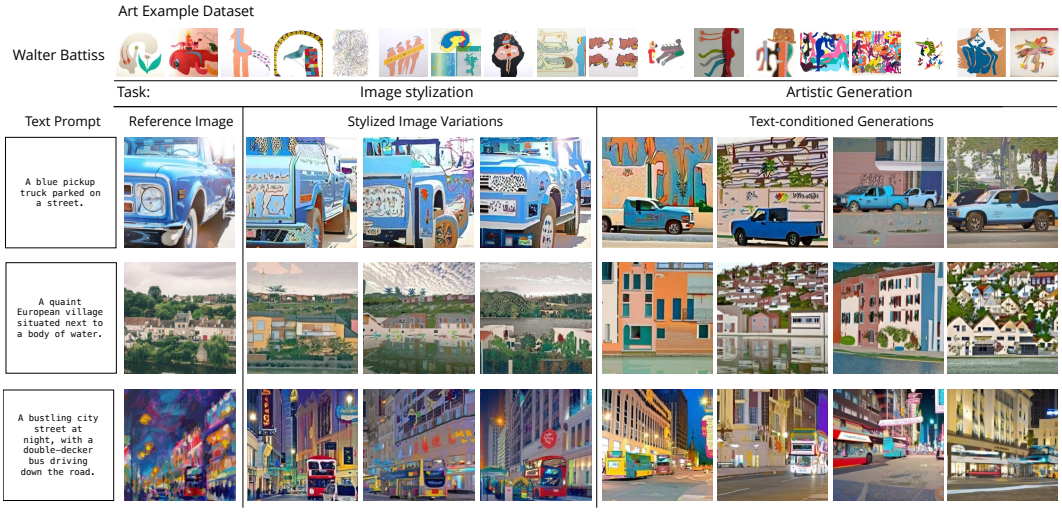



Figure 37: Additional qualitative experiments.