This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Style Aligned Image Generation via Shared Attention

Amir Hertz* [1], Andrey Voynov* [1], Shlomi Fruchter† [1], and Daniel Cohen-Or† [1,2]
[1] Google Research
[2] Tel Aviv University

Abstract

Large-scale Text-to-Image (T2I) models have rapidly gained prominence across creative fields, generating visually compelling outputs from textual prompts. However, controlling these models to ensure consistent style remains challenging, with existing methods necessitating fine-tuning and manual intervention to disentangle content and style. In this paper, we introduce StyleAligned, a novel technique designed to establish style alignment among a series of generated images. By employing minimal 'attention sharing' during the diffusion process, our method maintains style consistency across images within T2I models. This approach allows for the creation of style-consistent images using a reference style through a straightforward inversion operation. Our method's evaluation across diverse styles and text prompts demonstrates high-quality synthesis and fidelity, underscoring its efficacy in achieving consistent style across various inputs.

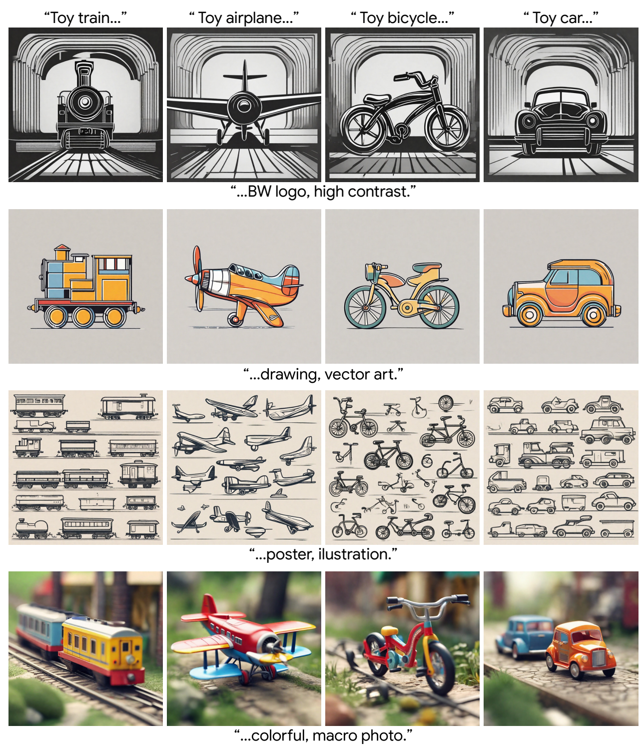

Figure 1. **Style aligned image set generation.** *By fusing the features of the toy train image (left) during the diffusion process, we can generate an image set of different content that shares the style.*

1. Introduction

Large-scale Text-to-Image (T2I) generative models [37, 39, 45] have emerged as an essential tool across creative disciplines, such as art, graphic design, animation, architecture, gaming and more. These models show tremendous capabilities of translating an input text into an appealing visual result that is aligned with the input description.

An envisioned application of T2I models revolves around the rendition of various concepts in a way that shares a consistent style and character, as though all were created by the same artist and method (see Fig. 1). While proficient in aligning with the textual description of the style, state-of-the-art T2I models often create images that diverge significantly in their interpretations of the same stylistic descriptor, as depicted in Fig. 2.

Recent methods mitigate this by fine-tuning the T2I model over a set of images that share the same style [13,49]. This optimization is computationally expensive and usually requires human input in order to find a plausible subset of images and texts that enables the disentanglement of content and style.

We introduce StyleAligned, a method that enables *consistent style interpretation* across a set of generated images (Fig. 1). Our method requires no optimization and can be applied to any attention-based text-to-image diffusion model. We show that adding minimal *attention sharing* operations along the diffusion process, from each generated image to the first one in a batch, leads to a style-consistent set. Moreover, using diffusion inversion, our method can be applied to generate style-consistent images given a reference style image, with no optimization or fine-tuning.

*Equal contribution.
†Equal Advising.

Figure 2. **Standard text-to-image vs. StyleAligned set generation.** *Given style description of "minimal origami", standard text-to-image generation (top) results with an unaligned image set while our method (bottom) can generate variety of style aligned content.*

We present our results over diverse styles and text prompts, demonstrating high-quality synthesis and fidelity to the prompts and reference style. We show diverse examples of generated images that share their style with a reference image that can possibly be a given input image. Importantly, our technique stands as a zero-shot solution, distinct from other personalization techniques, as it operates without any form of optimization or fine-tuning. For our code and more examples, please visit the project page `style-aligned-gen.github.io`

2. Related Work

Text-to-image generation. Text conditioned image generative models [7, 31, 38] show unprecedented capabilities of generating high quality images from text descriptions. In particular, T2I diffusion models [35, 38, 46] are pushing the state of the art and they are quickly adopted for different generative visual tasks like inpainting [2, 44], image-to-image translation [55, 59], local image editing [9, 24], subject-driven image generation [42, 51] and more.

Attention control in diffusion models. Hertz et al. [17] have shown how cross and self-attention maps within the diffusion process determine the layout and content of the generated images. Morover, they showed how the attention maps can be used for controlled image generation. Other studies have leveraged modifications in attention layers to enhance the fidelity or diversity of generated images [8, 34], or apply attention control for image editing [5, 12, 30, 32, 33, 53]. However, in contrast to prior approaches that primarily enable structure-preserving image editing, our method excels at generating images with diverse structures and content while maintaining a consistent style interpretation.

Style transfer. Transferring a style from a reference image to a target content image is well studied subject in computer graphics. Classic works [10, 11, 18, 26] rely on optimization of handcrafted features and texture resampling algorithms from an input texture image, combined with structure constrains of a content image. With the progress of deep learning research, another line of works utilizes deep neural priors for style transfer optimization using deep features of pre-trained networks [15, 52], or injecting attention features from a style image to a target one [1]. More related to our approach, Huang et al. [22] introduced a real time style transfer network based on Adaptive Instance Normalization layers (AdaIN) that are used to normalize deep features of a target image using deep features statistics of a reference style image. Follow-up works, employ the AdaIN layer for additional unsupervised learning tasks, like style-based image generation [25] and Image2Image translation [23, 29].

T2I personalization To generalize T2I over new visual concepts, several works developed different optimization techniques over a small collection of input images that share the same concept [13, 16, 42, 56]. In instances where the collection shares a consistent style, the acquired concept becomes the style itself, affecting subsequent generations. Most close to our work is StyleDrop [49], a style personalization method that relies on fine-tuning of light weight adapter layers [20] at the end of each attention block in a non-autoregressive generative text-to-image transformer [7]. StyleDrop can generate a set of images in the same style of by training the adapter layers over a collection of images that share the same style. However, it struggles to generate a consistent image set of different content when training on a single image.

Our method can generate a consistent image set without optimization phase and without relying on several images for training. To skip the training phase, recent works developed dedicated personalization encoders [14, 27, 47, 58, 59] that can directly inject new priors from a single input image to the T2I model. However, these methods encounter challenges to disentangle style from content as they focus on generating the same subject as in the input image.

3. Method overview

In the following section we start with an overview of the T2I diffusion process, and in particular the self–attention mechanism Sec. 3.1. We continue by present-

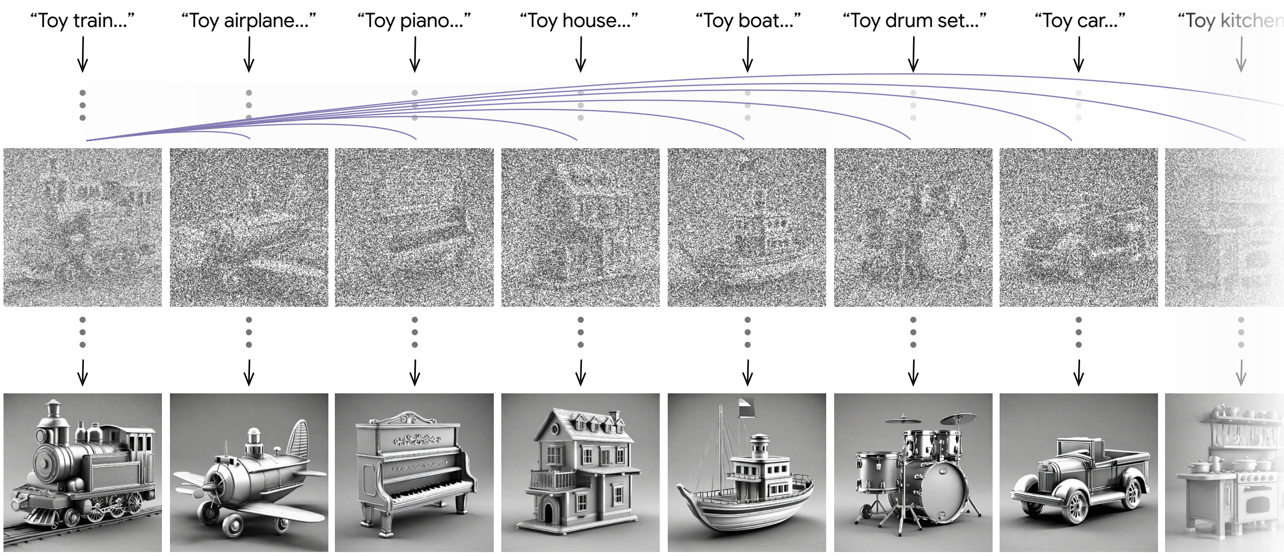

Figure 3. **Style Aligned Diffusion.** *Generation of images with a style aligned to the reference image on the left. In each diffusion denoising step all the images, except the reference, perform a shared self-attention with the reference image.*

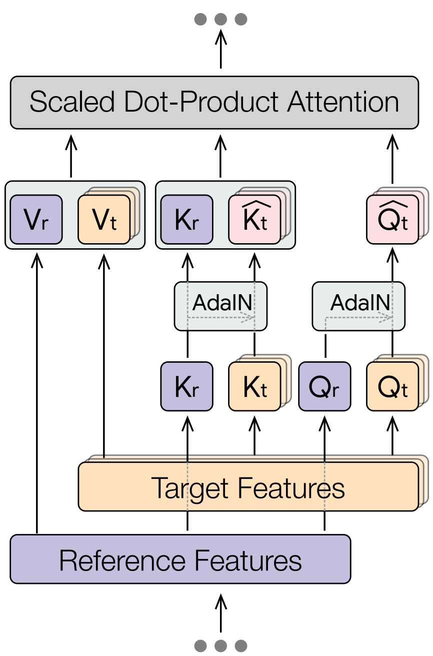

Figure 4. **Shared attention layer.** *The target images attends to the reference image by applying AdaIN over their queries and keys using the reference queries and keys respectively. Then, we apply shared attention where the target features are updated by both the target values V_t and the reference values V_r.*

ing our attention-sharing operation within the self–attention layers that enable style aligned image set generation.

3.1. Preliminaries

Diffusion models [19, 48] are generative latent variable models that aim to model a distribution $p_\theta(x_0)$ that approximates the data distribution $q(x_0)$ and are easy to sample from. Diffusion models are trained to reverse the diffusion "forward process":

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, \ \ \epsilon \sim N(0, I),$$

where $t \in [0, \infty)$ and the values of α_t are determined by a scheduler such that $\alpha_0 = 1$ and $\lim_{t\to\infty} \alpha_t = 0$. During inference, we sample an image by gradually denoising an input noise image $x_T \sim \mathcal{N}(0, I)$ via the reverse process:

$$x_{t-1} = \mu_{t-1} + \sigma_t z, \ \ z \sim N(0, I),$$

where the value of σ_t is determined by the sampler and μ_{t-1} is given by

$$\mu_{t-1} = \frac{\sqrt{\alpha_{t-1}} x_t}{\sqrt{\alpha_t}} + \left(\sqrt{1-\alpha_{t-1}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\right) \epsilon_\theta(x_t, t),$$

where $\epsilon_\theta(x_t, t)$ is the output of a diffusion model parameterized by θ.

Moreover, this process can be generalized for learning a marginal distribution using an additional input condition. That leads text-to-image diffusion models (T2I), where the output of the model $\epsilon_\theta(x_t, t, y)$ is conditioned on a text prompt y.

Self-Attention in T2I diffusion models. State-of-the-art T2I diffusion models [4,35,46] employ a U-Net architecture [40] that consists of convolution layers and transformer attention blocks [54]. In these attention mechanisms, deep

Figure 5. **Ablation study – qualitative comparison.** *Each pair of rows shows two sets of images generated by the same set of prompts "...in minimal flat design illustartion" using different configurations of our method, and each row in a pair uses a different seed. Sharing the self–attention between all images in the set (bottom) results with some diversity loss (style collapse across many seeds) and content leakage within each set (colors from one image leak to another). Disabling the queries–keys AdaIN opeartion results with less consistent image sets compared to our full method (top) which keeps on both diversity between different sets and consistency within each set.*

image features $\phi \in \mathbb{R}^{m \times d_h}$ attend to each other via self-attention layers and to contextual text embedding via cross-attention layers.

Our work operates at the self-attention layers where deep features are being updated by attending to each other. First, the features are projected into queries $Q \in m \times d_k$, keys $K \in m \times d_k$ and values $V \in m \times d_h$ through learned linear layers. Then, the attention is computed by the scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} V\right),$$

where d_k is the dimension of Q and K. Intuitively, each image feature is updated by a weighted sum of V, where the weight depends on the correlation between the projected query q and the keys K. In practice, each self-attention layer consists of several attention heads, and then the residual is computed by concatenating and projecting the attention heads output back to the image feature space d_h:

$$\hat{\phi} = \phi + \text{Multi-Head-Attention}(\phi).$$

3.2. Style Aligned Image Set Generation

The goal of our method is to generate a set of images $\mathcal{I}_1 \dots \mathcal{I}_n$ that are aligned with an input set of text prompts $y_1 \dots y_n$ and share a consistent style interpretation with each other. For example, see the garnered image set of toy objects in Fig. 3 that are style-aligned with each other and to the input text on top. A naïve way to generate a style aligned image set of different content is to use the same style description in the text prompts. As can be seen at the bottom of Fig. 2, generating different images using a shared style description of "in minimal origami style" results in an unaligned set, since each image is unaware of the exact appearance of other images in the set during the generation process.

The key insight underlying our approach is the utilization of the self-attention mechanism to allow communication among various generated images. This is achieved by sharing attention layers across the generated images.

Formally, let Q_i, K_i, and V_i be the queries, keys, and values, projected from deep features ϕ_i of $\mathcal{I}_i$ in the set, then, the attention update for ϕ_i is given by:

$$\text{Attention}(Q_i, K_{1 \dots n}, V_{1 \dots n}), \tag{1}$$

where $K_{1 \dots n} = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{bmatrix}$ and $V_{1 \dots n} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}$. However, we have noticed that by enabling full attention sharing, we may harm the quality of the generated set. As can be seen in Fig. 5 (bottom rows), full attention sharing results in content leakage among the images. For example, the unicorns got green paint from the garnered dino in the set. Moreover, full attention sharing results with less diverse sets of the same set of prompts, see the two sets in Fig. 5 in bottom rows compared to the sets above.

To restrict the content leakage and allow diverse sets, we share the attention to only one image in the generated set

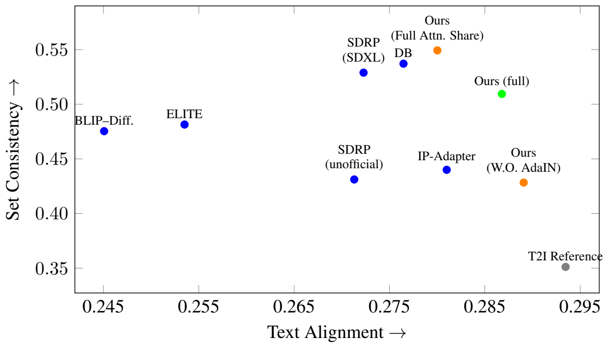

Figure 6. **Quantitative Comparison.** *We compare the results of the different methods (blue marks) and our ablation experiments (orange marks) in terms of text alignment (CLIP score) and set consistency (DINO embedding similarity).*

Table 1. **User evaluation for style aligned image set generation.** *In each question, the user was asked to select between two image sets, Which is better in terms of style consistency and match to the text descriptions (see Sec. 4). We report the percentage of judgments in favor of StyleAligned over 800 answers (2400 in total).*

StyleDrop (unofficial MUSE)	StyleDrop (SDXL)	DreamBooth–LoRA (SDXL)
85.2 %	67.1 %	61.3%

(the first in the batch). That is, *target* image features ϕ_t are attending to themselves and to the features of only one *reference* image in the set using Eq. 1. As can be seen in Fig. 5 (middle), sharing the attention to only one image in the set results in diverse sets that share a similar style. However, in that case, we have noticed that the style of different images is not well aligned. We suspect that this is due to low attention flow from the reference to the target image.

As illustrated in Fig. 4, to enable balanced attention reference, we normalize the queries Q_t and keys K_t of the target image using the queries Q_r and keys K_r of the reference image using the adaptive normalization operation (AdaIN) [22]:

$$\hat{Q}_t = \text{AdaIN}(Q_t, Q_r) \quad \hat{K}_t = \text{AdaIN}(K_t, K_r),$$

where the AdaIn operation is given by:

$$\text{AdaIN}\,(x, y) = \sigma\,(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu_y,$$

and $\mu(x), \sigma(x) \in \mathbb{R}^{d_k}$ are the mean and the standard deviation of queries and keys across different pixels. Finally, our shared attention is given by

$$\text{Attention}(\hat{Q}_t, K_{rt}^T, V_{rt}),$$

where $K_{rt} = \begin{bmatrix} K_r \\ \hat{K}_t \end{bmatrix}$ and $V_{rt} = \begin{bmatrix} V_r \\ V_t \end{bmatrix}$.

4. Evaluations and Experiments

We have implemented our method over Stable Diffusion XL (SDXL) [35] by applying our attention sharing overall 70 self-attention layers of the model. The generation of a four images set takes 29 seconds on a single $A100$ GPU. Notice that since the generation of the reference image is not influenced by other images in the batch, we can generate larger sets by fixing the prompt and seed of the reference image across the set generation.

For example, see the sets in Fig. 2 and 3.

Evaluation set. With the support of ChatGPT,we have generated 100 text prompts describing different image styles over four random objects. For example, “{A guitar, A hot air balloon, A sailboat, A mountain} in papercut art style.” For each style and set of objects, we use our method to generate a set of images. The full list of prompts is provided in the appendix.

Metrics. To verify that each generated image contains its specified object, we measure the CLIP similarity [36] between the image and the text description of the object. In addition, we evaluate the style consistency of each generated set, by measuring the pairwise average cosine similarity between DINO VIT-B/8 [6] embeddings of the generated images in each set. Following [41, 56], we used DINO embeddings instead of CLIP image embeddings for measuring image similarity, since CLIP was trained with class labels and therefore it might give a high score for different images in the set that have similar content but with a different style. On the other hand, DINO better distinguishes between different styles due to its self-supervised training.

4.1. Ablation Study

The quantitative results are summarized in Fig. 6, where the right–top place on the chart means better text similarity and style consistency, respectively. As a reference, we report the score obtained by generating the set of images using SDXL (T2I Reference) using the same seeds without any intervention. As can be seen, our method achieves a much higher style consistency score at the expense of text similarity. See qualitative comparison in Fig. 2

In addition, we compared our method to additional two variants of the shared attention as described in Sec. 3.2. The first variant uses full attention sharing (*Full Attn. Share*) where the keys and values are shared between each pair of images in the set. In the second variant (*W.A. AdaIN*) we omit the AdaIN operation over queries and keys. As expected, this *Full Attn. Share* variant, results with higher style consistency and lower text alignment. As can be seen in Fig. 5, *Full Attn. Share* harms the overall quality of the image sets and diversity across sets. Moreover, our method without the use of AdaIN results in much lower style consistency. Qualitative results can be seen in Fig. 5 and in the supplementary..

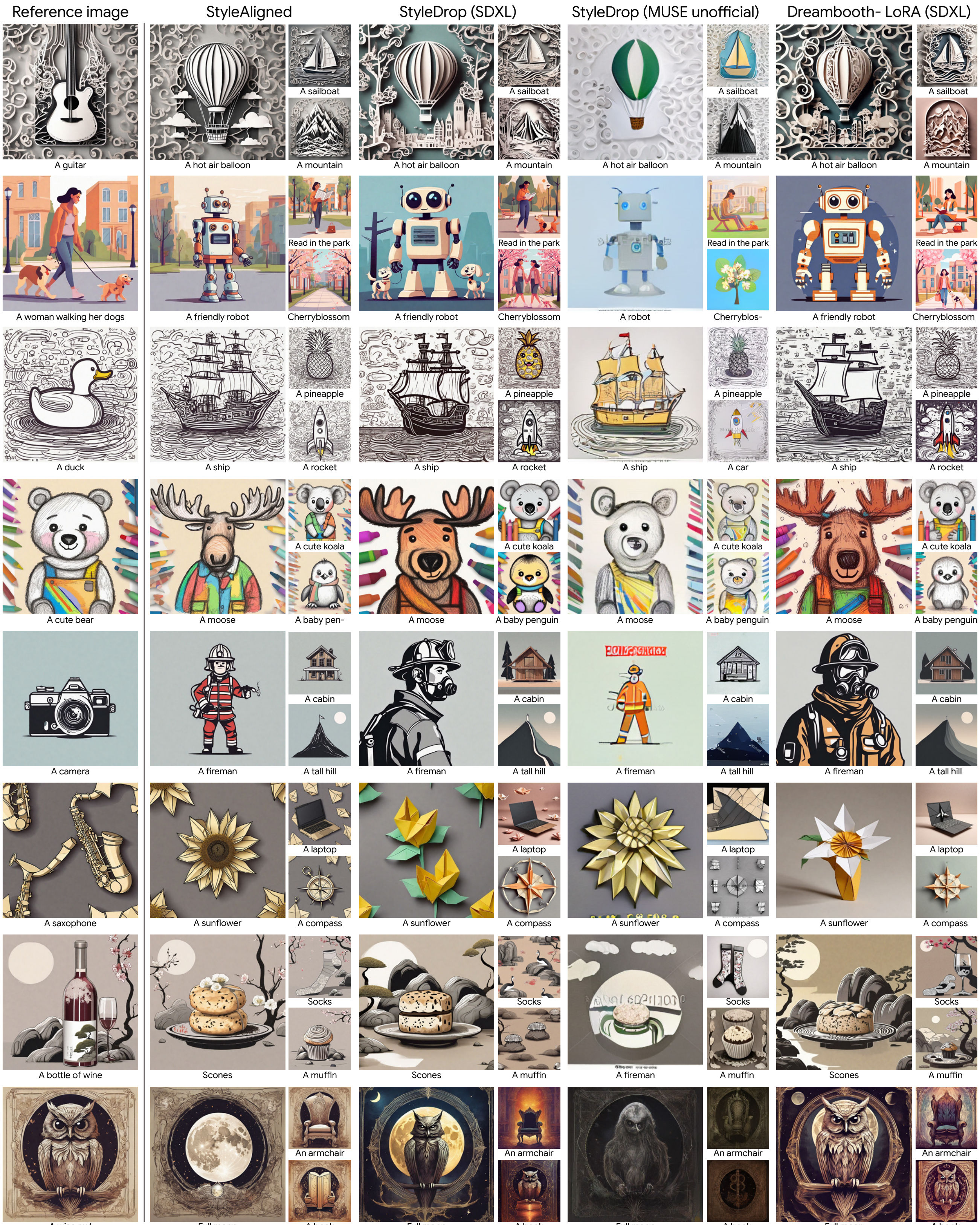

Figure 7. **Qualitative comparison to personalization based methods.**

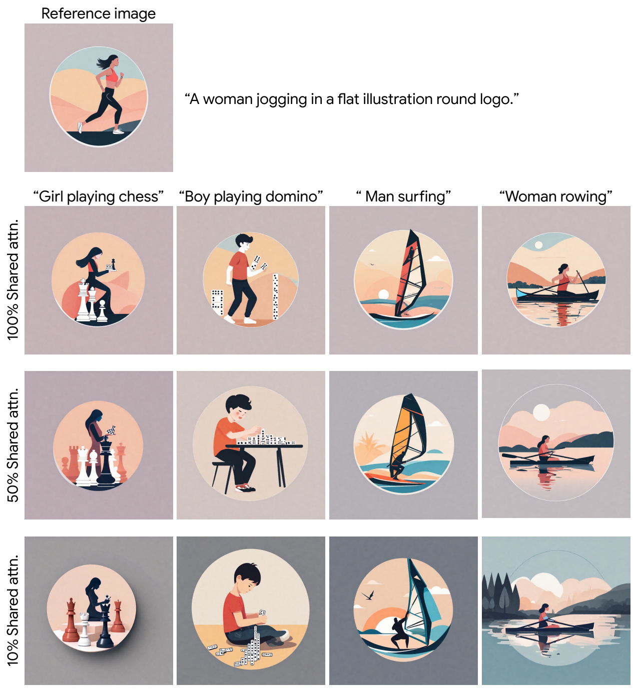

Figure 8. **Varying level of attention sharing.** *By reducing the number of shared attention layers, i.e., allowing only self-attention in part of the layers, we can get more varied results (bottom rows) at the expense of style alignment (top row).*

4.2. Comparisons

For baselines, we compare our method to T2I personalization methods. We trained StyleDrop [49] and DreamBooth [41] over the first image in each set of our evaluation data, and use the trained personalized weights to generate the additional three images in each set. We use a public unofficial implementation of StyleDrop * (SDRP–unofficial) over non-regressive T2I model. Due to the large quality gap between the unofficial MUSE model † to the official one [7], we follow StyleDrop and implement an adapter model over SDXL (SDRP–SDXL) * , where we train a low rank linear layer after each Feed-Forward layer at the model's attention blocks. For training DreamBooth, we adapt the LoRA [21, 43] variant (DB–LoRA) over SDXL using the public huggingface–diffusers implementation ‡. We follow the hyperparameters tuning reported in [49] and train both SDRP–SDXL and DB–LoRA for 400 steps to prevent overfitting to the style training image.

As can be seen in the qualitative comparison, Fig. 7, the image sets generated by our method, are more consistent across style attributes like color palette, drawing style, and composition. Moreover, the personalization-based methods may leak the content of the training reference image (on the left) when generating the new images. For example, see the repeated woman and dogs in the results of DB–LoRA and SDRP–SDXL at the second row or the repeated owl at the bottom row. Similarly, because of the content leakage, these methods obtained lower text similarity scores and higher set consistency scores compared to our method.

We also apply two encoder-based personalization methods ELITE [57], IP–Adapter [59], and BLIP–Diffusion [27] over our evaluation set. These methods receive as input the first image in each set and use its embeddings to generate images with other content. Unlike the optimization-based techniques, these methods operate in a much faster feed-forward diffusion loop, like our method. However, as can be seen in Fig. 6, their performance for style aligned image generation is poor compared to the other baselines. We argue that current encoder-based personalization techniques struggle to disentangle the content and the style of the input image. We supply qualitative results in the supplementary .

User study. In addition to the automatic evaluation, we conducted a user study over the results of our method, StyleDrop (unofficial MUSE), StyleDrop (SDXL), and DreamBooth–LoRA (SDXL). In each question, we randomly sample one of the evaluation examples and show

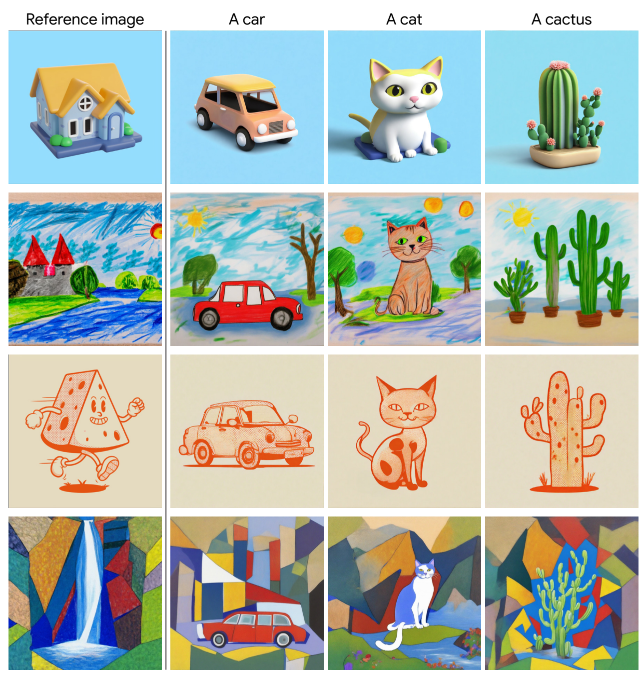

Figure 9. **Style aligned image generation to an input image.** *Given an input reference image (left column) and text description, we first apply DDIM inversion over the image to get the inverted diffusion trajectory* $x_T, x_{T-1} \dots x_0$. *Then, starting from* x_T *and a new set of prompts, we apply our method to generate new content (right columns) with an aligned style to the input.*

*github.com/aim-uofa/StyleDrop-PyTorch
†github.com/baaivision/MUSE-Pytorch
*github.com/amirhertz/sdxl_adapter
‡github.com/huggingface/diffusers

the user the 4 image set that resulted from our and another method (in a random order). The user had to choose which set is better in terms of style consistency, and text alignment. A print screen of the user study format is provided in the supplementary. Overall, we collected 2400 answers from 100 users using the Amazon Mechanical Turk service. The results are summarized in Tab. 1 where for each method, we report the percentage of judgments in our favor. As can be seen, most participants favored our method by a large margin. More information about our user study can be found in the supplementary.

4.3. Additional Results

Style alignment control. We provide means of control over the style alignment to the reference image by applying the shared attention over only part of the self-attention layers. As can be seen in Fig. 8, reducing the number of shared attention layers results with a more diverse image set, which still shares common attributes with the reference image.

StyleAligned from an input image. To generate style-aligned images to an input image, we apply DDIM inversion [50] using a provided text caption. Then, we apply our method to generate new images in the style of the input using the inverted diffusion trajectory $x_T, x_{T-1}, \ldots x_0$ for the reference image. Examples are shown in Fig. 9, where we use BLIP captioning [28] to get a caption for each input image. For example, we used the prompt "A render of a house with a yellow roof" for the DDIM inversion of the top example and replaced the word house with other objects to generate the style-aligned images of a car, a cat, and a cactus. Notice that this method does not require any optimization. However, DDIM inversion may fail [30] or results with an erroneous trajectory [24]. More results and analysis, are provided in the supplementary.

StyleAligned with other methods. Since our method doesn't require training or optimization, it can be easily combined on top of other diffusion based methods to generate style-consistent image sets. Fig. 10 shows several such examples where we combine our method with ControlNet [60], DreamBooth [42] and MultiDiffusion [3]. More examples and details about the integration of StyleAligned with other methods can be found in supplementary.

5. Conclusions

We have presented StyleAligned, which addresses the challenge of achieving style-aligned image generation within the realm of large-scale Text-to-Image models. By introducing minimal attention sharing operations with AdaIN modulation during the diffusion process, our method successfully establishes style-consistency and visual coherence across generated images. The demonstrated efficacy of StyleAligned in producing high-quality, style-consistent images across diverse styles and textual prompts underscores its potential in creative domains and practical applications. Our results affirm StyleAligned capability to faithfully adhere to provided descriptions and reference styles while maintaining impressive synthesis quality.

In the future we would like to explore the scalability and adaptability of StyleAligned to have more control over the shape and appearance similarity among the generated images. Additionally, due to the limitation of current diffusion inversion methods, a promising direction is to leverage StyleAligned to assemble a style-aligned dataset which then can be used to train style condition text-to-image models.

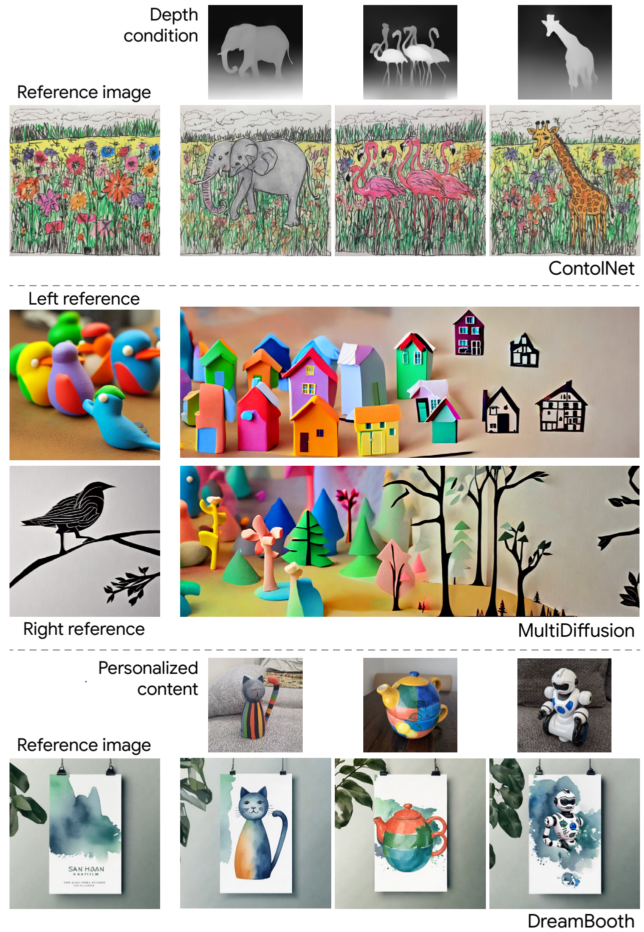

Figure 10. **StyleAligned with other methods.** *On top, StyleAligned is combined with ControlNet to generate style-aligned images conditioned on depth maps. In the middle, our method combined with MultiDiffusion to generate panorama images that share multiple styles. On the bottom, style consistent and personalized content created by combining our method with pre-trained personalized DreamBooth–LoRA models.*

6. Acknowledgement

We thank Or Patashnik, Matan Cohen, Yael Pritch, and Yael Vinker for their valuable inputs that helped improve this work.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023. 2

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), jul 2023. 2

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 8

[4] James Betker, Gabriel Goh, Li Jing, TimBrooks, Jianfeng Wang, Linjie Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. 2023. 3

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 5

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, 2023. 2, 7

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42:1 – 10, 2023. 2

[9] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[10] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 571–576. 2023. 2

[11] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999. 2

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 2

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

[14] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[16] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *ArXiv*, abs/2303.11305, 2023. 2

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[18] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 557–570. 2023. 2

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020. 3

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 7

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 5

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[24] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023. 2, 8

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[26] Hochang Lee, Sanghyun Seo, Seungtaek Ryoo, and Kyunghyun Yoon. Directional texture transfer. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering*, pages 43–48, 2010. 2

[27] Dongxu Li, Junnan Li, and Steven C. H. Hoi. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *ArXiv*, abs/2305.14720, 2023. 2, 7

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 8

[29] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 8

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2021. 2

[32] 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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 2

[33] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[34] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. *ArXiv*, abs/2303.11306, 2023. 2

[35] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. 2, 3, 5

[36] 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. In *International Conference on Machine Learning*, 2021. 5

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[38] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[41] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 5, 7

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2, 8

[43] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. `https://github.com/cloneofsimo/lora`, 2022. 7

[44] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 2021. 2

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[47] Jing Shi, Wei Xiong, Zhe L. Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *ArXiv*, abs/2304.03411, 2023. 2

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[49] 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. 1, 2, 7

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 8

[51] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[52] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 2

[53] 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*, pages 1921–1930, 2023. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[55] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. *arXiv preprint arXiv:2211.13752*, 2022. 2

[56] Andrey Voynov, Q. Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. *ArXiv*, abs/2303.09522, 2023. 2, 5

[57] Yuxiang Wei. Official implementation of ELITE. `https://github.com/csyxwei/ELITE`, 2023. 7

[58] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. *ArXiv*, abs/2302.13848, 2023. 2

[59] Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv*, abs/2308.06721, 2023. 2, 7

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 8