Real-World Image Variation by Aligning Diffusion Inversion Chain – Appendix

Anonymous Author(s)
Affiliation
Address
email

Contents

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

A Basic Background of Diffusion Models

This section uses a modified background description provided in [5]. We only consider the condition-free case for the diffusion model here. Diffusion Denoising Probabilistic Models (DDPMs) [3] are generative latent variable models designed to approximate the data distribution $q(x_0)$. The diffusion operation starts from the latent x_0, adding step-wise noise to diffuse data into pure noise x_T. It's important to note that this process can be viewed as a Markov chain starting from x_0, where noise is gradually added to the data to generate the latent variables $x_1, \dots, x_T \in X$. The sequence of latent variables follows the conditional distribution $q(x_1, \dots, x_t \mid x_0) = \prod_{i=1}^{t} q(x_t \mid x_{t-1})$. Each step in the forward process is defined by a Gaussian transition $q(x_t \mid x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-k_t}x_{t-1}, k_t I)$, which is parameterized by a schedule $k_0, \dots, k_T \in (0, 1)$. As T becomes sufficiently large, the final noise vector x_T approximates an isotropic Gaussian distribution.

The forward process allows us to express the latent variable x_t directly as a linear combination of noise and x_0, without the need to sample intermediate latent vectors.

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}w, \quad w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{11}$$

where $\alpha_t := \prod_{i=1}^{t}(1-k_i)$. To sample from the distribution $q(x_0)$, a reversed denoising process is defined by sampling the posteriors $q(x_{t-1} \mid x_t)$, which connects isotropic Gaussian noise x_T to the actual data. However, the reverse process is computationally challenging due to its dependence on the unknown data distribution $q(x_0)$. To overcome this obstacle, an approximation of the reverse process with a parameterized Gaussian transition network denoted as $p_\theta(x_{t-1} \mid x_t)$, where $p_\theta(x_{t-1} \mid x_t)$ follows a normal distribution with mean $\mu_\theta(x_t, t)$ and covariance $\Sigma_\theta(x_t, t)$. As an alternative approach, the prediction of the noise $\epsilon_\theta(x_t, t)$ added to x_0, which is obtained using equation 11, can replace the use of $\mu_\theta(x_t, t)$ as suggested in [3]. Bayes' theorem could be applied to approximate

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{k_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right). \tag{12}$$

Once we have a trained $\epsilon_\theta(x_t, t)$, we can using the following sample method

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, \quad z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{13}$$

In DDIM sampling [11], a denoising process could become deterministic when set $\sigma_t = 0$.

B Details of the Attention Pipeline

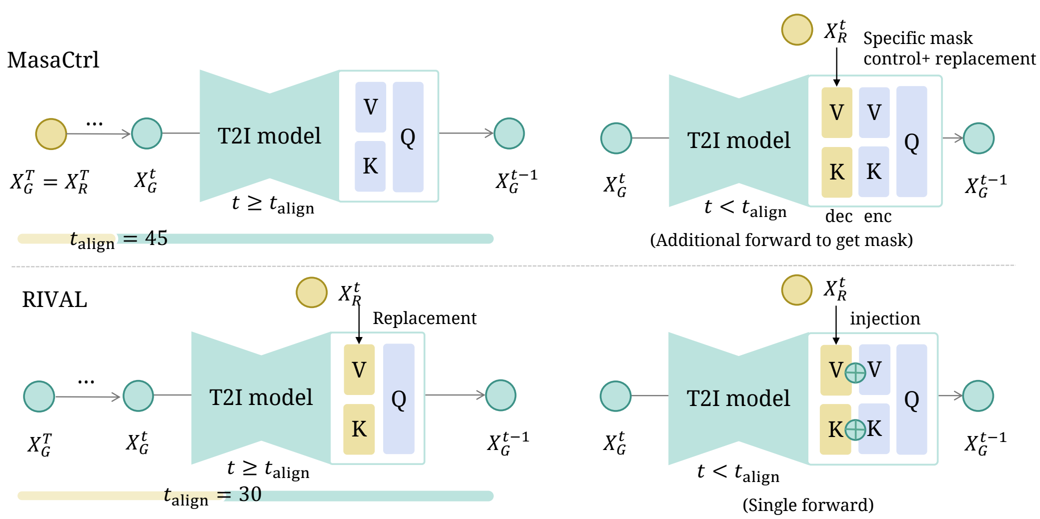

Figure 11: Self-attention control compare with MasaCtrl [2]. The default split of two stages is shown as a bar for each method.

We present a comparative analysis of attention injection methods. As depicted in Fig. 11, MasaCtrl [2], while also adopting a self-attention injection approach, employs a more complex control mechanism in its second stage. In the first stage of MasaCtrl, the inverted latent representation X_T^R is directly utilized by applying a modified prompt. In the second stage, a cross-attention mask is introduced to control specific word concepts modified in the prompt, which requires an additional forward pass. In contrast, our proposed method, RIVAL, primarily focuses on generating inconsistent variations. Consequently, we aim to guide feature interaction by replacing KV features with an aligned latent distribution. Unlike MasaCtrl, our approach does not limit content transfer through editing prompts with only a few words. Hence, in the second stage, we employ a single forward pass without calculating an additional cross-attention mask, allowing fast and flexible text-to-image generation with diverse text prompts.

In recent updates, ControlNet [15] has incorporated an attention mechanism resembling the second stage of RIVAL to address image variation. However, a notable distinction lies in using vanilla noised latents as guidance, leading to a process akin to the attention-only approach employed in RePaint [4] with the Stable Diffusion model. Consequently, this methodology is limited to generating images within the fine-tuned training data domain.

C More About Comparisons

Implementation Details. We compare our work with ELITE [14], Stable Diffusion image variation [6], and DALL·E 2 [8]. We utilize the official demo of ELITE to obtain results. To extract context tokens, we mask the entire image and employ the phrase *"A photo/painting of <S>."* based on the production method of each test image. Inference for ELITE employs the default setting with denoising steps set to $T = 300$. For Stable Diffusion's image variation version, we utilize the default configuration, CFG guidance $m = 3$, and denoising steps $T = 50$. In the case of DALL·E 2, we utilize the official image variation API, specifically requesting using the most advanced API available to generate images of size 1024×1024.

Comparison with UnCLIP. UnCLIP [8], also known as DALL·E 2, is an image generation framework trained using image CLIP features as direct input. Thanks to its large-scale training and image-direct conditioning design, it generates variations solely based on image conditions when adapted to image variation. However, when faced with hybrid image-text conditions, image-only UnCLIP struggles to produce satisfactory results, particularly when CLIP does not recognize the image content correctly. We provide comparative analysis in Fig. 12. Additionally, we demonstrate in the last two columns of Figure 12 that our approach can enhance the accuracy of low-level details in open-source image variation methods such as SD image variation [6].

Additional Visual Results. We showcase additional results of our techniques in variation generation, as illustrated in Fig. 13, and text-driven image generation with image condition, as shown in Fig. 14. The results unequivocally demonstrate the efficacy of our approach in generating a wide range of image variations that accurately adhere to textual and visual guidance.

D Additional Ablation Results

Ablation on early fusion step. In addition to Fig. 8 of the main paper, we present comprehensive early-step evaluation results based on a grid search analysis in Fig. 15. By decreasing the duration of the feature replacement stage (larger t_{align}), we observe an increase in the similarity of textures and contents in the generated images. However, excessively long or short early latent alignment durations (t_{early}) can lead to color misalignment. Users can adjust the size of the early fusion steps as hyperparameters to achieve the desired outcomes.

Ablation on different alignment designs. Fig. 16 illustrates ablations conducted on various alignment designs. Two latent initialization methods, as formulated in Eq. (7) and Eq. (8), exhibit comparable performance. Nevertheless, incorporating alignments in additional areas, such as hidden states within each transformer block, may harm performance. Hence, we opt for our RIVAL pipeline's simplest noise alignment strategy.

Ablation on different text conditions. We conduct ablations on text conditions in three aspects. First, we evaluate the impact of different CFG scales m for text prompt guidance. As shown in Fig. 17 (a), our latent rescaling technique enables control over the text guidance level while preserving the reference exemplar's low-level features. Second, we employ an optimization-based

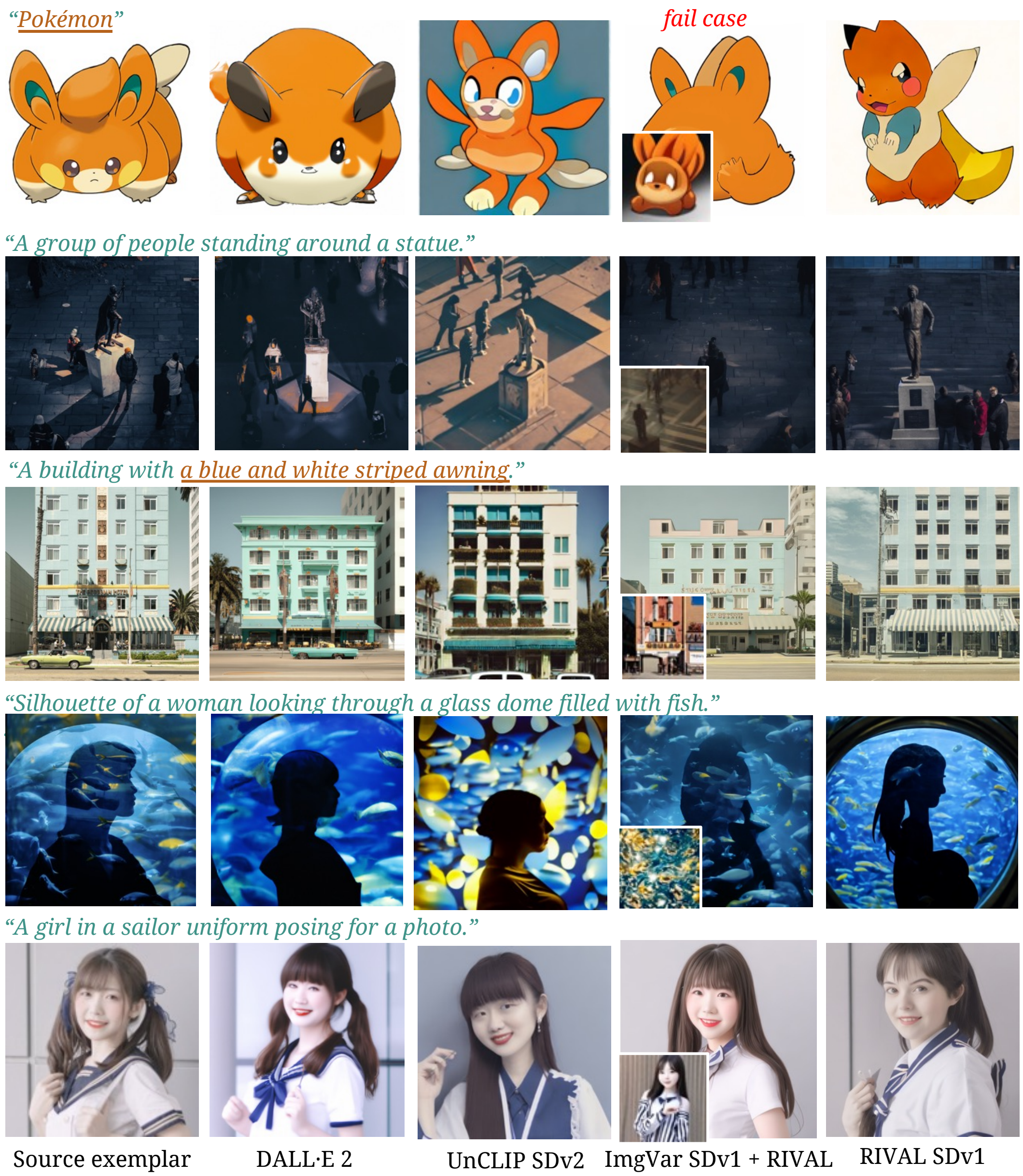

Figure 12: Comparision and adaptation with UnCLIP [8]. We highlight texts that enhance the image understanding for each case. Our inference pipeline is adapted to the image variation model depicted in the fourth column, in contrast to the variation achieved through vanilla inference in the bottom left corner of each image.

null-text inversion method [5] to obtain an inversion chain with improved reconstruction quality. However, this method is computationally intensive, and the optimized embeddings are sensitive to the guidance scale. Furthermore, when incorporating this optimized embedding into the unconditional inference branch, there is a variation in generation quality, as depicted in Fig. 17 (b). Third, we utilize empty text as the source prompt to obtain the latents in the inversion chain while keeping the target prompt unchanged. As depicted in Fig. 17 (c), the empty text leads to weak semantic content correspondence between the two chains but sometimes benefits text-driven generation. For example, if users do not want to transfer the "gender" concept to the generated robot.

Figure 13: Text-driven free-form image generation results. The image reference is in the left column. In the last row, we also present variations for one customized concept <sks> bag.

Figure 14: Text-driven free-form image generation results, with the image reference placed in the top left corner. The text prompts used are identical to those presented in Fig. 5 of the main paper. Every two rows correspond to a shared text prompt.

method base model	SD [9] V1-5	ImgVar [6] V1-3	ELITE [14] V1-4	UnClIP [8] V2-1	**RIVAL** V1-5
KID ↓	17.1	18.5	25.7	<u>13.5</u>	**13.2**

Table 2: Quantitative comparisons for KID ($\times 10^3$). All methods are Stable Diffusion based.

E Quantitative Evaluations

This section comprehensively evaluates our proposed method with various carefully designed metrics, including CLIP Score, color palette matching, user study, and KID.

CLIP Score. For evaluating the CLIP Score, we employ the official ViT-Large-Patch14 CLIP model [7] and compute the cosine similarity between the projected features, yielding the output.

Color palette matching. To perform low-level matching, we utilize the Pylette tool [12] to extract a set of 10 palette colors. Subsequently, we conduct a bipartite matching between the color palette of each generated image and the reference palette colors in the RGB color space. Before matching, each color is scaled to $[0, 1]$. The matching result is obtained by calculating the sum of L1 distances.

User study. To evaluate the effectiveness of our approach against other methods, we conducted a user study using an online form. The user study interface, depicted in Figure 18, was designed to elicit user rankings of image variation results. We collected 41 questionnaire responses, encompassing 16 cases of ranking comparisons.

KID evaluation. To provide a comprehensive assessment of the quality, we utilize Kernel Inception Distance (KID)[1] to evaluate the perceptual generation quality of our test set. As depicted in Table2, with Stable Diffusion V1-5, our method achieves the best KID score, which is slightly superior to the UnCLIP [8], employing the advanced Stable Diffusion V2-1.

Figure 15: Ablation results for alignment steps, with the reference exemplar at the bottom right. We fix each generation's initial latent X_G^T.

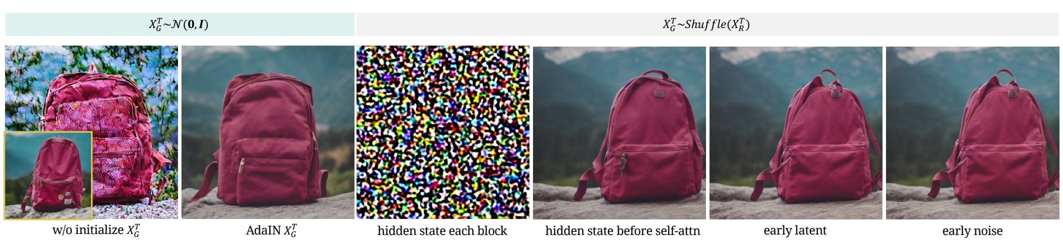

Figure 16: Ablation studies for different feature alignment strategies.

F Additional Considerations

Correction of an equation error in the main paper. In the main paper, it has been identified that an error exists in Equation (4). The residual should be applied after completing the entire self-attention process. Therefore, the updated output of the hidden state in the self-attention mechanism is expressed as follows:

$$\mathbf{v}_G^* = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{14}$$

We will correct this equation in the updated version of the main paper.

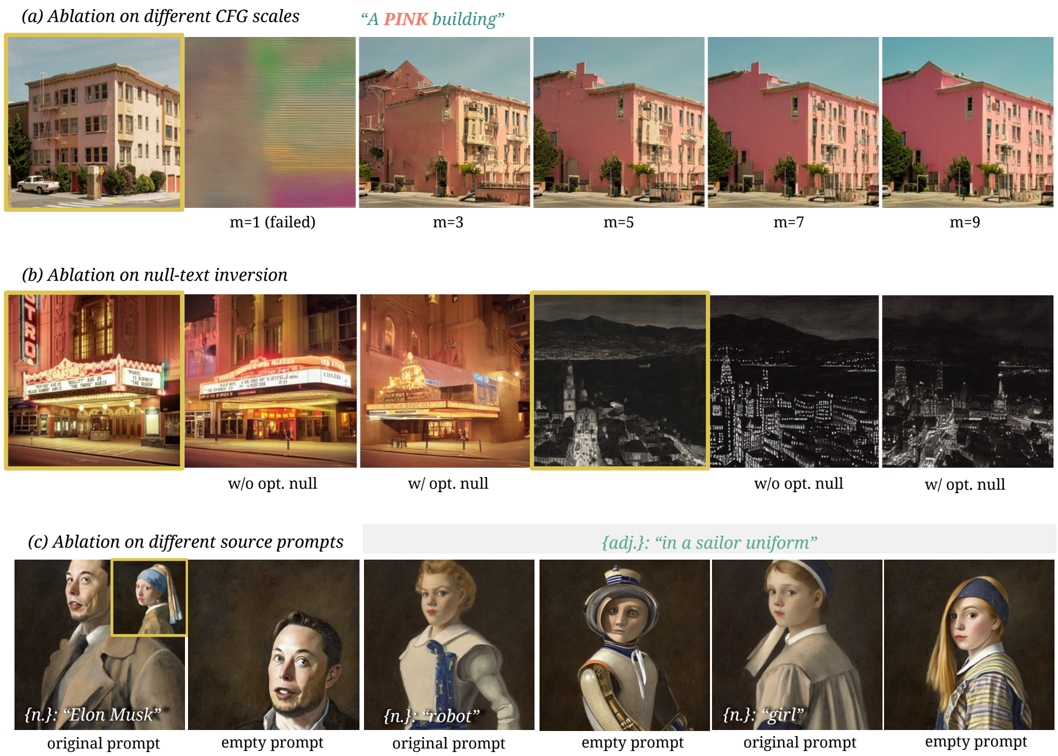

Figure 17: Ablation studies on different text conditions and guidance scales. Reference exemplars are highlighted with a golden border.

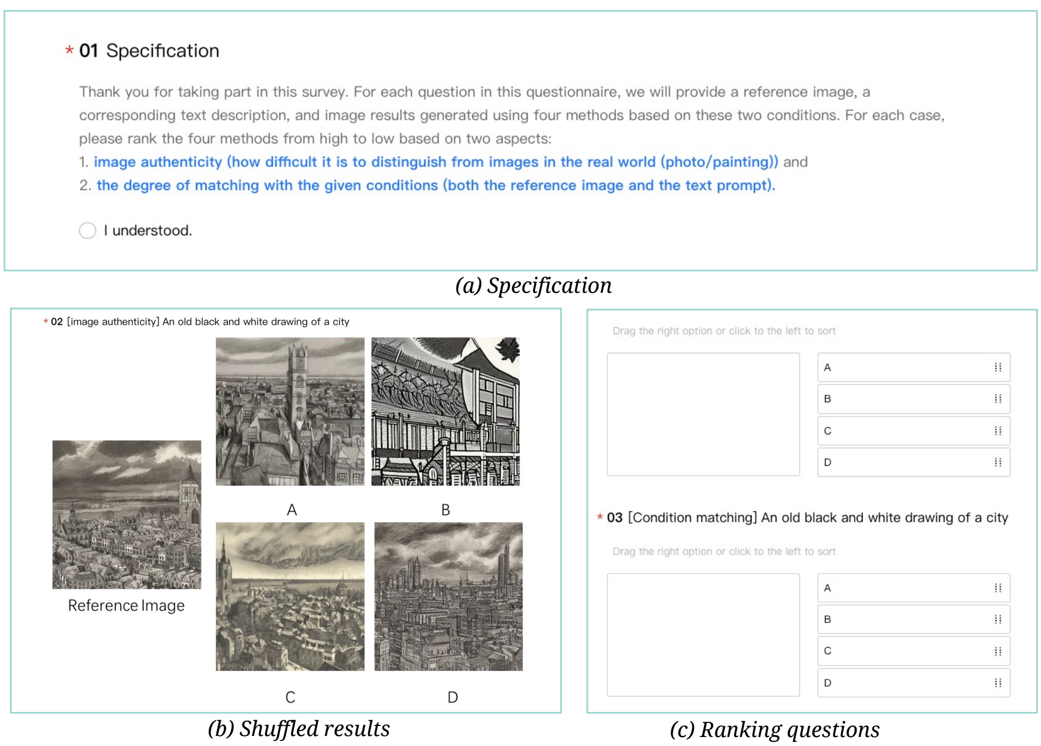

Figure 18: User study user interface. In this case, four methods are: (A). SD ImageVar [6], (B). ELITE [14], (C). DALL·E 2[8], (D). RIVAL (ours).

Data acquisition. To comprehensively evaluate our method, we collected diverse source exemplars from multiple public datasets, such as DreamBooth [10] and Interactive Video Stylization [13]. Some exemplars were obtained from Google and Behance solely for research purposes. We will not release our self-collected example data due to license restrictions.

Societal impacts. This paper introduces a novel framework for image generation that leverages a hybrid image-text condition, facilitating the generation of diverse image variations. Although this application has the potential to be misused by malicious actors for disinformation purposes, significant advancements have been achieved in detecting malicious generation. Consequently, we anticipate that our work will contribute to this domain. In forthcoming iterations of our method, we intend to introduce the NSFW (Not Safe for Work) test for detecting possible malicious generations. Through rigorous experimentation and analysis, our objective is to enhance comprehension of image generation techniques and alleviate their potential misuse.

References

[1] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[2] 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. *arXiv preprint arXiv:2304.08465*, 2023. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[4] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, June 2022. 3

[5] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 4

[6] Justin Pinkney. Experiments with stable diffusion. `https://github.com/justinpinkney/stable-diffusion`, 2023. 3, 6, 8

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 6

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 4, 6, 8

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 6

[10] 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 *CVPR*, 2023. 9

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[12] Ivar Stangeby. A python library for extracting color palettes from supplied images. `https://github.com/qTipTip/Pylette`, 2022. 6

[13] Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menclei Chai, Sergey Tulyakov, and Daniel Sỳkora. Interactive video stylization using few-shot patch-based training. *ACM Transactions on Graphics (TOG)*, 39(4):73–1, 2020. 9

[14] 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 preprint arXiv:2302.13848*, 2023. 3, 6, 8

[15] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3