

PhotoDoodle: Learning Artistic Image Editing from Few-Shot Examples

Shijie Huang^{1,5*} Yiren Song^{1,5*} Yuxuan Zhang^{2,5} Hailong Guo^{3,5}
 Xueyin Wang⁴ Mike Zheng Shou^{1†} Jiaming Liu^{5‡}
¹National University of Singapore ²Shanghai Jiao Tong University
³Beijing University of Posts and Telecommunications ⁴Byte Dance
⁵Tiamat

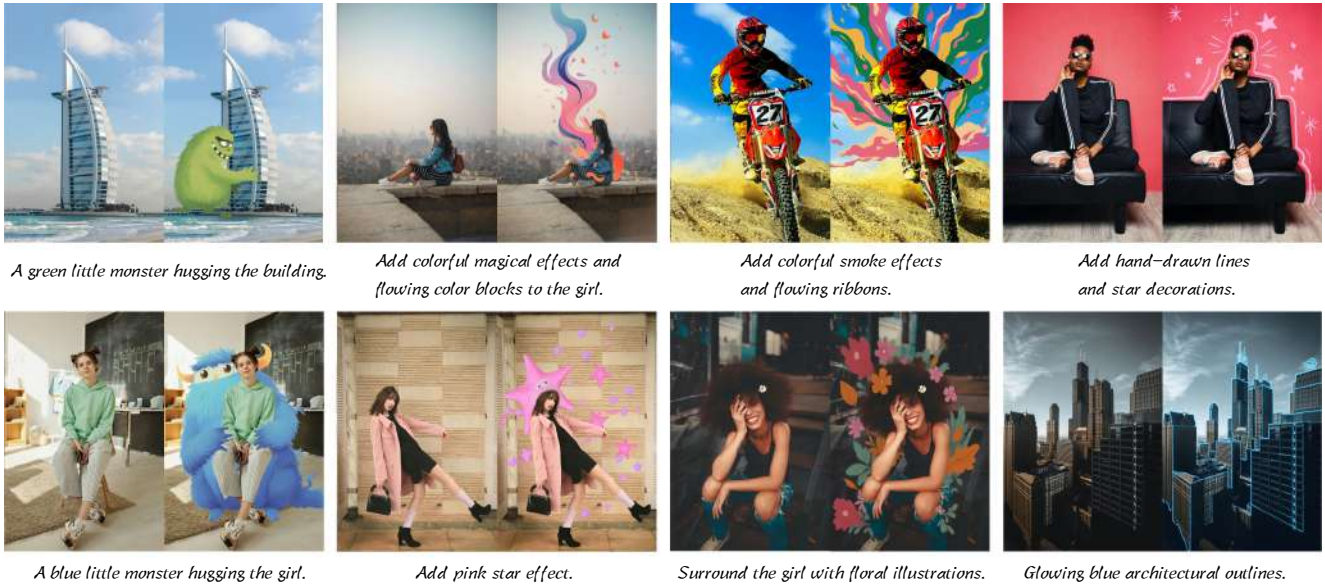


Figure 1. PhotoDoodle can mimic the styles and techniques of human artists in creating photo doodles, adding decorative elements to photos while maintaining perfect consistency between the pre- and post-edit states.

Abstract

We introduce PhotoDoodle, a novel image editing framework designed to facilitate photo doodling by enabling artists to overlay decorative elements onto photographs. Photo doodling is challenging because the inserted elements must appear seamlessly integrated with the background, requiring realistic blending, perspective alignment, and contextual coherence. Additionally, the background must be preserved without distortion, and the artist’s unique style must be captured efficiently from limited training data. These requirements are not addressed by previous methods that primarily focus on global style transfer or

regional inpainting. The proposed method, PhotoDoodle, employs a two-stage training strategy. Initially, we train a general-purpose image editing model, OmniEditor, using large-scale data. Subsequently, we fine-tune this model with EditLoRA using a small, artist-curated dataset of before-and-after image pairs to capture distinct editing styles and techniques. To enhance consistency in the generated results, we introduce a positional encoding reuse mechanism. Additionally, we release a PhotoDoodle dataset featuring six high-quality styles. Extensive experiments demonstrate the advanced performance and robustness of our method in customized image editing, opening new possibilities for artistic creation. Code is released at <https://github.com/showlab/PhotoDoodle>.

*Equal contribution.

†Corresponding author.

‡Project leader.

1. Introduction

The rise of diffusion models has started a new chapter for image creation and control. Using the pretrain-finetune approach, the community has achieved remarkable progress in customized image generation [9, 13, 27], with applications spanning identity preservation [20], artistic stylization [1, 30, 32, 47], and subject coherence [9, 13, 15, 17, 27, 28, 51]. These advancements have fueled applications ranging from digital art creation to commercial design.

Despite these successes, there remains a critical gap between customized image generation and image editing. Current methods primarily focus on content creation, while intelligent context-aware editing, particularly for artistic enhancement, remains underexplored. This imbalance stands in contrast to the growing demand for precision image editing tools, with current approaches focusing on content generation rather than context-aware editing. As a result, this underexplored frontier of customized image editing is both a technical challenge and a key to unlocking transformative applications.

PhotoDoodling, as a paradigmatic challenge in customized image editing, demands artists to enhance background photographs through strategic integration of decorative elements (e.g., stylized linework, ornamental patterns) and context-aware modifications for personalized aesthetics. Conventional workflows involve artistic techniques, such as: (1) local stylization, (2) decorative contour rendering, (3) semantic-aware object insertion, and (4) ornamental augmentation. While these processes showcase distinctive artistic signatures and strategic design logic, their manual execution incurs prohibitive time costs, fundamentally constraining both production scalability and the curation of large-scale paired training datasets required for data-driven approaches. However, automating these workflows introduces three interlocked technical barriers: First, harmonious integration demands generated decorations to simultaneously satisfy perspective alignment and semantic coherence with background contexts. Second, strict background preservation requires mechanisms to prevent unintended changes, such as color distribution shifting and texture pattern alternation. Third, efficient style distillation must extract artists’ unique editing patterns from sparse pairwise examples (30-50 image pairs). These compounded challenges cause existing methods to be incompetent in addressing the problem in a comprehensive way.

Prevailing image editing paradigms can hardly deal with these challenges altogether. Global editing methods (e.g., Prompt-to-Prompt[12], InstructP2P[4]), while effective for consistent style transfer, inadvertently distort background content during localized modifications. Inpainting-based approaches[44, 52], though capable of preserving unmasked regions through localized editing, impose impractical demands for pixel-perfect user-defined masks, funda-

mentally conflicting with the need for automatic PhotoDoodling. To bridge this gap, we introduce an instruction-guided image editing framework that discarded mask dependency, enabling precise and style-conscious decorative generation while ensuring background consistency.

Our proposed framework, PhotoDoodle, presents a few-shot artistic image editing framework built upon Diffusion Transformers (DiT), featuring a dual-stage training architecture. In the first phase, we evolve a pre-trained text-to-image DiT model into a universal image editor (OmniEditor) through two key innovations: (1) a Positional Encoding (PE) Cloning mechanism that preserves spatial fidelity by providing coordinate-aware hints, and (2) a noise-free conditioning paradigm that offers non-distorted information of the source image. This foundational stage is trained on 3.5M image editing pairs[10], establishing robust general editing capabilities. The second phase introduces an EditLoRA module that distills artist-specific editing patterns from merely 30-50 exemplar pairs through low-rank adaptation(LoRA), enabling efficient style customization while maintaining base model’s capability. This co-designed architecture ensures a balance between artistic flexibility and strict consistency.

In summary, our contributions are threefold:

- First framework for artistic photo-doodling: A DiT-based architecture to enable few-shot learning of style-specific editing operations while preserving background integrity.
- We propose a noise-free conditioning paradigm with positional encoding cloning for implicit feature alignment, enabling high-fidelity image editing through EditLoRA-enhanced Diffusion Transformers (DiTs) that efficiently learn customized operations while maintaining strict background consistency
- We collected the first publicly available curated photo-doodle dataset comprising 300+ high-quality pairs across 6 artistic styles, establishing a benchmark for reproducible research.

2. Related work

2.1. Diffusion Model and Conditional Generation

Recent advances in diffusion models have significantly advanced the state of text-conditioned image synthesis, achieving remarkable equilibrium between generation diversity and visual fidelity. Pioneering works such as GLIDE [22], hierarchical text-to-image models [25], and photorealistic synthesis frameworks [29] have systematically addressed key challenges in cross-modal generation tasks. The emergence of Stable Diffusion [26], which implements a Latent Diffusion Model with text-conditioned UNet architecture, has established new benchmarks in text-to-image generation and inspired subsequent innovations including SDXL [24], GLiGEN [19], and Ranni [8]. To enhance

domain-specific adaptation, parameter-efficient fine-tuning techniques like Low-Rank Adaptation (LoRA) [13] and DreamBooth [27] have demonstrated effective customization of pre-trained models. Concurrently, research efforts have focused on precise control over pictorial concepts through multi-concept customization [?], image prompt integration [48?], and identity-preserving generation [38]. The introduction of ControlNet [45] further extended controllable generation capabilities to spatial constraints and depth information, with subsequent extended applications on various scenarios [11, 39, 49, 50]. Some work [6, 34, 35, 37] also focuses on the security issues of customized generation. Building upon these foundations, our work investigates the novel application of pre-trained diffusion transformers for generating artistic photo-doodles. Unlike previous approaches focusing on photorealism or explicit control modalities, we explore the model’s potential in capturing freehand artistic expression while maintaining structural coherence.

2.2. Text-guided Image Editing

Recent advances in text-guided image editing have established this field as a critical research frontier in visual content manipulation, with current methodologies generally classified into three paradigms: global description-guided, local description-guided, and instruction-guided editing. Global description-guided approaches (e.g., Prompt2Prompt [12], Imagic [16], EditWorld [41], ZONE [18]), achieve fine-grained manipulation through cross-modal alignment between textual descriptions and image regions, yet demand meticulous text specification for target attributes. Local description-guided methods such as Blended Diffusion [2] and Imagen Editor [40] enable pixel-level control via explicit mask annotations and regional text prompts, though their practical application is constrained by the requirement for precise spatial specifications, particularly in complex editing scenarios like object removal. The emerging instruction-guided paradigm, exemplified by InstructPix2Pix [4] and HIVE [46], represents a paradigm shift through its natural language interface that accepts editing commands (e.g., ”change the scene to spring”). This approach eliminates the dependency on detailed textual descriptions or precise mask annotations, significantly enhancing user accessibility.

2.3. Image and video Doodles

Image and video doodling involve the creative process of adding hand-drawn elements or animations to static images or video content, blending aspects of graphic design, illustration, and animation to produce playful and engaging visual styles. In academic research, advanced techniques [3, 42, 43] have been developed to automate parts of this process. These methods enable users to generate in-

tricate doodle animations from simple textual descriptions, sketches, or keyframes. By preserving artistic intent and reducing the time and expertise required, these models make multimedia content creation more accessible and efficient. Video doodling extends these capabilities to dynamic video sequences, allowing static doodle elements to be seamlessly integrated and animated in response to video motion and context. This innovation not only enhances interactivity but also introduces greater complexity and realism, making it a powerful tool for creative applications in entertainment, education, and storytelling.

3. Method

In this section, we begin by exploring the preliminaries on diffusion transformer as detailed in Section 3.1. Next, Section 3.2 outlines our three-stage system design. We then detail two key innovations—OmniEditor Pre-training (Section 3.3) and EditLoRA for style adaptation (Section 3.4). Finally, Section 3.5 describes how we built the PhotoDoodle dataset.

3.1. Preliminary

The Diffusion Transformer (DiT) [23] powers modern image generators like Stable Diffusion 3 [7] and PixArt [5]. At its core, DiT uses a special transformer network that denoise the noisy image tokens step-by-step.

DiT operates on two categories of tokens: noisy image tokens $z \in \mathbb{R}^{N \times d}$ and text condition tokens $c_T \in \mathbb{R}^{M \times d}$, where d denotes the embedding dimension, and N and M represent the numbers of image and text tokens, respectively. These tokens retain their shapes consistently as they propagate through multiple transformer layers.

In FLUX.1, each DiT block incorporates layer normalization, followed by Multi-Modal Attention (MMA) [7], which employs Rotary Position Embedding (RoPE) [36] to encode spatial information. For image tokens z , RoPE applies rotation matrices based on the token’s position (i, j) in the 2D grid:

$$z_{i,j} \rightarrow z_{i,j} \cdot R(i, j), \quad (1)$$

where $R(i, j)$ represents the rotation matrix corresponding to position (i, j) . Similarly, text tokens c_T are transformed with their positions designated as $(0, 0)$.

The multi-modal attention mechanism projects these position-encoded tokens into query Q , key K , and value V representations, enabling attention computation across all tokens:

$$\text{MMA}([z; c_T]) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad (2)$$

where $Z = [z; c_T]$ denotes the concatenation of image and text tokens. This approach ensures bidirectional attention between the tokens.

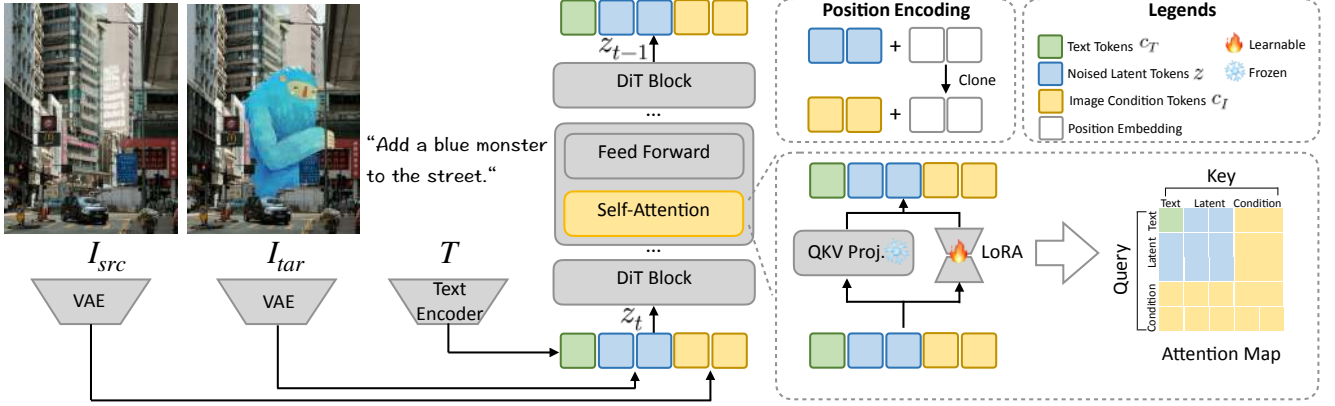


Figure 2. The overall architecture and training paradigm of photodoodle. The omniEditor and EditLoRA all follow the lora training paradigm. We use a high rank lora for pre-training the OmniEditor on a large-scale dataset for general-purpose editing and text-following capabilities, and a low rank lora for fine-tuning EditLoRA on a small set of paired stylized images to capture individual artists’ specific styles and strategies for efficient customization. We encode the source image into a condition token and concatenate it with a noised latent token, controlling the generation outcome through MMAttention.

3.2. Overall Architecture

The overall architecture of PhotoDoodle is illustrated in Fig. 2. It comprises the following stages:

Pre-training OmniEditor. The OmniEditor is pre-trained on a large-scale image editing dataset to acquire general-purpose image editing capabilities and strong text-following abilities. This stage ensures the model’s capability in diverse editing tasks.

Fine-tuning EditLoRA. After pre-training, EditLoRA is fine-tuned on a small set of pairwise stylized images (20–50 pairs) to learn the specific editing styles and strategies of individual artists. This stage enables efficient customization for personalized editing needs.

Inference. During inference, the input source image I_{src} is encoded as condition tokens c_I via VAE. We then randomly sample Gaussian noise as image tokens z , cloning the Position Encoding from condition tokens, and concatenate the tokens along the sequence dimension. Subsequently, we apply the flow matching method to predict the target velocity, iterating multiple steps to obtain the predicted image latent representation. Finally, the predicted image tokens are converted by the VAE decoder to achieve the final predicted image.

3.3. OmniEditor Pre-training

We denote the pre-edited image as the source image I_{src} and the post-edited image as the target image I_{tar} . Previous works, such as SDEdit, model image editing as an adding-denoising problem, they often altering unintended areas. Others like InstructP2P[4] redesign core model parts, significantly degrading the capacity of the pretrained t2i models. Unlike them, our approach, PhotoDoodle, conceptualizes image editing as a conditional generation problem and

minimizes the modification of pretrained text-to-image DiT.

Our model leverages the advanced capabilities of the DiT-based pretrained model, and extends it to function as an image editing tool. Both I_{src} and I_{tar} are encoded into their respective latent representations, c_I and z , via a VAE. After applying position encoding cloning, the latent tokens are then concatenated along the sequence dimension to perform joint attention. The Multi-modal attention mechanisms are used to provide conditional information for the denoising of the doodle image.

$$\text{MMA}([z; c_i; c_T]) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V, \quad (3)$$

where $Z = [z; c_i; c_T]$ denotes the concatenation of noised latent tokens, image condition tokens, and text tokens. Here, c_I corresponds to I_{src} . This formulation enables bidirectional attention, letting the conditional branch and denoise branch interact on demand.

Position Encoding Cloning. Existing approaches to conditional image editing often struggle with pixel-level misalignment between input I_{src} and edited output (I_{tar}) that undermine visual coherence. To address this fundamental challenge, we propose a novel Position Encoding(PE) Cloning strategy, motivated by the need for implicit spatial correspondence.

The PE Cloning is a simple yet powerful strategy, it simply apply the position encoding calculated on I_{src} on both I_{src} and I_{tar} . The identical positional encoding serves as a strong hint so that the DiT is capable of learning correct corresponding easily. By enforcing identical positional encodings between the latent representations c_I and z , our method establishes a pixel-perfect coordinate mapping that persists throughout the diffusion process. This geometric

consistency ensures that every edit respects the original image’s spatial structure, eliminating ghosting artifacts and misalignments that plague conventional approaches.

Noise-free Conditioning Paradigm. A critical innovation lies in our noise-free conditioning paradigm. We preserve c_I as a reference during the generation of I_{tar} . This design choice achieves two objectives through its operational duality. First, by maintaining c_I in a noise-free state, we ensure the retention of high-frequency textures and fine structural details from the original image, thereby preventing degradation during iterative denoising. This preservation mechanism acts as a safeguard against the blurring artifacts commonly observed in conventional approaches. Second, the MM attention mechanism is flexible enough to choose either to copy from the source or generate new content via instruction, making the model learn to manipulate only designated target regions.

Through the combined action of position encodings cloning and MMA mechanism, our framework achieves unprecedented precision in localized editing while maintaining global consistency, a balance previously unattainable in conditional image generation tasks.

Conditional flow matching loss. The conditional flow matching loss function is following SD3 [7], which is defined as follows:

$$L_{CFM} = E_{t, p_t(z|\epsilon), p(\epsilon)} \left[\|v_{\Theta}(z, t, c_I, c_T) - u_t(z|\epsilon)\|^2 \right] \quad (4)$$

Where $v_{\Theta}(z, t, c_I, c_T)$ represents the velocity field parameterized by the neural network’s weights, t is timestep, c_I and c_T are image condition tokens extracted from source image I_{src} and text tokens. $u_t(z|\epsilon)$ is the conditional vector field generated by the model to map the probabilistic path between the noise and true data distributions, and E denotes the expectation, involving integration or summation over time t , conditional z , and noise ϵ .

3.4. EditLoRA

LoRA [13] enhances model adaptation by freezing the pre-trained model weights W_0 and inserting trainable rank decomposition matrices A and B into each layer of the model. These matrices, $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$, where $r \ll \min(d, k)$, are used to fit the residual weights adaptively. The forward computation integrates these modifications as follows:

$$y' = y + \Delta y = W_0 x + B A x \quad (5)$$

where $y \in \mathbb{R}^d$ is the output and $x \in \mathbb{R}^k$ denotes the input. $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$ with $r \ll \min(d, k)$. Normally, matrix B is initialized with zero.

To learn an individual artist’s editing style and effectively transfer it from a small number of before-and-after image pairs, we introduce EditLoRA. Inspired by recent low-rank adaptation (LoRA) techniques [31, 33], EditLoRA

fine-tunes only a small set of trainable parameters, significantly reducing the risk of overfitting while preserving most of the pretrained model’s expressive power. In our work, the general-purpose OmniEditor is trained on a large-scale paired dataset with a higher-rank lora. The EditLoRAs are lower-rank loras that specifically focus on mimicking the style and strategies of single artists in creating photo doodles. EditLoRA’s training set consists of before-and-after pairwise data and corresponding text instruction, which differ from the conventional text-image pairs required for learning image generation models.

The EditLora steers the behaviour of the OmniEditor to the specified artist’s style. When a new image I_{src} is provided, along with the textual instructions, the model generates I_{tar} that reflects both the previously learned editing capabilities and the distinctive stylistic effects from the artist.

3.5. PhotoDoodle

In collaboration with professional artists and designers, we created the first PhotoDoodle dataset. We introduce the dataset containing six high-quality styles and over 300 photo doodle samples. The six styles include cartoon monster, hand-drawn outline, 3D effects, flowing color blocks, flat illustrator, and cloud sketch. Each sample in our dataset consists of a pre-edit photo (e.g., a real-world scene or portrait) and a post-edit photo doodle showing unique modifications by the artist, such as localized stylization, decorative lines, newly added objects, or modifications to existing elements. For each example, we store the raw input image I_{src} and the doodled version I_{tar} , along with textual instructions.

4. Experiment

4.1. Experiment Setting

Setup. During the OmniEditor pre-training stage, we take the parameters of Flux.1 dev model as the initialization of the DiT architecture, and trained it with the SeedEdit dataset. Images were resized to 768x512. We trained a LoRA rank of 256, a batch size of 128, and a learning rate of 1×10^{-4} , on 8 H100 GPUs for 330,000 steps. After merge the lora into the base DiT model, we acquired the OmniEditor model for further usage. In the EditLoRA training phase, we fine-tuned the merged model on a paired photo doodle dataset (50 pairs) using a single GPU for 10,000 steps, with a LoRA rank of 128, batch size of 2, and a learning rate of 1×10^{-4} .

Baseline Methods. The baselines compared in this paper include InstructP2P[4], MagicBrush[44], and SDEdit[21] based on Flux. For a fair comparison, tests were conducted in both general image editing and customized editing scenarios. For the general image editing tests, OmniEditor was evaluated against the aforementioned baselines. In the cus-

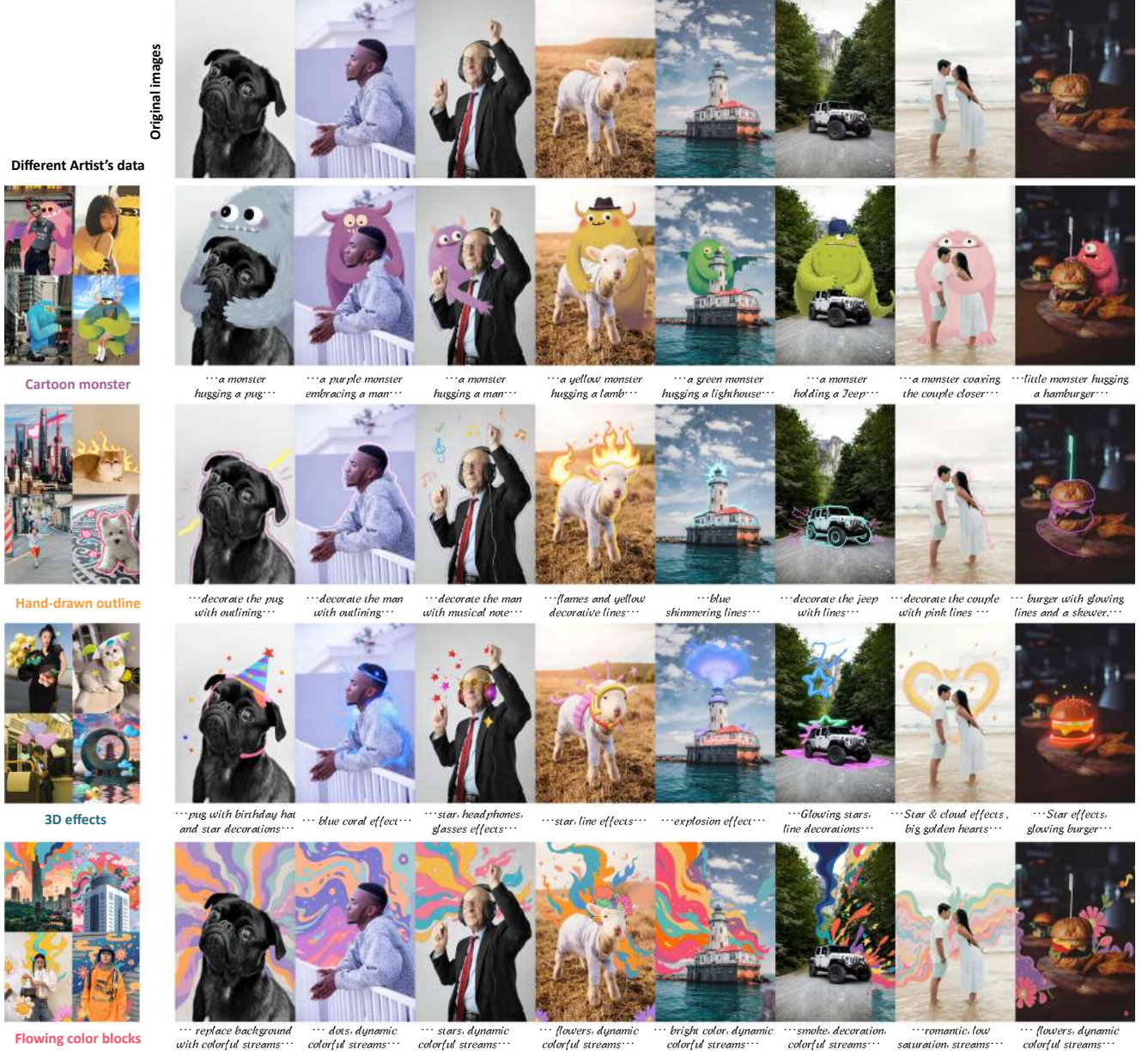


Figure 3. The generated results of PhotoDoodle. PhotoDoodle can mimic the manner and style of artists creating photo doodles, enabling instruction-driven high-quality image editing.

tomized editing scenario, we trained Flux LoRA using doodle results created by professional artists and used it alongside SDEdit as a baseline. For InstructP2P and MagicBrush, the attention layers were also fine-tuned with the same doodle dataset. Finally, all the trained LoRA models were compared with the proposed EditLoRA model.

Benchmarks. As with previous methods, we tested the performance of the proposed OmniEditor on the HQ-Edit benchmark[14]. For the customized generation tasks, this paper introduced a new benchmark collected from the Pex-

els website, consisting of 50 high-quality photographs of portraits, animals, and architecture.

4.2. Generation Results

Fig. 3 and Fig. 7 displays the image editing results of PhotoDoodle, which excels in text following due to training on a large dataset of before-after pairs. The generated doodles blend well with the original images. When trained on a limited dataset of artist-paired data using EditLoRA, PhotoDoodle consistently produces artistic doo-

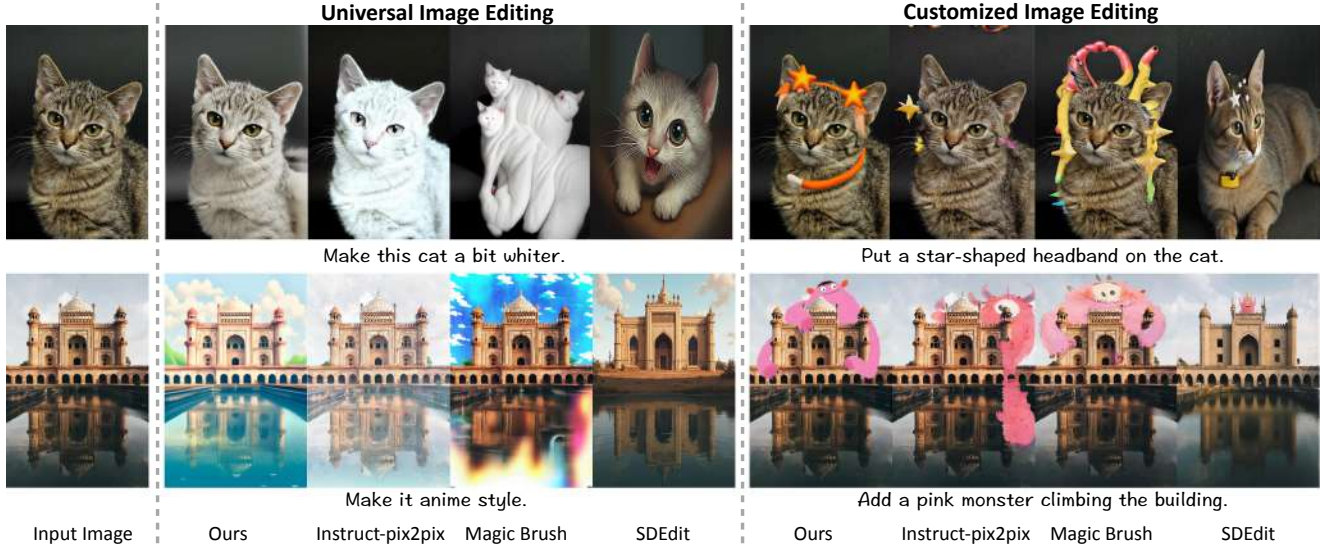


Figure 4. Compared to baselines, PhotoDoodle demonstrates superior instruction following, image consistency, and editing effectiveness.

dles while preserving image consistency and avoiding unwanted changes. Notably, our method maintains stable performance and a high success rate, making it suitable for production use and reducing the need for selective sampling.

Table 1. Comparison Results in General Image Editing Tasks. The best results are denoted as **Bold**.

Methods	$CLIP\ Score\uparrow$	$GPT\ Score\uparrow$	$CLIP_{img}\uparrow$
Instruct-Pix2Pix	0.237	38.201	0.806
Magic Brush	0.234	36.555	0.811
SDEdit(FLUX)	0.230	34.329	0.704
Ours	0.261	51.159	0.871

4.2.1. Qualitative Evaluation

In this section, we present the results of the qualitative analysis. As illustrated in Fig. 4, OmniEdito demonstrates superior consistency and minimizes unintended alterations in general image editing tasks compared to state-of-the-art (SOTA) methods. This performance is attributed to the use of high-quality datasets and a thoughtfully designed model architecture. For custom image editing tasks, our method significantly surpasses baseline methods, as evidenced by the high quality of generated outputs and the strong alignment of the editing style with the original artistic intent, while avoiding undesired modifications.

4.2.2. Quantitative Evaluation

In this section, we present the quantitative analysis results. Following InstructP2P[4], we compute the $CLIP\ Score$ and $CLIP_{img}$ metrics for both tasks. Furthermore, as proposed in HQ-Edit [14], we employ GPT4-o to evaluate the alignment between text instructions and editing outputs. As

Table 2. Comparison Results in Customized Image Editing Tasks. The best results are denoted as **Bold**.

Methods	$CLIP\ Score\uparrow$	$GPT\ Score\uparrow$	$CLIP_{img}\uparrow$
Instruct-Pix2Pix	0.249	36.359	0.832
Magic Brush	0.247	38.478	0.885
SDEdit(FLUX)	0.209	21.793	0.624
Ours	0.279	63.207	0.854

shown in Table 1, our method outperforms all baselines across all metrics in general image editing tasks, achieving the highest $CLIP\ Score$, $GPT\ Score$, and $CLIP_{img}$ Score. In custom image editing tasks, while some models fail to produce meaningful edits, which leads to high $CLIP_{img}$ scores, our method still holds a clear advantage over the baselines. This is evident in the substantial improvements in $GPT\ Score$ and $CLIP\ Score$, both of which evaluate the consistency and quality of the generated content in relation to the artist’s original work.

4.3. Ablation Study

To demonstrate the effectiveness of the key strategies and modules proposed in this paper, we conducted detailed ablation experiments. We evaluated OmniEditor Pre-training, Position Encoding Cloning, and EditLoRA. As shown in Fig. 5: Without OmniEditor Pre-training, directly training EditLoRA leads to reduced harmony between the generated sketches and photos, along with weaker text-following capabilities in the output. Removing Position Encoding Cloning results in decreased consistency in the generated outputs, with unwanted changes occurring in the background. When EditLoRA is not used, and only the pre-

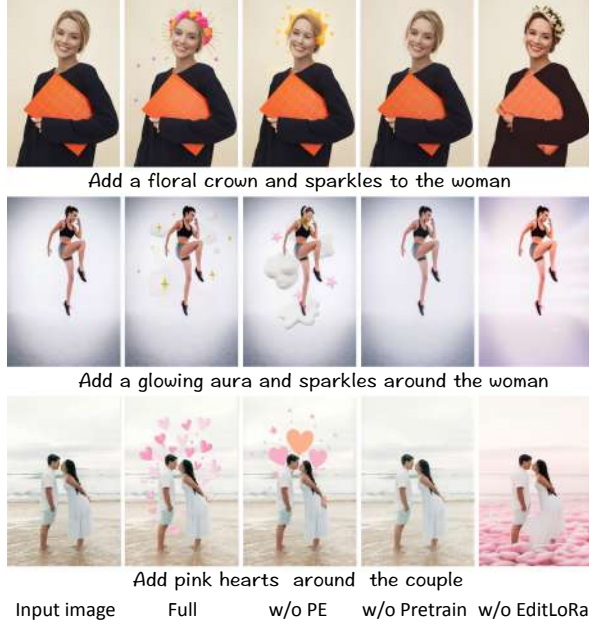


Figure 5. Ablation study results.

trained OmniEditor is employed for generation, the degree of stylization in the results is significantly reduced.

4.4. User Study

To further demonstrate the superiority of our proposed method, we conducted a user study with 30 participants via online questionnaires. We evaluated user preferences in both general and customized image editing scenarios. Participants were presented with PhotoDoodle’s outputs alongside baseline methods, and asked to evaluate which results they preferred based on three criteria: 1) Overall preference, 2) Instruction following, and 3) Consistency between the edited images and the original images. During the study, participants viewed the original unedited images, the edit instructions, and reference images edited by models. They were then asked to decide whether PhotoDoodle (Option A) or a baseline method (Option B) performed better, or if they were about equally effective. The results of this user study are collected in Fig. 6, where we reported the percentage scores of each criterion, highlighting our method’s effectiveness in aligning closely with artistic intentions and maintaining high consistency in edits without introducing unwanted changes.

5. Limitation and Future Work

One limitation of PhotoDoodle is its dependence on the collection of dozens of paired datasets (pre-edit and post-edit images) and the need for thousands of training steps using LoRA. This data collection process can be challenging, as paired images are not always readily accessible. In the fu-

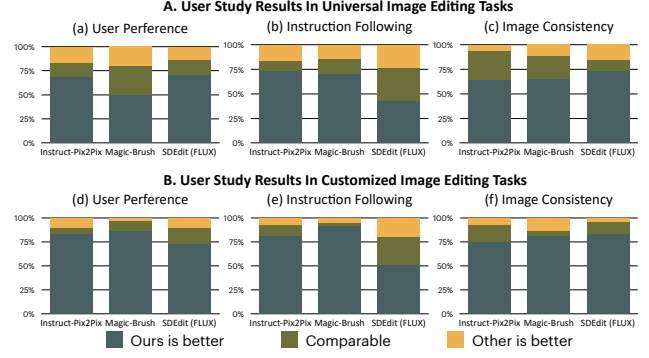


Figure 6. User study results. The scores demonstrate the percentage of users who prefer ours over others under three evaluation metrics. PhotoDoodle outweighs all other baselines in user study.



Figure 7. More photo doodling results: one adds lines to a photo of clouds, imagining them as animals; the other converts the photo into a monochrome version and decorates it with color blocks.

ture, we will attempt to learn doodling strategies from single image pairs using an Encoder structure.

6. Conclusion

In this paper, we present PhotoDoodle, a diffusion-based framework for artistic image editing that learns unique artistic styles from minimal paired examples. By combining large-scale pretraining of the OmniEditor with efficient EditLoRA fine-tuning, PhotoDoodle enables precise decorative generation while maintaining background integrity through positional encoding cloning. Key innovations, including a noise-free conditioning paradigm and parameter-efficient style adaptation requiring only 50 training pairs, significantly reduce computational barriers. We also contribute a new dataset with six artistic styles and 300+ curated samples, establishing a benchmark for reproducible research. Extensive experiments demonstrate superior performance in style replication and background harmony, outperforming existing methods in both generic and customized editing scenarios.

References

- [1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. 2024. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 674–681. 2
- [2] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18208–18218. 3
- [3] Alla Belova. 2021. Google doodles as multimodal storytelling. *Cognition, communication, discourse* 23 (2021), 13–29. 3
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402. 2, 3, 4, 5, 7
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023. PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv:2310.00426 [cs.CV]* 3
- [6] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. 2024. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*. Springer, 338–354. 3
- [7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*. 3, 5
- [8] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. 2023. Ranni: Taming Text-to-Image Diffusion for Accurate Instruction Following. *arXiv preprint arXiv:2311.17002* (2023). 2
- [9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022). 2
- [10] Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. 2024. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007* (2024). 2
- [11] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. 2025. Any2AnyTryon: Leveraging Adaptive Position Em-beddings for Versatile Virtual Clothing Tasks. *arXiv preprint arXiv:2501.15891* (2025). 3
- [12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022). 2, 3
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021). 2, 3, 5
- [14] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990* (2024). 6, 7
- [15] Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. 2024. Videobooth: Diffusion-based video generation with image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6689–6700. 2
- [16] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017. 3
- [17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941. 2
- [18] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xiuhui Liu, Jiaming Liu, Lin Li, Xu Tang, Yao Hu, Jianzhuang Liu, et al. 2024. Zone: Zero-shot instruction-guided local editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6254–6263. 3
- [19] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521. 2
- [20] Yang Liu, Cheng Yu, Lei Shang, Ziheng Wu, Xingjun Wang, Yuze Zhao, Lin Zhu, Chen Cheng, Weitao Chen, Chao Xu, et al. 2023. Facechain: A playground for identity-preserving portrait generation. *arXiv preprint arXiv:2308.14256* (2023). 2
- [21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon.

2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021). 5
- [22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021). 2
- [23] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205. 3
- [24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023). 2
- [25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv 2022. arXiv preprint arXiv:2204.06125* (2022). 2
- [26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695. 2
- [27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510. 2, 3
- [28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2024. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6527–6536. 2
- [29] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems* 35 (2022), 36479–36494. 2
- [30] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. Style-drop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983* (2023). 2
- [31] Yiren Song, Danze Chen, and Mike Zheng Shou. 2025. LayerTracer: Cognitive-Aligned Layered SVG Synthesis via Diffusion Transformer. *arXiv preprint arXiv:2502.01105* (2025). 5
- [32] Yiren Song, Shijie Huang, Chen Yao, Xiaojun Ye, Hai Ci, Jiaming Liu, Yuxuan Zhang, and Mike Zheng Shou. 2024. ProcessPainter: Learn Painting Process from Sequence Data. *arXiv preprint arXiv:2406.06062* (2024). 2
- [33] Yiren Song, Cheng Liu, and Mike Zheng Shou. 2025. MakeAnything: Harnessing Diffusion Transformers for Multi-Domain Procedural Sequence Generation. *arXiv preprint arXiv:2502.01572* (2025). 5
- [34] Yiren Song, Shengtao Lou, Xiaokang Liu, Hai Ci, Pei Yang, Jiaming Liu, and Mike Zheng Shou. 2024. Anti-Reference: Universal and Immediate Defense Against Reference-Based Generation. *arXiv preprint arXiv:2412.05980* (2024). 3
- [35] Yiren Song, Pei Yang, Hai Ci, and Mike Zheng Shou. 2024. IDProtector: An Adversarial Noise Encoder to Protect Against ID-Preserving Image Generation. *arXiv preprint arXiv:2412.11638* (2024). 3
- [36] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neuro-computing* 568 (2024), 127063. 3
- [37] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. 2023. Antidreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2116–2127. 3
- [38] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519* (2024). 3
- [39] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. 2024. StableGarment: Garment-Centric Generation via Stable Diffusion. *arXiv preprint arXiv:2403.10783* (2024). 3
- [40] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18359–18369. 3
- [41] Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. 2024. EditWorld: Simulating World Dynamics for Instruction-Following Image Editing. *arXiv preprint arXiv:2405.14785* (2024). 3

- [42] Emilie Yu. 2023. *Designing tools for 3D content authoring based on 3D sketching*. Ph.D. Dissertation. Université Côte d’Azur. 3
- [43] Emilie Yu, Kevin Blackburn-Matzen, Cuong Nguyen, Oliver Wang, Rubaiat Habib Kazi, and Adrien Bousseau. 2023. Videodoodles: Hand-drawn animations on videos with scene-aware canvases. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. 3
- [44] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* 36 (2024). 2, 5
- [45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847. 3
- [46] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. 2024. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9026–9036. 3
- [47] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10146–10156. 2
- [48] Yuxuan Zhang, Yiren Song, Jinpeng Yu, Han Pan, and Zhongliang Jing. 2024. Fast Personalized Text to Image Synthesis with Attention Injection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6195–6199. <https://doi.org/10.1109/ICASSP48485.2024.10447042> 3
- [49] Yuxuan Zhang, Lifu Wei, Qing Zhang, Yiren Song, Jiaming Liu, Huaxia Li, Xu Tang, Yao Hu, and Haibo Zhao. 2024. Stable-makeup: When real-world makeup transfer meets diffusion model. *arXiv preprint arXiv:2403.07764* (2024). 3
- [50] Yuxuan Zhang, Qing Zhang, Yiren Song, and Jiaming Liu. 2024. Stable-hair: Real-world hair transfer via diffusion model. *arXiv preprint arXiv:2407.14078* (2024). 3
- [51] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Li Xiu. 2024. MultiBooth: Towards Generating All Your Concepts in an Image from Text. *arXiv preprint arXiv:2404.14239* (2024). 2
- [52] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. 2021. Stylized neural painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15689–15698. 2