Magic Insert: Style-Aware Drag-and-Drop

Nataniel Ruiz[1] Yuanzhen Li[1] Neal Wadhwa[2] Yael Pritch[2]
Michael Rubinstein[1] David E. Jacobs[2] Shlomi Fruchter[1]
[1]Google DeepMind [2]Google

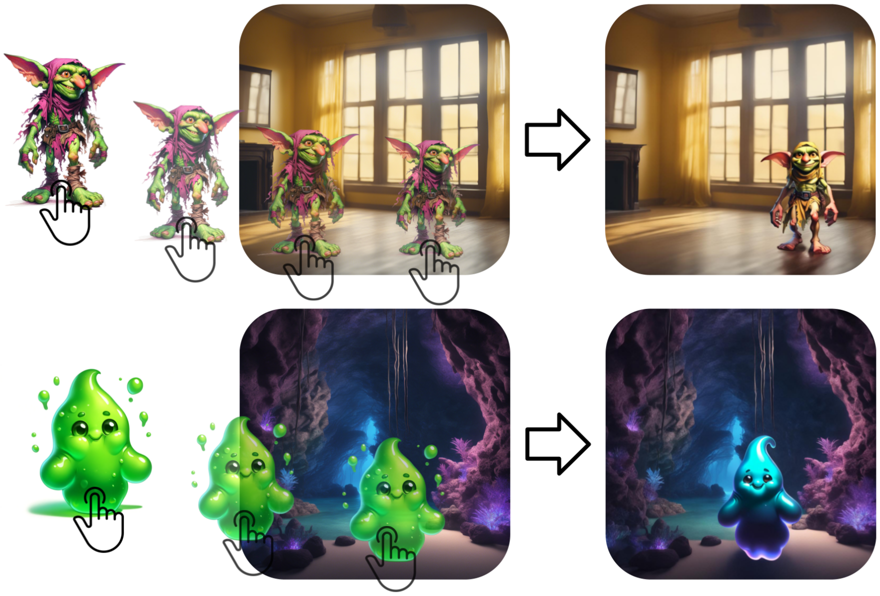

Figure 1. Using *Magic Insert* is a state-of-the-art method to drag-and-drop a subject from an image with an arbitrary style onto another target image with a vastly different style and achieve a style-aware and impressively realistic insertion of the subject into the target image.

Abstract

*We present **Magic Insert**, a method to drag-and-drop subjects from a user-provided image into a target image of a different style in a plausible manner while matching the style of the target image. This work formalizes our version of the problem of style-aware drag-and-drop and proposes to tackle it by decomposing it into two sub-problems: style-aware personalization and realistic object insertion in stylized images. For style-aware personalization, we cast our method as a weight-and-text-embedding finetuning method with inference-time module-targeted style injection. For subject insertion, we propose Bootstrapped Domain Adaption (BDA) to adapt a domain-specific photorealistic object insertion model to the domain of diverse artistic styles. Overall, the method significantly outperforms traditional and state-of-the-art approaches that struggle with quality, subject fidelity and harmonious stylization. Finally, we present a new dataset, SubjectPlop, to facilitate evaluation and future progress in this area.*

1. Introduction

Large text-to-image models have recently made significant progress in generating high-quality images. However, to make these models truly useful, controllability is essential. Users have diverse needs and want to interact with these models in different ways depending on their specific use case. Influential work has been done to enable controllability in these networks, robustly addressing foundational applications and controls such as subject personalization, style learning, layout controls, and semantic controls. Despite this progress, the full potential of these powerful large models has not been fully realized. Some applications that seemed clearly out of reach just a couple of years ago are now possible with careful approaches.

We explore one such application: *style-aware drag-and-drop*, where aesthetics and quality are of utmost importance and optimality in these directions has not yet been achieved. We formalize this problem and introduce *Magic Insert*, our method to tackle it. One might initially consider addressing style-aware drag-and-drop by trying to inpaint using a stylized subject, for example by combining Dreambooth [35], StyleDrop [43], and inpainting. This direction suffers from background erasure, incomplete subject generation and subpar results. Another approach would be to use composition-based methods [22, 54], which suffers from unrealistic insertion with insufficient shadows, reflections and missing contextual interactions, as well as from problems with style inheritance and identity preservation. Other methods such as realistic insertion methods [7, 55] cannot stylize or harmonize the inserted subject.

In developing Magic Insert, we decompose our problem into two interesting sub-problems: *style-aware personalization* and *realistic object insertion in stylized images*. For style-aware personalization, there have been attempts on adjacent problems, such as learning a style and then representing a specific subject in that style [13, 43], or combining pre-trained custom style and subject models [8, 41]. Recent style work suggests that fast style learning is possible, but fast learning of a subject, including all the intricacies of identity, is a much harder problem that has arguably not been solved yet [10, 36, 52, 59]. We propose leveraging learnings from both domains and settle on a solution that uses adapter injection of style paired with subject-learning in the embedding and weight space of a diffusion model.

One key idea we propose is to not attempt inpainting directly into an image after achieving style-aware personalization. Instead, for best results, we first generate a high-quality subject and then insert that subject into the target image. To achieve our results, we introduce an innovation called *Bootstrapped Domain Adaptation* (BDA), that allows progressive retargeting of a model's initial distribution to a target distribution. We apply this idea to adapt a subject insertion network that has been trained on real images to perform well on the stylized image domain, enabling the insertion of our generated stylized subject into the background image.

Our method allows the generated output to exhibit strong adherence to the target style while preserving the essence and identity of the subject, and for realistic insertion of the stylized subject into the generated image. The method also provides flexibility in terms of the degree of stylization desired and how closely to adhere to the original subject's specific details and pose (or to allow more novelty).

In summary, we propose the following contributions:

- We propose and formalize the problem of **style-aware drag-and-drop**, where a subject (a character or object) is dragged from one image into another. Specifically, in our problem formulation the subject reference image and the target image may be in vastly different styles, and the plausibility and realism of the subject insertion is of utmost importance - including shadows and reflections.
- In order to encourage exploration into this problem, we present **SubjectPlop**, a novel dataset of subjects and backgrounds that span widely different styles and overall semantics. We will release this dataset for public use, as well as our evaluation suite.
- We propose **Magic Insert**, a method to tackle the style-aware drag-and-drop problem. Our method is composed of a style-aware personalization component and a style-consistent drag-and-drop component.
- For **style-aware personalization**, we demonstrate strong and consistent results using subject-learning in the embedding and weight space of a pre-trained diffusion models, along with inference-time module-targeted style injection.
- For **drag-and-drop**, we propose *Bootstrapped Domain Adaptation* (BDA), a method that allows for progressive retargeting of a model's initial distribution unto a target distribution.

2. Related Work

Text-to-Image Models Recent text-to-image models such as Imagen [39], DALL-E 2 [31], Stable Diffusion (SD) [33], Muse [5] and Parti [60] have demonstrated remarkable capabilities in generating high-quality images from text descriptions. They leverage advancements in diffusion models [14, 42, 45] and generative transformers. Our work builds on top of SDXL [29] and the LDM architecture [33].

Image Inpainting The task of filling masked pixels of a target image has been explored using a wide range of approaches: Generative adversarial networks [11] e.g. [16, 20, 27, 28, 32, 61] and end-to-end learning methods [17, 19, 48, 56]. More recently, diffusion models enabled significant progress [3, 23, 25, 38, 53]. Such inpainting methods are a precursor to many object insertion approaches.

Generative Object Insertion The problem of inserting an object into an existing scene has been originally explored using Generative Adversarial Networks (GANs) [11]. [18] breaks down the task into two generative modules, one determines where the inserted object mask should be and the other determines what the mask shape and pose. ShadowGAN [63] addresses the need to add a shadow cast by the inserted object, leveraging 3D rendering for training data. More recent works use diffusion models. Paint-By-Example [58] allows inpainting a masked area of the target image with reference to the object source image, but it only preserves semantic information and has low fidelity to the original object's identity. Recent work also explores swapping objects in a scene while harmonizing, but focuses on swapping areas of the image which were previously populated [12]. There also exists an array of work that focuses on inserting subjects or concepts in a scene either by inpainting [21, 37] or by other means [40, 46] - these do not handle large style adaptation and inpainting methods usually suffer from problems with insertion such as background removal, incomplete insertion and low quality results. Recent advances in image harmonization [4, 26], cross-domain composition [22, 47, 54] show promise for this task, but suffer from limits in insertion capabilities including lack of shadows, reflections and contextual interactions, as well as some limits in style-adherence which are easier to handle by our style-aware personalization approach. AnyDoor [7] photo-realistically inserts objects using a purposefully trained feature extractor and UNet-based diffusion model. ObjectDrop [55] trains a diffusion model for object removal/insertion using a counterfactual dataset captured in the real world. The trained model can insert segmented objects into real images with contextual cues such as shadows and reflections. We build upon this novel and incredibly useful paradigm by tackling the challenging domain of stylized images instead.

Personalization, Style Learning and Controllability Text-to-image models enable users to provide text prompts and sometimes input images as conditioning input, but do not allow for fine-grained control over subject, style, layout, etc. Textual Inversion [9] and DreamBooth [35] are pioneering works that demonstrated personalization of such models to generate images of specific subjects, given few casual images as input. Textual Inversion [9] and follow-up techniques such as P+ [50] optimize text embeddings, while DreamBooth optimizes the model weights. This type of work has also been extended to 3D models [30], scene completion [49] and others. There also exists work on fast subject-driven generation [2, 6, 10, 36, 52]. Other work allows for conditioning on new modalities such as ControlNet [62] and on image features (IP-Adapter [59]). There is a body of work that dives more deeply into style learning and generating consistent style as well with StyleDrop [43] as a pioneer, with newer work that achieves fast stylization [13, 34, 41, 51], or combines subject models with style models like ZipLoRA [41] and others [8]. Our work leverages ideas from Textual Inversion, DreamBooth and IP-Adapter to unlock style-aware personalization prior and combine it with subject insertion.

3. Method

3.1. Style-Aware Drag-and-Drop Problem Formulation

We formalize the style-aware drag-and-drop problem as follows. Let $\mathcal{I}_s$ and $\mathcal{I}_t$ denote the space of subject and target images, respectively. The space of subject images consists of images of solely the subject in front of plain backgrounds. Given a subject image $x_s \in \mathcal{I}_s$ and a target image $x_t \in \mathcal{I}_t$, our goal is to generate a new image $\hat{x}_t \in \mathcal{I}_t$ such that:

1. The subject from x_s is inserted into $\hat{x}_t$ in a semantically consistent and realistic manner, accounting for factors such as occlusion, shadows, and reflections.
2. The inserted subject in $\hat{x}_t$ adopts the style characteristics of the target image x_t while preserving its essential identity and attributes from x_s.

Formally, we aim to learn a function $h : \mathcal{I}_s \times \mathcal{I}_t \to \mathcal{I}_t$ that satisfies:

$$h(x_s, x_t) = \hat{x}_t \quad \text{s.t.} \quad \hat{x}_t \sim p(\hat{x}_t | x_t, x_s) \tag{1}$$

where $p(\hat{x}_t | x_t, x_s)$ represents the conditional distribution of the output image given the subject and target images. This distribution encapsulates the desired properties of semantic consistency, realistic insertion, and style adaptation. To learn the function h, we decompose the problem into two sub-tasks: style-aware personalization and realistic object insertion in stylized images. Style-aware personalization

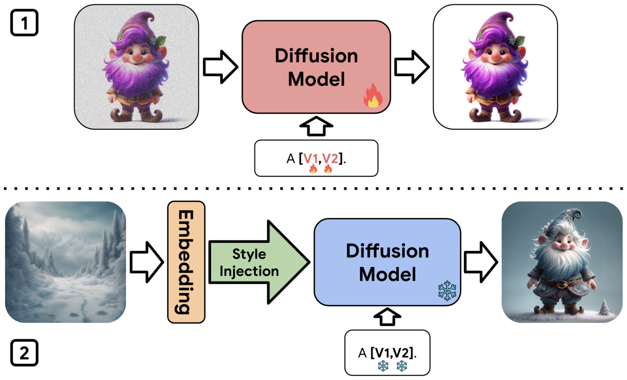

Figure 2. **Style-Aware Personalization:** To generate a subject that fully respects the style of the target image while also conserving the subject's essence and identity, we **(1)** personalize a diffusion model in both weight and embedding space, by training LoRA deltas on top of the pre-trained diffusion model and simultaneously training the embedding of two text tokens using the diffusion denoising loss **(2)** use this personalized diffusion model to generate the style-aware subject by embedding the style of the target image and conducting adapter style-injection into select upsampling layers of the model during denoising.

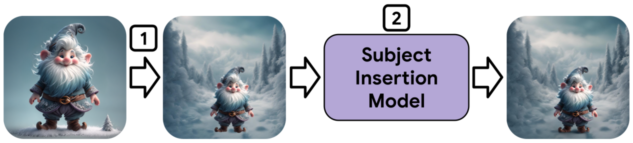

Figure 3. **Subject Insertion:** To insert the personalized generation, we (1) copy-paste a segmented version of the subject onto the target image (2) run our subject insertion model on the deshadowed image - this creates context cues and realistically embeds the subject into the image including shadows and reflections.

focuses on generating a subject that adheres to the target image's style while maintaining its identity. Realistic object insertion aims to seamlessly integrate the stylized subject into the target image, accounting for the scene's geometry and lighting conditions. By addressing these sub-tasks, we can effectively solve the style-aware drag-and-drop problem and generate visually coherent and compelling results. In the following sections, we present our dataset and the components of our proposed method.

3.2. SubjectPlop Dataset

To facilitate the evaluation of the style-aware drag-and-drop problem, we introduce the SubjectPlop dataset and make it publicly available. As this is a novel problem, a dedicated dataset is crucial for enabling the research community to make progress in this area.

SubjectPlop consists of a diverse collection of subjects generated using DALL-E3 [31] and backgrounds generated using the open-source SDXL model [29]. The dataset includes various subject types, such as animals and fantasy characters, and both subjects and backgrounds exhibit a wide range of styles, including 3D, cartoon, anime, realistic, and photographic. The diversity in color hues and lighting conditions ensures comprehensive coverage of different scenarios for evaluation. No real people are represented in the dataset.

The dataset comprises 20 distinct backgrounds and 35 unique subjects, allowing for a total of 700 possible subject-background pairs. The entire dataset is meant for evaluation of the task. This rich set of test cases enables the assessment of performance and generalization capabilities of style-aware drag-and-drop techniques. By introducing SubjectPlop, we aim to provide a standardized benchmark for evaluating and comparing different approaches to the style-aware drag-and-drop problem. We believe this dataset will serve as a valuable resource for researchers and practitioners working in image manipulation and generation, fostering further advancements in this area.

3.3. Style-Aware Personalization

Our style-aware personalization approach is illustrated in Figure 2. Let f_θ denote a pre-trained diffusion model with parameters θ. Given a subject image $x_s \in \mathcal{I}_s$, our method personalizes f_θ on x_s in both the weight and embedding space, similar to DreamBooth [35] and Textual Inversion [9].

In the first step, we train LoRA [15] (Low-Rank Adaptation) deltas Δ_θ to produce an efficiently fine-tuned adapted model $f_{\theta'}$ where $\theta' = \theta + \Delta_\theta$, while preserving the model's original capabilities. Simultaneously, we learn embeddings $e_1, e_2 \in \mathbb{R}^d$ for two personalized text tokens, where d is the embedding dimensionality. We use two learned embeddings since we found better empirical performance for both

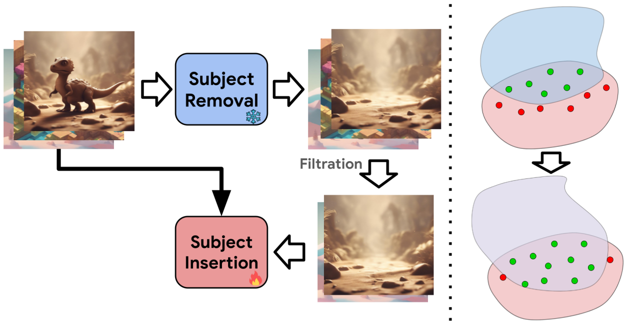

Figure 4. **Bootstrapped Domain Adaptation:** A diffusion model trained for subject insertion/removal on real world data has some limited generalization to new domains. We introduce *bootstrapped domain adaptation*, where a model's effective domain can be adapted by using a subset of its own outputs. **(left)** We use a subject removal model to first remove subjects and shadows from a dataset from our target domain. Then, we filter flawed outputs and use the new image set to retrain the insertion model. **(right)** We observe that, the initial distribution *(blue)* changes after training *(purple)* and initially incorrect outputs *(red samples)* become correctly treated *(green)*.

subject preservation and editability in this configuration, as opposed to one, or more than two embeddings. The LoRA deltas and and embeddings are jointly trained using the diffusion denoising loss:

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\theta'}(x_s^t, t, [e_1; e_2])\|_2^2\right] \tag{2}$$

where $t \sim \mathcal{U}(0,1)$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $x_s^t = \sqrt{\bar{\alpha}_t}x_s + \sqrt{1-\bar{\alpha}_t}\epsilon$, and $\epsilon_{\theta'}$ is the noise prediction of the adapted model $f_{\theta'}$. The joint optimization of Δ_θ, e_1, and e_2 is performed using the loss $\mathcal{L}_{\text{joint}}$. These personalized text tokens $[e_1; e_2]$ serve as a compact representation of the subject's identity. We find that performing embedding and weight-space learning simultaneously, with two text tokens, captures the subject's identity more strongly while allowing sufficient editability to introduce the target style.

In the second step, we leverage the personalized diffusion model $f_{\theta'}$ to generate the style-aware subject $\hat{x}_s$. To infuse the target image x_t's style into $\hat{x}_s$, we employ style injection. Specifically, we generate a style embedding $e_t = \text{CLIP}(x_t)$ of x_t using a frozen CLIP encoder CLIP. We then use a frozen IP-Adapter model ν to inject e_t into a subset of the UNet blocks of $f_{\theta'}$ during inference:

$$\hat{x}_s = f_{\theta'}([e_1; e_2], \nu(e_t)) \tag{3}$$

with injection into the upsample block that is adjacent to the midblock since it controls style more strongly than other blocks [51]. To the best of our knowledge, our central idea of combining adapter injection and personalized models works surprisingly well and remains unexplored in the published literature. Overall, this ensures that $\hat{x}_s$ maintains

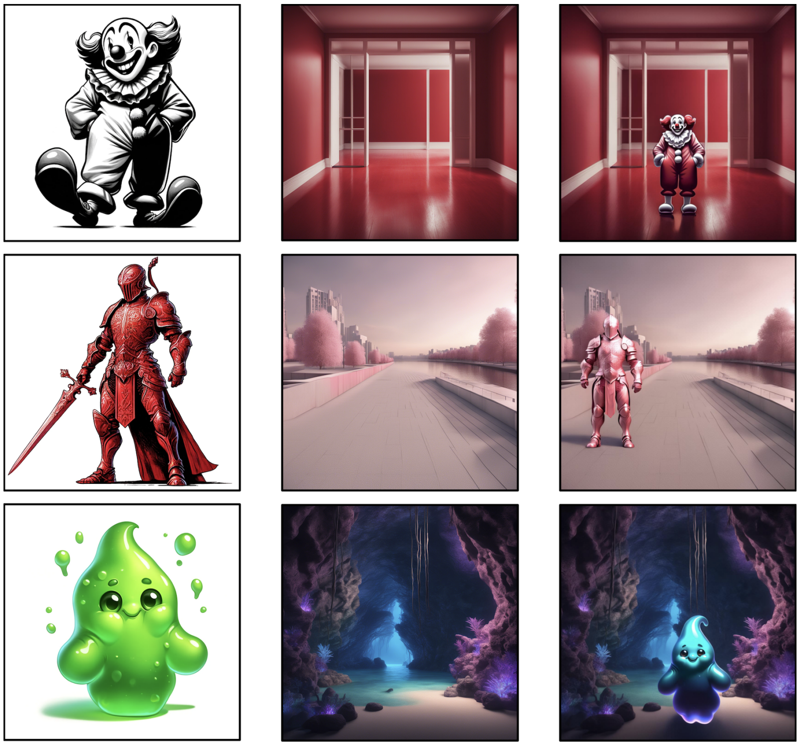

Figure 5. **Result Gallery:** Examples of our Magic Insert method for different subjects and backgrounds with vastly different styles.

the subject's identity while adopting x_t's style characteristics.

By combining style-aware personalization with style injection, our method generates subjects that harmoniously blend into the target image while retaining their essential identity, effectively tackling the first challenge of style-aware drag-and-drop and enabling the creation of visually coherent and style-consistent results.

3.4. Bootstrapped Domain Adaptation for Subject Insertion

In this section, we address the problem of subject insertion and propose a novel solution using bootstrapped domain adaptation. We formalize the concept of bootstrapped domain adaptation and describe the dataset used for this purpose. Subject insertion is a crucial component of the style-aware drag-and-drop problem, as it involves seamlessly integrating a stylized subject into a target background image. While diffusion-based inpainting approaches [24, 33, 39] can be used for this, they still face challenges such as generating content in smooth regions, producing incomplete figures, erasing objects behind inserted subjects, and having problems with boundary harmonization. We take a simpler and stronger approach, which is to insert the subject by copying and pasting it into the target image, and then subsequently generating contextual cues such as shadows and reflections [55] in a second step. Unfortunately, existing subject insertion models are trained on data captured in the real world, severely limiting their ability to generalize to images with diverse artistic styles.

Let $\mathcal{D}_r$ denote the distribution of real-world images and $\mathcal{D}_s$ denote the distribution of stylized images. Existing subject insertion models are trained on samples from $\mathcal{D}_r$, but our goal is to adapt them to perform well on samples from $\mathcal{D}_s$. To overcome this limitation, we introduce bootstrapped domain adaptation, a technique that enables a model to adapt its effective domain by leveraging a subset of its own outputs. As illustrated in Figure 4 (left), we employ a subject removal/insertion model g_θ trained on real-data ([55] in our case) to first remove subjects and shadows from a dataset $\mathcal{S} \sim \mathcal{D}_s$ belonging to our target domain. Subsequently, we filter out flawed outputs and obtain a filtered set of images $\mathcal{S}' \subseteq \mathcal{S}$, which we use to retrain the subject removal/insertion model. Filtering can be done using human feedback or automatically given a quality evaluation module.

The bootstrapped domain adaptation process can be formalized as follows:

$$\omega = \arg\min_{\omega} \mathbb{E}_{(x,y)\sim\mathcal{S}'} \mathcal{L}(g_\omega(x), y) \tag{4}$$

where ω denotes the adapted model parameters, $\mathcal{L}$ is the diffusion denoising loss, and (x, y) are pairs of input images and corresponding subject removal/insertion ground truths from the filtered set $\mathcal{S}_f$. The concept of bootstrapped do-

Table 1. **Subject Fidelity Comparisons.** We compare our method for subject fidelity (DINO, CLIP-I, CLIP-T Simple, CLIP-T Detailed) across different methods. Our method variants show high subject fidelity. (P) denotes a prompt-based approach and (CN) denotes the use of edge-based ControlNet.

Method	DINO ↑	CLIP-I ↑	CLIP-T Simple ↑	CLIP-T Detailed ↑	Overall Mean ↑
StyleAlign (P)	0.223	0.743	0.266	0.299	0.383
StyleAlign (CN)	0.414	0.808	0.289	0.294	0.451
InstantStyle P	0.231	0.778	0.283	0.300	0.398
InstantStyle (CN)	0.446	0.806	0.281	0.283	0.454
Ours	0.295	0.829	0.276	0.293	0.423
Ours (CN)	0.514	0.869	0.289	0.308	**0.495**

main adaptation is based on the surprising observation that a diffusion model trained for subject insertion/removal on real-world data can generalize to a wider stylistic domain to a limited extent. By retraining the model on its own filtered outputs, we can effectively adapt its domain to better handle stylized images.

Figure 4 (right) demonstrates the effect of bootstrapped domain adaptation on the model's distribution. The initial distribution, represented as $p_\omega(x)$, evolves after training, becoming $p_{\omega^*}(x)$. Images that were initially treated incorrectly, shown as samples from $\mathcal{D}_s \setminus \mathcal{S}'$, are subsequently handled correctly, as indicated by their inclusion in $\mathcal{S}'$. During the bootstrapped domain adaptation process, we train the model only on the initially correct samples from $\mathcal{S}'$ to further refine its performance on the target domain. Several steps of bootstrapped domain adaptation can be performed, further enhancing the model's performance. In our work we find that one step suffices, with a small set of samples (around 50). Figure 7 shows results with and without bootstrap domain adaptation.

To facilitate the bootstrapped domain adaptation process, we curate a dataset $\mathcal{S}$ specifically tailored to this task. The dataset comprises a diverse range of stylized images, selected to represent the target domain $\mathcal{D}_s$. In our case, this dataset is constructed by sampling from different text-to-image generative models with diverse prompts that elicit prominent subjects with shadows and reflections in a variety of global styles. By finetuning the subject removal/insertion model on this dataset using the bootstrapped domain adaptation technique, we enable it to effectively handle subject insertion in the context of style-aware drag-and-drop.

4. Experiments

In this section, we present experiments and applications. Our method enables insertion of arbitrary subjects into images with diverse styles and extensive text-guided semantic modifications. The subject retains its identity while inheriting the target image's style, and we can modify key characteristics like pose, accessories, appearance, shapes, or even create species hybrids (see appendix). These changes integrate with components like LLMs for automatic affordances and environment interactions (Figure 6).

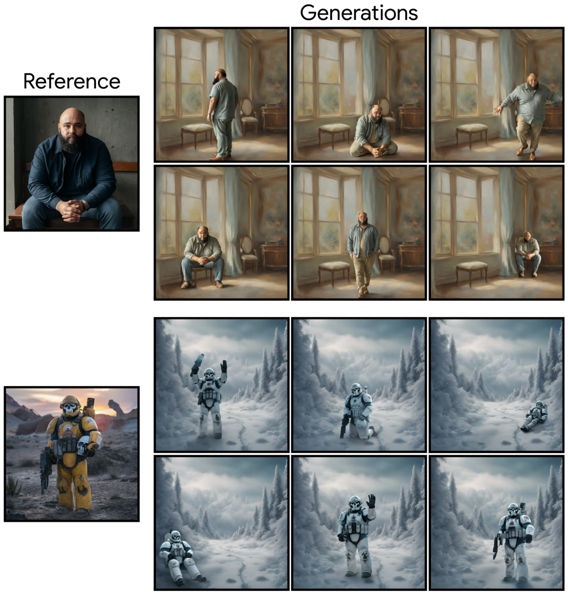

Figure 6. **LLM-Guided Affordances:** Examples of an LLM-guided pose modification for Magic Insert, with the LLM suggesting plausible poses and environment interactions for areas of the image and Magic Insert generating and inserting the stylized subject with the corresponding pose into the image.

Table 2. **Style Fidelity Comparisons.** We compare our method for style fidelity (CLIP-I, CSD, CLIP-T). Our method variants show strong style-following.

Method	CLIP-I ↑	CSD ↑	CLIP-T ↑	Overall Mean ↑
StyleAlign Prompt	0.570	0.150	0.248	0.323
StyleAlign ControlNet	0.575	0.188	0.274	0.345
InstantStyle Prompt	0.583	0.312	0.276	0.390
InstantStyle ControlNet	0.588	0.334	0.279	**0.400**
Ours	0.560	0.243	0.268	0.357
Ours ControlNet	0.575	0.294	0.274	0.381

Table 3. **ImageReward Metric Comparisons.** We compare different methods using the ImageReward metric, which correlates with human preference for aesthetics. Higher scores indicate better performance. Our variants outperform all benchmarks

Method	ImageReward Score ↑
StyleAlign Prompt	-1.1942
StyleAlign ControlNet	-0.5180
InstantStyle Prompt	-0.4638
InstantStyle ControlNet	-0.2759
Ours	-0.2108
Ours ControlNet	**-0.1470**

4.1. Style-Aware Drag-and-Drop Results

Magic Insert Results Figure 5 presents qualitative results demonstrating our method's effectiveness and versatility across diverse subjects and backgrounds—from photorealistic scenes to cartoons and paintings. We use SDXL [29]

Table 4. **User Study.** This study evaluates our method against two different baselines based on subject identity, style fidelity, and realistic insertion. Participants ranked each method by preference.

Method	User Preference ↑
Ours over StyleAlign ControlNet	**85%**
Ours over InstantStyle ControlNet	**80%**
Ours over AnyDoor	**90%**
Ours over TF-ICON	**95%**

for style-aware personalization and our trained latent diffusion model for subject insertion.

Our method successfully extracts subjects and blends them into target backgrounds, adapting appearance to match style. Inserted subjects adopt the target's colors, textures, and stylistic elements, with coherent shadows and reflections enhancing plausibility.

LLM-Guided Affordances Our style-aware personalization supports significant pose changes via diffusion model priors. Using ChatGPT 4o, we generate LLM-guided affordances by providing instruction prompts, the full background, and the insertion region. These suggestions enable character generation with appropriate poses and environment interactions. Figure 6 demonstrates this first attempt at automatic realistic subject insertion with scene interactions.

Bootstrapped Domain Adaptation (BDA) Figure 7 shows an ablation for BDA in subject insertion. Without BDA, our method exhibits failures including unrealistic shadow placement, missing shadows, hallucinations, artifacts, and missing pixels (marked in red boxes). BDA resolves these issues through proper domain understanding.

4.2. Comparisons

Here we introduce baselines, as well as quantitative and qualitative comparisons, as well as a user study. Specifically, our proposed baselines utilize the StyleAlign [13] and InstantStyle [51] stylization methods, which can generate images in reference styles given either inversion or embedding of the reference image. We combine these methods with either sufficiently detailed prompting guided by a VLM [1] or edge-conditioned ControlNet.

Baseline Comparisons We run studies in order to compare the performance of subject stylization for different baselines and our style-aware personalization method. We study the performance of these methods on subject fidelity, style fidelity, and human preference.

For subject fidelity (Table 1), our proposed variants achieve high scores across various subject fidelity metrics (DINO, CLIP-I, CLIP-T Simple, CLIP-T Detailed). DINO and CLIP-I metrics are identical to those presented in DreamBooth [35] and CLIP-T Simple / Detailed denotes the CLIP similarity between the output image CLIP embedding and the CLIP embedding of simple and detailed text prompts describing the subject, which are in turn generated by ChatGPT 4.

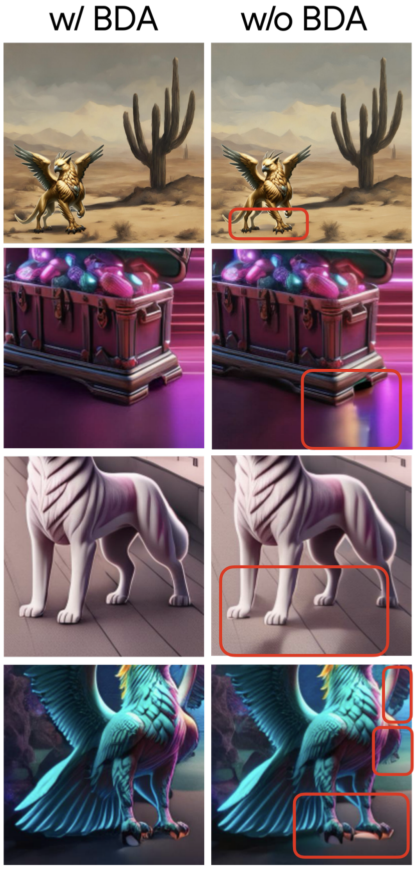

Figure 7. **Bootstrapped Domain Adaptation (BDA):** Inserting a subject with the pre-trained subject insertion module without BDA generates subpar results, with missing shadows and reflections, or added distortions and artifacts.

Regarding style fidelity (Table 2), our proposed variants demonstrate strong style-following performance using CLIP-I [35, 43], CSD [44], CLIP-T [35, 43] metrics. For aesthetic human-correlated preference, we compute ImageReward [57] scores in Table 3. We observe that our variants strongly outperform the other methods. For a deeper discussion about style metrics and qualitative examples of the superiority of our method on harmonized insertion with inherited style please check the supplementary material. We have included qualitative examples and a more fine-grained discussion on the goals and measurements.

User Study Following previous work [35, 36, 43, 49] we perform a robust user study to compare our full method with the state-of-the-art related work: StyleAlign [13], InstantStyle [51], ObjectDrop [55], TF-ICON [22] and Any-

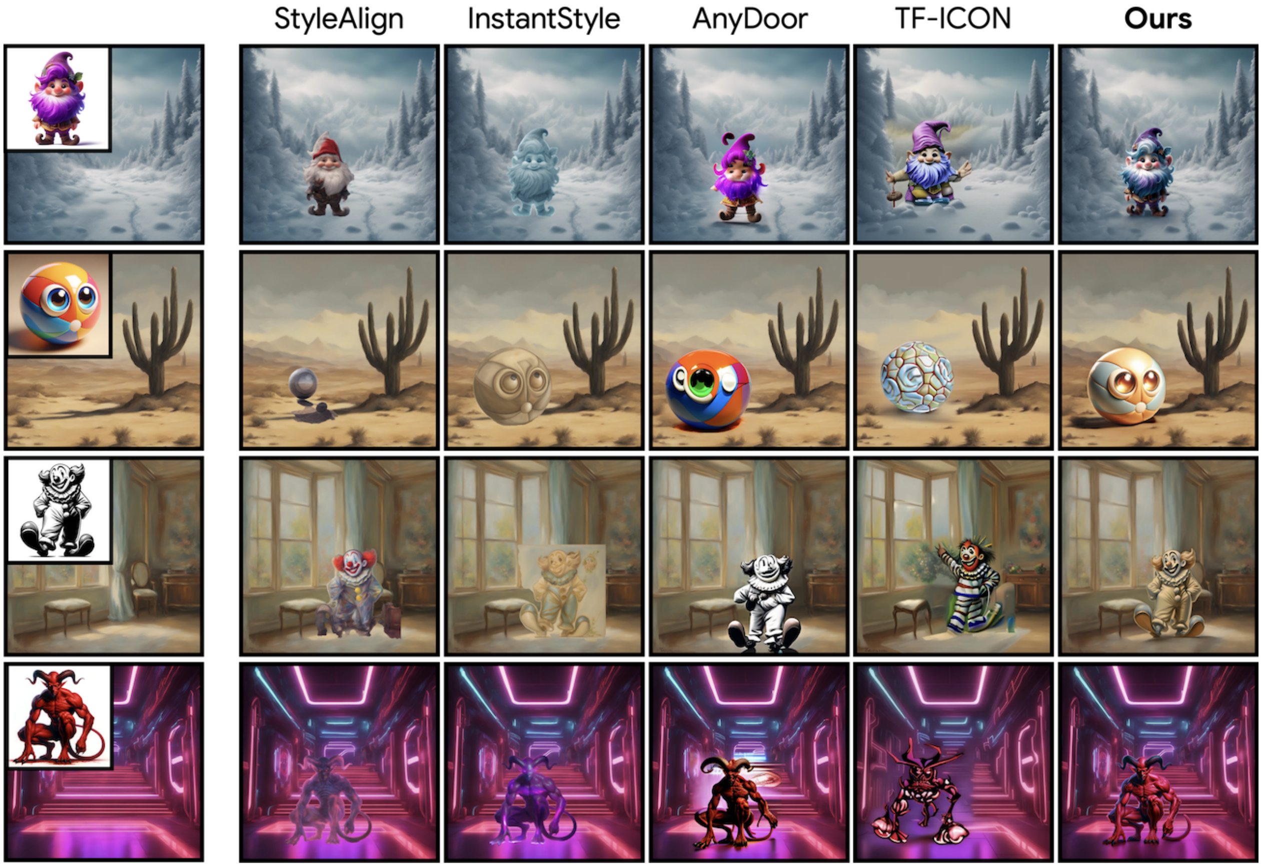

Figure 8. **Comparison w/ Related Work:** Harmonization-based methods like TF-ICON [22] tend to fail on subject fidelity, style adherence, background erasure and lack of shadows and reflections. Non-stylized insertion work like AnyDoor [7] **cannot** stylize the inserted subject and can fail with subject distortions. Our baselines of StyleAlign and InstantStyle struggle with capturing subject fidelity and correct harmonization. Our method performs well across the board in these elements.

Door [7]. We recruit a total of 60 users (4 sets of 15 users) to answer 40 evaluation tasks (2 sets of 20 tasks) for each baseline comparison (2 baseline comparisons). We collect a total of 1200 user evaluations. We ask users to rank their preferred methods with respect to subject identity preservation, style fidelity with respect to the background image, and realistic insertion of the subject into the background image. We show the results in Table 4. We observe a resounding preference of users for our generated outputs compared to all other methods.

Broad Qualitative Comparisons We show in Figure 8 a broad qualitative comparison between our method and state-of-the-art methods including StyleAlign, InstantStyle, TF-ICON, ObjectDrop and AnyDoor. We can see that our method outperforms competing methods on subject identity preservation, accurate stylization and realistic insertion. Other methods also heavily struggle with (1) artifacts (2) inexistent harmonization (3) erased objects (4) deformed subjects (5) no shadows or reflections and more.

5. Conclusion

We introduced the problem of style-aware drag-and-drop, a new challenge in image generation that enables intuitive subject insertion with style consistency. We proposed a method combining style-aware personalization and insertion through bootstrapped domain adaptation which outperforms baselines in style adherence and insertion realism. To support further research, we share the SubjectPlop dataset, featuring subjects and backgrounds across diverse styles and semantics.

Acknowledgements

We thank Daniel Winter, David Salesin, Yi-Hsuan Tsai, Robin Dua and Jay Yagnik for their invaluable feedback.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7

[2] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[4] Junyan Cao, Yan Hong, and Li Niu. Painterly image harmonization in dual domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 268–276, 2023. 3

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[6] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6593–6602, 2024. 2, 3, 8

[8] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. *arXiv preprint arXiv:2403.14572*, 2024. 2, 3

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3, 4

[10] 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, 3

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[12] Jing Gu, Yilin Wang, Nanxuan Zhao, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, and Xin Eric Wang. Swapanything: Enabling arbitrary object swapping in personalized visual editing. *arXiv preprint arXiv:2404.05717*, 2024. 2

[13] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*, 2023. 2, 3, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 2

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[16] Zheng Hui, Jie Li, Xiumei Wang, and Xinbo Gao. Image fine-grained inpainting. *arXiv preprint arXiv:2002.02609*, 2020. 2

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[18] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. *ArXiv*, abs/1812.02350, 2018. 2

[19] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 2

[20] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 725–741. Springer, 2020. 2

[21] Lingxiao Lu, Bo Zhang, and Li Niu. Dreamcom: Finetuning text-guided inpainting model for image composition. *arXiv preprint arXiv:2309.15508*, 2023. 3

[22] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2, 3, 7, 8

[23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 5

[25] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[26] Li Niu, Junyan Cao, Yan Hong, and Liqing Zhang. Painterly image harmonization by learning from painterly objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4343–4351, 2024. 3

[27] Evangelos Ntavelis, Andrés Romero, Siavash Bigdeli, Radu Timofte, Zheng Hui, Xiumei Wang, Xinbo Gao, Chajin Shin, Taeoh Kim, Hanbin Son, et al. Aim 2020 challenge on image

extreme inpainting. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 716–741. Springer, 2020. 2

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 4, 6

[30] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2349–2359, 2023. 3

[31] 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. 2, 4

[32] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 181–190, 2019. 2

[33] 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. 2, 5

[34] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 3

[35] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510. IEEE, 2023. 1, 3, 4, 7

[36] 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, 3, 7

[37] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6933, 2024. 3

[38] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[39] 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, 5

[40] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4208–4217, 2024. 3

[41] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. 2023. 2, 3

[42] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. 2015. 2

[43] 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. In *37th Conference on Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems Foundation, 2023. 2, 3, 7

[44] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 7

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2022. 2

[46] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*, 2022. 3

[47] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, He Zhang, Wei Xiong, and Daniel Aliaga. Imprint: Generative object compositing by learning identity-preserving representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8048–8058, 2024. 3

[48] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[49] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668*, 2023. 3, 7

[50] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3

[51] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 3, 4, 7

[52] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 3

[53] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 2

[54] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. Primecomposer: Faster progressively combined diffusion for image composition with attention steering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10824–10832, 2024. 2, 3

[55] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. *arXiv preprint arXiv:2403.18818*, 2024. 2, 3, 5, 7

[56] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022. 2

[57] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[58] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2

[59] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2, 3

[60] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[61] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1486–1494, 2019. 2

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[63] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019. 2