

A^TA: Adaptive Transformation Agent for Text-Guided Subject-Position Variable Background Inpainting

Yizhe Tang^{1*} Zhimin Sun^{1,2*} Yuzhen Du^{1,2} Ran Yi^{1†} Guangben Lu²
 Teng Hu¹ Luying Li² Lizhuang Ma¹ Fangyuan Zou²
¹Shanghai Jiao Tong University ²Tencent

Abstract

Image inpainting aims to fill the missing region of an image. Recently, there has been a surge of interest in foreground-conditioned background inpainting, a sub-task that fills the background of an image while the foreground subject and associated text prompt are provided. Existing background inpainting methods typically strictly preserve the subject’s original position from the source image, resulting in inconsistencies between the subject and the generated background. To address this challenge, we propose a new task, the “Text-Guided Subject-Position Variable Background Inpainting”, which aims to dynamically adjust the subject position to achieve a harmonious relationship between the subject and the inpainted background, and propose the Adaptive Transformation Agent (A^TA) for this task. Firstly, we design a PosAgent Block that adaptively predicts an appropriate displacement based on given features to achieve variable subject-position. Secondly, we design the Reverse Displacement Transform (RDT) module, which arranges multiple PosAgent blocks in a reverse structure, to transform hierarchical feature maps from deep to shallow based on semantic information. Thirdly, we equip A^TA with a Position Switch Embedding to control whether the subject’s position in the generated image is adaptively predicted or fixed. Extensive comparative experiments validate the effectiveness of our A^TA approach, which not only demonstrates superior inpainting capabilities in subject-position variable inpainting, but also ensures good performance on subject-position fixed inpainting.

1. Introduction

Image inpainting [37] aims to fill the missing region in an image, which is an important research topic in computer vision. With the development of text-to-image (T2I) diffusion models [18, 25, 27], text-guided inpainting meth-

A car is parked on the road, with a waterfall cascading in the distance, surrounded by lush vegetation.



Figure 1. For foreground-conditioned background inpainting, (a) fixing the object position specified by the input image (left-top) may contradict the generated background; (b) while our model achieves subject-position variable background inpainting, adaptively determines a suitable location for the subject, and generates an image with a harmonious subject-background relationship.

ods [1, 2, 10, 11, 40, 50] have achieved promising progress, which fill a user-specified region under the inputs of an image, a masked region, and a text prompt, and achieve inpainting results consistent with the text prompt. In most inpainting scenarios, the inpainted regions are foreground regions, and some parts of the foreground human or objects are restored or edited according to the text guidance [1, 11, 27, 40, 47, 50]. However, in this paper, we focus on the inpainting scenarios where the foreground region is given and the whole background region is to be inpainted [23, 41, 46], and aim to fill the background region according to the text prompt, which has wide applications in advertisement design, product promotional design, etc.

When filling the specified masked regions, existing inpainting methods [11, 40, 50] emphasize the *preservation of the unmasked (given) region* as much as possible, requiring pixel-level strong consistency for the unmasked region. However, in the case of foreground-conditioned background inpainting [23, 41, 46], where the foreground is given and the background is to be inpainted, this unmasked region preservation constraint can sometimes be overpowering and result in problems. Given a subject image, a back-

*Equal contribution. {tangyizhe, zhimin.sun}@sjtu.edu.cn

†Corresponding author. ranyi@sjtu.edu.cn

ground mask, and a text prompt describing the desired background, filling the content in the background while requiring the foreground strictly unchanged will limit the subject’s position to the same position in the original image. However, in background inpainting scenarios in advertisement and art designs, the scale and position of the subject in the generated image often cannot be determined before generation. More importantly, directly fixing the position of the subject to that in the original image may contradict the generated background. *E.g.*, given an image of a car in the center, and a text prompt “A car is parked on the road, with a waterfall cascading in the distance...”, if fixing the car position to that in the original image, the inpainted image may not fully draw the desired background (the car is on mud instead of on road in Fig. 1(a)).

Taking these considerations into account, we propose a new task for this scenario, the **Text-Guided Subject-Position Variable Background Inpainting** task, where the position of the subject can be adaptively varied. The core of this task is the ability to adaptively determine a suitable location of subject, based on the subject information in the original image and the text prompt describing the desired background, thereby generating an inpainted image that has a harmonious positional relationship between the subject and the inpainted background.

For the Text-Guided Subject-Position Variable Background Inpainting task, we propose the **Adaptive Transformation Agent ($A^T A$)**. Firstly, to achieve variable subject-position, we design a **PosAgent Block** to adaptively predict an appropriate displacement based on given features, to move the subject to the appropriate position. Specifically, the spatial transformation is applied at the feature level, where the PosAgent predicts a pair of displacement transformation parameters, which are then used to transform multi-scale subject features by spatial feature transform (SFT). Secondly, to transform hierarchical feature maps from deep to shallow based on semantic information, we design the **Reverse Displacement Transform (RDT) module**, which arranges multiple PosAgent blocks in a reverse structure, *i.e.*, the output of the first PosAgent block transforms the deepest feature map, while the last PosAgent block transforms the shallowest feature map. This reverse transform structure effectively alleviates the subject deformation problems of separate or sequential structures, while achieving subject-background harmonious position prediction. Thirdly, to allow the model to both adaptively move the subject and maintain the subject’s position, we equip the $A^T A$ with a **Position Switch Embedding**, which controls whether the position of the subject in the generated image is adaptively predicted or fixed. By setting the position switch embedding, users can flexibly switch between subject-position variable and subject-position fixed inpainting. To train $A^T A$ with the position switch embedding, we

use a hybrid training strategy, where half of the data uses variable position samples and half uses fixed position ones.

Extensive comparative experiments validate the effectiveness of our $A^T A$ approach, which not only demonstrates superior inpainting capabilities in subject-position variable inpainting, but also ensures good performance on subject-position fixed inpainting.

We summarize our contributions as follows:

1. We propose a novel Text-Guided Subject-Position Variable Background Inpainting task, which aims at generating inpainted images with a harmonious positional relationship between the subject and the inpainted background.
2. We propose Adaptive Transformation Agent ($A^T A$), which adaptively shifts and scales the subject to an appropriate position through the Reverse Displacement Transform (RDT) module. RDT consists of multiple PosAgent blocks that predict the displacement transformation parameters, which are then applied to transform hierarchical features by SFT, and arranges the PosAgent blocks in a novel reverse structure, effectively alleviating subject deformation.
3. We equip the $A^T A$ with a Position Switch Embedding, which controls whether the position of the subject is adaptively predicted or fixed, making the model both can adaptively move the subject and maintain the subject’s position. An end-to-end hybrid training strategy is designed to train $A^T A$ with the position switch embedding.

2. Related Work

2.1. Image Inpainting

With the development of T2I diffusion models [27], great progress has been made in image inpainting [1, 2, 7, 10, 11, 27, 37, 40, 42, 50]. Blended Latent Diffusion [1, 2] uses separate denoising for masked and unmasked areas to achieve inpainting, and SD-Inpainting [27] and ControlNet-Inpainting [47] further refine by fine-tuning StableDiffusion [27] models. SmartBrush [40] enhances model understanding of complex scenes by training with pairs of descriptive objects and corresponding masks. PowerPaint [50] combines image inpainting with image removal for better text-image alignment. BrushNet [11] employs a versatile two-branch network architecture to inject features from masked images and ensure consistent inpainting results.

Existing inpainting methods, though effective, sometimes fail to integrate backgrounds seamlessly, due to difficulties in matching the fixed subject with the varying background context during generation. In comparison, our $A^T A$ framework overcomes this by using the Reverse Displacement Transform module to dynamically adjust the position and scale of the subject, ensuring a natural integration of subject and background in the inpainted images.

2.2. Controllable Image Generation

As the image generation model continues to evolve, there is a growing emphasis on controllable image generation [32, 33, 45] within the research community. Works that involve adapter networks [24, 44, 47] control content by integrating side-branch adapter networks, which can be used for inpainting tasks after fine-tuning. However, these works can hardly achieve the appropriate subject position, requiring manual attempts at different positions and zoom levels.

Personalized Image Generation [15, 17, 28, 29, 36] generates new images by preserving the identity information from the original image, yielding artistic and stylized images with a high degree of creative freedom. Nevertheless, these personalized models generate images based solely on textual content, lacking the capability for controlled generation in conjunction with a subject image.

Layout-to-Image, a multi-stage task for controlled image generation using layouts, also meets challenges when applied to our task. In the first stage, users must specify the positions of objects and backgrounds to generate a layout [3, 9, 48], and controllable T2I methods [5, 16, 39] are used in the second stage to generate the image. Although these methods have achieved satisfactory results, they can not realize pixel-level control of the subject, which is crucial for our task. Additionally, the layout generation process is complicated, resulting in unreliable generation of inpainting images, further highlighting the need for our A^TA.

3. Preliminaries

3.1. Diffusion Models

Diffusion models [27] encompass forward and backward processes. In the forward diffusion process, a clean sample x_0 is converted into a noise sample x_t by adding Gaussian noise ϵ , which can be expressed as follows:

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

where x_t represents the noised feature at step t and α_t denotes a hyper-parameter of the noise level. In the backward denoising process, starting with the noise x_T , a trainable network ϵ_θ predicts the noise at each step t leveraging conditional c , and derives original data x_0 after T iterations.

The diffusion model training process aims to optimize the denoising network ϵ_θ so that it can accurately predict the noise given the condition c . The training loss function is formulated as:

$$\mathcal{L} = \mathbb{E}_{x_0, c, \epsilon \sim \mathcal{N}(0, 1), t} [\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2]. \quad (2)$$

This loss function quantifies the discrepancy between the predicted noise and the actual noise, thereby guiding the optimization process.

3.2. Hunyuan-DiT

Hunyuan-DiT [18] is the base model of Adaptive Transformation Agent, which is an open-source enhanced iteration of DiT [25], a transformer-based diffusion model operating on latent patches. Hunyuan-DiT integrates the text condition with the diffusion model through cross-attention [27], extends Rotary Positional Embedding (RoPE) [30] to a 2D format, which encodes absolute and relative positional dependencies, and expands its capabilities to the image domain. For text encoding, Hunyuan-DiT merges mT5 [43] and CLIP [26], leveraging their complementary strengths to enhance the precision and creativity of the T2I generation.

4. Text-Guided Subject-Position Variable Background Inpainting

4.1. Task Definition

A variety of tasks characterize the domain of image inpainting, which can be broadly categorized into Text-guided object inpainting, Shape-guided object inpainting, Context-aware image inpainting, and Outpainting, as detailed in previous studies [37, 50]. Recently, there has been a surge of interest in *foreground-conditioned background inpainting*, a sub-task that involves filling in the background of an image while the foreground subject and associated text prompt are provided [23, 41, 46]. This foreground-conditioned inpainting task typically fixes the position of the subject to match that in the original subject image. However, in applications such as advertising and artistic design, the position and scale of the subject within the generated image are often undetermined before the generation process. Moreover, strictly adhering to the original image’s subject position can conflict with the newly generated background, resulting in inconsistencies between the subject and background.

To address this issue, we introduce a novel task: **Text-Guided Subject-Position Variable Background Inpainting**. Unlike the conventional foreground-conditioned background inpainting task [23, 41, 46], this subject-position variable task demands intelligently adjusting the position and scale of the foreground subject to achieve a harmonious relationship with the generated background, while maintaining the original shape and details of the subject.

4.2. Adaptive Transformation Agent

To address the challenges in the Text-Guided Subject-Position Variable Background Inpainting task, we propose a novel framework **Adaptive Transformation Agent** (A^TA), as shown in Fig. 2, comprising 4 modules: Feature extraction, Reverse displacement transform, Feature fusion, and Diffusion denoising. We use Hunyuan-DiT [18] as the base model, and mainly develop the mechanisms of subject feature extraction, displacement prediction, and feature injection to achieve subject-position variable inpainting.

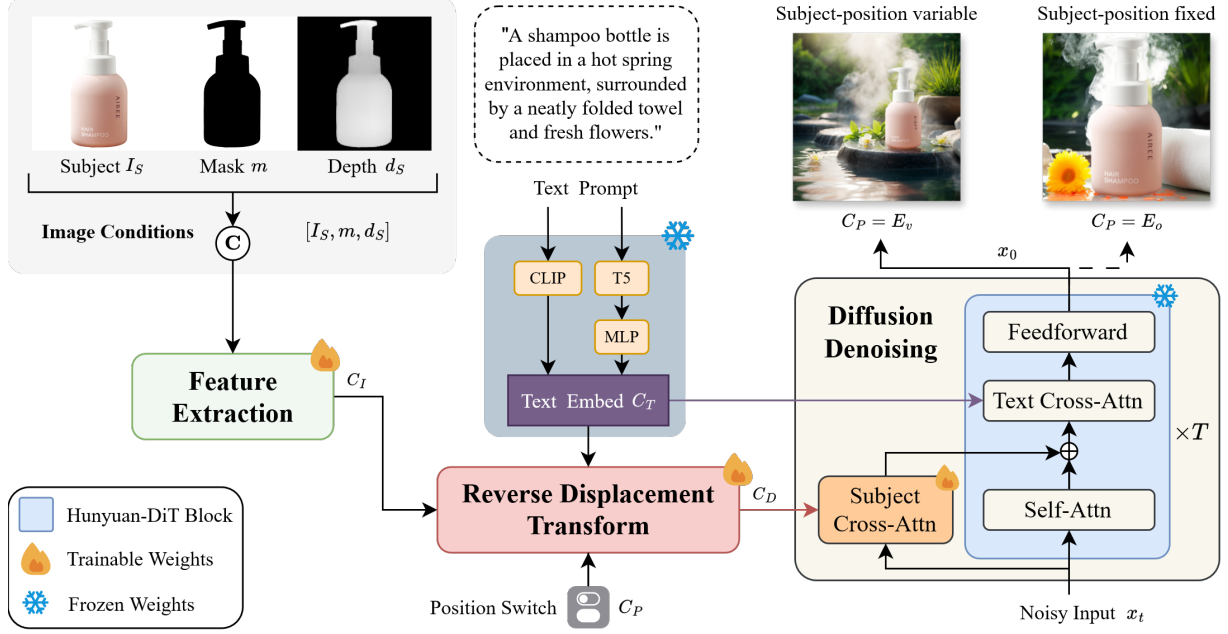


Figure 2. **Adaptive Transformation Agent (A^TA)** comprises 4 modules: Feature extraction, **Reverse displacement transform**, Feature fusion, and Diffusion denoising. We use Hunyuan-DiT [18] as the base model, and mainly develop the subject feature extraction, displacement transformation prediction, and displaced feature injection mechanisms to achieve *subject-position variable inpainting*. We also design a **position switch embedding** to control whether the position of the subject in the generated image is adaptively predicted or fixed.

Feature extraction. The inputs to our A^TA framework consists of: a subject image I , a background mask m , and a text prompt T (describing the desired background). A^TA will preprocess the subject image I for feature extraction. Specifically, we obtain the masked subject image by:

$$I_S = I \odot (1 - m). \quad (3)$$

Then, we consider the mask m and depth map d_S (extracted by [4]) of the subject image as the condition maps, and use Swin-Transformer [21] as an image encoder Φ to extract multi-scale subject image feature C_I :

$$C_I = \Phi([I_S, m, d_S]), \quad (4)$$

which can effectively capture the global and local features of the subject through its hierarchical structure and sliding window mechanism. Moreover, since our base model is a transformer-based structure (a text-to-image DiT), extracting subject features using Swin-Transformer can better align with the base model and allow for better fusion of the subject features through attention operations [23]. Meanwhile, the text features C_T are extracted by the CLIP [26] and mT5 [43] from the input text prompt T .

Reverse displacement transform. Next, A^TA adopts our Reverse Displacement Transform (RDT) module, which consists of multiple PosAgent blocks that predict displacement transformation parameters and arrange them in a reverse structure (detailed in Sec. 4.3). The RDT module can effectively generate the multi-level displaced features C_D :

$$C_D = \text{RDT}(C_I; C_T, C_P, t), \quad (5)$$

where C_P refers to the position switch embedding, which controls whether the position of the subject in the generated image is adaptively predicted or fixed (detailed in Sec. 4.4); and t represents the time embedding in diffusion model.

Feature fusion by subject cross-attention. Afterward, A^TA injects the displaced features C_D into the base T2I DiT model and strengthens the model’s adaptation to the subject’s location through decoupled cross-attention mechanism [44]. Specifically, an additional **subject cross-attention** layer is designed to insert the subject features, which calculates relationships between displaced subject features and latent. Then, different from conventional IP-Adapter, we follow the Self-IPA from [23] and integrate the subject cross-attention with the self-attention layer by:

$$x_t = \alpha \odot \text{Self-Attention}(x_t W_q, x_t W_k, x_t W_v) + \beta \odot \text{Cross-Attention}(x_t W'_q, C_D W'_k, C_D W'_v), \quad (6)$$

where α, β are learnable weights for attention blocks, and W_q, W_k, W_v are trainable parameters. This decoupled attention strategy maintains the stability of the original model through the injection of adapters. After this module, text features C_T are injected by the text cross-attention module, which is consistent with the base model [18]:

$$x_t = \text{Cross-Attention}(x_t W_q, C_T W_k, C_T W_v). \quad (7)$$

Diffusion denoising. After all these features are fused, A^TA performs a T-step denoising process as in conventional diffusion models and obtains the inpainted image.

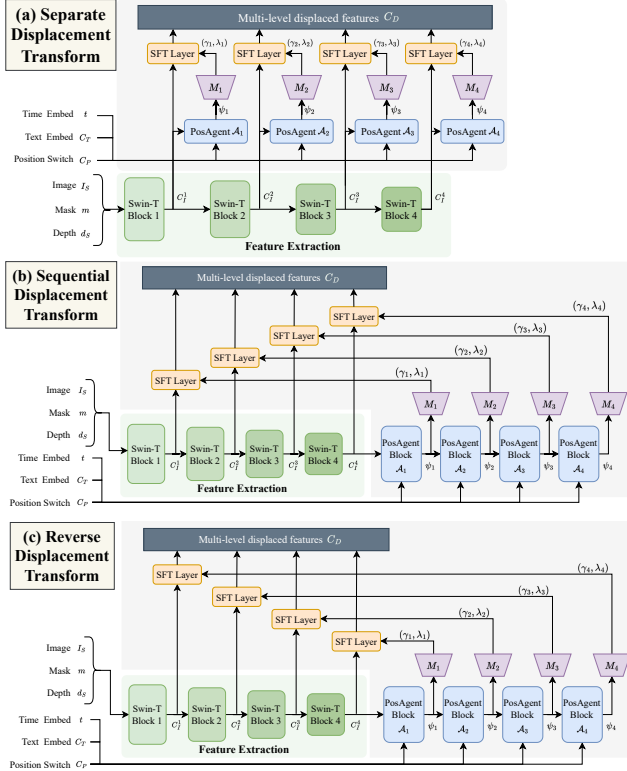


Figure 3. Comparison of different structures for Displacement Transform. To transform the hierarchical feature maps from deep to shallow based on semantic information, we propose a novel Reverse Displacement Transform module.

4.3. Reverse Displacement Transform

Existing works [11, 23, 50] use VAE [12] to extract conditions from unmasked images for injection into the inpainting model. Although effective for subject-position fixed inpainting tasks, VAE is impractical for our subject-position variable inpainting task, which cannot simultaneously extract sufficient image information and guarantee position independence. Therefore, apart from the subject feature extractor, a **PosAgent Block** is designed to predict a suitable displacement for the subject to move it to a suitable position, and perform SFT at the feature level, and multiple consecutive PosAgent blocks are arranged to cope with multi-level features. Different from straightforward ways to inject position offset by the Separate or Sequential Displacement Transform (Fig. 3a, 3b), which suffer from subject deformation, our **Reverse Displacement Transform** module (Fig. 3c) transforms hierarchical feature maps from deep to shallow based on the semantic information, effectively alleviating subject deformation and achieving subject-background harmonious position prediction.

PosAgent blocks. Firstly, to achieve variable subject position, we design a PosAgent block $\mathcal{A}$ to adaptively predict a suitable displacement based on the given features. The inputs of $\mathcal{A}$ consist of subject image feature C_I (Eq. 4),

text feature C_T , position switch embedding C_P , and time embedding t , where C_I contain multi-scale feature maps $\{C_I^1, \dots, C_I^N\}$. Given these inputs, $\mathcal{A}$ predicts a displacement transformation feature ψ , which will be mapped into transformation parameters. To cope with features at different scales, we design multiple consecutive PosAgent blocks $\{\mathcal{A}_1, \dots, \mathcal{A}_N\}$ to predict multiple displacement transformation features $\{\psi_1, \dots, \psi_N\}$. Specifically, the deepest feature map C_I^N serves as the input to $\mathcal{A}_1$, capturing the most refined texture information [21], while the output displacement feature ψ_{i-1} from $\mathcal{A}_{i-1}$ serves as input to $\mathcal{A}_i$:

$$\psi_i = \begin{cases} \mathcal{A}_i(C_I^N; C_T, C_P, t), & \text{if } i = 1, \\ \mathcal{A}_i(\psi_{i-1}; C_T, C_P, t), & \text{otherwise.} \end{cases} \quad (8)$$

For the output displacement feature ψ_i of each block $\mathcal{A}_i$, we then apply a learnable mapping function M_i that generates a pair of displacement transformation parameters, denoted as (γ_i, λ_i) , where γ_i represents scale and λ_i represents shift:

$$(\gamma_i, \lambda_i) = M_i(\psi_i). \quad (9)$$

To perform better spatial transformation of hierarchical feature maps, we introduce a Spatial Feature Transform (SFT) layer [35]. After obtaining the displacement transformation parameters $\{(\gamma_i, \lambda_i)\}_{i=1}^N$ from the multiple PosAgent blocks, SFT layer transforms the hierarchical feature maps $\{C_I^1, \dots, C_I^N\}$ through scaling and shifting.

Reverse-structure transformation. To transform the hierarchical feature maps from deep to shallow based on the semantic information, we propose a novel reverse transformation module, which arranges multiple PosAgent blocks in a **reverse structure**. *I.e.*, (γ_1, λ_1) are used to transform the deepest feature map C_I^N , while (γ_N, λ_N) will transform the shallowest feature map C_I^1 (shown in Fig. 3c). The process of using the i -th PosAgent’s transformation parameters to transform the $(N+1-i)$ -th feature map is formulated as:

$$\text{SFT}(C_I^{N+1-i}; \gamma_i, \lambda_i) = \gamma_i \odot C_I^{N+1-i} + \lambda_i. \quad (10)$$

The reverse structure ensures that the deepest features that capture the most semantically rich information are processed first, and to compensate for the detailed information, by gradually combining deeper features with shallower ones that capture finer-grained details, the model can better understand the entire scene.

Finally, we obtain the adaptively displaced multi-level features C_D from the RDT module, which will be subsequently injected into the base model:

$$\text{RDT}(C_I; C_T, C_P, t) = \{\gamma_i \odot C_I^{N+1-i} + \lambda_i, i \in [1, N]\}. \quad (11)$$

4.4. Hybrid Training with Position Switch

To provide users with more choices, we require ATA to implement both inpainting tasks in the form of adaptive subject position and fixed subject position. To achieve this, we

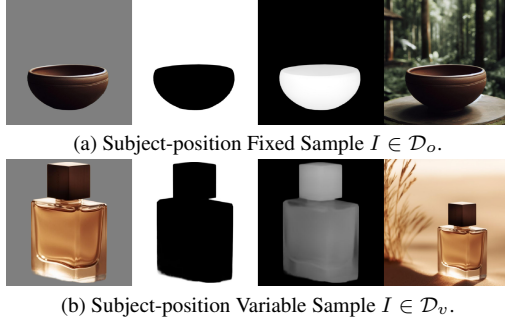


Figure 4. Training samples and corresponding image conditions. From left to right: masked subject image I_S , mask m , depth map d_S , ground truth image I .

equip our $A^T A$ with a **Position Switch Embedding**, enabling flexible switch between subject-position variable and fixed inpainting. To train the $A^T A$ with the position switch embedding, we use an end-to-end **Hybrid Training Strategy**, where the training dataset includes samples from two tasks, mixed together in a 1 : 1 ratio.

Position switch. To allow the model to both adaptively move the subject and maintain the subject’s position, we equip our $A^T A$ framework with a position switch embedding C_P , which serves as a “switch” to control whether the position of the subject is adaptively predicted or fixed. C_P has 2 states, E_v for subject-position variable inpainting, while E_o for subject-position fixed inpainting, where both E_o and E_v are learnable. Along with text embedding C_T and image embedding C_I , C_P is injected into the condition side of the RDT (Eq. 5), serving as an input to the PosAgent when predicting the transformation parameters.

Hybrid training. We use an end-to-end hybrid training strategy to train the $A^T A$ with the position switch embedding, where half of the training data uses variable position samples, and half uses fixed position samples. Besides the original training samples $\mathcal{D}_o$ (Fig. 4a) for the fixed-position inpainting, we adopt training samples $\mathcal{D}_v$ for our variable-position inpainting (Fig. 4b). To construct $\mathcal{D}_v$, we process the subject image to center and enlarge the subject, use the centered image as the input, and use the original image (subject in a suitable position, away from the center) as the ground truth, to make the model adaptively learn a suitable subject position. $\mathcal{D}_v$ and $\mathcal{D}_o$ are mixed in a 1 : 1 ratio. For variable-position inpainting, we set the position switch embedding to E_v and use $I \in \mathcal{D}_v$ as the input image, while encountering fixed-position samples $I \in \mathcal{D}_o$, we set the position switch embedding to E_o .

5. Experiments

5.1. Experimental Settings

Datasets. The datasets [13, 19, 34] typically employed for inpainting tasks, have limitations due to their low-quality,

Table 1. Quantitative Comparisons for the Text-Guided Subject-Position Variable Background Inpainting task.

Methods	Image Quality FID ↓	Extension Ratio BiRefNet ↓	Text-alignment VQA Score ↑	Multi-subject FlorenceV2 ↓	Rationality GPT4o ↑
LayerDiffusion	100.29	21.09%	0.862	26.40%	85.6
PowerPaint	89.31	21.16%	0.805	25.04%	91.6
BrushNet-SDXL	90.72	14.27%	0.869	20.32%	92.1
BrushNet-SD1.5	90.79	16.26%	0.857	25.40%	88.1
SD3-ControlNet	81.14	24.23%	0.908	14.75%	89.8
FLUX-ControlNet	82.82	22.38%	0.903	15.09%	87.8
GLIGEN	213.70	46.82%	0.662	9.37%	88.9
ELITE	160.14	58.36%	0.654	19.06%	81.6
BLIP-Diffusion	191.53	44.36%	0.524	8.03%	87.9
$A^T A$	79.66	7.92%	0.883	4.73%	92.8

cluttered real-world images or imprecise annotated mask boundaries [11], while our task requires training datasets with foreground subjects that can be completely isolated, allowing for adaptive movement of the subject. Thus, we follow Pinco [23] and collect 79K high-quality images with varying aspect ratios across diverse subject categories, including products, vehicles, people, and animals. We use BiRefNet [49] to generate subject segmentation masks, and ZoeDepth [4] to produce depth maps. When constructing the training samples for variable-position inpainting, to deal with samples with different subject positions, we process the images to move the subject to the center of the image, and enlarge them isometrically to 95% of the maximum scale (Fig. 4b). The mask m and depth map d_S are scaled up with the same proportions. This process results in data pairs where the input subject (centered image) and the reference subject (original image) differ in position and scale. We use Hunyuan-Vision [31] to label the subject category and generate detailed text prompts.

Implementation Details. Our training is conducted on 16 NVIDIA A100 GPUs for 100 epochs, with a batch size of 4 per GPU. We use AdamW [22] optimizer, and the learning rate is set to 1e-4, using a 300-steps warmup. We choose DiT [25] as the architecture for the PosAgent block. The image aspect ratios for training include 1 : 1, 9 : 16, 16 : 9. To ensure the full consistency of the unmasked region, previous works [11, 50] usually adopt a copy-and-paste operation. For $A^T A$, since the subject position and scale in generated images are adaptively variable, we first use Florence-2 [38] to detect the accurate bounding box of the subject based on its text prompt. Then we rescale the subject image to fit the bounding box and conduct the blending operation [11].

Evaluation Metrics. We evaluate the model performance by the Image Quality, Extension Ratio, Text-alignment, Multi-subject Rate, and Position Rationality: 1) We use FID to assess the image quality; 2) To measure the subject consistency, we utilize BiRefNet [49] to generate an accurate subject mask and calculate the OER [6] to measure the subject extension ratio; 3) For text alignment, we use VQA Score [20] to assess the subject-text alignment, and use FlorenceV2 [38] to detect the multi-subject rate based on the given subject prompt. Following Pinco [23], We leverage

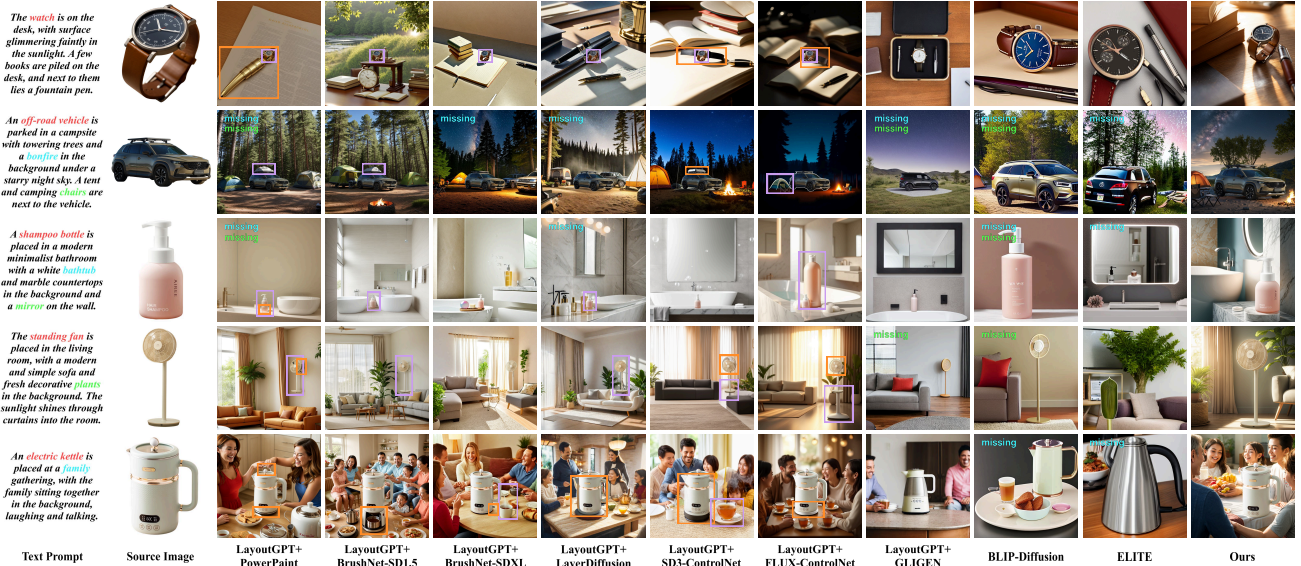


Figure 5. Comparison between our $A^T A$ and the baseline methods. We highlight the unreasonable extension parts with orange boxes and the unreasonable layouts with purple boxes, and label the missing objects with corresponding colors. Please zoom in for more details.

GPT4o to evaluate the rationality of the subject position (full score 100). 4) We further conduct a user study to assess the rationality of the subject position and overall quality.

5.2. Comparisons

We construct baselines for the Text-Guided Subject-Position Variable Background Inpainting task by coupling existing text-guided inpainting methods with a layout generation method. Existing T2I inpainting methods only achieve inpainting with subject-position fixed, but our task requires the subject to adaptively change its position. To enable a fair comparison, we use LayoutGPT [9] to generate a plausible subject-position layout, and then combine LayoutGPT with those existing inpainting methods for comparison, including PowerPaint [50], BrushNet1.5, BrushNetXL [11], LayerDiffusion [46], SD3-ControlNet [8] and FLUX-ControlNet [14]. Specifically, we first use LayoutGPT to generate the bounding box of the subject based on the given text, then scale and shift the subject into the given region¹. We also combine LayoutGPT with GLIGEN [16], which can generate an image based on the given bounding box using text and conditional image. Finally, we compare with 2 personalized image generation methods, ELITE [36] and BLIP-Diffusion [15].

Qualitative Comparisons. Fig. 5 shows the results of qualitative comparisons with different methods². As shown in Fig. 5, LayoutGPT tends to produce unreasonable layouts (the 1st and 4th rows) or layouts with inappropriate ratio (the 3rd row). In the 2nd row, although a suitable layout

¹We also compare with these methods using user-provided subject layout instead of LayoutGPT, shown in #Suppl Sec. S4, S5.

²More images of different aspect ratios are shown in #Suppl Sec. S7.

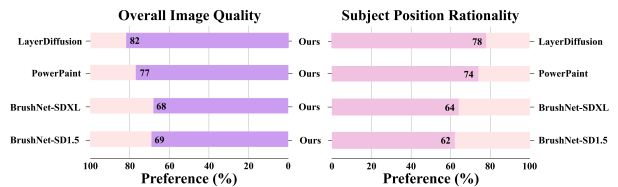


Figure 6. User study between $A^T A$ and other inpainting methods.

is given, these inpainting models cannot understand the position information of the subject, and the generated background results conflict with the text, resulting in inharmonious image content layout or lacking objects in the text. Meanwhile, customized methods struggle with subject consistency and text-image alignment due to inadequate control over image features, leading to poor performance.

Quantitative Comparisons. Tab. 1 shows the quantitative comparisons among different methods, where our $A^T A$ achieves the best FID score, demonstrating high generation quality. Additionally, our extension ratio and multi-subject rate are less than 60% of other methods, indicating $A^T A$'s ability to prevent subject expansion and redundancy. Finally, we achieve the best position rationality score, indicating a better positional relationship. Overall, these metrics demonstrate our superiority in terms of generation quality, extension rate, and subject position rationality.

User Study. We invite 31 participants for our user study, each receiving 40 sets of test questions, with each set consisting of 2 images: one generated by $A^T A$ and the other from a different method. Participants are asked to choose the better image based on *rationality of subject position*, and overall image quality. As shown in Fig. 6, our $A^T A$ receives the most preference from the participants, demonstrating our superior position rationality.

Table 2. Ablation study on different structures for Displacement Transform (DT) module. Here, HT refers to Hadamard Transform.

DT Module	Image Quality FID ↓	Expand Ratio BiRefNet ↓	Text-alignment VQA Score ↑	Multi-subject FlorenceV2 ↓
w/o DT	81.74	10.89%	0.873	4.89%
Separate DT	83.83	10.25%	0.880	4.84%
Sequential DT	83.36	9.73%	0.875	5.34%
Reverse HT	85.88	11.26%	0.877	5.79%
Reverse DT ($A^T A$)	79.66	7.92%	0.883	4.73%

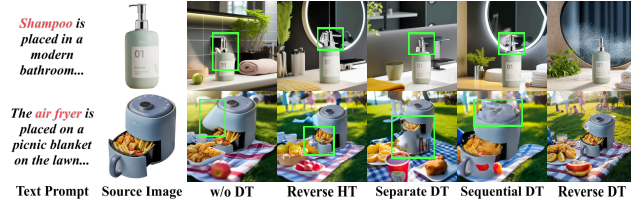


Figure 7. Qualitative ablation study on different structures for Displacement Transform (DT) module. We highlight the deformation part of the subject with green boxes.

5.3. Ablation Study

We conduct extensive ablation studies to validate the effectiveness of each module. We compare with different Displacement Transform structures and analyze condition embedding injection in the PosAgent block.

Displacement Transform modules. We perform ablation studies by comparing various Displacement Transform (DT) structures to assess their impact (Tab. 2 and Fig. 7). Initially, we remove the entire DT module, relying solely on the Swin-Transformer to process condition data, but the generated results have less plausible subject positions and layouts. Next, we examine the Separate DT module, comprising N distinct PosAgent blocks, each tasked with altering a specific attention map C_L^i . This approach results in disjointed feature transformations across levels, causing subject distortions. We also assess the Sequential DT module, featuring consecutive PosAgent blocks with the output of the i -th block θ_i affecting the i -th attention map C_L^i . This module also induces subject deformation and performs inferior to our Reverse DT. Additionally, we replace the SFT with the Hadamard transform and modify the hierarchical features using the Hadamard product, resulting in noticeable subject deformation. All these findings underscore the superiority of our Reverse Displacement Transform in maintaining subject integrity and overall image quality.

Condition Injected to PosAgent blocks. We perform ablation studies on the conditions input to the PosAgent block, *i.e.*, time, text, and position switch embeddings, to verify their roles, as illustrated in Tab. 3 and Fig. 8. 1) Given that the Position Switch is integral to our hybrid training strategy, we train the PosAgent block w/o Position Switch input using the subject-position variable images $\mathcal{D}_v$ only, and we unexpectedly observed unstable performance and worse quantitative metrics than w/ the switch.

Table 3. Ablation study on the conditions of the PosAgent module. Here, PS refers to Position Switch.

PosAgent Text Time PS	Image Quality FID ↓	Expand Ratio BiRefNet ↓	Text-alignment VQA Score ↑	Multi-subject FlorenceV2 ↓
✓ ✓	83.71	9.68%	0.879	5.95%
✓ ✓ ✓	82.20	8.41%	0.879	4.89%
✓ ✓ ✓	82.36	8.08%	0.881	4.61%
✓ ✓ ✓	79.66	7.92%	0.883	<u>4.73%</u>

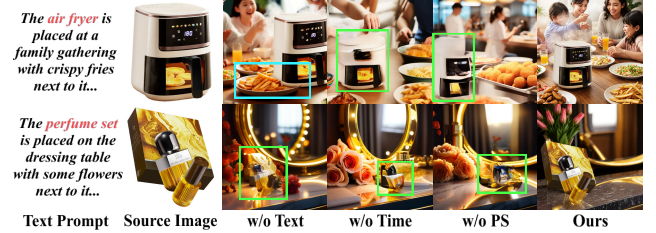


Figure 8. Qualitative ablation study on the conditions of the PosAgent module. We highlight the deformation part of the subject with green boxes and the misplacement part with the blue box.

We attribute this to the hybrid training strategy’s ability to effectively separate feature extraction and displacement transform processes, and enhance the ability in each task (position-variable/fixed) by multi-task learning. 2) When the PosAgent lacks time embedding, the resulting images occasionally exhibit deformation, and the metrics indicate a decline in performance. Moreover, omitting time embedding in PosAgent leads to an unstable training process. 3) In the case of the PosAgent without text embedding, the image quality, extension ratio, and text-alignment are worse, whereas we notice a slight improvement in the multi-subject rate. The potential reason is that, without text embedding, the transformed features retain more subject information at the image level, allowing the base model to recognize the injected data more effectively. However, the absence of text guidance in PosAgent makes it challenging to position subjects appropriately in line with textual prompts, potentially leading to missing objects or misplacement due to conflicts between subject positioning and the text.

6. Conclusion

This paper introduces a novel Text-Guided Subject-Position Variable Background Inpainting task, aiming to generate inpainted images with a harmonious positional relationship between the subject and the background. For this task, we propose the Adaptive Transformation Agent ($A^T A$) framework, which dynamically adjusts the subject’s position relative to the background. $A^T A$ leverages multiple PosAgent blocks for displacement prediction, the RDT module for feature transform, and a Position Switch Embedding for flexible positioning control. Experiments confirm the effectiveness of our $A^T A$ in both variable and fixed subject-position inpainting tasks.

Acknowledgments

This work was supported by Shanghai Sailing Program (22YF1420300), National Natural Science Foundation of China (No. 62302297, 72192821, 62272447, 62472282, 62472285), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), the Fundamental Research Funds for the Central Universities (YG2023QNB17, YG2024QNA44), Beijing Natural Science Foundation (L222117).

References

- [1] 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. 1, 2
- [2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023. 1, 2
- [3] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 3
- [4] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4, 6
- [5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3
- [6] Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Virtualmodel: Generating object-id-retentive human-object interaction image by diffusion model for e-commerce marketing. *arXiv preprint arXiv:2405.09985*, 2024. 6
- [7] Yuzhen Du, Teng Hu, Ran Yi, and Lizhuang Ma. Ld-bfr: Vector-quantization-based face restoration model with latent diffusion enhancement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2852–2860, 2024. 2
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yan-nik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 7
- [9] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7
- [10] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8526–8534, 2024. 1, 2
- [11] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 1, 2, 5, 6, 7
- [12] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5
- [13] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 6
- [14] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 7
- [15] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7
- [16] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3, 7
- [17] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 3
- [18] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xincheng Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1, 3, 4
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [20] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 6
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 5
- [22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [23] Guangben Lu, Yuzhen Du, Zhimin Sun, Ran Yi, Yifan Qi, Yizhe Tang, Tianyi Wang, Lizhuang Ma, and Fangyuan

- Zou. Pinco: Position-induced consistent adapter for diffusion transformer in foreground-conditioned inpainting. *arXiv preprint arXiv:2412.03812*, 2024. 1, 3, 4, 5, 6
- [24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 3
- [25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1, 3, 6
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4
- [27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3
- [28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3
- [29] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 3
- [30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3
- [31] Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*, 2024. 6
- [32] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. Contrastive pseudo learning for open-world deepfake attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20882–20892, 2023. 3
- [33] Zhimin Sun, Shen Chen, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Rethinking open-world deepfake attribution with multi-perspective sensory learning. *International Journal of Computer Vision*, pages 1–24, 2024. 3
- [34] 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. 6
- [35] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 5
- [36] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3, 7
- [37] Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, and Hongkai Yu. Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046, 2023. 1, 2, 3
- [38] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 6
- [39] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3
- [40] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 1, 2
- [41] Tianyidan Xie, Rui Ma, Qian Wang, Xiaoqian Ye, Feixuan Liu, Ying Tai, Zhenyu Zhang, and Zili Yi. Anywhere: A multi-agent framework for reliable and diverse foreground-conditioned image inpainting. *arXiv preprint arXiv:2404.18598*, 2024. 1, 3
- [42] Zhekai Xu, Haohong Shang, Shaoze Yang, Ruiqi Xu, Yichao Yan, Yixuan Li, Jiawei Huang, Howard C Yang, and Jianjun Zhou. Hierarchical painter: Chinese landscape painting restoration with fine-grained styles. *Visual Intelligence*, 1(1): 19, 2023. 2
- [43] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020. 3, 4
- [44] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 3, 4
- [45] Ran Yi, Teng Hu, Mengfei Xia, Yizhe Tang, and Yong-Jin Liu. FEditNet++: Few-shot editing of latent semantics in gan spaces with correlated attribute disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(12):9975–9990, 2024. 3
- [46] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 1, 3, 7
- [47] 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. [1](#), [2](#), [3](#)
- [48] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. [3](#)
- [49] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. [6](#)
- [50] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)