# Locate, Assign, Refine: Taming Customized Image Inpainting with Text-Subject Guidance

Yulin Pan[1], Chaojie Mao[1], Zeyinzi Jiang[1], Zhen Han[1], and Jingfeng Zhang[1]

Alibaba Group
{yanwen.pyl, chaojie.mcj, zeyinzi.jzyz, hanzhen.hz, zhangjingfeng.zjf}@alibaba-inc.com

**Abstract.** Prior studies have made significant progress in image inpainting guided by either text or subject image. However, the research on editing with their combined guidance is still in the early stages. To tackle this challenge, we present **LAR-Gen**, a novel approach for image inpainting that enables seamless inpainting of masked scene images, incorporating both the textual prompts and specified subjects. Our approach adopts a coarse-to-fine manner to ensure subject identity preservation and local semantic coherence. The process involves (i) **Locate**: concatenating the noise with masked scene image to achieve precise regional editing, (ii) **Assign**: employing decoupled cross-attention mechanism to accommodate multi-modal guidance, and (iii) **Refine**: using a novel RefineNet to supplement subject details. Additionally, to address the issue of scarce training data, we introduce a novel data construction pipeline. This pipeline extracts substantial pairs of data consisting of local text prompts and corresponding visual instances from a vast image dataset, leveraging publicly available large models. Extensive experiments and varied application scenarios demonstrate the superiority of LAR-Gen in terms of both identity preservation and text semantic consistency. Project page can be found at https://ali-vilab.github.io/largen-page/.

**Keywords:** Image inpainting · Diffusion model · Text-subject-guided

## 1 Introduction

Thanks to the exhilarating advancements achieved by diffusion models [15,16,32,36], image generation [7,10,31,33,34] is witnessing an exuberant proliferation and has found its application in various scenarios [6,20,54]. Image inpainting, which aims to fill in missing regions of an image based on various guiding information, has emerged as a significant application and is exhibiting particularly attractive characteristics recently. Commonly, plain image inpainting technologies [43,57,58] leverage the inherent semantics of the original image to appropriately fill in masked regions. Additionally, some auxiliary information could be incorporated to facilitate customized edits. For example, text-guided image inpainting methods [1,46] adjust images based on user-provided text descriptions. However,

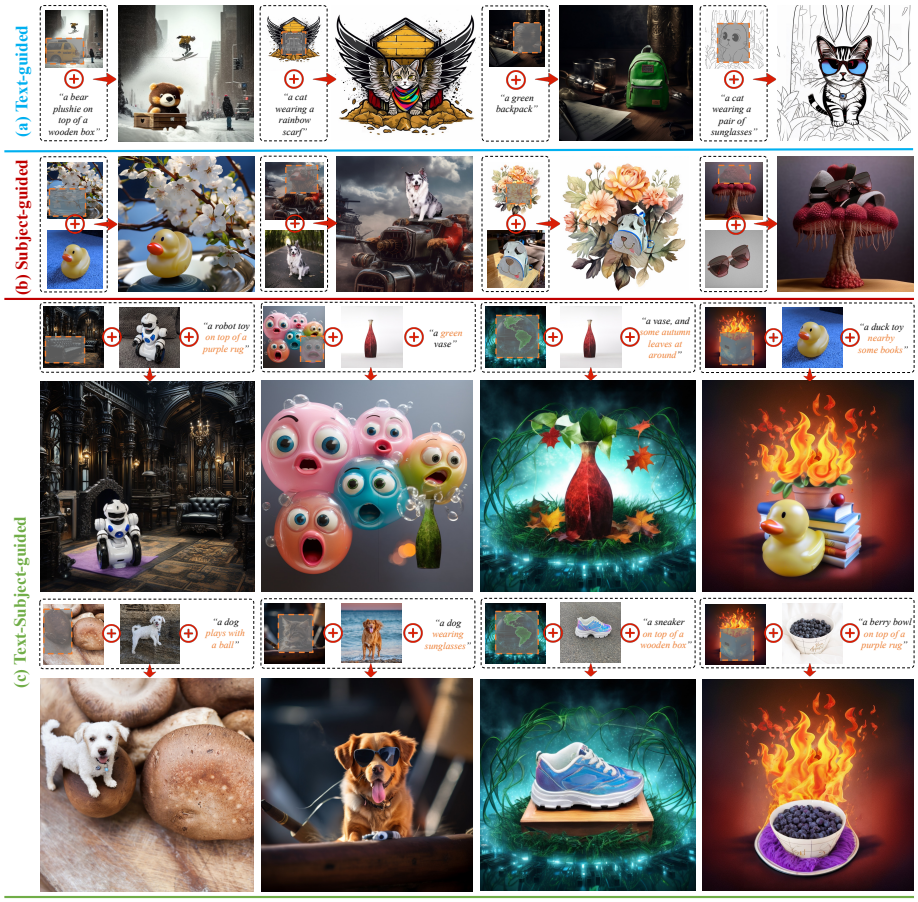


**Fig. 1: Image inpainting results of our method.** Given a quadruplet comprising a scene image, a scene mask which marked in gray, a subject image, and a text prompt, our proposed method can accurately inpaint the masked area within the scene image as specified by the scene mask. It does so according to guidance that may be (a) text-only, (b) subject-only, or (c) a combination of text and subject, all within the same model. Best viewed when zoomed in.

such methods often depend on global descriptions of the image, potentially limiting the fidelity of local semantic correspondence within the masked areas. Conversely, subject-guided image inpainting approaches [8,49,55] reconstruct the masked region using a subject image. Despite this, the prevalent use of a collage strategy [8,55] can result in issues resembling a simple copy-paste effect. Besides, both text-only and image-only guided methods lack the precision required for fine control over the reconstruction of missing regions.

To address the aforementioned drawbacks, we first present **text-subject-guided image inpainting**, a novel task that seamlessly integrates an arbitrary

customized object into the desired location within a scene image, and allows for auxiliary text prompt to achieve fine-grained control. This task is elaborately designed to process with a maximum input of quadruplet, including a scene image, a scene mask, a subject image, and a local textual description, adhering to the following distinct principles:

- The background area of the scene image should remain unchanged, focusing solely on the editing of the specific masked area.
- The subject identity and details should be preserved as much as possible.
- The inpainted content should semantically correspond to the local text.
- The reconstructed image should achieve high-quality and seamless integration within the scene.

We propose a tuning-free method, termed as **LAR-Gen**, which follows a "**L**ocate, **A**ssign, **R**efine" pipeline to achieve the above objectives and enable creative **Gen**eration. Specifically, the masked scene image is first encoded into latent space and concatenated with the noise input, along with the mask image. This stage compels the model to seamlessly inpaint the masked region while keeping the background unaltered. Then, a decoupled cross-attention mechanism is employed to effectively guide the diffusion process under the joint control of the text and the subject, ensuring that the guidance process conforms to the semantics of the local description and the coarse-grained subject reference. Finally, an auxiliary U-Net [37] termed RefineNet is introduced to supplement subject details. It encodes the noisy subject image at each sampling step and injects the detail features into the main branch at multi-scale self-attention layers, facilitating the subject detail preservation. To address the scarcity of public datasets providing subject images paired with localized textual prompts for the text-subject-guided image inpainting task, we introduce an innovative data construction strategy for assembling the necessary quadruple data. The pipeline operates by automatically categorizing [56], detecting [30], instance segmenting [23], and visual captioning [29], thereby producing region-level quadruples comprising a scene image, a scene mask, a subject image and a text prompt. With the help of the collected quadruple data and the tailored pipeline, LAR-Gen not only excels in text-subject-guided image inpainting task, but also supports text-only and image-only guided inpainting within the same framework, as illustrated in Fig. 1. This indicates that our LAR-Gen is a unified framework for image inpainting and can be employed in various scenarios. Extensive experiments demonstrate the superiority of LAR-Gen in terms of both subject identity consistency and text semantic consistency.

## 2 Related Work

**Text-to-image Generation.** In the past few years, diffusion-based image generation [5, 15, 25, 32, 40, 41] has emerged as a burgeoning research trend. This is due to the diffusion models' superior ability to generate high-quality images compared with GANs [14, 21, 22] and auto-regressive models [11, 12]. Large-scale

diffusion models have been proven to be excellent starting points for various downstream tasks, such as image inpainting [1,46,48], image super-resolution [28, 44, 50], and video generation [17, 19]. Our LAR-Gen builds upon Stable Diffusion [2, 3, 36] to fully leverage its power on generating a high-fidelity image.

**Subject-driven Image Generation.** Subject-driven image generation [18, 26, 45, 53] aims to generate images conditioned on a customized subject and a text prompt that describes the context. Existing works can be categorized in terms of the necessity of test-time tuning. One line of them requires one more image to conduct training optimization for specific subjects. Textual Inversion [13] is the first work to inverse a subject into textual representation space. It fixes the U-Net [37] backbone and only tunes the newly added embedding. DreamBooth [38] tunes the entire U-Net together with the registered embedding and introduces an auxiliary loss to prevent generative ability degradation. Custom Diffusion [24] performs parameter updating only at cross-attention layers and is able to combine multiple subjects into one image. While they excel at capturing subject detail, the time-consuming nature of the tuning process renders these methods less efficient than zero-shot approaches, which in turn limits their practical application. Another line of them encodes the subject with a visual encoder, avoiding test-time tuning. InstantBooth [39] and ELITE [47] combine both global and local visual features and inject the fine-grained local features with an extra cross-attention layer to preserve the high-level details of the subject. CustomNet [52] proposes a novel view generation method for the subject, which introduces the camera extrinsic as an extra condition to control the generated view of the subject. IP-Adapter [51] introduces a decoupled cross-attention mechanism, which achieves multi-modal prompts control by weighted mixing of two attention layer outputs. Despite the significant advancements, these methods are still at the early stage of maintaining the subject's identity. Moreover, their inability to perform localized editing constrains their applicability.

**Image Inpainting.** Given a masked scene image, image inpainting aims to inpaint the masked region guided by visual or textual prompts. While text-guided image inpainting technology [1, 46] has reached a stage of maturity, subject-driven image inpainting remains in its nascent phase. Recently many works have been proposed to address subject-driven image inpainting. Paint-by-Example [49] replaces the text embedding with global image embedding in the cross-attention layer to inject the subject information into the diffusion process. ObjectStitch [42] introduces a content adapter to mapping the image embedding to textual embedding space and applies the mask to constrain the denoising area, thereby achieving image conditional editing. AnyDoor [8] and PhD [55] apply a ControlNet [54] to control the diffusion process with a subject-scene collage. Although these approaches have achieved significant advancements in subject- or text-conditioned image inpainting, none of them supports multi-modal prompt control. This paper addresses the text-subject-guided image inpainting task, which enables the joint control of visual and textual prompts.

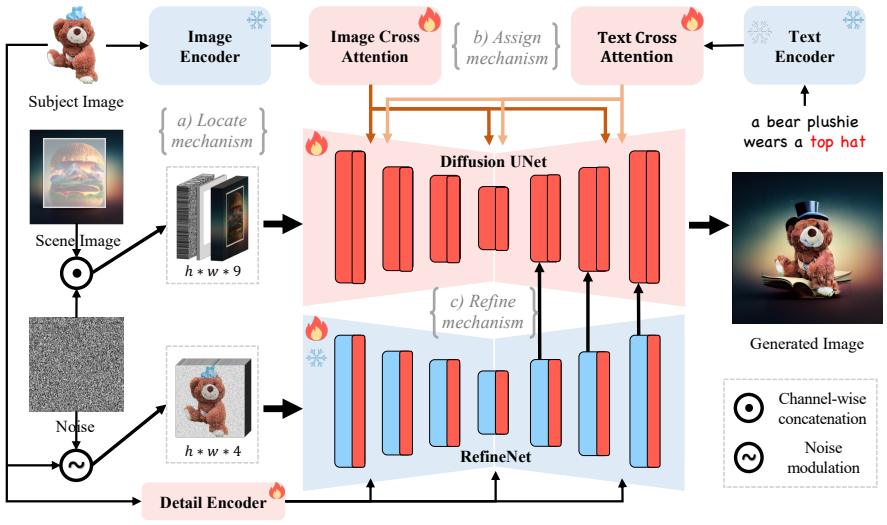


**Fig. 2: Overall pipeline of LAR-Gen.** The whole framework consists of three key mechanisms: the *Locate mechanism* concatenates the noise with masked scene image and mask to compel the model to seamlessly inpaint the masked region while keeping the background unaltered; the *Assign mechanism*, indeed a decoupled cross-attention mechansim, achieves text compatible subject guidance on denoising process; the *Refine mechanism* utilizes a RefineNet to gradually supplement the subject details, facilitating subject identity preservation. The fire icon indicates that the module parameter requires tuning, while the snow icon signifies that no tuning is needed.

## 3 Method

**Overview.** Given a scene image $x_s \in \mathbb{R}^{H\times W\times 3}$, a binarized mask $m \in \mathbb{R}^{H\times W}$, a subject image $x_{\text{obj}}$ and a text prompt $s$, LAR-Gen aims to inpaint a local region of $x_s$ specified by $m$, under the joint control of $x_{\text{obj}}$ and $s$. The inpainted content inside region should preserve the subject identity referring to $x_{\text{obj}}$, and also align with the semantic of text prompt. The overall pipeline of LAR-Gen is depicted in Fig. 2. LAR-Gen reconciles three mechanisms to achieve high-fidelity and seamless composition: **L**ocate, **A**ssign and **R**efine. We will describe each mechanism in detail next.

### 3.1 Locate Mechanism

The text-to-image diffusion U-Net [37] receives the latent code of a noisy image $z \in \mathbb{R}^{h\times w\times 4}$ and predicts the noise added to the image. To support local image editing, we concatenate the masked image with noise at channel dimension, following SD-Inpainting [1]. It first encodes the masked scene image, formulated as:

$$z_s = \text{Enc}(x_s \odot (1-m)), \tag{1}$$

| | | | | |
|---|---|---|---|---|
| **Image** |  |  |  |  |
| **Region Caption** | a large elephant with a big trunk, standing next to another elephant. | a large brown cardboard box sitting on a wooden floor. The box is wrapped in plastic, and there is a white label on it. | a large black stone monument located in a park. | a colorful bird stuffed animal, which is knitted and has a purple and pink body. The bird is wearing a sweater and is positioned in front of a white background. |
| **Global Caption** | a group of three elephants standing next to each other. The elephants are of different sizes, with one being larger than the other two, and the smallest elephant being the one on the left. | a man and a woman sitting next to each other on a couch in a living room. | a statue of a man riding a horse, prominently displayed in front of a large white building. | a playful and whimsical scene with three stuffed birds, each holding a toy gun. The birds are positioned next to each other, with one on the left, one in the middle, and one on the right. |

**Fig. 3: Comparison between global caption and regional caption.** Global captions can overlook specific subjects in complex images, while regional captions offer more precise semantic alignment with local visual content.

where Enc represents the encoder of VAE of stable diffusion. $\odot$ represents the element-wise product operation. Then we concatenate $z_s$ with $z$ and resized mask $m^* \in \mathbb{R}^{h \times w}$ to form the U-Net input:

$$\tilde{z} = [z; m^*; z_s], \tag{2}$$

where $\tilde{z} \in \mathbb{R}^{h \times w \times 9}$ and $[\cdot;\cdot]$ denotes the concatenation operation. Furthermore, we utilize a blend strategy during inference to ensure the background unaltered. It is formulated as:

$$\hat{z}^t = z_s^t \odot (1 - m^*) + z^t \odot m^* \tag{3}$$

where $z^t$ and $z_s^t$ denote the model output and the noisy latent code of scene image at sampling step $t$, respectively.

**Training Data.** The standard inpainting method trains the model with (image, prompt) pair data, where the prompt is a global description of image. However, we find that global caption may ignore the target subject or fail to describe it in detail. This leads to semantic misalignment on our training stage, and sometimes fail to generate subject. To address this, we propose a novel data process method to collect the needed quadruplets. Fig. 3 presents several examples to demonstrate the distinctions between global captions and region-specific captions. We will introduce the data collection strategy in Sec. 3.4.

### 3.2 Assign Mechanism

While the pursuit of image inpainting conditioned independently on text descriptions or reference images has garnered considerable attention in recent years, mixing the two type of conditions presents a non-trivial challenge. Inspired by IP-adapter [51], we implement a decoupled cross-attention mechanism to enable

compatible multi-modal prompt control. Specifically, given the subject image $x_{\text{obj}}$, we adopt a projection module $f$ attached to CLIP image encoder to encode the subject image to a sequence of features $c_i$ with length $N$. It is formulated as:

$$c_i = f(\text{CLIP}(x_{\text{obj}})), \tag{4}$$

In our experiments, we adopt IP-Adapter-Plus [51] as the projection module $f$ and freeze both CLIP encoder and the projection module during training. The feature sequence length $N$ is 16. Subsequently, two extra parameters $\mathbf{W}_k^{'}$, $\mathbf{W}_v^{'}$ are introduced to insert image feature to cross-attention layer:

$$\mathbf{K}^{'} = c_i\mathbf{W}_k^{'}, \mathbf{V}^{'} = c_i\mathbf{W}_v^{'}, \tag{5}$$

where $\mathbf{K}^{'}, \mathbf{V}^{'}$ are the key and value matrices of the image-conditioned attention operation. At each cross-attention layer, we first perform attention on image and text separately, which formulated as:

$$\mathbf{Z} = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}})\mathbf{V}, \tag{6}$$

$$\mathbf{Z}^{'} = \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}^{'})^{\top}}{\sqrt{d}})\mathbf{V}^{'}, \tag{7}$$

where $\mathbf{Z}$ and $\mathbf{Z}^{'}$ represent the text-conditioned and image-conditioned attention outputs, respectively. $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the query, key, value matrices of the text-conditioned attention operation. The final cross attention output is a weighted sum of $\mathbf{Z}$ and $\mathbf{Z}^{'}$:

$$\mathbf{Z}^{\text{out}} = \mathbf{Z} + \beta\mathbf{Z}^{'}, \tag{8}$$

where $\beta$ is a hyper-parameter used to modulate the image control strength.

### 3.3 Refine Mechanism

To achieve text-subject-compatible guidance, the hyper-parameter $\beta$ in the decoupled cross-attention layer is generally set to a small value to ensure text consistency. However, this compromises image control and may result in the loss of subject details. To counteract this, we introduce an auxiliary U-Net, termed RefineNet, which enhances the object's details even when $\beta$ is small. The RefineNet performs denoising process on the noisy subject image, which is modulated with the same level of Gaussian noise. It utilizes a detail encoder, comprised of a CLIP image encoder and an MLP, to encode the subject's patch features:

$$c_i^{'} = \text{MLP}(\text{CLIP}(x_{\text{obj}})) \tag{9}$$

During training, the CLIP encoder is frozen and only the parameters of MLP are updated. The feature $c_i^{'}$ is then used in the cross-attention layers of RefineNet to guide the diffusion process. Hence the fine-grained details of subject can be captured by RefineNet at each sampling step $t$. In each self-attention layer of

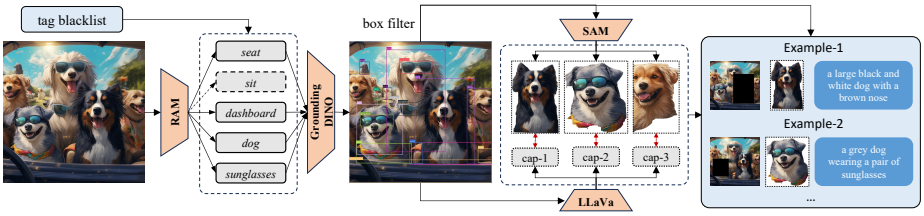


**Fig. 4: Data construction pipeline.** The pipeline involves a sequence of automated processes including categorization, detection, segmentation, and captioning. As a result, it creates region-level quadruplets that include a scene image, a scene mask, a subject image, and a text prompt.

the RefineNet's decoder, the input features that represent the subject details are stored. These features are then concatenated with the corresponding layer's features of the main U-Net, thereby encouraging the model to reconstruct the subject by leveraging both contextual information and detailed subject features. The self-attention operation in the main U-Net is formulated as:

$$\mathbf{O}_s = \text{Softmax}(\frac{\mathbf{Q}_s\mathbf{K}_s^\top}{\sqrt{d}})\mathbf{V}_s, \tag{10}$$

where $\mathbf{Q}_s$, $\mathbf{K}_s$ and $\mathbf{V}_s$ represent the query, key and value matrices respectively. Given the context feature $c_{\text{ctx}}$ and the subject detail feature $c_{\text{obj}}$, the $\mathbf{Q}_s$, $\mathbf{K}_s$ and $\mathbf{V}_s$ can be calculated as:

$$\mathbf{Q}_s = c_{\text{ctx}}\mathbf{W}_{qs}, \mathbf{K}_s = [c_{\text{ctx}}; c_{\text{obj}}]\mathbf{W}_{ks}, \mathbf{V}_s = [c_{\text{ctx}}; c_{\text{obj}}]\mathbf{W}_{vs}, \tag{11}$$

where $\mathbf{W}_{qs}$, $\mathbf{W}_{ks}$ and $\mathbf{W}_{vs}$ are the weight matrices of the self-attention layer in the main U-Net.

### 3.4 Training and inference

**Data Collection.** To our best knowledge, there is no available public dataset that contains both regional text prompts and subject images. So we leverage the power of pretrained large models to produce necessary training data. Specifically, as depicted in Fig. 4, starting from a still image, we first employ RAM [56] to get the included tags within the image. We exclude non-entity tags such as "sky", "nature" and "skin". Grounding-DINO [30] and SAM [23] are then utilized in turn to get bounding box and segmentation mask corresponding to each tag. Objects too big or too small are also excluded. Subsequently, we crop the object region according to bounding box and send it to LLaVa [29] to produce the regional caption. By means of this process, we are able to extract quadruplets consisting of *(scene image, scene mask, subject image, text prompt)* from a vast repository of image data.

**Objectives.** LAR-Gen is trained with the mean square error loss that is a conventional choice among diffusion-family models. Given the concatenated latent

**Table 1: Quantitative results of different approaches on test set**. Here $\beta$ represents the image prompt control strength. The condition $s = \emptyset$ indicates that the text condition is omitted, and only the fidelity of the subject is considered.

| Method | CLIP-I | CLIP-T |
|---|---|---|
| SD-Inpainting | 68.68 | **30.37** |
| Paint-by-Example | 79.02 | 24.70 |
| Anydoor | 82.10 | 25.52 |
| LAR-Gen ($\beta = 0.3$) | 79.15 | **29.94** |
| LAR-Gen ($\beta = 1.0, s = \emptyset$) | **83.39** | 26.07 |

code $\tilde{z}$, the reference subject image $x_{\text{obj}}$ and the text prompt $s$, the loss function is formulated as:

$$\mathcal{L} = \mathbb{E}_{\tilde{z},t,x_{\text{obj}},s,\epsilon\in\mathcal{N}(0,\mathbf{I})}\|\epsilon - \epsilon_\theta(\tilde{z}, t, x_{\text{obj}}, s)\|_2 \tag{12}$$

**Training.** A two-stage training strategy is adopted. Specifically, during the first stage, the RefineNet component is omitted, permitting focused training of the main U-Net, augmented by the locate and assign mechanism. The hyper-parameter $\beta$ is currently set to 1. Upon transitioning to the secondary stage, $\beta$ is typically set to a constant value smaller than 1 to ensure the textual semantic alignment, while the RefineNet is integrated to supplement the object detail. At this stage, the main U-Net is frozen, with updates being confined solely to the parameters within the cross-attention layers of the RefineNet.

**Inference.** We adopt classifier-free guidance [16] during inference, a popular technique that has been proven effective in conditional image generation. Specifically, during training, both the image and text prompt are independently assigned a null value $\emptyset$ with a predetermined probability, thereby enabling the model to concurrently train on both conditional and unconditional denoising capabilities. At inference, it shifts the score estimate towards the conditional direction and away from the unconditional direction:

$$\tilde{\epsilon}_\theta(\tilde{z}, t, x_{\text{obj}}, s) = \epsilon_\theta(\tilde{z}, t, \emptyset, \emptyset) + w(\epsilon_\theta(\tilde{z}, t, x_{\text{obj}}, s) - \epsilon_\theta(\tilde{z}, t, \emptyset, \emptyset)), \tag{13}$$

where $w$ is the classifier-free guidance scale.

## 4 Experiments

### 4.1 Experimental Setup

**Evaluation Dataset**. To evaluate the capability of our method on both subject identity consistency and text semantic consistency, we construct a benchmark that contains 2,000 *(scene image, scene mask, subject image, text prompt)* samples, using 20 scene images, 10 customized objects, and 10 pre-defined text prompts. The subjects are provided by DreamBooth [38] dataset, and we manually pick 5 non-live subjects and 5 live subjects for fair comparison. For the

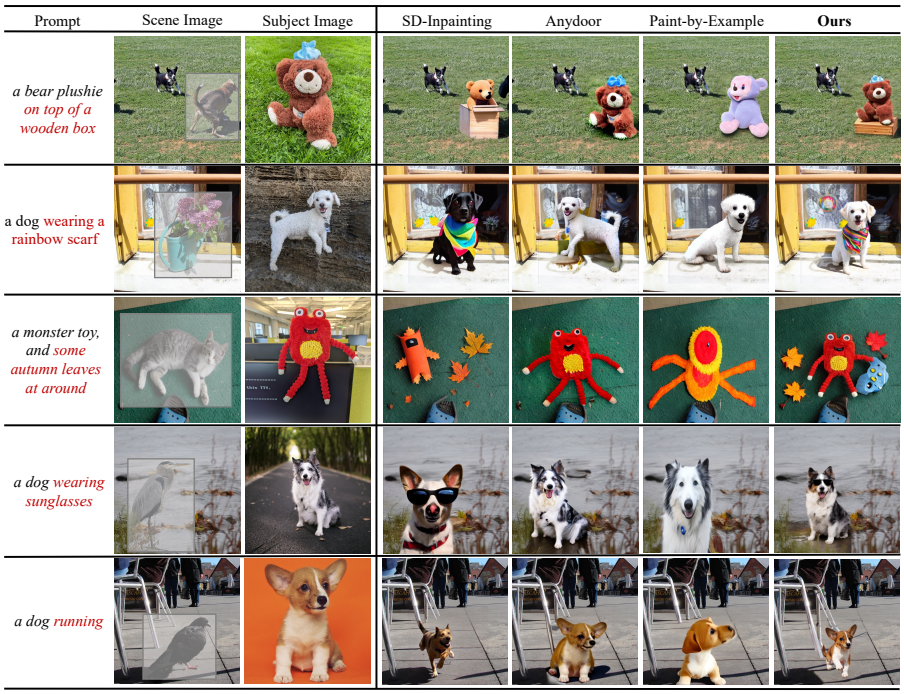


**Fig. 5: Comparison with existing alternatives.** Unlike other methods that depend solely on textual or visual prompts, LAR-Gen achieves customized image inpainting by leveraging both the subject image and text prompt.

scene images and scene masks, we randomly pick 20 images with corresponding boxes from the COCO-val dataset [27]. Then, for each scene-subject pair, we bind it with 10 pre-defined prompts to form the final test dataset.

**Evaluation Metrics**. Following Dreambooth, we evaluate the subject identity consistency by calculating the clip [35] score between the inpainted region and the background-free subject image, denoted as CLIP-I [38]. Additionally, the text semantic consistency is also evaluated by calculating the clip score between the inpainted region and text prompt, denoted as CLIP-T [38].

**Implementation Details.** We choose Stable Diffusion V1.5 [2] as the base architecture to conduct experiments, while other UNet-based architecture is also compatible with our method, *e.g.*, Stable Diffusion XL. Qualitative results for both architectures are included in the appendix. We use the zoom-in strategy introduced in Anydoor [8], which involves cropping a subimage from the scene image around the masked region to serve as the input for the diffusion model. The model is trained at a batch size of 128 for 20k steps with a learning rate of 1e-5 for both two training stages. The hyper-parameter $\beta$ is set to 1 during the first training stage, and adjusted to 0.3 for the second. The classifier-free guidance scale $w$ is set to 7.5 during inference. All experiments are conducted on A100 gpu of 80G memory.

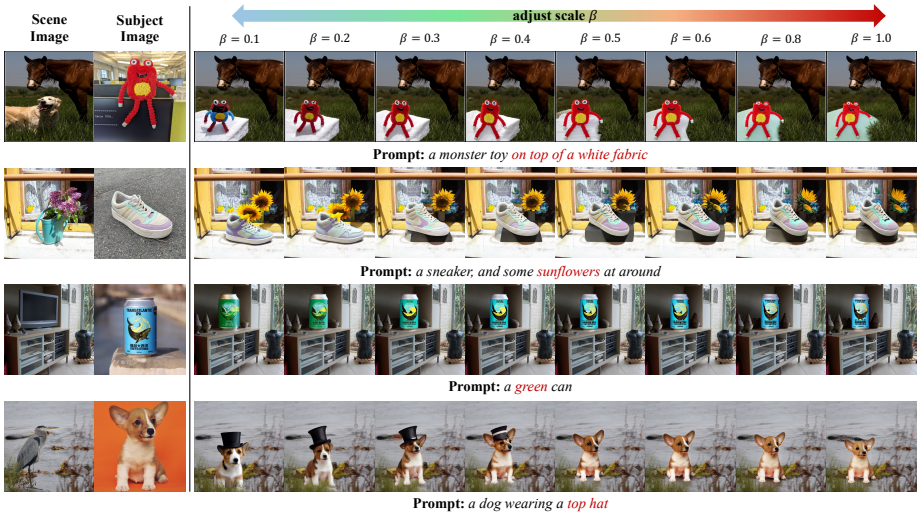


**Fig. 6: Ablation studies on varying hyper-parameter $\beta$.**

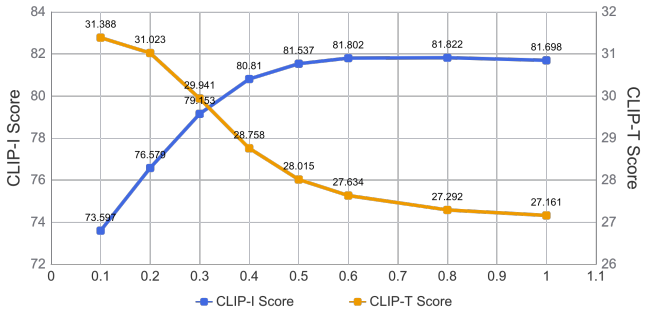


**Fig. 7: Quantitative analysis of varying the value of hyper-parameter $\beta$.** The horizontal axis represents the value of the hyper-parameter $\beta$.

### 4.2 Comparison with Existing Alternatives

We first compare our method with some existing zero-shot inpainting approaches, including SD-Inpainting [1], Paint-by-Example [49], and Anydoor [8]. Note that SD-Inpainting supports only text-based conditions, whereas Paint-by-Example and Anydoor are limited to using a reference image as the condition. Unlike them, our LAR-Gen is able to inpaint image by simultaneously leveraging a subject image and a text prompt for joint guidance. As depicted in Fig. 5, the qualitative outcomes demonstrate that our LAR-Gen not only ensures that the inpainted region meets the semantic demands of the input prompt but also accurately aligns the generated content with that of the reference image.

We present the quantitative results in Tab. 1. We observe that when $\beta$ is set to 0.3, our method achieves a CLIP-T score comparable to that of SD-Inpainting,

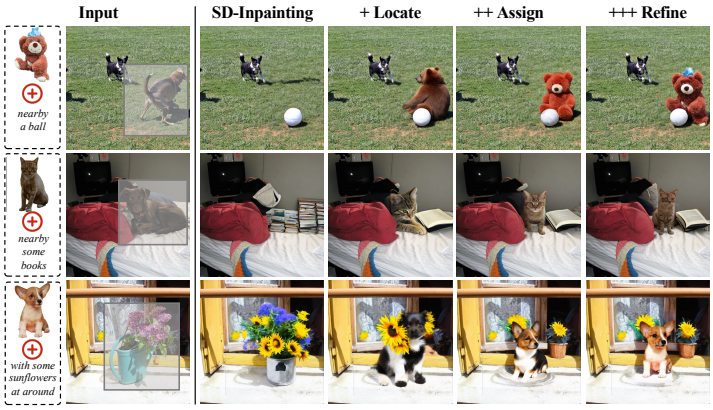


**Fig. 8: Qualitative ablation studies on each proposed components.** Starting with the standard SD-Inpainting model, we incrementally incorporate each component to evaluate its impact on performance.

greatly surpassing Paint-by-Example and Anydoor, thereby demonstrating its superior capability for maintaining text semantic consistency. Furthermore, our LAR-Gen surpasses SD-Inpainting and Paint-by-Example by 10.47% and 0.13% in terms of CLIP-I score under $\beta = 0.3$. However, it trails Anydoor by a margin of 2.94% in terms of CLIP-I scores, due to the fact that the generated object is supposed to be deformed to align the text prompt. We also evaluate the performance of LAR-Gen under the setting of $\beta = 1.0$ and no text condition. LAR-Gen achieves the highest CLIP-I score among all approaches, surpassing Anydoor by 1.29%. This result demonstrate its superior capability for preserving the subject identity consistency.

### 4.3 Ablation Studies

**Image Control Strength $\beta$.** We conduct ablation experiments to clarify the contribution of hyper-parameter $\beta$. To this end, the first stage model is trained as described previously. Subsequently, in the second stage of training, we adjust $\beta$ from 0.1 to 1.0, which yields a series of second-stage models with varying control strengths. We present the visualization results and quantitative results in Fig. 6 and Fig. 7, respectively. With the increase of scale $\beta$, we observe a greater retention of object details, albeit at the cost of reduced fidelity to the textual descriptions. From Fig. 7 we observe that the CLIP-I score progressively increases with the augmentation of $\beta$, while the CLIP-T score correspondingly diminishes as $\beta$ escalates. This observation corroborates the conclusion that $\beta$ plays a pivotal role in balancing the control between textual semantics and subject identity.

**Effectiveness of Each Component.** We conduct experiments to verify the effectiveness of core components proposed in this paper. LAR-Gen consists of

**Table 2: Quantitative ablation studies on core components of LAR-Gen**. The Locate mechanism here refers to the use of region-specific captions instead of global captions during training. We assess the performance of Assign mechanism and Refine mechanism by setting $\beta = 0.3$.

| Module | CLIP-I | CLIP-T |
|---|---|---|
| SD-Inpainting | 68.68 | 30.37 |
| + Locate mechanism | 68.60 | 31.01 |
| ++ Assign mechanism | 75.00 | 30.53 |
| +++ Refine mechanism | 79.15 | 29.94 |

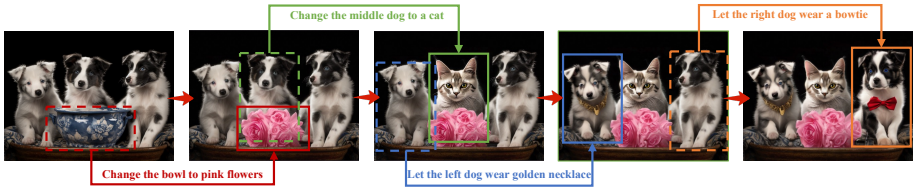


**Fig. 9: Multi-turn image editing results of our method.** Thanks to the tailored framework, our model is capable of performing inpainting with various types of guidance. This enables applications that require multi-turn, multi-type inpainting.

three key mechanisms: Locate, Assign and Refine. Starting with the vanilla SD-Inpainting model, we incrementally incorporate each mechanism and assess its impact on text and subject identity consistency by calculating the CLIP-I and CLIP-T scores. In this context, the Locate mechanism specifically refers to the use of regional captions that are associated with particular image regions during training, as opposed to relying on global captions. This is because the SD-Inpainting [1] model already includes the operation of concatenating masked scene images. As shown in Tab. 2, this approach yields a 0.65% improvement in the CLIP-T score over the SD-Inpainting model, suggesting that region-specific captions enhance text semantic consistency due to their more accurate descriptions compared to global captions. Upon this, the addition of the Assign mechanism results in a significant improvement in the CLIP-I score (*i.e.*, 68.60% $\rightarrow$ 75.00%) with only a marginal decrease in the CLIP-T score. This suggests that the Assign mechanism can effectively incorporate the condition from the reference image without substantially compromising text semantic consistency. Lastly, the inclusion of the Refine mechanism leads to a 4.15% increase in CLIP-I score, underscoring its exceptional efficacy in enhancing the high-frequency details of the subject.

### 4.4 Unified Image Inpainting Framework

Although LAR-Gen is tailored for text-subject-guided image inpainting, it could serve as a unified image inpainting framework that supports text-only and image-only guided inpainting as well. On one hand, to achieve text-only guided in-

**Fig. 10: Qualitative results on virtual try-on.** The cases shown are selected from the test set of VITON-HD.

painting, we set $\beta = 0$ and remove RefineNet, thereby our model degenerates to standard SD-Inpainting architecture enhanced by regional caption and performs denoising under text-only guidance. On the other hand, Tab. 1 indicates that by setting the text prompt to $\emptyset$, our model achieves the best fidelity compared to other tuning-free subject-only guided inpainting methods, *i.e.*, Paint-by-Example and Anydoor. It demonstrates the effectiveness of our method on subject-only guided inpainting. Benefiting from the unified framework, our LAR-Gen is capable of performing multi-turn, multi-type inpainting, as depicted in Fig. 9, and show immense potential in a variety of applications, such as virtual try-on, object reshaping, instruct-based image editing, and so on. We present several examples in Fig. 10 to illustrate its capabilities on virtual try-on, which randomly selected from the VITON-HD [9] test set. Note that our model has not been carefully optimized with try-on data. It requires only a single clothing image as input and does not need human pose or any other auxiliary information that generally required on traditional try-on methods. We observe that LAR-Gen achieves seamless cloth composition while preserving fine details of the clothing, such as color, texture, and pattern.

## 5 Discussion and Conclusion

This paper proposes LAR-Gen, a diffusion-based image inpainting framework that could generate high-fidelity image with the joint guidance of text and subject image. It follows a coarse-to-fine pipeline, which first generates the target subject by introducing the image embedding into the decoupled cross-attention mechanism, then refines the subject details with the proposed RefineNet, avoiding the copy-paste issue caused by collage-based strategy. A hyper-parameter $\beta$ is introduced to balance the strength of image and text condition, thereby enabling fine-grained tuning to seek the optimal setting of preserving both subject identity and text semantic consistency.

Nevertheless, our method still suffers from two limitations. Firstly, the reliance on a single subject image as the conditional input may lead to inaccuracies in subject deformation. This is primarily due to the inherent ambiguity associated with generating deformations that have not been encountered previously, rendering the task inherently ill-posed. Secondly, the denoising process is influenced by multiple conditional factors, which may lead to the neglect of certain conditions, particularly when they conflict with others.

## References


1. AI, R.: Stable Diffusion Inpainting Model Card, https://huggingface.co/runwayml/stable-diffusion-inpainting (2022) 1, 4, 5, 11, 13
2. AI, R.: Stable Diffusion v1.5 Model Card, https://huggingface.co/runwayml/stable-diffusion-v1-5 (2022) 4, 10, 19
3. AI, S.: Stable Diffusion v2-1 Model Card, https://huggingface.co/stabilityai/stable-diffusion-2-1 (2022) 4
4. AI, S.: Stable Diffusion XL Model Card, https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0 (2022) 19
5. Avrahami, O., Lischinski, D., Fried, O.: Blended Diffusion for Text-driven Editing of Natural Images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18208–18218 (2022) 3
6. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning To Follow Image Editing Instructions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18392–18402 (2023) 1
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-$\alpha$: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv preprint arXiv:2310.00426 (2023) 1
8. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: AnyDoor: Zero-shot Object-level Image Customization. arXiv preprint arXiv:2307.09481 (2023) 2, 4, 10, 11
9. Choi, S., Park, S., Lee, M., Choo, J.: VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14131–14140 (2021) 14
10. Cloud, A.: Tongyi Wanxiang, https://tongyi.aliyun.com/wanxiang/ (2023) 1
11. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., Tang, J.: CogView: Mastering text-to-image generation via Transformers. In: Adv. Neural Inform. Process. Syst. (2021) 3
12. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for high-resolution image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12868–12878 (2020) 3
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In: Int. Conf. Learn. Represent. (2022) 4
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. arXiv preprint arXiv:1406.2661 (2014) 3
15. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Adv. Neural Inform. Process. Syst. Curran Associates, Inc. (2020) 1, 3
16. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance. In: Adv. Neural Inform. Process. Syst. (2021) 1, 9
17. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. arXiv preprint arXiv:2311.17117 (2023) 4
18. Hua, M., Liu, J., Ding, F., Liu, W., Wu, J., He, Q.: DreamTuner: Single Image is Enough for Subject-Driven Generation. arXiv preprint arXiv:2312.13691 (2023) 4
19. Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: VideoBooth: Diffusion-based Video Generation with Image Prompts. arXiv preprint arXiv:2312.00777 (2023) 4

20. Jiang, Z., Mao, C., Pan, Y., Han, Z., Zhang, J.: SCEdit: Efficient and Controllable Image Diffusion Generation via Skip Connection Editing. arXiv preprint arXiv:2312.11392 (2023) 1
21. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: Int. Conf. Learn. Represent. (2018) 3
22. Karras, T., Laine, S., Aila, T.: A Style-Based Generator Architecture for Generative Adversarial Networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4401–4410 (2019) 3
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023) 3, 8
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-Concept Customization of Text-to-Image Diffusion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1931–1941 (2022) 4
25. Li, D., Li, J., Hoi, S.C.H.: BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In: Adv. Neural Inform. Process. Syst. (2023) 3
26. Li, T., Ku, M., Wei, C., Chen, W.: DreamEdit: Subject-driven Image Editing. arXiv preprint arXiv:2306.12624 (2023) 4
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755 (2014) 10, 19
28. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv preprint arXiv:2308.15070 (2023) 4
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Adv. Neural Inform. Process. Syst. (2023) 3, 8
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint arXiv:2303.05499 (2023) 3, 8
31. Midjourney: Midjourney, https://www.midjourney.com/ (2023) 1
32. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv preprint arXiv:2112.10741 (2022) 1, 3
33. OpenAI: DALL·E 2, https://openai.com/dall-e-2 (2022) 1
34. OpenAI: DALL·E 3, https://openai.com/dall-e-3 (2023) 1
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020 (2021) 10
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022) 1, 4
37. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Med. Image Comput. Computer-Assisted Interv. (2015) 3, 4, 5
38. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven

Generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22500–22510 (2023) 4, 9, 10, 19
39. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. arXiv preprint arXiv:2304.03411 (2023) 4
40. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: Int. Conf. Learn. Represent. (2021) 3
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: Int. Conf. Learn. Represent. (2021) 3
42. Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Aliaga, D.: ObjectStitch: Generative Object Compositing. arXiv preprint arXiv:2212.00932 (2023) 4
43. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-Robust Large Mask Inpainting With Fourier Convolutions. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 2149–2159 (2022) 1
44. Wang, J., Yue, Z., Zhou, S., Chan, K.C.K., Loy, C.C.: Exploiting Diffusion Prior for Real-World Image Super-Resolution. arXiv preprint arXiv:2305.07015 (2023) 4
45. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: InstantID: Zero-shot Identity-Preserving Generation in Seconds. arXiv preprint arXiv:2401.07519 (2024) 4
46. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., Baldridge, J., Norouzi, M., Anderson, P., Chan, W.: Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18359–18369 (2023) 1, 4
47. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. arXiv preprint arXiv:2302.13848 (2023) 4
48. Xie, S., Zhao, Y., Xiao, Z., Chan, K.C.K., Li, Y., Xu, Y., Zhang, K., Hou, T.: DreamInpainter: Text-Guided Subject-Driven Image Inpainting with Diffusion Models. arXiv preprint arXiv:2312.03771 (2023) 4
49. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by Example: Exemplar-based Image Editing with Diffusion Models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18381–18391 (2022) 2, 4, 11
50. Yang, T., Ren, P., Xie, X., Zhang, L.: Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. arXiv preprint arXiv:2308.14469 (2023) 4
51. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv preprint arXiv:2308.06721 (2023) 4, 6, 7
52. Yuan, Z., Cao, M., Wang, X., Qi, Z., Yuan, C., Shan, Y.: CustomNet: Zero-shot Object Customization with Variable-Viewpoints in Text-to-Image Diffusion Models. arXiv preprint arXiv:2310.19784 (2023) 4
53. Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W., Niu, L.: ControlCom: Controllable Image Composition using Diffusion Model. arXiv preprint arXiv:2308.10040 (2023) 4
54. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: Int. Conf. Comput. Vis. pp. 3836–3847 (2023) 1, 4

55. Zhang, X., Guo, J., Yoo, P., Matsuo, Y., Iwasawa, Y.: Paste, Inpaint and Harmonize via Denoising: Subject-Driven Image Editing with Pre-Trained Diffusion Model. arXiv preprint arXiv:2306.07596 (2023) 2, 4
56. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., Guo, Y., Zhang, L.: Recognize Anything: A Strong Image Tagging Model. arXiv preprint arXiv:2306.03514 (2023) 3, 8
57. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In: Int. Conf. Learn. Represent. (2021) 1
58. Zhu, M., He, D., Li, X., Li, C., Li, F., Liu, X., Ding, E., Zhang, Z.: Image Inpainting by End-to-End Cascaded Refinement with Mask Awareness. IEEE Trans. Image Process. pp. 4855–4866 (2023) 1

## A Training and Evaluation Data

In this section, we showcase some training samples that are generated using our proposed data construction strategy. This strategy can create quadruplet data consisting of a scene image, a scene mask, a subject image, and a text prompt from a vast repository of image data. Fig. 11 displays examples of these generated quadruplets. The caption generated for the region accurately details the subject depicted in the subject image, demonstrating its effectiveness in terms of regional semantic alignment.

Additionally, we present the scene and subject images used during the evaluation stage, as shown in Fig. 12. The 20 scene images are randomly selected from COCO-val [27] dataset and the 10 subjects are manually selected from DreamBooth [38] dataset, containing 5 non-live subjects and 5 live subjects. The textual descriptions are also presented in Tab. 3. Following DreamBooth, we design different prompts for non-live and live subjects. Each subject is associated with 10 prompts.

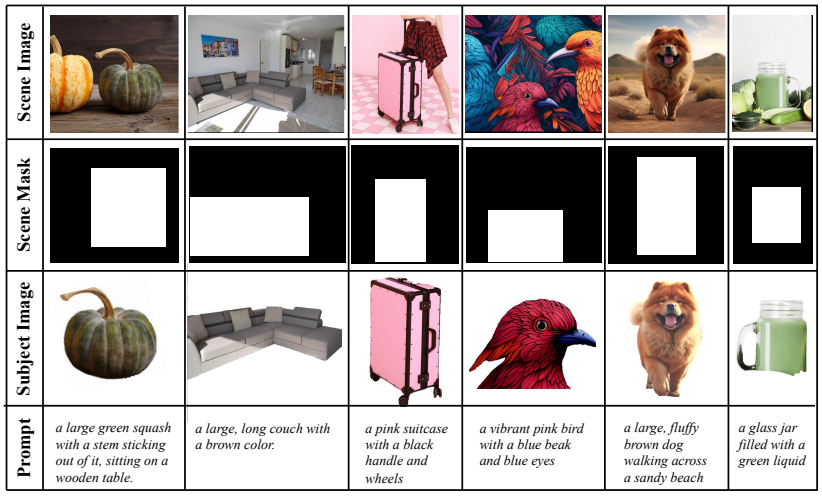


**Fig. 11: Visualization of some training samples.** Note that these are produced by the proposed data construction strategy.

## B Qualitative Results of Different Backbones

To demonstrate the generalizability of our method, we conduct experiments on Stable Diffusion V1.5 [2] and Stable Diffusion XL [4], and present the qualitative results in Fig. 13 and Fig. 14, respectively. The observations indicate that our method performs well on both versions of Stable Diffusion, thereby confirming its effectiveness across different model architectures.


**Fig. 12: Visualization of scene images and subject images for evaluation.** The 20 scene images are randomly selected from COCO-val dataset and the 10 subjects are manually selected from DreamBooth dataset, containing 5 non-live subjects and 5 live subjects.

**Table 3: Text prompt list for quantitative evaluation.** At inference time, the placeholder S* is replaced with the category label of the specific subject.

| Text prompts for non-live subjects | Text prompts for live subjects |
|---|---|
| "a S* on top of a white fabric" | "a S* running" |
| "a S* on top of a purple rug" | "a S* on top of a purple rug" |
| "a S* nearby some books" | "a S* nearby some books" |
| "a S* on top of a wooden box" | "a S* wearing a bowtie" |
| "a red S*" | "a S* wearing a top hat" |
| "a green S*" | "a S* plays with a ball" |
| "a S*, and some sunflowers at around" | "a S*, and some sunflowers at around" |
| "a S*, and some autumn leaves at around" | "a S*, and some autumn leaves at around" |
| "a S* nearby a ball" | "a S* wearing sunglasses" |
| "a S* in front of a cube-shaped metal" | "a S* wearing a rainbow scarf" |

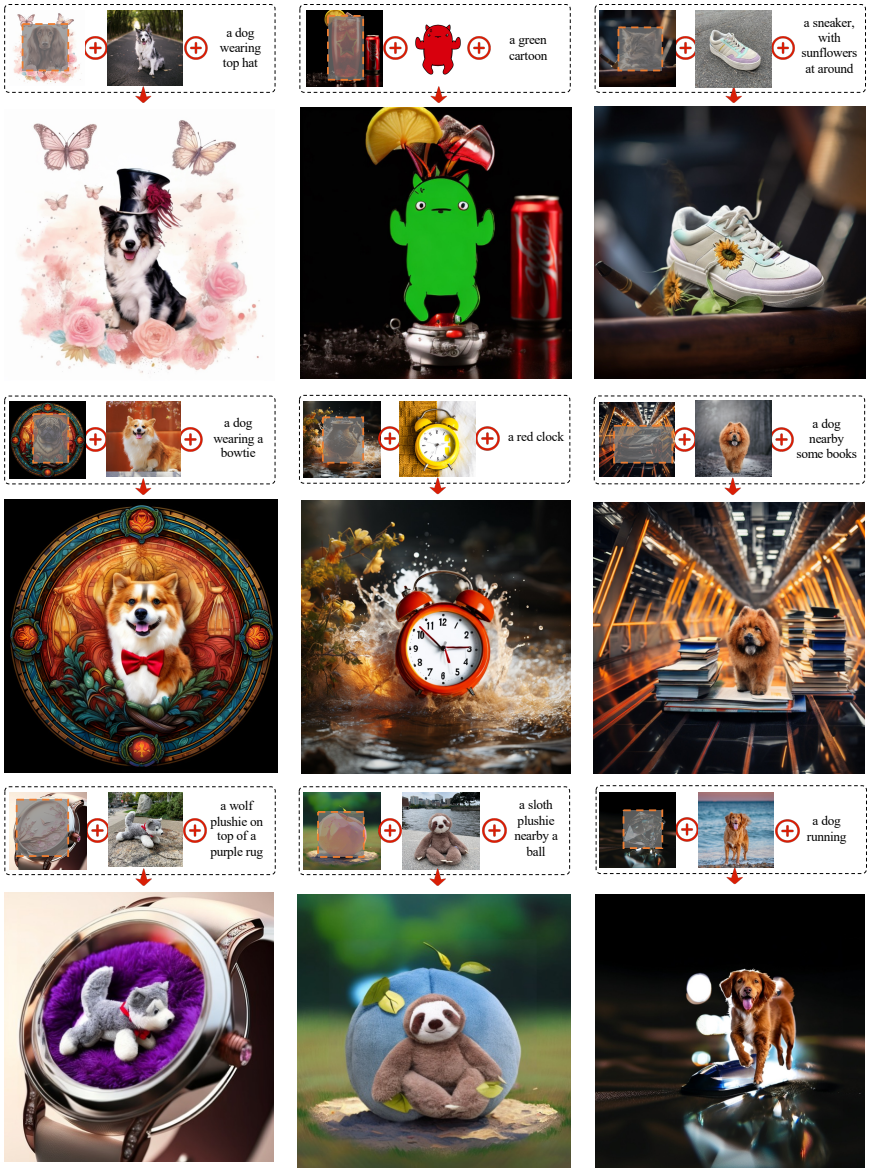


**Fig. 13: Qualitative results of Stable Diffusion V1.5 backbone.**

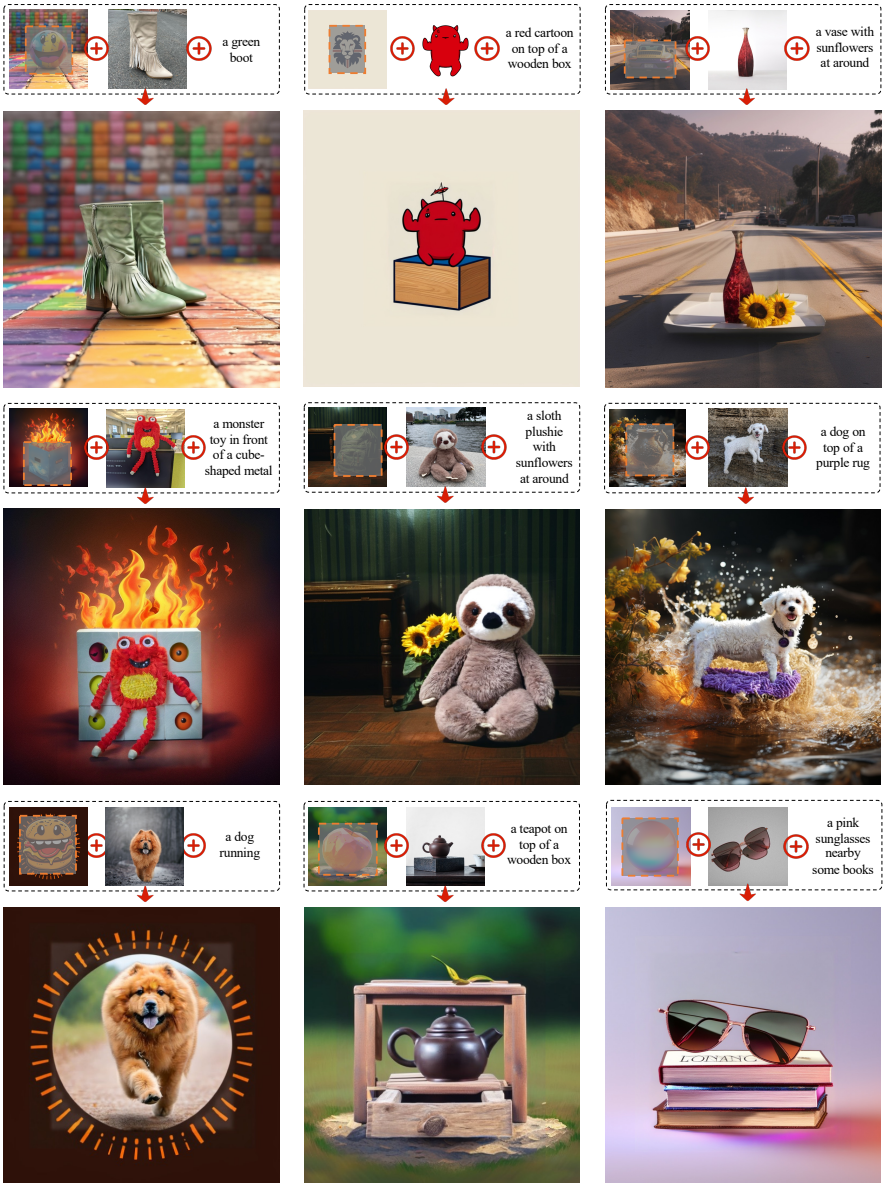


**Fig. 14: Qualitative results of Stable Diffusion XL backbone.**