FreeCompose: Generic Zero-Shot Image Composition with Diffusion Prior

Zhekai Chen[1*], Wen Wang[1,2*], Zhen Yang[1], Zeqing Yuan[1], Hao Chen[1], and Chunhua Shen[1,2]

[1] Zhejiang University, China [2] Ant Group

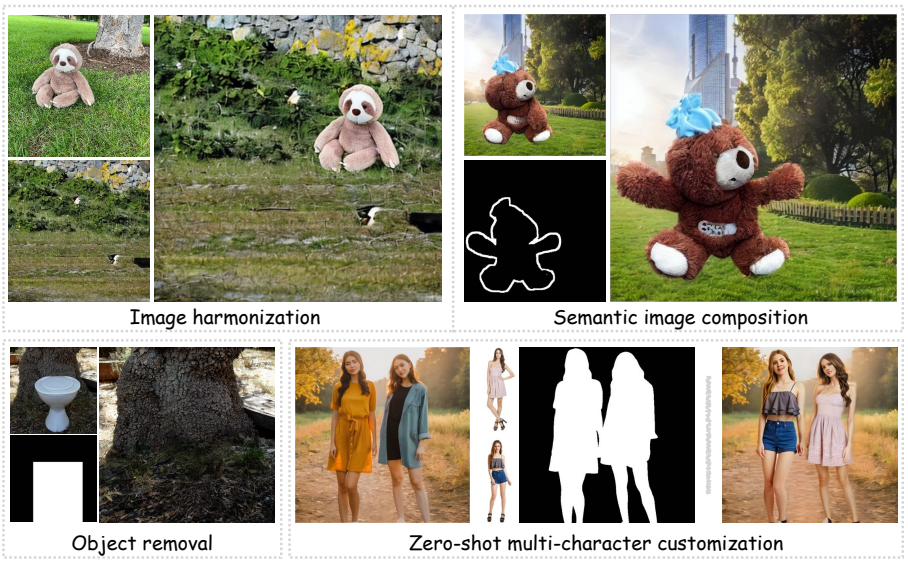

Fig. 1: FreeCompose harnesses the generative prior of pre-trained diffusion models to achieve versatile image composition, such as appearance editing (image harmonization) and semantic editing (semantic image composition). Furthermore, it can be extended to various downstream applications, including object removal and multi-character customization.

Abstract. We offer a novel approach to image composition, which integrates multiple input images into a single, coherent image. Rather than concentrating on specific use cases such as appearance editing (image harmonization) or semantic editing (semantic image composition), we showcase the potential of utilizing the powerful generative prior inherent in large-scale pre-trained diffusion models to accomplish generic image composition applicable to both scenarios. We observe that the pre-trained diffusion models automatically identify simple copy-paste boundary areas as low-density regions during denoising. Building on this insight, we propose to optimize the composed image towards high-density regions

*Equal contribution. HC and CS are the corresponding authors. Part of this work was done when WW was an intern at Ant Group.

guided by the diffusion prior. In addition, we introduce a novel mask-guided loss to further enable flexible semantic image composition. Extensive experiments validate the superiority of our approach in achieving generic zero-shot image composition. Additionally, our approach shows promising potential in various tasks, such as object removal and multi-concept customization.
Project webpage: `https://github.com/aim-uofa/FreeCompose`

Keywords: Image composition · Zero-shot · Diffusion prior

1 Introduction

Image Composition is a fundamental task in computer vision [46, 48, 55], which aims to fuse the foreground object from one image with the background of another image to generate a smooth natural image. It has a wide range of applications in many fields, such as image restoration, art design, game development, virtual reality, and so on.

For this reason, a large amount of research has been conducted on image composition [8, 46, 48, 55]. Considering only how the object is composed with the background, image composition can be broadly categorized as image harmonization [48, 55] and semantic image composition [4, 51], depending on whether there is a change in the semantic structure of the composite image. The former modifies only the statistical information of the local area after pasting the foreground pixels into the background image, to obtain an image with a smooth transition between the front and background. In contrast, the latter fine-tunes the structure of the image according to the global image context and semantically blends the foreground and background.

As deep learning [24] gains its popularity, mainstream solutions for image composition adopt the learning-based pipeline [8, 48]. They require model training on data triplet of foreground, background, and composite images to achieve image combination. However, due to the difficulty in obtaining the triplets, these models can only be trained on a limited amount of training data with a specific data distribution, making it difficult to generalize to various scenarios in real-world applications.

In contrast, recent text-to-image diffusion models [35, 37, 39] have achieved large-scale pre-training using simple graphical data pairs, demonstrating strong generalization over open-world data distributions. Inspired by this, we attempt to utilize the image prior of the pre-trained diffusion model to realize generic image composition, in zero shot. Our key assumption is that the pre-trained diffusion model can accurately predict the noise component in natural images, while inaccurately for unnatural image regions that deviate from the pre-training data distribution. Based on this, we can localize the unnatural regions in a composite image after simply copying and pasting.

To validate this hypothesis, we conduct preliminary explorations on composite images, as shown in Figure 2. Based on the above observations, we propose

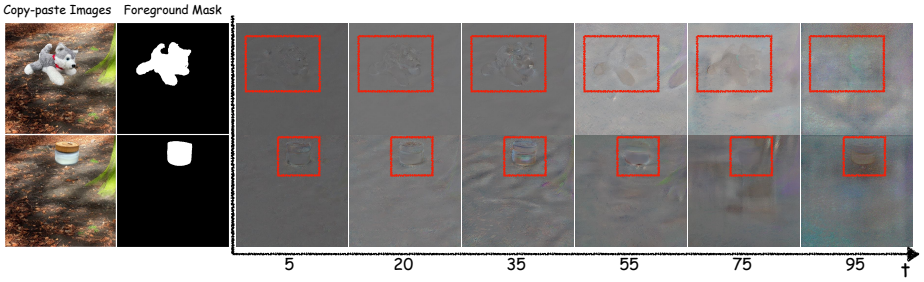

Fig. 2: Observations on the diffusion prior. The images on the left, denoted as copy-paste images, are obtained by simply pasting the foreground object to the background image. The frozen diffusion model takes the noisy copy-paste images from varying diffusion forward steps as input, and predicts the gradient to update the images (visualized on the right). Low-density regions with larger gradient updates are highlighted by red boxes. The low-density regions are highly consistent with the inharmonious regions caused by naive copy-paste.

FreeCompose, which optimizes the pixels in the image such that it can be consistent with the image prior of the pre-trained diffusion model.

In our method, we aim to use the prior of the diffusion model to combine the object with the background without having to train the diffusion model itself (referred to as Training-free in this field). We propose a generic pipeline for composition that consists of three phases: object removal, image harmonization, and semantic image composition. Unlike current works [45,51] that rely on task-specific training for image harmonization or semantic image composition, our FreeCompose can directly utilize a pre-trained diffusion model and achieve composition in zero-shot. During the object removal phase, our pipeline eliminates the foreground in the original image by manipulating the K, V values of the diffusion UNet's self-attention layer. In the image harmonization phase, the new object is combined with the background to create a harmonious scene. If additional conditions for semantic image composition are provided, the composition is guided by the difference between the conditions, while preserving the object's identity through an additional replacement of the K, V in the self-attention.

Based on these phases and techniques, FreeCompose can be effectively used for various tasks with promising results. These tasks include basic object removal, image harmonization, and semantic image composition. Moreover, FreeCompose demonstrates the ability to stylize objects by utilizing prompts during the image harmonization phase. Additionally, when combined with existing works, it can be applied to a wide range of tasks, such as multi-character customization.

To summarize, our contributions are listed as follows.

- Our findings indicate that the diffusion prior can automatically identify and focus on regions in the composite image that appear unnatural.
- Developing from the vanilla DDS loss, we explore and prove the possibility of additional designs for specific tasks including mask-guided loss and

operations on K, V embeddings. These enhancements expand the range of applications for this loss format.
- FreeCompose achieves competitive results on both image harmonization and semantic image composition. Moreover, it facilitates broad applications including object removal and multi-character customization.
- In contrast to existing methods that train separate models for individual image composition problems, the diffusion prior that we use offers a generalized natural image prior that can effectively perform both image harmonization and semantic image composition in a zero-shot manner.

2 Related Work

Image Harmonization Image harmonization aims to generate a realistic combination of foreground and background contents from different images. It focuses on adjusting low-level appearances, like the global and local color distribution change caused by light and shadows, while maintaining the content structure unchanged. Early works on image harmonization [6, 32, 36, 43, 46] rely on handcrafted priors on color [32], gradient [46], or both [43]. With the advance of deep learning [24], recent methods [5, 7, 8, 10, 19, 25, 41, 48, 55] explore learning-based methods for image harmonization. For example, Zhu *et al.* [55] train a discriminative model to judge the realism of a composited image, and leverage the model to guide the appearance adjustment of a composed image. Tsai *et al.* [48] propose the first end-to-end network for image composition. Subsequently, DoveNet [8] leverages a domain verification discriminator to migrate the domain gap between the foreground and background images. Recently, Tan *et al.* [45] proposed a new end-to-end net named DocuNet by leveraging the channels of images and achieved excellent success. While effective, these image harmonization models are trained on domain-specific datasets, and struggle to generalize to open-world images. By contrast, we leverage the natural image prior preserved in large-scale pre-trained diffusion models for zero-shot image harmonization in the wild. Chen *et al.* [16] also attempted to use diffusion model as a base model for harmonization by a method called Diff-harmonization composed of inversion and re-denoising, but limited to harmonization.

Image Editing Text editing is a broad area that encompasses many research topics, including image-to-image translation [18, 22, 29, 54, 56], inpainting [9, 17, 23, 26, 27, 31, 52], text-driven editing [2, 12, 30, 47, 50], etc. We refer readers to [16, 53] for more comprehensive review. Here we focus on the image inpainting task. Traditional image inpainting takes the masked image as input, and predicts the masked pixels from the image context. For example, LaMa [44] enlarges the receptive fields from the perspective of both modeling and losses, thus achieving inpainting in large masks and complex scenarios. Recently, benefiting from large-scale pre-trained text-to-image generative models [35, 37], researchers explore additional text input to guide the inpainting process [1, 3]. For example, Blended Latent Diffusion [1] proposes to smoothly blend the latent of the foreground region and the background areas to achieve text-guided

inpainting. Another line of work [4, 51] inpaints the masked image with an example image, which is also known as semantic image composition [51]. Different from image harmonization which only alters low-level statistics, semantic image composition semantically transfers the foreground object (often with structural changes) during composition. A representative work Paint-by-Example [51] finetunes the pre-trained Stable Diffusion model to take additional exemplar images as input for inpainting. AnyDoor [4] improves the semantic image composition pipeline to preserve the texture details in exemplar images and leverage the multi-view information in video datasets for effective training.

Diffusion Models Diffusion models [13, 42] have emerged as powerful generative models for images. Large-scale pre-trained diffusion models, like DALLE-2 [35], Imagen [39], Stable Diffusion [37], and SDXL [33], demonstrate unprecedented text-to-image generation capacities in terms of both realism and diversity. Motivated by the success of diffusion models, attempts have been made to leverage pre-trained image diffusion models as the prior for other generative tasks [21, 34, 49]. Considering the data scarcity of 3D assets, DreamFusion [34] uses Imagen [39] as a generative prior, and proposes a novel Score Distillation Sampling (SDS) loss for optimizing the implicit representation of a 3D object. Subsequently, ProlificDreamer [49] models the parameters of 3D assets as a random variable and proposes the variational score distillation to alleviate the oversaturation and over-smoothness in DreamFusion. Different from these works that focus on text-to-3D generation, DDS [11] tackles the task of text-guided image editing, and identifies the editing region by referencing the original image and its corresponding prompt. In this work, we also leverage diffusion models as the generative prior (diffusion prior). Our key observation is that diffusion prior helps locate unnatural areas in simple copy-paste image composition. Based on this, a masked guided loss is proposed to enable generic smooth image composition.

3 Preliminaries

3.1 DDS Loss

The Delta Denoising Score (DDS) [11] is developed froma modification of the diffusion loss and Score Distillation Sampling [38] for image editing. Given an input image I, the diffusion model encodes it into a latent variable $\mathbf{z}$. Using a prompt P for the to generation of a text embedding y, a timestep t is randomly chosen from a uniform distribution $\mathcal{U}(0, 1)$, and noise ϵ is sampled from a normal distribution $\mathcal{N}(0, \mathbf{I})$. A noised latent variable $\mathbf{z_t}$ can then be represented as $\mathbf{z_t} = \sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon$, where α_t is determined by a noise scheduler based on t.

Given a pre-trained diffusion model ϵ_ϕ with parameter set ϕ, a modified predicted noise according to classifier-free guidance [14] can be expressed as

$$\epsilon_\phi^w(\mathbf{z_t}, y, t) = (1 + w)\epsilon_\phi(\mathbf{z_t}, y, t) - w\epsilon_\phi(\mathbf{z_t}, t),$$

where $\epsilon_\phi(\mathbf{z_t}, y, t)$ is the raw noise predicted by the diffusion model conditioned on y, $\epsilon_\phi(\mathbf{z_t}, t)$ is unconditioned noise, and w is a weight for balance.

Using two image-text pairs I_i, P_o and I_t, P_t, the DDS loss with respect to parameter θ can be expressed in gradient form as:

$$\nabla_\theta \mathcal{L}_{\text{DDS}} = (\epsilon_\phi^w(\mathbf{z_t}, y, t) - \epsilon_\phi^w(\hat{\mathbf{z}}_\mathbf{t}, \hat{y}, t))\frac{\partial \mathbf{z_t}}{\partial \theta}, \tag{1}$$

where $\epsilon_\phi^w(\mathbf{z_t}, y, t)$ is predicted from I_i, P_o and $\epsilon_\phi^w(\hat{\mathbf{z}}_\mathbf{t}, \hat{y}, t)$ is predicted from I_t, P_t with the same t and ϵ. For simplicity, this loss is denoted as $\mathcal{L}_{\text{DDS}}(I_i, I_t, P_o, P_t)$.

3.2 Perceptual Loss

The perceptual loss [20] is proposed to measure the perceptual similarity of images based on the features of VGG-16 [40]. Although originally designed for the super-resolution task by maintaining the features of the original image, it also allows for the preservation of selected regions. We denote the perceptual loss between I_i and I_t as $\mathcal{L}_{\text{per}}(I_i, I_t)$.

4 Method

Given a target image I_t with the object's mask M_t and a background image I_s with a designated region M_s for placing the object, our goal is to compose a new coherent image that retains the background from I_s while incorporating the target image's object as the foreground.

To achieve generic image composition, our method comprises three phases: **object removal**, **image harmonization**, and **semantic image composition**. This design allows for the composition of various foreground object and background images. In Figure 3, we illustrate the pipeline with special segments of different phases. The overview of the pipeline is presented in § 4.1, followed by details of object removal in §4.2, image harmonization in §4.3, and semantic image composition in §4.4.

4.1 Overall pipeline

The removal stage takes I_s and M_s as inputs to generate a background image I_b with the object in M_s removed. Subsequently, the composition stage produces a coherent image I_c given I_b, M_s, I_t, M_t. Furthermore, if conditions are provided to transfer the object from the original condition C_o to the target condition C_t, the editing stage can integrate these conditions onto I_c to synthesize image I_{res}. The conditions can take the form of text or other formats accepted by T2I-Adapter [28]. Each stage is optimized with different loss functions: $\mathcal{L}_{\text{rmv}}, \mathcal{L}_{\text{ham}}$, and $\mathcal{L}_{\text{com}}$.

The method follows a general pipeline across all three phases, as depicted in Figure 3. With inputs including an image I_i, an original prompt P_o, and a target prompt P_t, the pipeline initializes with an optimized image I_t and guides its progression to the output image I_T using a phase-specific loss function.

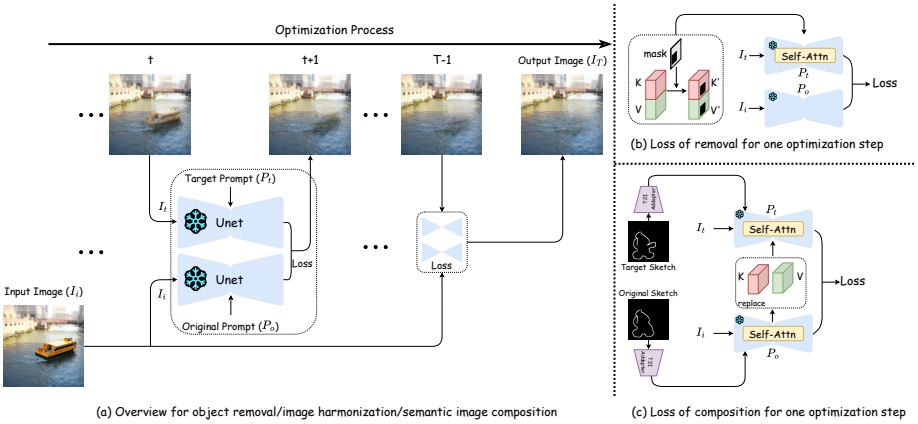

Fig. 3: Pipeline overview. Our FreeCompose pipeline consists of three phases: object removal, image harmonization, and semantic image composition. In each phase, the pipeline takes an input image and two text prompts to calculate the loss. In the object removal phase, an additional mask is required to select K, V values. In the semantic image composition phase, text prompts can be replaced by other formats, and an additional K, V replacement is implemented for identity consistency.

P_o and P_t are set as general prompts for object removal and image harmonization as elaborated in §4.2 and §4.3, whereas for semantic image composition, they are taken as input conditions.

In general, the DDS loss can modify images but may also distort them during optimization. Meanwhile, the perceptual loss helps maintain object identity. When used together, these losses can create a balanced loss function that forms the backbone of the pipeline. At the same time, minimum adjustment to the loss function enables other specific tasks, as detailed in the following sections.

4.2 Object Removal

In this phase, we take I_s as the input image I_i, and the object region mask M_s is required. P_o and P_t are set as placeholder prompts, such as "Something in some place" and "Some place," when no prompt is provided. These prompts are partially effective, but they do not have the capability to directly eliminate the object, as shown in the ablation study (see Figure 7).

We add an extra segment during the calculation of the DDS loss to enhance the ability of removal, as shown in Figure 3(b). The diffusion model is based on a UNet architecture, composed of residual, self-attention and cross-attention blocks. In the self-attention blocks, features are projected into quires Q, keys K and values V, and the output can be represented as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\big(\frac{QK^T}{\sqrt{d}}\big)V, \tag{2}$$

where d is the dimension of the hidden states.

Based on previous work such as [15], the K, V values of the self-attention layer during the denoising step are observed to have an effect on the semantic result. Guided by this discovery, we use M_s to discard some K, V values partially. Specifically, for a K or V value of shape $B \times\ l \times\ d$, where B, l, d represent the batch size, sequence length, and input dimension, respectively, we resize the mask to shape $h \times w = l$ and flatten it to a sequence with length l. By selecting indices from $v_i >$ threshold, where v_i represents the value of index i in the sequence, the semantic information of the masked region is guided by its surroundings' feature values, thereby achieving the objective of removal. This mask guided loss can be represented as $\mathcal{L}_{\text{DDS}}^{\text{rmv}}(I_i, I_t, P_o, P_t, M)$ with the gradient form:

$$\nabla_\theta \mathcal{L}_{\text{DDS}}^{\text{rmv}} = (\epsilon_\phi^w(\mathbf{z_t}, y, t) - \epsilon_\phi^w(\hat{\mathbf{z}}_\mathbf{t}, \hat{y}, t, M)) \frac{\partial \mathbf{z_t}}{\partial \theta},$$

The only difference with Eq. 1 is the $\epsilon_\phi^w(\hat{\mathbf{z}}_\mathbf{t}, \hat{y}, t, M)$, which means that the K, V values of the self-attention layers masked by M are excluded during noise prediction.

The overall loss function, thus, comprises two terms:

$$\mathcal{L}_{\text{rmv}} = \mathcal{L}_{\text{DDS}}^{\text{rmv}}(I_s, I_t, P_o, P_t, M_s) + \lambda_{\text{per}} \mathcal{L}_{\text{per}}(I_s \otimes M_s', I_t \otimes M_s'). \tag{3}$$

Here, M_s' is the reversed mask of M_s, $\otimes$ denotes the Hadamard production of two images, and λ_{per} is a hyperparameter used to balance the two losses.

4.3 Image Harmonization

Applying the bounding box of M_s, a copy-paste image I_p and its corresponding object mask M_p can be obtained from I_b, I_t, M_t. This image is used as input image I_i ($I_i = I_p$) in this phase. Without designated prompts, an empty prompt and "A harmonious scene" are initialized as P_o, P_t for the DDS loss. The perceptual loss is used separately for the background and the foreground to preserve background appearance and object identity. The overall loss consists of three terms:

$$\begin{aligned} \mathcal{L}_{\text{har}} = & \mathcal{L}_{\text{DDS}}(I_p, I_t, P_o, P_t) \\ & + \lambda_{\text{bak}} \mathcal{L}_{\text{per}}(I_p \otimes M_p', I_t \otimes M_p') + \lambda_{\text{for}} \mathcal{L}_{\text{per}}(I_p \otimes M_p, I_t \otimes M_p), \end{aligned} \tag{4}$$

where M_p' is the revered mask of M_p, λ_{bak} is a hyperparameter used to balance the perceptual loss related to the background and λ_{for} is a hyperparameter used to balance the perceptual loss related to the target object.

4.4 Semantic Image Composition

This phase accepts either the copy-paste image I_p or the composition result I_c and requires two additional conditions: C_o and C_t. If the conditions are in text form, they will be directly used as P_o and P_t for the DDS loss. Conditions in

Table 1: Results of User Study on Object Removal. The participants are requested to evaluate the results based on two aspects: (1) the level of image harmony after the object has been removed, and (2) the extent to which the object has been completely removed. Each criterion is rated from 1 (worst) to 5 (best) without additional explanation.

	Image Harmony ↑	Object Removal ↑
Repaint [27]	3.24 ± 1.23	3.82 ± 1.35
SD Inpainting [23]	2.99 ± 1.37	3.55 ± 1.34
Lama [44]	3.47 ± 1.16	4.14 ± 0.94
FreeCompose (ours)	$\mathbf{3.85 \pm 1.01}$	$\mathbf{4.47 \pm 0.73}$

other forms will be translated by T2I-Adapter [28] and added to the diffusion UNet as shown in Figure 3(c).

An additional design is employed to maintain the identity of the object during the editing procedure. As displayed in Figure 3(c), FreeCompose replaces the optimized image I_t's K, V values with I_i's K, V values during the calculation of DDS loss. Specifically, for a DDS loss with K_i, V_i, K_t, V_t, where K_i, V_i represent the K, V values of I_i, and K_t, V_t represent the K, V values of I_t, we modify the calculation of self-attention in the diffusion UNet concerning I_t as follows:

$$\begin{cases} \text{Attention}(Q, K_i, V_i), & \text{if } t > T \text{ and } l > L \\ \text{Attention}(Q, K_t, V_t), & \text{otherwise,} \end{cases}$$

where t is the count of optimization, l is the layer index of the self-attention layer, T and L are hyperparameters indicating the count number and layer index of self-attention to start such replacement. Because the background is also preserved along with the replacement, no perceptual loss is required. Therefore, the complete loss has the same format as the DDS loss:

$$\mathcal{L}_{\text{com}} = \mathcal{L}_{\text{DDS}}^{\text{com}}(I_c, I_t, C_o, C_t),$$

where $\mathcal{L}_{\text{DDS}}^{\text{com}}(I_c, I_t, C_o, C_t)$ represents the DDS loss using C_o, C_t as substitutes for conditions in forms besides text, with an additional design of K, V replacement during calculation.

5 Experiments

5.1 Implementation Details

Global Hyperparameters. We use Stable Diffusion V2.1[1] as the pre-trained model for real images, and AnyLoRA[2] as the pre-trained model for anime and cartoon images. We align the resolution of input images with the diffusion model to 512×512. The Adam optimizer is adopted with a fixed learning rate of $5e^{-2}$.

[1] https://huggingface.co/stabilityai/stable-diffusion-2-1
[2] https://huggingface.co/Lykon/AnyLoRA

Hyperparameters. In the object removal phase, the DDS loss outside the mask resized from M_s to the latent size is multiplied by 0.2 to limit the transformation of the background. Additionally, $\lambda_{\text{per}} = 0.3$. In the image harmonization phase, $\lambda_{\text{bak}} = 0.3$ and $\lambda_{\text{for}} = 0.1$. The semantic image composition only uses the DDS loss with $T = 400$ and $L = 10$ for the replacement design.

Prompt Usage. Two prompts, P_o and P_t, are required for every calculation of the DDS loss. Providing specific prompts will improve the optimization procedure. Our FreeCompose does not rely on user-provided text prompts for image composition. Instead, we predefined the prompts for different phases. Specifically, in the object removal phase, we set P_o as "Something in some place." and P_t as "Some place.", respectively. Similarly, we adopt empty prompts for P_o and "A harmonious scene." for P_t in the image harmonization phase. These prompts have proven to be effective.

5.2 Main Results

Object removal In Figure 4, we present the results of object removal, comparing them with previous work on removal and inpainting. When using the default prompts in §5.1, Lama [44], Stable Diffusion Inpainting [23], and Repaint [27] require the same input as our method. This includes one original image along with a corresponding mask for the region that needs to be removed. As shown, SD Inpainting and Repaint struggle to completely remove the object, leaving some parts unchanged or replaced by something that doesn't fit well, like the outline of the dog in the second case and the unknowns in the fourth case. Although Lama performs better in removing the object and reconstructing the background, it fails to remove certain attachments of the object, such as the shadow in the third case. In general, our method demonstrates a stronger capability in removing the object and seamlessly filling the resulting areas, as can be observed in the third case where other methods perform poorly.

Image harmonization As shown in Figure 5, Diff Harmonization successfully generates primary shadow as surface variation in the first candle case. However, it struggles to retain identity features such as the color of the second emoji case

Table 2: Results of User Study on Image Harmonization. Participants are asked to rate the results based on (1) image harmony after the composition of the object, and (2) how well the identity of the object is preserved.

	Image Harmony↑	Object Identity Preserving↑
Diff Harmonization [16]	3.11 ± 1.04	3.83 ± 1.10
DucoNet [45]	3.14 ± 1.17	$\mathbf{4.16 \pm 1.04}$
FreeCompose (ours)	$\mathbf{3.69 \pm 1.07}$	4.11 ± 0.92

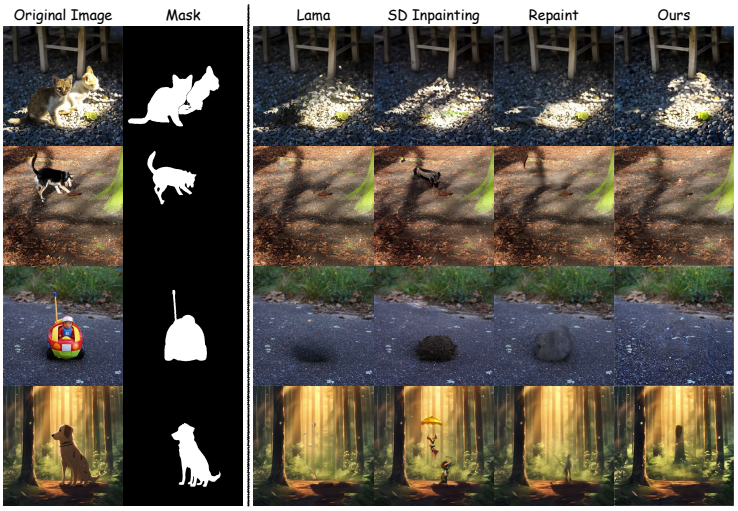

Fig. 4: Qualitative comparison on object removal. We compare with Lama [44], Stable Diffusion Inpainting [23], and Repaint [27].

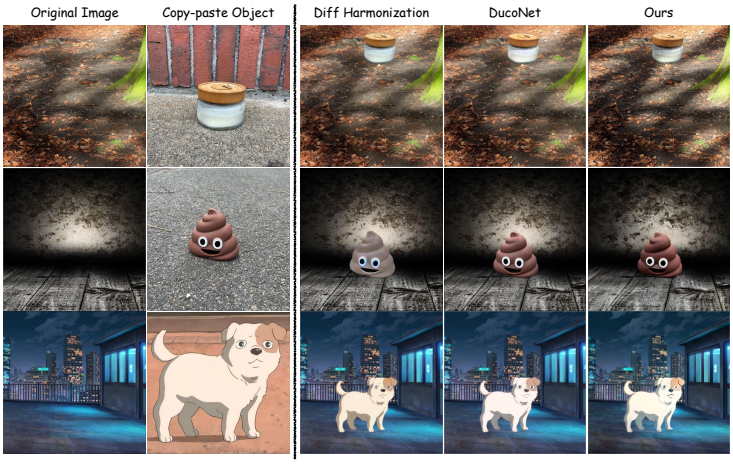

Fig. 5: Qualitative comparison on image harmonization. We compare our method with zero-shot Diff Harmonization [16] and training-based DucoNet [45].

and the shape of the third dog case's eye. On the other hand, DucoNet preserves these features well but lacks realistic shadow and light effect under certain environments. For instance, in the first case, DucoNet simply illuminates the entire object, without accurately transforming the dark and bright sections according to the original image. In contrast, our method is capable of both preserving the object's identity and generating realistic lighting effects. For example, in the first case, FreeCompose enables the object to be covered by the shadow of the

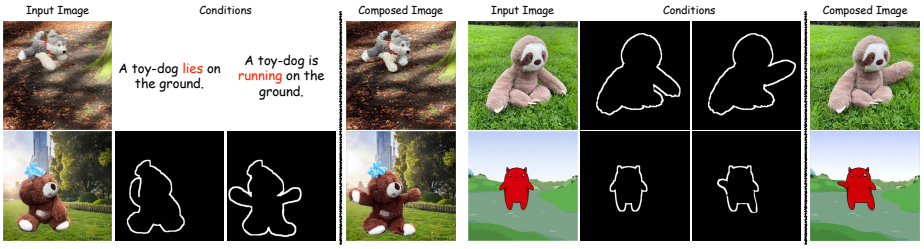

Fig. 6: Results on semantic image composition. Our method accepts various conditions as guidance, including text and sketches. The case in the top-left corner uses different prompts as guidance for editing, while the other cases are guided by different sketches with identical prompts.

surroundings, resulting in the corresponding dark section while maintaining the object's identity.

Semantic image composition Figure 6 illustrates the results of our semantic image composition. By using an input image (either a copy-paste image or an image after harmonization), FreeCompose is able to generate a composed image that maintains semantic consistency, guided by the disparity between two input conditions. As shown, the top-left case makes use of the difference between two prompts to transfer the dog from a lying posture to a running posture. In other cases, with the same prompt during calculation, the features extracted by T2i-Adapter [28] from different sketch images serve as guidance for semantic composition, proving the feasibility of wider usage.

Quantitative comparison. Since our primary focus is on open domain questions, we believe that evaluating performance through user studies is more appropriate. We have planned a user study to assess the results of object removal and image harmonization, comparing them with previous works with five cases respectively. The study involves more than twenty volunteers. To evaluate the effectiveness of object removal, participants are asked to assess the outcomes based on two criteria: (1) the level of image harmony achieved after the object is removed, and (2) the extent to which the object removal is executed. In terms of image harmonization, participants are instructed to assess the results based on two aspects: (1) the level of visual coherence achieved after integrating the object into the composition, and (2) the degree to which the object's identity is preserved. Each metric is rated on a scale ranging from 1 to 5.

The results are shown in Table 1 and Table 2. As demonstrated, our method excels in both aspects for object removal. Our method in image harmonization received the highest rating for "Image Harmony", but it lagged behind DucoNet in terms of "Object Identity Preservation." One possible reason is that our method employs a stronger composition strategy by restricting the weight of the foreground loss, resulting in partial degradation of the object's identity.

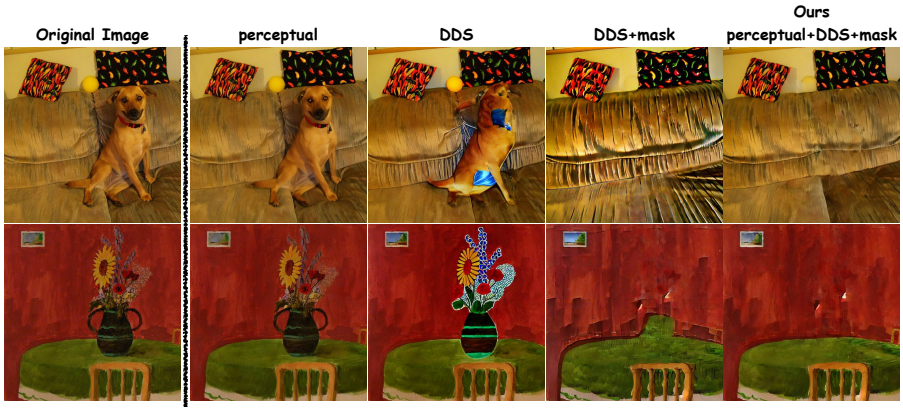

Fig. 7: Qualitative ablation studies on loss sections of the object removal phase. "perceptual" refers to the perceptual loss section in Eq. (3) alone, "DDS" refers to the vanilla DDS loss, and "DDS+mask" refers to the mask-guided DDS loss in §4.2.

5.3 Ablation Study

We conducted an ablation study to validate our designs and analyze their functions. by disassembling and visualizing each design to clearly demonstrate their effects.

Object Removal Phase. In Figure 7, designs of the object removal phase are disassembled for analysis. The perceptual loss alone maintains the original image without any changes as the the "perceptual" column displayed. When using a raw DDS loss with default prompts, the object cannot be completely eliminated, resulting in some variations in the object in line with the "DDS" column. The introduced mask design in §4.2, which selectively discards specific KV values based on the mask of the object, overcomes this limitation and enables the loss to successfully remove the object. However, such mask guided loss affects the background which should be preserved, as presented in the "DDS+mask" column. The last addition of perceptual loss section helps preserve the background while calculating the mask guided loss and generates the background image independently from the original foreground as demonstrated in the "Ours" column.

Image Harmonization Phase. In Figure 8, different sections of the loss are ablated for observation of their respective functions. The perceptual loss, comprising the background perceptual loss and the foreground perceptual loss, ensures the consistency with the original copy-paste image, as seen in the "perceptual" column. When using the raw DDS loss, it allows for seamless blending of the object with the background, but may unintentionally remove certain features from both the foreground and the background, compatible with the "DDS" column. By employing distinct perceptual loss functions for the foreground and the

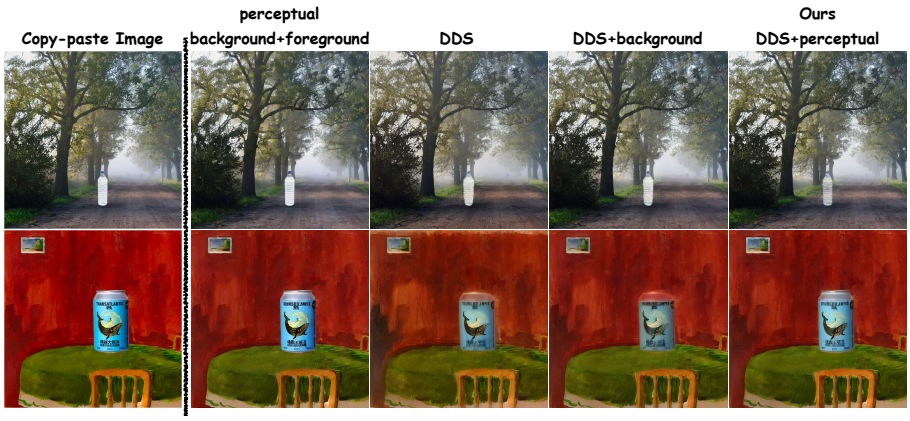

Fig. 8: Qualitative ablation studies on loss sections of the image harmonization phase. "background" refers to the background perceptual loss in Eq. (4), "foreground" refers to the foreground perceptual loss in Eq. (4), "perceptual" represents the sum of "background" and "foreground", and "DDS" represents the DDS loss.

background, the trade-off among the degree of integration, the identity of the object and the features of the background is achieved, enabling the generation of a harmonious image as shown in the "Ours" column.

6 Conclusion

We present FreeCompose, a generic zero-shot image composition method that utilizes diffusion prior. In this work, we noticed that pre-trained diffusion models are capable of detecting inharmonious portions in copy-paste images. Building on this observation, we successfully apply this prior to both image harmonization and semantic image composition. FreeCompose is a zero-shot method, allowing easy usage without additional training. Moreover, it's suitable for various applications, showcasing the potential of the diffusion model prior.

We believe that the prospect of diffusion prior extends beyond what we have achieved thus far. In the future, we plan to explore additional uses for other composition tasks and to apply our method to video, capitalizing on its full capabilities.

Acknowledgement

This work is partially supported by the National Key R&D Program of China (NO. 2022ZD0160101). The authors would like to thanks Hangzhou City University for accessing its GPU cluster.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Trans. Graphics (2023)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2022)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2022)
4. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2024)
5. Chen, X., Fang, L., Ye, L., Zhang, Q.: Deep video harmonization by improving spatial-temporal consistency. J. Mach. Learn. Res. (2024)
6. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.Q.: Color harmonization. ACM Trans. Graphics (2006)
7. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: Bargainnet: Background-guided domain translation for image harmonization. arXiv: Comp. Res. Repository (2020)
8. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2020)
9. Gao, R., Grauman, K.: On-demand learning for deep image restoration. In: Proc. IEEE Int. Conf. Comp. Vis. (2017)
10. Hao, G., Iizuka, S., Fukui, K.: Image harmonization with attention-based deep feature modulation. In: Trans. Mach. Learn. Res. (2020)
11. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proc. IEEE Int. Conf. Comp. Vis. (2023)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv: Comp. Res. Repository (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Proc. Advances in Neural Inf. Process. Syst. (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv: Comp. Res. Repository (2022)
15. Huang, J., Liu, Y., Qin, J., Chen, S.: Kv inversion: Kv embeddings learning for text-conditioned real image action editing. arXiv: Comp. Res. Repository (2023)
16. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Chen, S., Cao, L.: Diffusion model-based image editing: A survey. arXiv: Comp. Res. Repository (2024)
17. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graphics (2017)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2017)
19. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. arXiv: Comp. Res. Repository (2021)
20. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. arXiv: Comp. Res. Repository (2016)
21. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv: Comp. Res. Repository (2023)

22. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. arXiv: Comp. Res. Repository (2017)
23. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2023)
24. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Pattern Recogn. (2015)
25. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2021)
26. Liu, G., Reda, F., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. Proc. Eur. Conf. Comp. Vis. (2018)
27. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2022)
28. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv: Comp. Res. Repository (2023)
29. Mustafa, A., Mantiuk, R.K.: Transformation consistency regularization- a semi-supervised paradigm for image-to-image translation. arXiv: Comp. Res. Repository (2020)
30. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. arXiv: Comp. Res. Repository (2021)
31. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
32. Pitie, F., Kokaram, A.C., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 (2005)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv: Comp. Res. Repository (2023)
34. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv: Comp. Res. Repository (2022)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv: Comp. Res. Repository (2022)
36. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications (2001)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2022)
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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2023)
39. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Proc. Advances in Neural Inf. Process. Syst. (2022)

40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv: Comp. Res. Repository (2014)
41. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021)
42. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proc. Int. Conf. Mach. Learn. (2015)
43. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Trans. Graphics (2010)
44. 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: Proceedings of the IEEE/CVF winter conference on applications of computer vision (2022)
45. Tan, L., Li, J., Niu, L., Zhang, L.: Deep image harmonization in dual color spaces. arXiv: Comp. Res. Repository (2023)
46. Tao, M.W., Johnson, M.K., Paris, S.: Error-tolerant image compositing. Int. J. Comput. Vision (2013)
47. Tao, M., Bao, B., Tang, H., Wu, F., Wei, L., Tian, Q.: De-net: Dynamic text-guided image editing adversarial networks. Proc. AAAI Conf. Artificial Intell. (2022)
48. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2017)
49. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Proc. Advances in Neural Inf. Process. Syst. (2024)
50. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. arXiv: Comp. Res. Repository (2020)
51. 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: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2023)
52. Yang, C., Lu, X., Lin, Z.L., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
53. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C., Xing, E.: Multimodal image synthesis and editing: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
54. Zhang, R., Che, T., Ghahramani, Z., Bengio, Y., Song, Y.: Metagan: an adversarial approach to few-shot learning. In: Proc. Advances in Neural Inf. Process. Syst. (2018)
55. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Learning a discriminative model for the perception of realism in composite images. In: Proc. IEEE Int. Conf. Comp. Vis. (2015)
56. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. IEEE Int. Conf. Comp. Vis. (2017)