HomeDiffusion: Zero-Shot Object Customization with Multi-View Representation Learning for Indoor Scenes

Guoqiu Li, Jin Song, Yiyun Fei

Alibaba Group
{liguoqiu.lgq, songjin.song, yunhun.fyy}@alibaba-inc.com

Abstract

Recently, zero-shot object customization generation methods have rapidly developed and shown tremendous potential for applications. For instance, in the e-commerce domain, consumers can observe the visual effect of furniture placed within their personal living spaces or clothes worn on their own bodies. Many existing approaches perform object customization generation based on diffusion models and extracted reference object features. However, the generated object significantly diverges from the original reference object in details such as patterns and curves. Particularly for asymmetrical reference objects, the absence of comprehensive multi-viewpoint information prevents the generation of object poses that harmonize with the background scene. To address these shortcomings, we have constructed a novel dataset comprising multi-angle images of furniture and indoor scenes. Based on diffusion models, we introduce HomeDiffusion, which can leverage multi-viewpoint images of the same reference object to accurately generate visually harmonious object poses within specified areas of the background scene. During the diffusion process, we further extract high-fidelity details of the reference object and perform cross-attention with the noise latents in the latent space, thereby ensuring the preservation of details in the customized object generation. Extensive qualitative and quantitative experiments demonstrate that our method achieves superior performance over other existing zero-shot as well as few-shot object customization approaches.

Introduction

The goal of object customization in image editing is to seamlessly integrate objects from reference images into a specific area of the target edited image, ensuring that they harmoniously align with the background in layout, lighting, perspective, and spatial relationships, while maintaining the object's identity and intrinsic properties, including texture, shape, and distinctive features. Object customization image editing method opens up a world of possibilities, such as the ability to visualize how a new sofa looks in your living room with just a few clicks, without ever stepping into a store. The field's potential makes it a significant research interest.

Many state-of-the-art methods (Ruiz et al. 2023; Yang et al. 2023; Chen et al. 2024b) utilize large pre-trained text-to-image models such as Stable Diffusion (Rombach et al. 2021) as their backbone networks, leveraging the rich prior knowledge of object categories contained within. The customized objects typically harmonize with the background in terms of lighting and color tone, while still retaining a considerable amount of their identity features. However, two key challenges remain highly formidable. The first challenge is the acquisition of multi-view object representations. While various approaches attempt to blend the reference object into an edited image by subtly adjusting the edges' tone and texture for coherence, these methods can sometimes struggle with perspective harmony. This can result in a mismatch between the object and its environment, leading to a visually awkward appearance. Particularly for asymmetric objects, it becomes impossible to generate an object with a completely different perspective using a single-view object representation when the orientation of the background is entirely dissimilar. The second challenge is to preserve the high-fidelity object details. Although some methods (Yang et al. 2023; Chen et al. 2024b) utilize powerful pre-trained multi-modal or self-supervised image encoders to encode the identity features of the target, ensuring consistency throughout generation. The lack of maintenance of detail information results in generated images that do not meet expectations (refer to Figure 5).

To address the above challenges, we propose a novel zero-shot object customization image editing framework for indoor scenes, named HomeDiffusion. HomeDiffusion adeptly focuses on embedding high-fidelity objects into edited images using reference images of objects from various viewpoints, ensuring the generated object's perspective is aligned with the background scene. Compared to outdoor scenes, indoor scenes have a more defined spatial structure. The placement of objects follows certain patterns and knowledge, making it an ideal and essential scenario for object customization capabilities. Specifically, we build a multi-view indoor dataset, which comprises a collection of indoor scene images and furniture images captured from multiple viewpoints. We propose a powerful HD visual encoder that can extract high-resolution details of objects from both global and local perspectives. Then, we introduce the Multi-view Object Representation Learning (MORL) process, wherein multiple viewpoint images of a furniture item serve as input, and the model is trained to pre-

Figure 1: HomeDiffusion enables users to virtually place e-commerce furniture in any scene, ensuring a high degree of fidelity and harmonious perspective with the background.

dict another viewpoint of this furniture. Through an extensive self-generated training process utilizing images of objects from multiple viewpoints, the model acquires the capability to learn multi-view representations of the objects.

We further propose the Background-driven Object Customization Learning (BOCL) process. After the MORL stage, we preserve the model's ability to extract multi-view object representations. During the BOCL stage, the model is trained to insert the reference object into a designated area of the background image, ensuring that the object's perspective is in harmony with the surrounding scene. To maintain the object's high-fidelity details, we create a composited image that fuses the reference object with scene information. This image, containing high-fidelity object details, will be used to guide the image generation. Given the perspective deviation between the reference object in the composite image and the target in the generated image, we further introduce pixel-aligned cross-attention calculations within the latent space, thereby more effectively extracting high-fidelity details from the composite image. Our main contributions are:

- we propose HomeDiffusion, a novel zero-shot object customization method for indoor scenes, which ensures high-fidelity object detail and harmonious perspective blending with the background scene.
- We propose an HD visual encoder that can extract fine details of objects from both global and local views.
- we present a MORL process that learns robust multi-view object representation through self-generative training using images from multiple viewpoints.
- we introduce a BOCL method that harmoniously integrates reference objects into designated areas within background scenes. Furthermore, by executing pixel-aligned cross-attention calculations in the latent space, we effectively preserve the high-fidelity object details.
- we carry out experiments on our collected ZOC-Indoor-Eval benchmark and the publicly Viton-HD benchmark. The qualitative and quantitative results show that our approach achieves superior performance compared to many zero-shot or few-shot object customization methods.

Related Works

Diffusion Based Text-to-Image Generation. Recently, Diffusion models have achieved significant success in the text-to-image generation field (Rombach et al. 2021; Nichol et al. 2022; Saharia et al. 2022; Podell et al. 2023; Dai et al. 2023; Balaji et al. 2023; Ho et al. 2022; Kumari et al. 2023a), yielding better results compared to methods such as GAN (Dhariwal and Nichol 2021; Goodfellow et al. 2020). Diffusion models create images by learning to progressively recover images from pure Gaussian noise (Ho, Jain, and Abbeel 2020). Stable Diffusion (Rombach et al. 2021) employs a VAE to compress the image from pixel space to latent space, reducing the computational power required for image generation. However, in T2I models, it is not possible to implement object embedding through textual guidance.

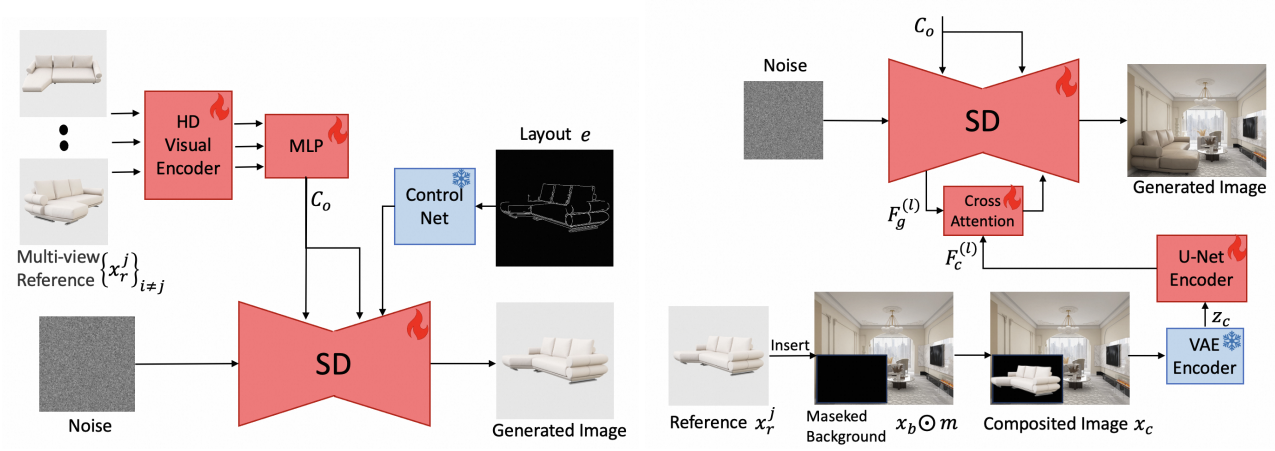
Diffusion Based Image Editing. There has been some impressive work in image editing using diffusion models (Lugmayr et al. 2022; Meng et al. 2022; Kawar et al. 2023; Kim, Kwon, and Ye 2022; Hertz et al. 2022; Zhang et al. 2023). These methods typically use text to guide specific areas of images. However, it's clear that text descriptions alone are not enough to capture all the small, important details needed to describe objects accurately. To solve this problem, some methods use photos of the object as guidance, combining text input to generate images. A standout example is Paint by Example (Yang et al. 2023), the method utilizes the CLIP model (Radford et al. 2021) to extract image embeddings as the image condition injected into the inpainting process; however, the features extracted based on CLIP are not sufficient for preserving the fine details of objects. Even with these approach, preserving all details of the object in the final edited image remains a challenging task.

Subject-Driven Image Generation. The goal of personalized image generation is to retain as much detail of the object as possible in the generated images, given one or more photos of the object. Recently, several approaches based on pre-trained text-to-image models have been proposed (Ruiz et al. 2023; Gal et al. 2022; Xiao et al. 2023; Li et al. 2023; Voynov et al. 2023; Shi et al. 2023; Chen et al. 2022; Kumari et al. 2023b; Chen et al. 2024a). DreamBooth (Ruiz et al. 2023) finetunes a pre-trained text-to-image model by training a LoRA model, which works by associating a special word in the prompt with the example images. Works closely related to ours are AnyDoor (Chen et al. 2024b). AnyDoor (Chen et al. 2024b), utilizing DINO-V2 (Oquab et al. 2024) for image feature extraction, is trained a single time and seamlessly generalizes to various object-scene combinations at inference, producing images that richly preserve object details. However, these methods struggle with achieving high similarity in detail when generating images from multiple viewpoints. Conversely, our method can deliver consistency across various angles and high-fidelity details of objects without the need for test-time training.

Methodology

Preliminaries

Our proposed method is based on the state-of-the-art Stable Diffusion (SD) model, a latent diffusion model that synthesizes high-quality images by modeling the progressive denoising diffusion process in a lower-dimensional latent



(a) Stage 1: Multi-view Object Representation Learning.

(b) Stage 2: Background-driven Object Customization Learning.

Figure 2: The overall pipeline of HomeDiffusion. In Stage 1 (MORL), the multi-view object representation C_o is learned from a set of multi-view reference images via an HD visual encoder and an MLP layer. Additionally, a ControlNet is utilized for viewpoint guidance. Stage 2 (BOCL) focuses on integrating the learned multi-view object representation into a background scene at a user-specified location, utilizing a masked background and a composited image to guide the diffusion model in object placement and object details. **Flames** and **snowflakes** refer to learnable and frozen parameters, respectively.

Figure 3: A sample of the training data.

space. Specifically, for an input RGB image $x_0 \in \mathbb{R}^{H\times W\times 3}$, SD first utilize a variational autoencoder (VAE) $\mathcal{E}$ to compress x_0 into a smaller low-dimensional latent representation $z_0 \in \mathbb{R}^{h\times w\times c}$. Then, during the forward diffusion process, Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$ is progressively added to z_0 to obtain the latent representation z_t at time step t. After that, during the reverse denoising process, a conditional U-Net parameterized by θ predicts the noise ϵ_θ added in the forward process, step by step subtracting the noise ϵ_θ until a final latent representation $\hat{z_0}$ is obtained. Finally, the VAE decoder $\mathcal{D}$ maps $\hat{z_0}$ back to the pixel space, thereby obtaining the generated image $\hat{x_0}$. During the denoising process, the U-Net can be guided on various forms of condition C such as text prompts or image embeddings through the cross-attention mechanism. The overall training objective of the SD model can be mathematically expressed as:

$$\mathcal{L}_{SD} = \mathbb{E}_{z_0,C,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_\theta(z_t,t,C)\|_2^2\right] \quad (1)$$

Dataset Collection

A major challenge in object customization image editing is ensuring that the generated object's pose harmonizes with the scene. This requires the model to understand the 3D spatial information inherent in 2D images and to adjust the object's perspective relationships appropriately during the generation process. Compared to outdoor scenes, indoor scenes have a more defined spatial structure; the presence of walls directly conveys depth and viewpoint information in the images, and the placement and orientation of furniture generally correspond to the layout of the indoor space. Therefore, the intrinsic spatial structure and complexity of indoor scenes pose significant challenges for object customization generative models. High-quality 3D indoor scene data is scarce, which poses challenges for research. After extensive searching, we found that the 3D-FRONT (Fu et al. 2021) dataset, commonly used for 3D indoor scene design tasks (Hu et al. 2024; Yang et al. 2024), is a suitable dataset.

3D-FRONT is a large-scale, and comprehensive repository of synthetic indoor scenes designed by professional designers. We utilized the rendering tool provided by 3D-FRONT to render scene images from different camera angles, where the corresponding furniture will also have different viewpoints. We will segment the target furniture from the images, remove the background, and obtain images of the same furniture from different viewpoints. Then, we can remove the furniture and render a pure background image with the same camera angle, restoring the changes in lighting and shadow caused by placing the furniture. A sample of the constructed training data is shown in Figure 3.

Specifically, our training dataset contains more than 180k scene images in total, with targeted furniture including 8 categories: bed, chair, coffee table, floor lamp, nightstand, sofa, TV cabinet, and wardrobe. Statistics show that over 60% of the furniture has images with more than 5 different viewpoints.

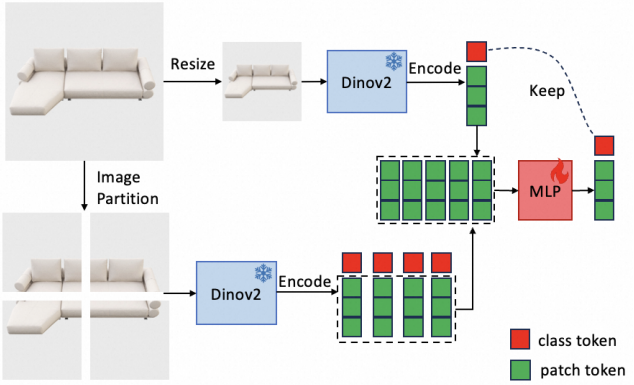


Figure 4: The illustration of HD Visual Encoder.

HomeDiffusion

Given N multi-view images $\{x_r^i\}_{i=1}^N \in \mathbb{R}^{h\times w\times 3}$ of a reference object, a background image $x_b \in \mathbb{R}^{H\times W\times 3}$, and a user-specified location binary mask $m \in \mathbb{R}^{H\times W}$ (where 0 pixels represent the background and 1 pixels denote the editable area). Object customization image editing aims to inject the reference object into the specified area of the background image, ensuring it harmonizes with the background scene. Facing this task, there are two key points: firstly, the model must be capable of generating any desired viewpoint of the reference object (especially when the object is asymmetric) from a limited set of N reference images $\{x_r^i\}_{i=1}^N$; secondly, the model should generate the reference object with an appropriate pose in the specified area, in accordance with the background layout, ensuring that the generated image is harmonious. The overall pipeline of HomeDiffusion is shown in Figure 2. HomeDiffusion consists of two stages: Multi-view Object Representation Learning (MORL) and Background-driven Object Customization Learning (BOCL), which correspond to the two key points mentioned above respectively.

HD Visual Encoder. To preserve the details of the object, we need a powerful visual encoder to extract more discriminative features. According to the AnyDoor (Chen et al. 2024b) method, we can employ pre-trained DINO-V2 (Oquab et al. 2024) model as the base visual encoder, which encodes a 224×224 resolution image as a class token $T_c^{1\times 1536}$ and patch tokens $T_p^{256\times 1536}$. Naturally, we need to be able to encode higher resolution object images, but the number of tokens will sharply increase, leading to a huge demand for GPU memory. Drawing inspiration from some VLM methods (Dong et al. 2024), we propose an HD visual encoder that extracts fine visual details from global and local views. As shown in Figure 4, given a 448×448 object image, we process it from both global and local views. For the global view, we resize the image to 224×224 size and then extract features using DINO-V2. This provides a macro understanding of the image. For the local view, we divide the image into 4 non-overlapping 224×224 patches, and then extract features for each patch, providing high-resolution local details of the image. We then need to merge the global and local features. We concatenate the patch tokens $T_p^{5\times 256\times 1536}$ from all 5 images and then map them to new patch tokens $\hat{T}_p^{256\times 1536}$ through an MLP layer. Finally, we retain the class token $T_c^{1\times 1536}$ from the global view, which provides macro information of object, and together with $\hat{T}_p^{256\times 1536}$, forms the final high-definition visual encoding. This significantly reduces the number of tokens.

Multi-view Object Representation Learning. Figure 2a illustrates the process of MORL. Given N multi-view reference images $\{x_r^i\}_{i=1}^N$, we randomly select one viewpoint image x_r^j as the source image for the diffusion process, and then use the remaining images $\{x_r^i\}_{i\neq j}$ as the condition to guide the diffusion process. Each reference image $\{x_r^i\}_{i\neq j}$ will be encoded by the HD visual encoder, yielding multiple object features of size 257×1536. We stack these object features and then map them to a 257×1024 token C_o through an MLP layer. This token, which encapsulates information from multiple reference viewpoints, is subsequently injected into the diffusion process via the cross-attention layers within the U-Net. But lacking any supplementary information about the target viewpoint, how does the diffusion model generate the target image x_r^j ? We employ the ControlNet (Zhang, Rao, and Agrawala 2023) method to provide extra layout information to guide the image generation process. Specifically, we extract the Canny edge map e of the target viewpoint x_r^j. Then, we use a pre-trained ControlNet model to inject the edge map into the decoder of the U-Net via residual connections, using the layout as a conditional guide for the image generation process. Therefore, the training objective of MORL can be expressed as follows:

$$\mathcal{L}_{MORL} = \mathbb{E}_{z_0,C_o,e,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, C_o, e)\|_2^2\right] \quad (2)$$

Background-driven Object Customization Learning. Figure 2b shows the BOCL process. In this stage, we need to train the model to insert the reference object into the specified area of the background image, ensuring that the object's perspective is harmonious with the background scene. After the MORL stage, we freeze the weights of the MLP layer to preserve the multi-view object representation capabilities of token C_o, while also reducing the difficulty of training. We combine the background image x_b with the user-specified location mask m to create a masked background image $x_b \odot m$, which provides the scene layout and location information necessary to guide the diffusion model in generating the object with an appropriate viewpoint. We further paste the reference object image x_r^j onto the editable area in masked background image to create a composited image x_c, which contains both scene information and high-fidelity details of the reference object. We then use the VAE encoder $\mathcal{E}$ to compress the composite image x_c into a latent embedding z_c, thus ensuring that the composite image and the generated image are pixel-aligned within the latent space.

Similarly to ControlNet, we utilized a new U-Net encoder to extract multi-level feature maps $F_c^{(l)}$ from the latent embedding z_c, where $l \in \{1, \ldots, 13\}$. Correspondingly, in the SD model, the U-Net encoder will generate multi-level feature maps $F_g^{(l)}$ with varying resolutions. As Equation

Figure 5: Qualitative comparison with other zero-shot object customization methods. including Paint-by-Example (2023) and AnyDoor (2024b). **All methods use only a single-view image with the same resolution as a reference**.

3 shows, we perform cross-attention calculations between $F_c^{(l)}$ and $F_g^{(l)}$ at corresponding levels:

$$\text{CrossAttention}(F_g^{(l)}, F_c^{(l)}) = \text{softmax}\left(\frac{(W_Q^{(l)} F_g^{(l)} (W_K^{(l)} F_c^{(l)})^T}{\sqrt{d}}\right) W_V^{(l)} F_c^{(l)}, \quad (3)$$

where $W_Q^{(l)}$, $W_K^{(l)}$, and $W_V^{(l)}$ represent three learnable projection matrices, and d is the output dimension of the key and query vectors. The calculated results will be passed to the corresponding layer of the U-Net decoder. In this manner, the model can integrate high-fidelity details from the composite image into the generated image at various layer levels. The training objective of BOCL stage can be expressed as Equation 4:

$$\mathcal{L}_{BOCL} = \mathbb{E}_{z_0, x_c, \epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, C_o, x_c)\|_2^2\right] \quad (4)$$

Training Strategies

In the MORL stage, for the multi-view reference image set $\{x_r^i\}_{i \neq j}$, we adopted sampling with replacement, which means that the sampled image set may contain duplicate views. Additionally, with a 10% probability, we randomly select one view image from the reference image set $\{x_r^i\}_{i \neq j}$ and replace all other view images in the set with that one. Through this approach, the model is able to learn a robust multi-view object representation. Another advantage is that at the inference stage, users can input one or several object reference images, making the model more flexible and convenient to use. To facilitate classifier-free guidance sampling (Ho and Salimans 2022), with a 10% probability, we set the entire reference image set $\{x_r^i\}_{i \neq j}$ to zero pixel images. In the BOCL stage, to drive the model focus on restoring high-fidelity details of the object within the editable area, there is a 50% probability that we will discard the background area when calculating the loss, called Background Drop.

Experiments

Implementation Details and Evaluation

Implementation Details. We use Stable Diffusion V2.1 as our backbone, and choose DINO-V2 giant version as the base image encoder. The resolution of the training images is 512×512. We use the Adam optimizer (2017) with the initial learning rate of $1e^{-5}$ and a weight decay of $1e^{-2}$ to train our model. We train our model on 8 NVIDIA Tesla V100 GPUs using PyTorch (Paszke et al. 2019).

Method	CLIP Score	DINO Score
Paint-by-Example (2023)	83.4	78.5
AnyDoor (2024b)	87.1	83.3
Ours	**89.4**	**86.2**

Table 1: Quantitative comparison of different zero-shot methods on ZOC-Indoor-Eval benchmark using single viewpoint reference image.

Method	CLIP Score	DINO Score
Paint-by-Example (2023)	80.1	56.7
AnyDoor (2024b)	81.8	59.3
Ours	**82.2**	**60.3**

Table 2: Quantitative comparison of different zero-shot methods on the Viton-HD test benchmark using single viewpoint reference image.

Benchmarks. For quantitative results, we construct a new benchmark, named the ZOC-Indoor-Val benchmark. It comprises 8 categories of furniture (described in the Dataset Collection Section). Each category includes 20 different scenes, where a scene consists of a multi-view reference image set and a scene file. The multi-view reference image set contains 5 images of the furniture from different views, while the scene file includes a scene image and a binary mask indicating the furniture's location. The view of the furniture in the scene image is different from all reference images, providing a means to validate whether the model can understand the spatial and perspective relationships between the object and its surroundings. We plan to publicly release the ZOC-Indoor-Val benchmark. To facilitate comparison with other relevant methods, we also performed quantitative analysis on the Viton-HD test (Choi et al. 2021) benchmark to measure the performance of different models in virtual try-on. **In all qualitative and quantitative experiments, the resolution of the reference images is the same.**

Evaluation Metrics. Following DreamBooth (Ruiz et al. 2023) and AnyDoor (Chen et al. 2024b) methods, we use the CLIP Score and DINO Score to calculate the similarity between the generated object and the actual target object in the scene image, where higher scores indicate better performance.

Comparisons with Related Methods

Comparison with Methods Using a Single-view Object Image. The results from Table 1 clearly demonstrate the superior performance of our method in comparison to existing zero-shot methods, as evaluated on the ZOC-Indoor-Eval benchmark. Our approach outperforms the AnyDoor and Paint-by-Example methods, achieving a CLIP Score of

Method	One view	Three views	Five views
DreamBooth (2023)	69.8	81.3	85.5
Ours	86.2	87.5	88.1

Table 3: Quantitative comparison of our method with the DreamBooth method on the ZOC-Indoor-Eval benchmark. The terms 'one,' 'three,' and 'five' views correspond to using one, three, and five different viewpoint reference images, respectively. The reported results are the DINO Scores.

Figure 6: Qualitative comparison with DreamBooth. Our method achieves higher fidelity in details such as the color and pattern of the headboard.

89.4 and a DINO Score of 86.2. In Figure 5, we present the visualization results compared with Paint-by-Example and AnyDoor. It is evident that while Paint-by-Example can generate the correct perspective for the target bed, it fails to preserve details. Because Paint-by-Example employs CLIP to encode reference images and utilizes only the class token portion as the condition. As stated in (Yang et al. 2023): "This tends to ignore the high-frequency details". AnyDoor verified that using DINO-V2 as the image encoder and taking all tokens as the condition significantly improves fidelity. However, due to the lack of multi-view object representation information, AnyDoor appears to simply "paste" the target bed into the image, resulting in a discordant perspective. Our method, on the other hand, not only generates the appropriate perspective but also retains the details well. Although we add an encoding network, it requires only a single-step computation, with the extracted multi-level features being utilized in each subsequent diffusion step. Specifically, for the 20-step diffusion sampling at a 512 resolution, it adds only an extra 5% to the inference time.

Comparison with Methods Using Multi-view Object Images. To validate the effectiveness of our method in leveraging multi-viewpoint object information, we compare our method with the few-shot DreamBooth (Ruiz et al. 2023) approach. DreamBooth accepts one or more images of the same object and then undergoes extensive training to link the object concept to a specific word for generation during inference. Our method, conversely, requires no such training and can efficiently utilize unseen images from multiple viewpoints to generate the desired object. Table 3 presents the quantitative results. Our method outperforms the DreamBooth method under the conditions of "one viewpoint," "three viewpoints," and "five viewpoints," achieving DINO Scores of 86.2, 87.5, and 88.1, respectively. As the number of received viewpoints increases, the DINO Score also

Method	Fidelity (↑)	Harmony (↑)
DreamBooth (2023)	2.18	2.57
AnyDoor (2024b)	2.70	2.46
Ours	3.45	3.51

Table 4: The average results of the small-scale human evaluation study. "Fidelity" measures object identity preservation, and "Harmony" evaluates the object's consistency with its surroundings in terms of viewpoint and lighting.

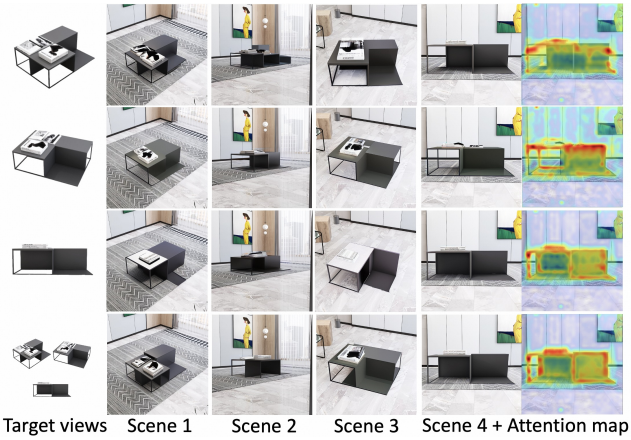


Figure 7: The impact of target viewpoints ablation on HomeDiffusion performance. The first to third rows show the generation results under different single viewpoint of the furniture, while the fourth row depicts the results when all three viewpoints are provided. In Scene 4, we specifically visualize the attention map within the U-Net.

progressively improves, indicating the effectiveness of our method in utilizing multi-view object information. Figure 6 shows our method's fidelity advantage over DreamBooth.

User Study. We invited 6 participants, including interior design experts and novices, to score images generated by different methods. A total of 900 images were evaluated on "Fidelity" and "Harmony". "Fidelity" measures the maintenance of object identity, and "Harmony" assesses congruence with surrounding viewpoints and lighting. Scores ranged from 1 (lowest) to 5 (highest), with the average score for each method calculated. As shown in Table 4, our method demonstrated significant advantages in both fidelity and harmony compared to AnyDoor and DreamBooth.

Ablation Study

Analysis on Multi-view Object Representation Learning. To verify the effectiveness of our proposed MORL method, we conduct ablation experiments reported in Figure 7 and Table 5. In Figure 7, we conducted an ablation test on a complex structured piece of furniture, which cannot be accurately reconstructed from a single viewpoint alone. We then chose four scenes with different orientations to observe the generation results when only a single viewpoint is provided, finally comparing these with the results when all three

	One view	Five views
w/o MORL	84.6	86.9
Full	86.2	88.1

Table 5: Ablation studies on the MORL method. The results shown in the table are the DINO Scores measured on the ZOC-Indoor-Eval benchmark.

Figure 8: Ablation on the impact of HD Visual Encoder.

viewpoint images are available. It can be observed in detail that when there is a significant difference between the furniture's viewpoint and the scene's orientation, reconstruction becomes challenging, leading to various issues such as color changes and panels turning into hollow spaces. However, when HomeDiffusion receives all three viewpoints, it can effectively understand the furniture's multi-view information and generate accurate results in different scenes. We further visualized the attention map within the U-Net for the results of Scene 4. It is noticeable that the attention map in the third row is most similar to the one when multiple viewpoints are used, and the furniture view in the third row closely matches the orientation of Scene 4. This demonstrates HomeDiffusion's capability in correlating the viewpoint of the objects with the scene. The quantitative results in Table 5 further suggest that MORL learns effective multi-view object information through a self-generative training process. The absence of this initial information is likely to increase the difficulty of subsequent BOCL training.

Analysis on High-fidelity Object Details Extraction. Table 6 presents the ablation studies on the high-fidelity detail extraction. The baseline contains the complete MORL stage, but for the BOCL process, we used the masked image $x_b \odot m$ alone as the guiding condition. Compared to the baseline, we introduced the HD visual encoder to replace the basic DINO-V2 encoder, which resulted in significant score increases of 1.4 and 1.3 in one and five views, respectively. This gain can be visually seen in Figure 8, where the HD visual encoder can extract more fine details, such as the correct number of pillows and the linings on the sofa. The composited image guidance strategy means that we paste the reference object x_r^j into the masked image and then use this composite image as the guiding condition, thereby transferring the rich high-fidelity details from the reference image into the diffusion model. With this strategy, the DINO scores for one view case and five views case increased by 0.4 and 0.6 points, respectively. When the composite image is encoded into the latent space, because the pasted reference object's viewpoint does not align with the target viewpoint in the generated image, directly adding the feature maps from each level could have a negative effect on the results. Instead, applying pixel-aligned cross-attention computations for each level of feature maps will be more effective. This method brings 0.6 and 1.2 DINO score gains for cases with one view and five views, respectively.

	One view	Five views
Baseline	83.8	85.0
+ HD visual encoder	85.2	86.3
+ Composited image guidance	85.6	86.9
+ Pixel-aligned cross-attention	86.2	88.1

Table 6: Ablation studies on the high-fidelity object details extraction. The results shown in the table are the DINO Scores measured on the ZOC-Indoor-Eval benchmark.

Figure 9: The generalization ability of HomeDiffusion.

The Generalization Ability of HomeDiffusion

Although the training data for HomeDiffusion all come from indoor spaces, it can also generalize to be used in outdoor natural scenes. As shown in the first row of Figure 9, chairs and beds can be naturally placed in outdoor scenes with high fidelity and harmony. We also trained our network on the Viton-HD dataset. The results, as shown in the second row of Figure 9, demonstrate that our method performs well in virtual try-on scenarios. Furthermore, when comparing our method to others using the Viton-HD-test dataset, as Table 2 shows, it is evident that our method outperforms AnyDoor and Paint-by-Example. This underscores the potential of our method to be applied across a broader range of fields.

Conclusion

In this work, we present HomeDiffusion, a novel zero-shot object customization method that maintains detailed fidelity of objects while seamlessly blending them into background scenes. We propose an HD visual encoder to extract fine image details from global and local views. We also introduce a MORL method, which learns multi-view object representations through self-generative training from various views. Moreover, we propose a BOCL approach that skillfully incorporates objects into specific areas within scenes. Our experimental results demonstrated the good performance of HomeDiffusion in customizing objects within indoor scenes, as well as its potential to generalize to other scenes.

References

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M.-Y. 2023. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv:2211.01324.

Chen, W.; Hu, H.; Li, Y.; Ruiz, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2024a. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

Chen, W.; Hu, H.; Saharia, C.; and Cohen, W. W. 2022. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. arXiv:2209.14491.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6593–6602.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Dai, X.; Hou, J.; Ma, C.-Y.; Tsai, S.; Wang, J.; Wang, R.; Zhang, P.; Vandenhende, S.; Wang, X.; Dubey, A.; Yu, M.; Kadian, A.; Radenovic, F.; Mahajan, D.; Li, K.; Zhao, Y.; Petrovic, V.; Singh, M. K.; Motwani, S.; Wen, Y.; Song, Y.; Sumbaly, R.; Ramanathan, V.; He, Z.; Vajda, P.; and Parikh, D. 2023. Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack. arXiv:2309.15807.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Zhang, S.; Duan, H.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Chen, Z.; Zhang, X.; Li, W.; Li, J.; Wang, W.; Chen, K.; He, C.; Zhang, X.; Dai, J.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD. arXiv:2404.06512.

Fu, H.; Cai, B.; Gao, L.; Zhang, L.; Li, J. W. C.; Xun, Z.; Sun, C.; Jia, R.; Zhao, B.; and Zhang, H. 2021. 3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics. arXiv:2011.09127.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1): 2249–2281.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.

Hu, S.; Arroyo, D. M.; Debats, S.; Manhardt, F.; Carlone, L.; and Tombari, F. 2024. Mixed Diffusion for 3D Indoor Scene Synthesis. arXiv:2405.21066.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023a. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023b. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2023. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. arXiv:2312.04461.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. arXiv:2108.01073.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. arXiv:2112.10741.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning. arXiv:2304.03411.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. P+: Extended Textual Conditioning in Text-to-Image Generation. arXiv:2303.09522.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. arXiv:2305.10431.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yang, Y.; Lu, J.; Zhao, Z.; Luo, Z.; Yu, J. J. Q.; Sanchez, V.; and Zheng, F. 2024. LLplace: The 3D Indoor Scene Layout Generation and Editing via Large Language Model. arXiv:2406.03866.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.