AnyDoor: Zero-shot Object-level Image Customization

Xi Chen[1] Lianghua Huang[2] Yu Liu[2] Yujun Shen[3] Deli Zhao[2] Hengshuang Zhao[1*]
[1]The University of Hong Kong [2]Alibaba Group [3]Ant Group

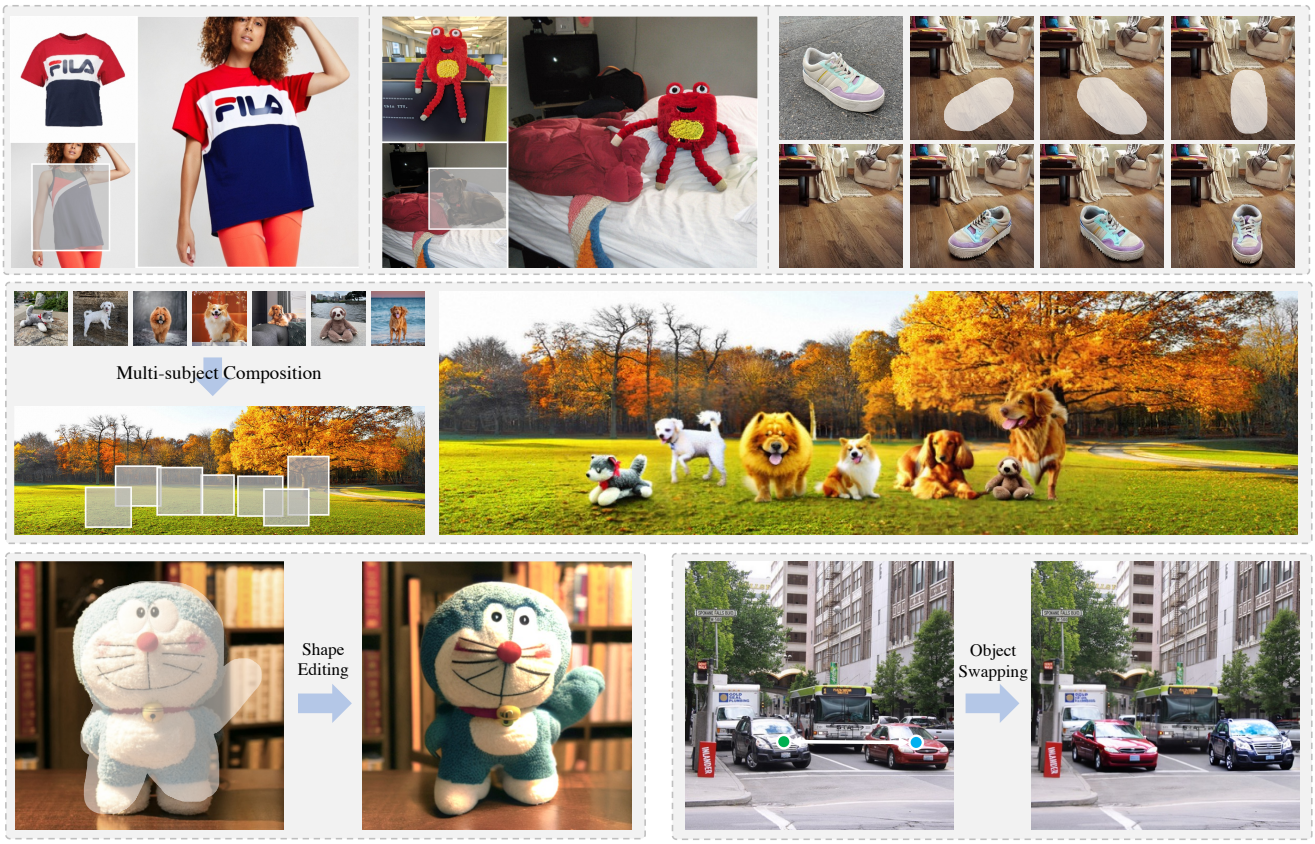

Figure 1. **Fantastic applications** of our proposed AnyDoor *without* any parameter tuning. Our model allows users to customize an image by placing an object at specific locations, with optional shape controls (top row). It can be extended to handle multiple objects (middle row) and also supports seamlessly editing the shape of the object or swapping objects within real scenes (bottom row).

Abstract

*This work presents **AnyDoor**, a diffusion-based image generator with the power to teleport target objects to new scenes at user-specified locations with desired shapes. Instead of tuning parameters for each object, our model is trained only once and effortlessly generalizes to diverse object-scene combinations at the inference stage. Such a challenging zero-shot setting requires an adequate characterization of a certain object. To this end, we complement the commonly used identity feature with detail features, which are carefully designed to maintain appearance details yet allow versatile local variations (e.g., lighting, orientation, posture, etc.), supporting the object in favorably blending with different surroundings. We further propose to borrow knowledge from video datasets, where we can observe various forms (i.e., along the time axis) of a single object, leading to stronger model generalizability and robustness. Extensive experiments demonstrate the superiority of our approach over existing alternatives as well as its great potential in real-world applications, such as virtual try-on, shape editing, and object swapping. Code is released at github.com/ali-vilab/AnyDoor.*

*Corresponding author.

1. Introduction

Image generation is flourishing with the booming advancement of diffusion models [22, 37, 40, 41, 43, 62]. Humans could generate favored images by giving text prompts, scribbles, skeleton maps, or other conditions. The power of these models also brings the potential for image editing. For example, some works [5, 24, 63] learn to edit the posture, styles, or content of an image via instructions. Other works [53, 59] explore re-generating a local image region with the guidance of text prompts.

In this paper, we investigate "object teleportation", which means accurately and seamlessly placing the target object into the desired location of the scene image. Specifically, we re-generate a box/mask-marked local region of a scene image by taking the target object as the template. This ability is a significant requirement in practical applications, like image composition, effect-image rendering, poster-making, virtual try-on, *etc*.

Although strongly in need, this topic is not well explored by previous researchers. Paint-by-Example [56] and Objectstitch [47] take a target image as the template to edit a specific region of the scene image, but they could not generate ID (identity)-consistent contents, especially for untrained categories. Customized synthesis methods [18, 27, 33, 34, 42] are able to conduct generations for the new concepts but could not be specified for a location of a given scene. Besides, most customization methods need finetuning on multiple target images for nearly an hour, which largely limits their practicability for real applications.

We address this challenge by proposing AnyDoor. Different from previous methods, AnyDoor is able to generate ID-consistent compositions with high quality in zero-shot. To achieve this, we represent the target object with identity- and detail-related features, then composite them with the interaction of the background scene. Specifically, we use an ID extractor to produce discriminative ID tokens and delicately design a frequency-aware detail extractor to get detail maps as a supplement. We inject the ID tokens and the detail maps into a pre-trained text-to-image diffusion model as guidance to generate the desired composition. To make the generated content more customizable, we explore leveraging additional controls (*e.g.* user-drawn masks) to indicate the shape/poses of the object. To learn customized object generation with high diversities, we collect image pairs for the same object from videos to learn the appearance variations, and also leverage large-scale statistic images to guarantee the scenario diversity.

Equipped with these techniques, AnyDoor demonstrates extraordinary abilities for zero-shot customization. As in Fig. 1, AnyDoor shows promising performance for the synthesis of the new concept with shape controls (top row). Besides, since AnyDoor owns the high controllability for editing the specific local regions of the scene image, it is easy to be extended to multi-subject composition (middle row), which is a hot and challenging topic explored by many customized generation methods [3, 19, 27, 34]. Moreover, the high generation fidelity and quality of AnyDoor unlock the possibilities for more fantastic applications like object moving and swapping (bottom row). We hope that AnyDoor could serve as a foundation solution for various image generation and editing tasks with image input, and act as the basic ability to energize more fancy applications.

2. Related Work

Local image editing. Most of the previous works focus on editing local image regions with text guidance. Blended Diffusion [2] conducts multi-step blending in the masked region to generate more harmonized outputs. Inpaint Anything [59] involves SAM [26] and Stable Diffusion [41] to replace any object in the source image with text described target. Paint-by-Example [56] uses CLIP [39] image encoder to convert the target image as an embedding for guidance, thus painting a semantic consistency object on the scene image. ObjectStitch [47] proposes a similar solution with [56], which trains a content adaptor to align the outputs of the CLIP image encoder to the text encoder to guide the diffusion process. However, those methods could only give coarse guidance for generation and often fail to synthesize ID-consistent results for untrained new concepts.

Customized image generation. Customized or termed subject-driven generation aims to generate images for specific objects given several target images and relevant text prompts. Some works [9, 18, 42] fine-tune a "vocabulary" to describe the target concepts. Cones [33] finds the corresponding neurons for the referred object. Although they could generate high-fidelity images, the user could not specify the scenario and the location of the target object. Besides, the time-consuming finetuning impedes them from being used in large-scale applications. Recently, BLIP-Diffusion [28] leverages BLIP-2 [29] to align images and text for zero-shot customization. Fastcomposer [52] binds the image representation with certain text embeddings to do multiple-person generation. Some concurrent works [30, 58, 61] also explore using one reference image to customize the generation results but fail to keep the fine details.

Image harmonization. A classical image composition pipeline is cutting the foreground object and pasting it on the given background. Image harmonization [7, 14, 20, 48] could further adjust the pasted region for more reasonable lighting and color. DCCF [55] designs pyramid filters to better harmonize the foreground. CDTNet [15] leverages dual transformers. HDNet [8] proposes a hierarchical structure to consider both global and local consistency and reaches the state-of-the-art. Nevertheless, these methods only explore the low-level changes, editing the structure, view, and pose of the foreground objects, or generating the shadows and reflections are not taken into consideration.

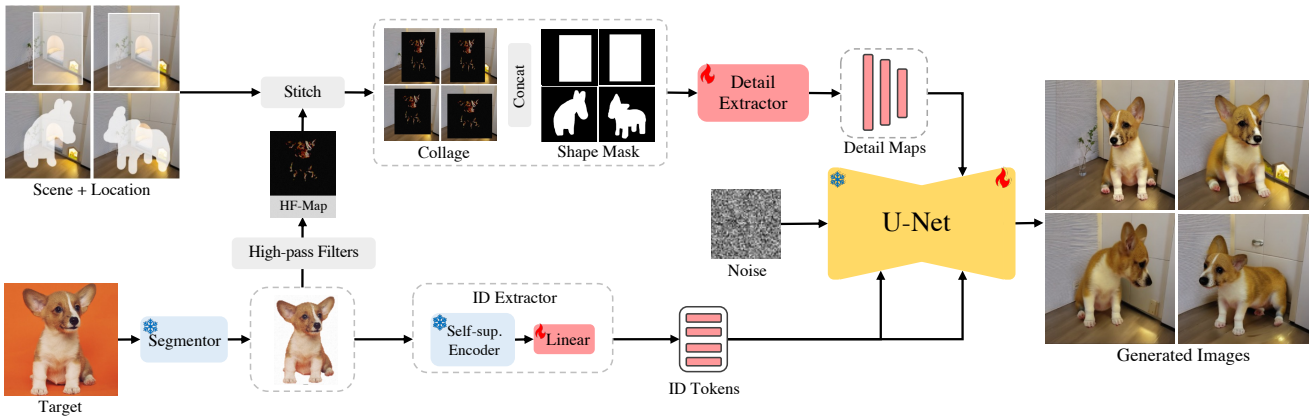

Figure 2. **Overall pipeline** of AnyDoor, which is designed to teleport an object to a scene with the desired location and shape. A segmentation module first removes the object background, followed by an ID extractor to obtain its identity information (Sec. 3.1). We then apply high-pass filters to the "clean" object, stitch the resulting high-frequency map (HF-Map) with the scene at the desired location, and concatenate the collage and shape mask. A detail extractor is designed to complement the ID extractor with appearance details (Sec. 3.2). Finally, the ID tokens and detail maps are injected into a pre-trained diffusion model to produce the final synthesis, where the target object favorably blends with its surroundings (Sec. 3.3). **Flames** and **snowflakes** refer to learnable and frozen parameters, respectively.

3. Method

The pipeline of AnyDoor is demonstrated in Fig. 2. Given the target object, the scene, and the location, AnyDoor generates the object-scene composition with high fidelity and diversity. The core idea is representing the object with identity- and detail-related features, and recomposing them in the given scene by injecting those features into a pre-trained diffusion model. To learn the appearance changes, we leverage large-scale data including both videos and images for training.

3.1. Identity Feature Extraction

We leverage the pre-trained visual encoders to extract the identity information of the target object. Previous works [47, 56] choose CLIP [39] image encoder to embed the target object. However, as CLIP is trained with text-image pairs with coarse descriptions, it could only embed semantic-level information but struggles to give discriminative representations that preserve the object identity. To overcome this challenge, we make the following updates.

Background removal. Before feeding the target image into the ID extractor, we remove the background with a segmentor and align the object to the image center. The segmentor model could be either automatic [26, 38] or interactive [11, 12, 32]. This operation has proven helpful in extracting more neat and discriminative features.

Self-supervised representation. In this work, we find the self-supervised models show a strong ability to preserve more discriminative features. Pretrained on large-scale datasets, self-supervised models are naturally equipped with the instance-retrieval ability and could project the object into an augmentation-invariant feature space. We choose the currently strongest self-supervised model DINOv2 [36] as the backbone of our ID extractor, which encodes image as a global token $\mathbf{T}_{\text{g}}^{1\times1536}$, and patch tokens $\mathbf{T}_{\text{p}}^{256\times1536}$. We concatenate the two types of tokens to preserve more information. We find that using a single linear layer as a projector could align these tokens to the embedding space of the pre-trained text-to-image UNet. The projected tokens $\mathbf{T}_{\text{ID}}^{257\times1024}$ are noted as our ID tokens.

3.2. Detail Feature Extraction

Considering that the ID tokens are represented in low resolution (16×16) , it would be hard for them to maintain the low-level details adequately. Thus, we need extra guidance for the detail generation in complementary.

Collage representation. Inspired by [6, 44], using collage as controls could provide strong priors, we attempt to stitch the "background removed object" to the given location of the scene image. With this collage, we observe a significant improvement in the generation fidelity, but the generated results are too similar to the given target which lacks diversity. Facing this problem, we explore setting an information bottleneck to prevent the collage from giving too many appearance constraints. Specifically, we design a high-frequency map to represent the object, which could maintain the fine details yet allow versatile local variants like the gesture, lighting, orientation, *etc*.

High-frequancy map. We extract the high-frequency map of the target object with

$$\mathbf{I}_h = (\mathbf{I}_{\text{gray}} \otimes \mathbf{K}_h + \mathbf{I}_{\text{gray}} \otimes \mathbf{K}_v) \odot \mathbf{I} \odot \mathbf{M}_{\text{erode}}, \quad (1)$$

where $\mathbf{K}_h, \mathbf{K}_v$ denote horizontal and vertical Sobel [23] kernels, acting as high-pass filters. $\otimes, \odot$ refer to convolution and Hadamard product. Given an Image $\mathbf{I}$, we first

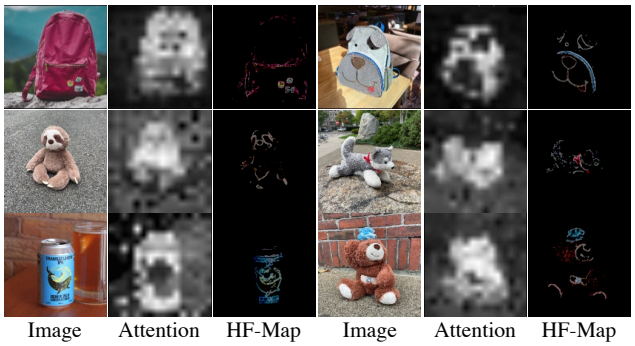

Figure 3. **Visualization of the focus region** of *ID extractor* and *detail extractor*. "Attention" refers to the attention map of the ID extractor backbone (DINOv2 [36]), while "HF-Map" refers to the high-frequency map used in the detail extractor. These two modules focus on global and local information in complementary.

extract the high-frequency regions using these high-pass filters, then extract the RGB colors using the Hadamard product. We also add an eroded mask $\mathbf{M}_{\text{erode}}$ to filter out the information near the outer contour of the target object.

As visualized in Fig. 3, the tokens produced by DINOv2 focus more on the overall structure, leaving it hard to encode the fine details like the logos of the backpack in the first row. In contrast, the high-frequency map could help take care of these details as a complementary.

Shape control. We use a shape mask to indicate the object's gestures. To simulate the user input, we downsample the ground truth masks with different ratios and apply random dilation/erosion to remove the details. To keep the ability to tackle a single box input, we set a probability of 0.3 to use the inner box region as the mask. During training, the object counter would be aligned with the shape mask. Thus, the users could control the target object's shape by drawing coarse shape masks during inference.

After getting the collage and the contour map, we concatenate them and feed them into the detail extractor. The detail extractor is a ControlNet-style [62] UNet encoder, which produces a series of detail maps with hierarchical resolutions.

3.3. Feature Injection

After getting the ID tokens and detail maps, we inject them into a pre-trained text-to-image diffusion model to guide the generation. We pick Stable Diffusion [41], which projects the images into latent space and conducts the probabilistic sampling using a UNet. We note the pre-trained UNet as $\hat{\mathbf{x}}_\theta$, it starts denoising from an initial latent noise $\epsilon \sim \mathcal{U}([0, 1])$ and takes the text embedding $\mathbf{c}$ as the condition to generate new image latent $\mathbf{z}_t = \alpha_t \hat{\mathbf{x}}_\theta(\epsilon, \mathbf{c}) + \sigma_t \epsilon$. The training supervision is a mean square error loss as:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,t}(\|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x}\|_2^2). \quad (2)$$

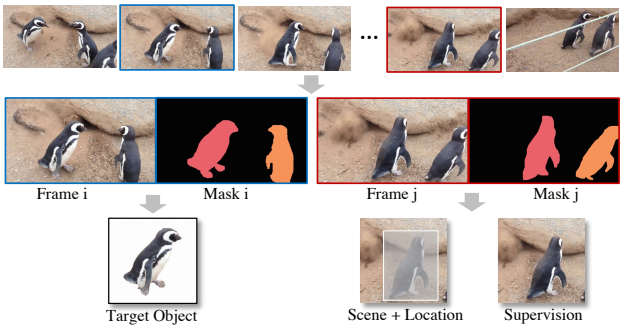

Figure 4. **Data preparation pipeline** for videos. Given a clip, we first sample two frames and get the masks for the instances within each frame. Then, we select an instance from one frame as the target object and treat the same instance on the other frame as the supervision (*i.e.*, the desired model output).

Table 1. **Statistics of datasets** used for training. "Variation" refers to whether an object enjoys local variations (*e.g.*, lighting, viewpoint, posture, *etc.*) within a data entry, while "quality" particularly refers to image resolution.

Dataset	Type	# Samples	Variation	Quality
YouTubeVOS [54]	Video	4,453	✓	Low
YouTubeVIS [57]	Video	2,883	✓	Low
UVO [51]	Video	10,337	✓	Low
MOSE [17]	Video	1,507	✓	High
VIPSeg [35]	Video	3,110	✓	High
BURST [1]	Video	1,493	✓	Low
MVImgNet [60]	Multi-view Image	104,261	✓	High
VitonHD [13]	Multi-view Image	11,647	✓	High
FashionTryon [64]	Multi-view Image	21,197	✓	High
MSRA-10K [4]	Single Image	10,000	✗	High
DUT [49]	Single Image	15,572	✗	High
HFlickr [14]	Single Image	4,833	✗	High
LVIS [21]	Single Image	118,287	✗	High
SAM (subset) [26]	Single Image	100,864	✗	High

$\mathbf{x}$ is the ground-truth image latent, t is the diffusion timestep, α_t, σ_t are denoising hyperparameters.

In this work, we replace the text embedding $\mathbf{c}$ as our ID tokens, which are injected into each UNet layer via cross-attention. For the detail maps, we concatenate them with UNet decoder features at each resolution. During training, we freeze the pre-trained parameters of the UNet encoder to preserve the priors and tune the UNet decoder to adapt it to our new task.

3.4. Training Strategies

Image pair collection. The ideal training samples are image pairs for "the same object in different scenes", which are not directly provided by existing datasets. As alternatives, previous works [47, 56] leverage single images and apply augmentations like rotation, flip, and elastic transforms. However, these naive augmentations could not well represent the realistic variants of the poses and views.

To deal with this problem, in this work, we utilize video datasets to capture different frames containing the same object. The data preparation pipeline is demonstrated in

Figure 5. **Qualitative comparison with reference-based image generation methods**, including Stable Diffusion [41], IP-Adapter [58], Paint-by-Example [56], and Graphit [16], where our AnyDoor better preserves the identity of the target object. Note that all approaches do *not* fine-tune the model on the test samples.

Fig. 4, where we leverage video segmentation/tracking data as examples. For a video, we pick two frames and take the masks for the foreground object. Then, we remove the background for one image and crop it around the mask as the target object. This mask could be used as the mask control after perturbation. For the other frame, we generate the box and remove the box region to get the scene image, and the unmasked image could serve as the training ground truth. The full data used is listed in Tab. 1, which covers a large variety of domains like nature scenes, virtual try-on, saliency, and multi-view objects.

Adaptive timestep sampling. Although the video data would be beneficial for learning the appearance variation, the frame qualities are usually unsatisfactory due to the low resolution or motion blur. In contrast, images could provide high-quality details and versatile scenarios but lack appearance changes. To take advantage of both video data and image data, we develop adaptive timestep sampling to make different modalities of data to benefit different stages of denoising training. The original diffusion model [41] evenly samples the timestep (T) for each training data. However, it is observed that the initial denoising steps mainly focus on generating the overall structure, the pose, and the view, and the later steps cover the fine details like the texture and colors. Thus, for the video data, we increase the possibility by 50% of sampling early denoising steps (500-1000) during training to better learn the appearance changes. For images, we increase 50% probabilities of the late steps (0-500) to learn how to cover the fine details.

4. Experiments

4.1. Implementation Details

Hyperparameters. We choose Stable Diffusion V2.1 [41] as the base generator. During training, we process the image resolution to 512×512. We choose Adam [25] optimizer with an initial learning rate of $1e^{-5}$. We train two versions of models, the original version only takes the box to indicate the location, and the plus version tasks shape masks. In this paper, if not specified with a shape mask, the results are produced by the original version.

Zoom-in strategy. During inference, given a scene image and a location box, we expand the box into a square with an amplifier ratio of 2.0. Then, we crop the square and resize it to 512×512 as the input for our diffusion model. Thus, we could deal with scene images with arbitrary aspect ratios and boxes for extremely small or large areas.

Benchmarks. For quantitative results, we construct a new benchmark with 30 new concepts provided by DreamBooth [42] for the target images. For the scene image, we manually pick 80 images with boxes in COCO-Val [31]. Thus we generate 2,400 images for the object-scene combinations. We also make qualitative analysis on VitonHD-test [13] to validate the performance for virtual try-on.

Evaluation metrics. On our constructed DreamBooth dataset, we follow DreamBooth [42] to calculate the CLIP-Score and DINO-Score, as these metrics could reflect the similarity between the generated region and the target object. In addition, we organize user studies with a group of 15 annotators to rate the generated results from the perspective of fidelity, quality, and diversity.

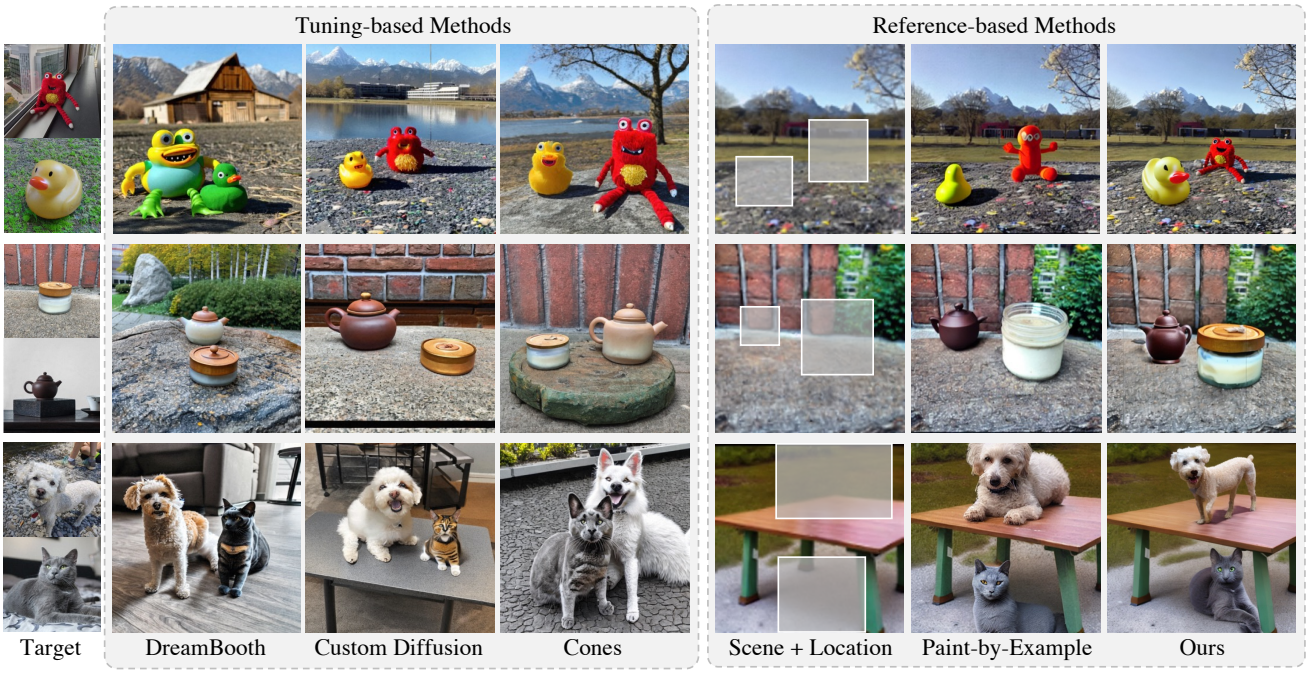

Figure 6. **Qualitative comparisons with existing alternatives for multi-subject composition**, including DreamBooth [42], Custom Diffusion [27], Cones [33], and Paint-by-Example [56], where our AnyDoor better preserves the object identity and harmoniously blends into the surroundings *without* any parameter tuning.

4.2. Comparisons with Existing Alternatives

Reference-based methods. In Fig. 5, we present the visualization results compared with previous reference-based methods. Paint-by-Example [56] and Graphit [16] support the same input format as ours, and they take a target image as input to edit a local region of a scene image without parameter tuning. IP-Adapter [58] is a universal method supporting image prompt, and we use its inpainting model for comparison. We also compare Stable Diffusion [41], which is a text-to-image model, and we use its inpainting version and give detailed text descriptions as the condition to conduct the generation for the text-described target.

Results show that previous reference-based methods could only keep the semantic consistency with distinguishing features like the dog face on the backpack, and coarse granites of patterns like the color of the sloth toy. However, as those new concepts are not included in the training category, their generation results are far from ID-consistent. In contrast, our AnyDoor shows promising performance for zero-shot image customization with highly-faithful details.

Tuning-based methods. Customized generation is extensively explored. Previous works [10, 18, 33, 42, 45] usually fine-tune a subject-specific text inversion to present the target object, thus making generations with arbitrary text prompts. They could better preserve the fidelity compared with previous reference-based methods, but have the following drawbacks: first, the fine-tuning usually requires 4-5 target images and takes nearly an hour; second, they could not specify the background scene and target locations; third, when it comes to multi-subject composition, the attributes of different subjects often mix together.

Table 2. **User study** on the comparison between our AnyDoor and existing reference-based alternatives. "Quality", "Fidelity", and "Diversity" measure synthesis quality, object identity preservation, and object local variation (*i.e.*, across four proposals), respectively. Each metric is rated from 1 (worst) to 4 (best).

	Quality (↑)	Fidelity (↑)	Diversity (↑)
Paint-by-Example [56]	2.71	2.10	**3.04**
Graphit [16]	2.65	2.11	2.84
AnyDoor (ours)	**3.04**	**3.06**	2.88

In Fig. 6, we include tuning-based methods for comparisons and also use Paint-by-Example [56] as the representative for previous reference-based methods. Results show that Paint-by-Example [56] performs well for trained categories like dog and cat (in row 3) but performs poorly for new concepts (row 1-2). DreamBooth [42], Custom Diffusion [27], and Cones [33] give better fidelity for new concepts but still suffer from the problem of "multi-subject confusion". In contrast, AnyDoor owns the advantages of both reference- and tuning-based methods, which could generate high-fidelity results for mult-subject composition without the need for parameter tuning.

User study. We organize a user study to compare Paint-by-Example [56], Graphit [16], and our model. We let 15

Figure 7. **Qualitative ablation studies** on the core components of AnyDoor. "HF-Map" stands for the high-frequency map in the detail extractor, while "ATS" refers to adaptive timestep sampling.

annotators rate 30 groups of images. For each group, we provide one target image and one scene image, and make each of the three models generates four predictions. We prepare detailed regulations and templates to rate the images for scores of 1 to 4 from three perspectives: "Fidelity", "quality", and "diversity". "Fidelity" measures the ability of ID preserving, and "Quality" counts for whether the generated image is harmonized without considering fidelity. As we do not encourage "copy-paste" style generation, we use "diversity" to measure the differences among the four generated proposals. The user-study results are listed in Tab. 2. It shows that our model owns obvious superiorities for fidelity and quantity, especially for fidelity. However, as [16, 56] only keeps the semantic consistency, but our methods preserve the instance identity. They naturally have a larger space for diversity. In this case, AnyDoor still gets higher rates than [16] and competitive results with [56], which verifies the effectiveness of our method.

4.3. Ablation Studies

We carry out extensive ablation studies to verify the effectiveness of our designs. We first validate the core components, then we dive into the details of the ID extractor and detail extractor to give an in-depth analysis.

Core components. As demonstrated in Fig. 7, given the same target object, scene, and location, we analyze the generated results with different model designs. We demonstrate the generation results of AnyDoor in the last column and remove each core component individually to observe the influences. We first change the backbone of our ID extractor from the DINOv2 [36] to CLIP image encoder [39], which is widely used in previous counterparts like [47, 56]. We find the generated results lose the identity features, and could only keep the semantic consistency. Then, we set the collage region from the high-frequency map to an all-zero map like the inpainting baselines [41, 62]. We find that the fine details degenerate compared with our full model (last column), like the logo of the bag (row 1), and the eye shape of the toy sloth (row 2). It shows that our frequency map effectively guides the generation of fine

Figure 8. **Qualitative analysis of using different backbones for the ID extractor.** "DINOv2*" refers to removing the background of the target object with a frozen segmentation model before feeding it into the DINOv2 model.

Table 3. **Quantitative analysis of using different backbones for the ID extractor.** Here, "G" refers to the global token, "P" refers to patch tokens, and "Seg" refers to removing the background of the target object with a frozen segmentation model.

	CLIP Score (↑)	DINO Score (↑)
VGG	71.7	27.7
CLIP (G+P)	73.8	31.5
DINOv2 (G)	73.1	35.4
DINOv2 (G+P)	81.0	64.1
DINOv2 (G+P) + Seg	82.1	67.8

structural details. We also make ablation for our adaptive timestep sampling (ATS) strategy. We replace ATS with an even distribution sampler and find the results present better diversity but are inferior for both image quality and fidelity.

ID extractor. We explore the key factors for designing the ID extractor. In Fig. 8, we compare VGG [46], CLIP [39] and DINOv2 [36] to extract the ID tokens. We conclude that DINOv2 [36] shows a dominant superiority for keeping the target identity. We also verify that it is significant to filter out the background information for the target object, and DINOv2 could extract cleaner and more discriminative features. Quantitative results are listed in Tab. 3, which are consistent with our visual analysis.

Detail extractor. We make multiple explorations for the collaged image. The CLIP and DINO scores are reported in Tab. 4, compared with non-collage, all these collaging methods bring notable improvements. To make better comparisons, we give visualization results in Fig. 9, which shows comparisons for no collage, pasting of the original target object, the noised inversion of the target object, the shuffled patches, and our high-frequency map. We observe a trade-off between fidelity and diversity. "Original image" presents the highest fidelity for both the robot and the dog, but the generated images seem like a copy-paste of the target. "None" shows the best diversity for the poses of the dog, but it lacks details like the badge of the dog and the

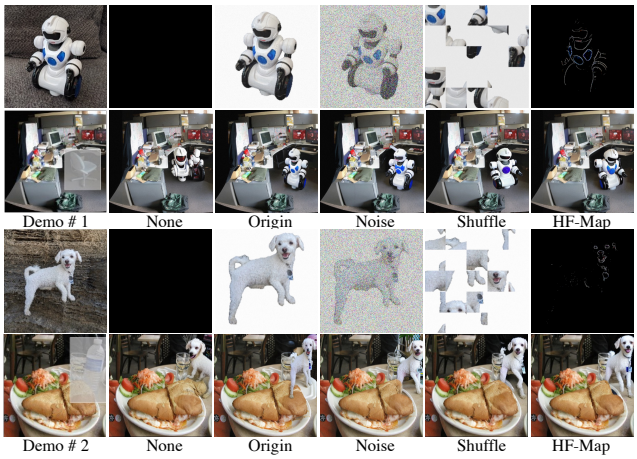

Figure 9. **Qualitative analysis of using different collages to extract details.** "None" means stitching the surroundings with an all-zero map. "Origin", "Noise", "Shuffle", and "HF-Map" refer to the original image with no background, noised image, patch-shuffled image, and the high-frequency map, respectively.

Table 4. **Quantitative analysis of using different collages to extract details.** It is noteworthy that, even the "Original Image" strategy best preserves the object identity, the object is with highly limited variation (*i.e.*, almost with the same form as the target) in the synthesis. Hence, we pick HF-map as our standard setting.

Strategy	CLIP Score (↑)	DINO Score (↑)
None (*i.e.*, all-zero map)	80.4	63.2
Original Image	82.2	68.8
Noise Image	81.6	68.1
Patch-shuffled Image	82.0	66.9
High-frequency Map	82.1	67.8

whole shape of the robots. Among those methods, the high-frequency map shows a satisfactory trade-off, which keeps the majority of the details but adjusts the dog and robot with proper poses and views.

4.4. More Applications

Virtual try-on. As shown in Fig. 10, without task-specific tuning, AnyDoor could give satisfactory performance for virtual try-on on VitonHD-test [13] and real-world scenarios for human with different sexes, ages, and races. Besides, AnyDoor supports users to draw coarse contour maps to control the style like tuck in or untuck (second row left) and shows strong generalization abilities for real-life scenarios with complex backgrounds.

Extensible controls. As demonstrated in Fig. 11, it is easy to extend AnyDoor to realize more fantastic functions like object moving, swapping, and reshaping. When taking a pose skeleton map as an additional control, AnyDoor could even serve as a strong baseline for human pose transfer.

The pipeline of object moving, swapping, and reshaping incorporates an additional inpainting model [41] and an interactive segmentation model [26]. We first get the mask of the object by clicking. Then, we use the inpainting model to fill the object's original position according to the scene background and apply the AnyDoor to re-generate it at the new location with optional shape control.

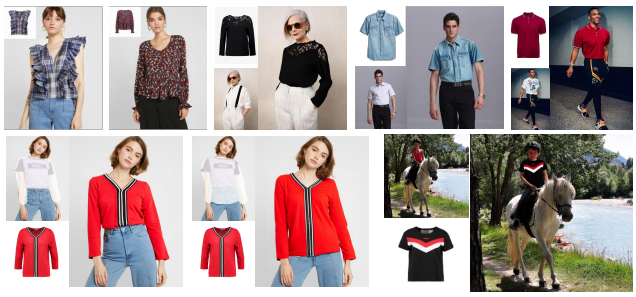

Figure 10. **Performance of AnyDoor on virtual try-on** on VitonHD-test [13] and real-life scenarios. AnyDoor could preserve the color, texture, and patterns of the target clothes and customize the garment shape (bottom left) with mask control.

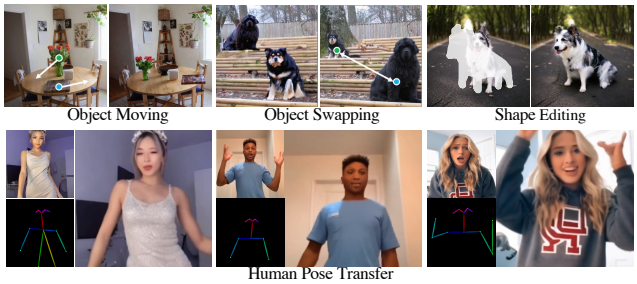

Figure 11. **Demonstrations for more applications of AnyDoor**, such as object moving, object swapping, shape editing, and human pose transfer. In row 2, the reference images and the novel poses are presented on the left with the generated results on the right.

In the second row of Fig. 11, when conducting human pose transfer, we add an extra ControlNet-copy on AnyDoor to control the human pose. Then, we train the model with the same configuration of Disco [50], a state-of-the-art human pose transfer method. The results are impressive that AnyDoor keeps the identity well for both the human faces and garments.

5. Conclusion

We present AnyDoor for object teleportation. The core idea is to use a discriminative ID extractor and a frequency-aware detail extractor to characterize the target object. Trained on a large combination of video and image data, we composite the object at the specific location of the scene image with effective shape control. AnyDoor provides a universal solution for general region-to-region mapping tasks and could be profitable for various applications.

Limitations. AnyDoor shows impressive results for keeping the object identification. However, it still struggles with fine details like the small characters or logos. This issue might be solved by collecting related training data, enlarging the resolution, and training better VAE decoders.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 2

[3] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 2

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *TIP*, 2015. 4

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 2

[6] Arantxa Casanova, Marlène Careil, Adriana Romero-Soriano, Christopher J Pal, Jakob Verbeek, and Michal Drozdzal. Controllable image generation via collage representations. *arXiv:2304.13722*, 2023. 3

[7] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *CVPR*, 2019. 2

[8] Haoxing Chen, Zhangxuan Gu, Yaohui Li, Jun Lan, Changhua Meng, Weiqiang Wang, and Huaxiong Li. Hierarchical dynamic image harmonization. In *ACMMM*, 2022. 2

[9] Hong Chen, Yipeng Zhang, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. *arXiv:2305.03374*, 2023. 2

[10] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. 6

[11] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *ICCV*, 2021. 3

[12] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: towards practical interactive image segmentation. In *CVPR*, 2022. 3

[13] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 4, 5, 8

[14] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, 2020. 2, 4

[15] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, 2022. 2

[16] Graphit Contributors. Graphit: A unified framework for diverse image editing tasks. https://github.com/navervision/Graphit, 2023. 5, 6, 7

[17] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 4

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2, 6

[19] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 2

[20] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, 2021. 2

[21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4

[22] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *ICML*, 2023. 2

[23] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *JSSC*, 1988. 3

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 3, 4, 8

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 6

[28] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 2

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[30] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv:2306.12624*, 2023. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[32] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *ICCV*, 2023. 3

[33] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023. 2, 6

[34] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023. 2

[35] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 4

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 3, 4, 7

[37] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023. 2

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 2020. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 7

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 5, 6, 7, 8

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 5, 6

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[44] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. In *WACV*, 2024. 3

[45] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv:2304.03411*, 2023. 6

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 7

[47] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, 2023. 2, 3, 4, 7

[48] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. In *SIGGRAPH*, 2010. 2

[49] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 4

[50] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv:2307.00040*, 2023. 8

[51] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 4

[52] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv:2305.10431*, 2023. 2

[53] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023. 2

[54] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv:1809.03327*, 2018. 4

[55] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV*, 2022. 2

[56] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023. 2, 3, 4, 5, 6, 7

[57] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 4

[58] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023. 2, 5, 6

[59] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv:2304.06790*, 2023. 2

[60] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 4

[61] Ziyang Yuan, Mingdeng Cao, Xintao Wang, Zhongang Qi, Chun Yuan, and Ying Shan. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv:2310.19784*, 2023. 2

[62] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 4, 7

[63] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 2023. 2

[64] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *ACMMM*, 2019. 4