

# Integrating View Conditions for Image Synthesis

Jinbin Bai<sup>1</sup>, Zhen Dong<sup>2</sup>, Aosong Feng<sup>3</sup>, Xiao Zhang<sup>1</sup>, Tian Ye<sup>4</sup>, Kaicheng Zhou<sup>1\*</sup>

<sup>1</sup>Collov Labs

<sup>2</sup>University of California, Berkeley

<sup>3</sup>Yale Univeristy

<sup>4</sup>The Hong Kong University of Science and Technology (Guangzhou)  
jinbin.bai@u.nus.edu, zhendong@berkeley.edu, aosong.feng@yale.edu,  
{xiao,caseyz}@collov.com, owentianye@hkust-gz.edu.cn

## Abstract

In the field of image processing, applying intricate semantic modifications within existing images remains an enduring challenge. This paper introduces a pioneering framework that integrates viewpoint information to enhance the control of image editing tasks, especially for interior design scenes. By surveying existing object editing methodologies, we distill three essential criteria — consistency, controllability, and harmony — that should be met for an image editing method. In contrast to previous approaches, our framework takes the lead in satisfying all three requirements for addressing the challenge of image synthesis. Through comprehensive experiments, encompassing both quantitative assessments and qualitative comparisons with contemporary state-of-the-art methods, we present compelling evidence of our framework’s superior performance across multiple dimensions. This work establishes a promising avenue for advancing image synthesis techniques and empowering precise object modifications while preserving the visual coherence of the entire composition.

## 1 Introduction

Applying intricate semantic modifications to existing images is a longstanding and fascinating endeavor in image processing. The primary objective of image manipulation is to synthesize an image that retains most of the original semantic content while altering specific elements. In recent years, the landscape of image-to-image models has witnessed a proliferation of methodologies, spanning the spectrum of Generative Adversarial Network (GAN)-based and diffusion-based approaches, encompassing both zero-shot and fine-tuned strategies, all dedicated to addressing this complex task. Faced with this multitude of approaches, a natural inquiry arises: does a given method genuinely fulfill the requirements of precise object modification, and by which criteria is a commendable solution for entity manipulation characterized?

To answer the question, we investigate various image editing applications and make some observations. First, an ex-

\*Corresponding Author

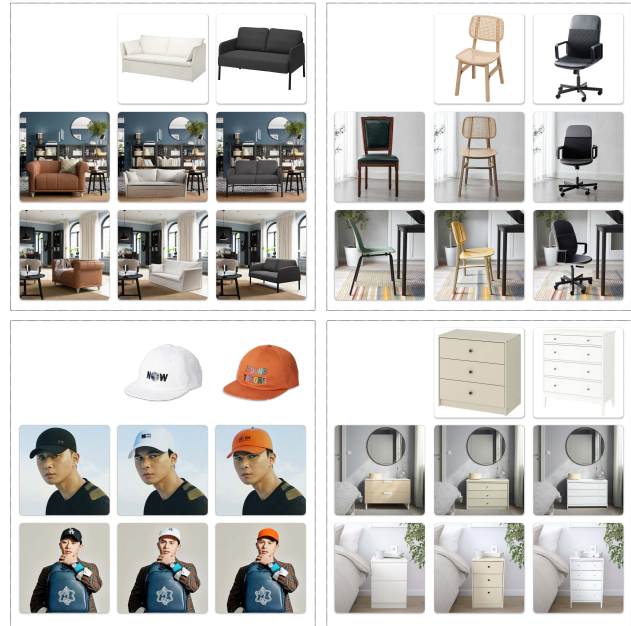


Figure 1: Applications of the proposed method. Our method can replace each object in the left column with the one in the upper row, ensuring not only consistency in the synthesized object but also, by introducing viewpoint conditions to the model, enabling precise control over the object’s pose and thus enhancing visual harmony.

cellent framework for object modification needs to satisfy the consistency in both the shape and color of the object. Approaches such as Paint-by-Example [Yang *et al.*, 2023] and Paint-by-Sketch [Kim *et al.*, 2023], where a reference image is utilized as input for the CLIP model, unfortunately falter in maintaining this object consistency. Conversely, Dream-Booth [Ruiz *et al.*, 2023] and its successors [Kumari *et al.*, 2023], exhibit competence in synthesizing objects while preserving their shape and color. Nevertheless, these approaches struggle with precise synthesis of the object’s spatial position and orientation, making them difficult to apply in entity replacement.

Second, in the pursuit of image editing tasks, despite the presence of textual or visual guidance, numerous intricacies

Aspect	PBE	DreamBooth	ControlNet	GLIGEN	ViewControl (Ours)
Consistency		✓			✓
Controllability	✓		✓	✓	✓
Harmony		✓	✓	✓	✓

Table 1: Evaluation Criteria for Image Synthesis Methods: Consistency, Controllability, and Harmony. Consistency refers to the synthesized object being consistent with the reference object. Controllability refers to the ability to manipulate the shape, color, angle, and position of the synthesized object through input. Harmony refers to the coherence between the synthesized object and the original image in terms of lighting, shadow, angle, and positional logic.

often evade direct control and depend on random seed values. For instance, variables like the precise position and orientation of the synthesis object tend to be stochastic. The issue of object position during synthesis can be effectively mitigated through the application of bounding box constraints, as exemplified by GLIGEN [Li *et al.*, 2023], but with bounding box constraints alone, the object’s synthesis remains location-specific without specifying its orientation. Recently, ControlCom [Zhang *et al.*, 2023a], AnyDoor [Chen *et al.*, 2023b], and DreamPaint [Seyfioglu *et al.*, 2023] have also made significant advancements in consistency and controllability. However, without prior reference to the object’s corresponding camera viewpoint, synthesizing specific object directions remains a persistent challenge even given a bounding box.

Third, the resulting synthesis must meet certain quality standards, characterized by harmony in terms of illumination, shading, and logical consistency. Concerning illumination and shading, it is vital that the shadow cast by the synthesized object conforms to the prevailing directional cues within the image. And the reflections displayed by the synthesized object should harmonize with its intrinsic attributes. Furthermore, logical consistency encompasses aspects such as the object’s angle, position, and quantity. In summary, the synthesized object must be harmoniously integrated with its surroundings, thereby establishing an optimal state of coordination.

This paper presents a novel framework that enhances existing models with awareness of viewpoint information, thereby enabling improved control over text-to-image diffusion models, such as Stable Diffusion. This advancement leads to a more controllable approach for image editing tasks. Our proposed pipeline aptly meets all the previously mentioned requirements, with a particular focus on the aspect of controlled pose adjustment, as detailed in Table 1.

To comprehensively evaluate our framework, we assess its performance across various applications, including entity replacement and angle adjustments. This comprehensive evaluation encompasses a wide range of scenarios, such as virtual try-on and interior design. Notably, we demonstrate that our method yields favorable results across multiple dimensions, even in cases where extensive training is not a prerequisite.

## 2 Related Work

In this section, we first introduce the work related to consistency, controllability, and harmony, and then introduce the work related to novel pose synthesis.

### 2.1 Few Shot Personalization and Customization

In the context of utilizing just a few reference images, several methods have been proposed to grasp the underlying concept, whether it’s a particular theme, style, object, or character. These methods include LoRA [Hu *et al.*, 2021], DreamBooth [Ruiz *et al.*, 2023], Textual Inversion [Gal *et al.*, 2022], HyperNetworks [Ha *et al.*, 2016], and their successors. While these methods and their combinations have opened up avenues for personalized or customized applications with minimal training data, they still rely on having multiple images at their disposal, making it challenging for them to envision different perspectives of an object using just a single image. Furthermore, achieving fine-grained, angle-controllable generation remains a formidable task for them.

Few-shot personalization and customization technologies will be popular in e-commerce, as every product in e-commerce catalogs typically has multiple images taken from different angles, which is naturally a good fit for these approaches.

### 2.2 Conditional and Controllable Image Editing and Generation

In addition to text-to-image translation [Radford *et al.*, 2021; Bai *et al.*, 2022; Rombach *et al.*, 2022a; Udandarao, 2022], image-to-image translation [Isola *et al.*, 2017; Zhang *et al.*, 2023b; Ye *et al.*, 2023; Chen *et al.*, 2023a; Jiang *et al.*, 2023a; Jiang *et al.*, 2023b; Feng *et al.*, 2024] is a kind of image-conditioned image synthesis, which has been instrumental in image editing and generation, allowing for the preservation of most of the existing semantic content while making specific alterations to particular elements within the source image. These elements can be categorized as style, object, background, and more.

In terms of style, style transfer techniques have played a significant role in advancing controllable image editing and generation. Initially rooted in artistic style transfer, neural style transfer methods, as demonstrated by [Gatys *et al.*, 2015], have evolved to grant users greater control over the degree of stylization and the independent manipulation of content and style [Johnson *et al.*, 2016]. These developments have facilitated more controlled artistic transformations.

More recently, diffusion models [Ho *et al.*, 2020; Ho and Salimans, 2022; Zhao *et al.*, 2023] have emerged as the new state-of-the-art family of deep generative models. Representative models such as Stable Diffusion [Rombach *et al.*, 2022a], yield impressive performance on conditional image generation, enabling control over various aspects of the gen-

erated content, surpassing those GAN-based [Goodfellow *et al.*, 2020] methods which dominated the field for the past few years. Diffusion models are equipped to accommodate various conditions, whether in the form of textual [Radford *et al.*, 2021] or visual [Zhang *et al.*, 2023b] inputs, making the process of image editing and generation more controllable. However, none of these approaches offer fine-grained control over certain image details, such as lighting, shadows, and object poses.

### 2.3 Image Harmonization

Image harmonization, as explored in previous work [Tsai *et al.*, 2017], focuses on adjusting the illumination and shading between the foreground and background. While several approaches have succeeded in appearance adjustments, they still struggle to address the geometric inconsistencies that may arise between the foreground and background. To handle issues related to inconsistent camera viewpoints, various methods [Chen and Kae, 2019; Lin *et al.*, 2018] have been proposed to estimate warping parameters for the foreground, aiming for geometric correction. However, these methods typically predict affine or perspective transformations, which may not effectively address more complex scenarios, such as synthesizing foreground objects with novel views or generalizing them to non-rigid objects like humans or animals.

### 2.4 Single Image to 3D

Before the emergence of CLIP [Radford *et al.*, 2021] and large-scale 2D diffusion models [Rombach *et al.*, 2022a], the conventional approach involved learning 3D priors using either synthetic 3D data or real scans [Yi *et al.*, 2023; Yi *et al.*, 2022]. Unlike 2D images, 3D data can be represented in various formats and numerous representations.

DTC123 [Yi *et al.*, 2024] collaborates the teacher diffusion model and 3D student model with the time-step curriculum to significantly improve the photo-realism and multi-view consistency of Image-to-3D generation. Zero123 [Liu *et al.*, 2023c] is a view-conditioned 2D diffusion model used to synthesize multiple views for object classes lacking 3D assets. It demonstrates that rich geometric information can be extracted directly from a pre-trained Stable Diffusion model, eliminating the need for additional depth information. Building on this, One-2345 [Liu *et al.*, 2023b] utilizes Zero123 to achieve single-image-to-3D mesh conversion.

## 3 Method

Given an input image of dimensions  $H$  by  $W$ , a reference image  $\mathbf{x}_r \in \mathbb{R}^{H_r \times W_r \times 3}$  containing the reference object, and a prompt description  $\mathbf{c}$  (e.g., "Adjust the hat up 10 degrees" or "Replace the laptop on the desk with an apple" as shown in Fig. 1), our objective is to synthesize an output image  $\mathbf{y}$  using the information from  $\{\mathbf{x}_s, \mathbf{x}_r, \mathbf{c}\}$ . The goal is to maintain the visual consistency between the output image  $\mathbf{y}$  and the source image  $\mathbf{x}_s$ , while only modifying the object mentioned in  $\mathbf{c}$  by the specified pose. Furthermore, when introducing a new object, it is crucial to harmoniously integrate it with the overall composition of the entire image.

This task is particularly intricate due to several inherent challenges. Firstly, the model needs to comprehend the object

in the reference image, capturing both its shape and texture while disregarding background noise. Secondly, it is essential to generate a transformed version of the object (varied pose, size, illumination, etc.) that seamlessly integrates into the source image. Furthermore, the synthesized object must align with the original object’s angle as specified in  $\mathbf{c}$ . Lastly, the model must inpaint the surrounding region of the object to produce a realistic image, ensuring a smooth transition at the merging boundary.

Therefore, we adopt the Divide and Conquer principle<sup>1</sup>, and break down this intricate problem into easier sub-problems and solve them one by one. To be specific, we address this challenge by combining various generative models, with our combination model is conditioned on the source image  $\mathbf{x}_s$ , text prompt  $\mathbf{c}$ , and reference image  $\mathbf{x}_r$ .

We mathematically formulate our approach as follows:

$$P(\mathbf{y}|\mathbf{x}_s, \mathbf{c}, \mathbf{x}_r) = P(O_s, A_s|\mathbf{x}_s, \mathbf{c}) \cdot P(A_c|\mathbf{c}, A_s) \cdot P(O_r|\mathbf{x}_r, A_c) \cdot P(\mathbf{y}|\mathbf{x}_s, O_s, O_r) \quad (1)$$

In this equation, we use  $O_s$  to represent the object within the source image  $\mathbf{x}_s$ ,  $A_s$  to denote the angle of this object in  $\mathbf{x}_s$ ,  $O_r$  to signify the reference object in the reference image  $\mathbf{x}_r$ , and  $A_c$  to indicate the specific angle extracted from the text prompt  $\mathbf{c}$ .

The four probabilistic models on the right side of the equation encompass various essential processes within our framework. These processes include object and angle extraction from the source image, angle extraction from the text prompt (as elaborated in Sec. 3.1), synthesis of the reference object (as elaborated in Sec. 3.2), and the ultimate image synthesis procedure (as elaborated in Sec. 3.3). We will now delve into each of these parts and explain how they are integrated.

### 3.1 LLM Planner

With in-context-learning and chain-of-thoughts reasoning capabilities, large language models (LLMs) have demonstrated remarkable proficiency in following natural language instructions and completing real-world tasks [Xie *et al.*, 2023; Peng *et al.*, 2023; Liu *et al.*, 2023a]. Given a text prompt, our LLM Planner is capable of generating a sequence of commands and prompts related to our task.

### 3.2 Pose Estimation and Synthesis

In this stage, we present pose representation, pose estimation from a single object image, and the synthesis of an object image given a specific pose.

#### Pose Representation

To effectively represent the pose of an object within an image, we employ two fundamental components: the relative camera rotation matrix ( $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ ) and the relative camera translation vector ( $\mathbf{T} \in \mathbb{R}^3$ ). These elements collectively encapsulate essential information regarding the object’s viewpoint and orientation relative to the camera’s perspective.

<sup>1</sup><https://llyasviel.github.io/Style2PaintsResearch/>

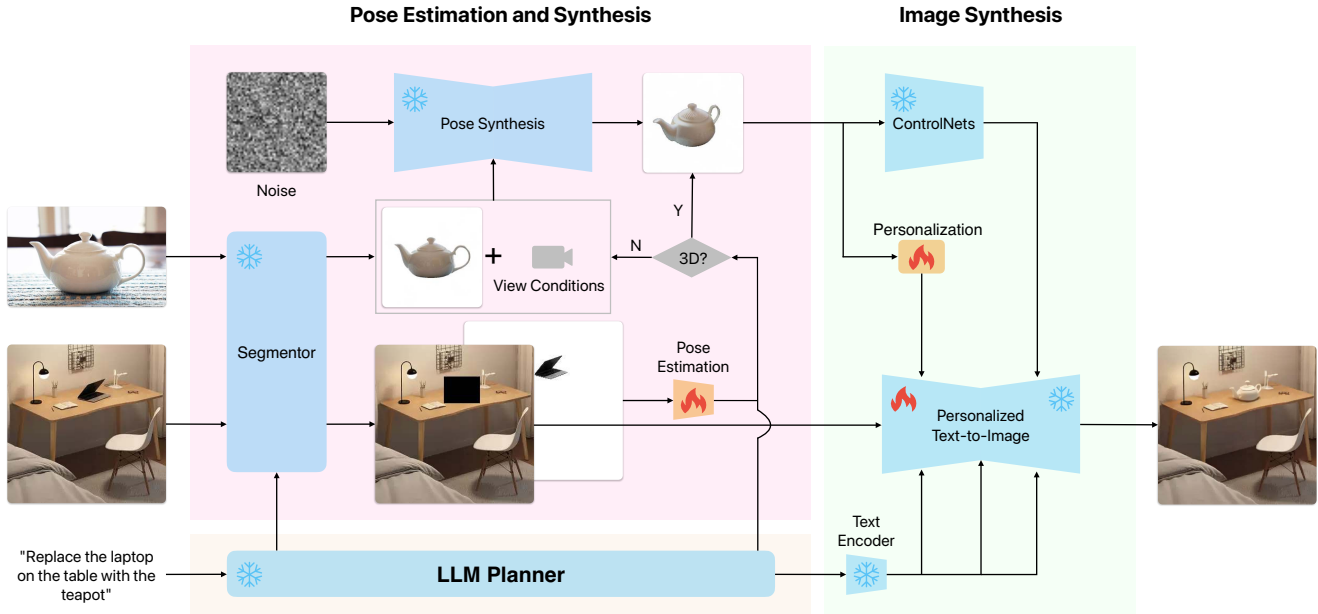


Figure 2: An illustrative overview of our method, which is designed for synthesizing an object with a user-specified view into a scene. "3D?" denotes whether a 3D model is available. Our approach consists of three components: Large Language Model (LLM) Planner (Sec. 3.1), Pose Estimation and Synthesis (Sec. 3.2), and Image Synthesis (Sec. 3.3). First, the LLM Planner is adopted to obtain the objects' names and pose information based on the user's input. Second, a segmentation module is adopted to remove the background from the specific object, followed by a pose estimation module to obtain its accurate pose. A pose synthesis module is then applied to synthesize the reference object respecting specific view conditions. Third, a personalized pre-trained diffusion model and ControlNets are adopted to produce the final synthesis. They ensure that the target object harmoniously melds with its surroundings, aligning with the user-specified view, while maintaining consistency in the object's representation. **Flames** and **snowflakes** refer to learnable and frozen parameters, respectively.

**Relative Camera Rotation ( $\mathbf{R}$ ).** The matrix  $\mathbf{R}$  characterizes the rotation transformation that aligns the object's coordinate system with that of the camera. It encompasses the angular changes required to transition from the object's intrinsic orientation to the camera's frame of reference.

**Relative Camera Translation ( $\mathbf{T}$ ).** The vector  $\mathbf{T}$  denotes the translation in three-dimensional space necessary to position the camera viewpoint with respect to the object. It signifies the displacement along the  $x$ ,  $y$ , and  $z$  axes, allowing the object's placement within the scene to be determined.

Together, the relative camera rotation ( $\mathbf{R}$ ) and translation ( $\mathbf{T}$ ) form a comprehensive pose representation, providing a detailed description of the object's spatial orientation and location within the image.

### Pose Estimation

We train a pose estimation model, building upon the foundation of the current image understanding model. The training supervision is

$$\Theta = \arg \min_{\Theta} \mathbb{E}_{\mathbf{x}} \left[ \left\| \hat{\mathbf{R}}_{\Theta}(\mathbf{x}) - \mathbf{R} \right\|_2^2 + \left\| \hat{\mathbf{T}}_{\Theta}(\mathbf{x}) - \mathbf{T} \right\|_2^2 \right]$$

Here,  $\mathbf{x}$  represents the image,  $\mathbf{R}$  and  $\mathbf{T}$  denote the relative camera rotation and translation, respectively.  $\Theta$  corresponds to the network parameters of our pose estimation model.

Given an object image, our pose estimation model predicts the corresponding relative camera rotation and translation based on the default camera view.

### Pose Synthesis

We use a view-conditioned diffusion model, Zero123 [Liu *et al.*, 2023c], to generate multi-view images and corresponding pose images. The input to Zero123 consists of a single RGB image  $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$  that encompasses the object requiring alignment, and a relative camera transformation rotation  $\mathbf{R} \in \mathbb{R}^{3 \times 3}$  and translation  $\mathbf{T} \in \mathbb{R}^3$ , which is the viewpoint condition control. The output of Zero123 is a synthesized image  $\hat{\mathbf{x}}_{\mathbf{R}, \mathbf{T}}$  capturing the same object from the perspective defined by the transformed camera view:

$$\hat{\mathbf{x}}_{\mathbf{R}, \mathbf{T}} = \mathbf{f}(\mathbf{x}, \mathbf{R}, \mathbf{T})$$

where  $\mathbf{f}$  denotes the freezing model Zero123.

Constrained by its limited generalization capacity, Zero123 excels primarily in a select few categories. Therefore, when a reference 3D object is available, we can directly specify view conditions for the reference object to obtain the desired image perspective. All images presented in this paper are synthesized by Zero123.

### 3.3 Image Synthesis

Although pose alignment has been achieved, it's possible that the object in the synthesized reference image may have a different size and position compared to the mask in the source



Figure 3: **Qualitative comparison with reference-based image synthesis methods**, where "PbE" denotes Paint-by-Example and "PbS" denotes Paint-by-Sketch.

image. Therefore, our initial step is to apply padding, either on the left and right or on the top, to the bounding box region of the object in the synthesized reference image. This ensures that the aspect ratio of the object mask in the synthesized reference image matches that of the mask in the source image. Following this, we resize the region that we just padded, ensuring the resized region aligns precisely with the bounding box part in the source image. As a result, we obtain the reference object image  $O_r$ .

With a source image  $I_s$  containing a bounding box mask of the object to be edited and the reference object image  $O_r$  with the corresponding camera view, we employ the personalized Stable Diffusion Inpaint Model, controlled by edge and color information, to synthesize the target image.

Why not simply overlay the synthesized object onto the original image? The reason lies in the fact that synthesized objects are typically not perfect, they may exhibit some degree of artifacts or distortion. Consequently, during the image synthesis process, we can only refer to the synthesized object rather than relying on it entirely.

### 3.4 All in One

We integrate all previously mentioned modules to establish an image synthesis framework that allows for view control, as illustrated in Fig. 2. First, we obtain essential object details via the LLM Planner, including angle and object name. Second, we synthesize an appropriate target object image through pose estimation and synthesis. Finally, off-the-shelf diffusion models and associated plugins are employed to achieve pose-controllable image editing.

## 4 Experiment

### 4.1 Implementation Details

The following components are utilized in our implementation: GPT-4 [Achiam *et al.*, 2023] as our LLM Planner, Segment-Anything [Kirillov *et al.*, 2023] as our segmentation model, Zero-123 [Liu *et al.*, 2023c] as our pose alignment model, and Stable Diffusion v1.5 [Rombach *et al.*, 2022b]

<sup>2</sup> and ControlNet 1.1 [Zhang *et al.*, 2023b] as our synthesis models. Additionally, we have developed a pose estimation model, which is trained on DINO-v2 [Oquab *et al.*, 2023].

In terms of training data, we initially curated product images spanning various categories from publicly available sources on the internet, all captured from a consistent viewpoint, which we have designated as the default camera perspective. Subsequently, employing existing zero-shot novel view synthesis models, we synthesized batches of images, each image batch corresponding to different relative camera viewpoints of each object. In total, our dataset comprises approximately 48.6k images, along with their corresponding relative camera view labels, and we've split them into training and test sets, following an 8:2 ratio. Furthermore, it's important to note that the test set is reserved only for testing.

### 4.2 Comparisons

In our comparisons, we have selected recently published open-source state-of-the-art image-driven image editing methods, namely Paint-by-Example [Yang *et al.*, 2023], Paint-by-Sketch [Kim *et al.*, 2023], as our baselines. Figure 3 provides qualitative comparisons and Table 2 provides quantitative comparisons of these methods. We can see that our method consistently achieves superior evaluation results in consistency and harmony.

Why don't we employ methods like CLIP Score for quantitative analysis of consistency? Our rationale is rooted in the belief that feature extractors like CLIP often result in the loss of fine-grained image details, which also explains why PbE struggles to achieve consistency. Consequently, evaluating fine-grained generation with a coarse-grained feature extractor may not yield meaningful results. Furthermore, numerous studies have indicated that quantitative evaluation metrics may not consistently align with human perceptual judgments. Given these considerations, we primarily rely on human evaluations to quantitatively assess the performance of our approach and only evaluate the aesthetics score with feature extractors <sup>3</sup>.

### 4.3 Ablation Study

In this section, we will begin by discussing the selection process for the pose estimation module backbone, and then demonstrate the essentiality of each component within our image synthesis module. Subsequently, we will show the necessity and robustness of our view conditions. Lastly, we will explain the reason behind our decision not to opt for a two-stage synthesis approach.

#### Effects of Using Different Backbones for Pose Estimation

We report the prediction error (MAE, mean absolute error) of our pose estimation module with different backbones, including ResNet-50, CLIP Image Encoder, ViT, DINO-v2 [Oquab *et al.*, 2023]. All the experiments are trained on one A100

<sup>2</sup>It should be noted that while we believe SDXL [Podell *et al.*, 2023] performs better in terms of harmony, its adoption has been temporarily withheld in the current version due to its limited community support and lack of widespread use. We will move to SDXL once related ControlNets are done.

<sup>3</sup><https://github.com/kenjiqqa/aesthetics-scorer>

Methods	Consistency ( $\uparrow$ )	Harmony ( $\uparrow$ )	Aesthetics ( $\uparrow$ )	Controllability ( $\uparrow$ )
Paint-by-Example	2.67	2.61	4.92	1.93
Paint-by-Sketch	2.79	2.21	3.93	1.87
ViewControl (Ours)	<b>4.44</b>	<b>4.54</b>	<b>5.37</b>	<b>4.53</b>

Table 2: **Quantitative Comparisons.** "Consistency" evaluates the similarity between the reference object and the synthesized object; "Controllability" evaluates the pose and view of the synthesized object; "Harmony" evaluates the uniformity of pose and view relationships between the background and foreground elements, and "Aesthetics" denotes the machine evaluation with an aesthetics-scorer. For "Consistency", "Controllability" and "Harmony" evaluations, we collect 15 reviews for each of the 30 sets of synthesized images, with each set comprising three different synthesis methods. Scores were assigned on a scale from 1 to 5, with 1 denoting "terrible", 2 denoting "poor", 3 denoting "average", 4 denoting "good", and 5 denoting "excellent". Besides, the aesthetics-scorer rates each image on a scale from 1 to 10 (higher is better).

Methods	#Params	GFLOPs	MAE ( $\downarrow$ )	RMSE ( $\downarrow$ )
ResNet-50	26.20 M	4.13 G	4.31	7.45
CLIP	87.88 M	4.37 G	3.28	10.59
ViT	86.34 M	16.86 G	1.65	6.56
DINO-v2	85.61 M	21.96 G	<b>0.80</b>	<b>5.01</b>

Table 3: **Quantitative ablation studies on the effects of using different backbones for pose estimation**, where MAE and RMSE denote mean absolute error and root mean squared error, respectively.



Figure 4: **Qualitative ablation studies on the effects of core components in image synthesis**, where "Personal" denotes the personalization module, "Color CN" denotes the ControlNet which controls the color, "Edge CN" denotes the ControlNet which controls the edge, "CNs" denotes all the ControlNets, and "Full Model" denotes with all components.

40GB GPU. As seen in Table 3, DINO-v2 achieves better performance with a prediction error of less than 1.

### Effects of Image Synthesis Core Components

It's evident from Figure 4 that each component of image generation is indispensable. The personalization module plays a pivotal role in determining the overall object condition, while multiple ControlNets govern the precise object-specific details.

### Effects of View Conditions

From Figure 5, we can make two key observations:

**Necessity of View Conditions.** If the provided view conditions display substantial error, or if no view conditions are provided, the object generation process tends to favor a semantic orientation within the source image (such as backing against a wall) or the direction most frequently observed during training (typically the front).



Figure 5: **Qualitative ablation studies on the effects of view conditions**, where "Slight" denotes an error range of 0-20 degrees viewing conditions, "Moderate" denotes an error range of 20-40 degrees viewing conditions, "Severe" denotes an error range of 40-90 degrees viewing conditions, and "Perfect" denotes there are no errors.

**Robustness of View Conditions.** View conditions exhibit a certain degree of robustness. Specifically, predictions remain relatively stable within an error range of 20 degrees.

These observations further underscore the dual significance of view conditions, emphasizing both their necessity and robustness.

### Effects of 2-Stage Synthesis

Although a two-stage synthesis approach, involving the initial removal of the original object and subsequent addition of the new object, may mitigate the impact on the original image in certain scenarios, as exemplified by the eyes under the hat in Figure 1. Our framework adheres to more general principles. These principles allow for the possibility of significant disparities in shape between the original object and the new object.

In our experiments, the act of removing the original object often results in the generation of redundant information at the inpaint position. Consequently, when incorporating a new object, if the mask area for it is not adequately large, this redundant information cannot be effectively eliminated.



Figure 6: **Qualitative ablation studies on the effects of 2-stage synthesis**, where "2-Stage-Mid" denotes the initial inpainting result of the 2-stage synthesis, "2-Stage-Final" denotes the subsequent inpainting result of the 2-stage synthesis, and "1-Stage" denotes the approach that we choose, which involves using only one inpainting step per synthesis.

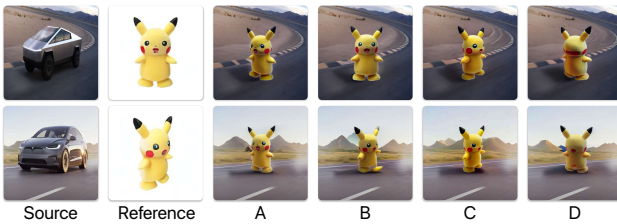


Figure 7: Visualization results on various poses of Pikachu synthesis with user input A = "Replace the car with Pikachu from the same view as the car", B = "Replace the car with Pikachu from the same view as Input Pikachu", C = "Replace the car with Pikachu from the same view as the car, and then turn Pikachu left with 90 degrees", D = "Replace the car with Pikachu from the same view as the car, and then turn Pikachu left with 180 degrees".

As a remedy, we employ a larger mask, the bounding box, and opt for a one-stage synthesis approach. Figure 6 visually illustrates such scenarios.

## 5 Discussion

### 5.1 Additional Visualization Results

We present three groups of visualization results in Figures 7, 8, and 9, to show that our method is also applicable in other scenarios. The first group aims to validate the effectiveness of various views in reference images, and the effectiveness of synthesizing various views of target images. The second group aims to demonstrate the effectiveness of intricate object synthesis and multiple object editing. The final group provides another visualization result on Mona Lisa portrait.

### 5.2 Applications

**Virtual Try-on.** Customers can use our approach to try on different shoes, clothing items, accessories, hats, or even hairstyles virtually. By providing a reference image of themselves and a description like "replace shoes/hats with Object\_B," users can see how different fashion items would look on them with specific viewpoint, without physically wearing

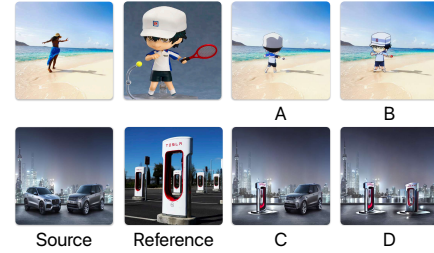


Figure 8: Visualization results on intricate object synthesis and multiple object editing with user input A = "Replace the girl on the beach with prince of tennis from the same view as the girl", B = "Replace the girl on the beach with prince of tennis from the same view as reference prince of tennis", C = "Replace the right car with Tesla Supercharger", D = "Replace the car with Tesla Supercharger".

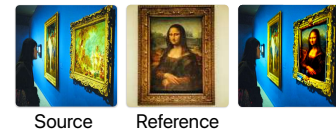


Figure 9: Visualization result on Mona Lisa portrait with user input "Replace the right portrait with Mona Lisa portrait"

them. Additionally, our framework enables precise adjustments to customize the placement and orientation of fashion items.

**Interior Design.** The framework allows users to specify actions like "replace Object\_A with Object\_B" to switch out furniture items, enabling them to visualize how different pieces would fit into their rooms. Moreover, users can make detailed adjustments to fine-tune the placement of furniture and decor items to create harmonious and aesthetically pleasing interiors.

### 5.3 Future Work

**End-to-End Integration with Latent Space View Control.** While our current approach relies on explicit pose estimation and pose synthesis steps, a promising direction for future work is to further streamline our framework into an end-to-end solution [Kumari *et al.*, 2024] that seamlessly integrates view conditions control within a latent space. Such latent space view control approach has the potential to significantly enhance inference efficiency and elevate the user experience, making image editing even more accessible and efficient.

## 6 Conclusion

We present a novel framework that integrates view conditions for image synthesis, which enhances the controllability of image editing tasks. Our framework effectively addresses crucial aspects of image synthesis, including consistency, controllability, and harmony. Through both quantitative and qualitative comparisons with recently published open-source state-of-the-art methods, we have showcased the favorable performance of our approach across various dimensions.

## Acknowledgements

This work is supported by Collov Inc. Kaicheng Zhou is the corresponding author.

## References

- [Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [Bai *et al.*, 2022] Jinbin Bai, Chunhui Liu, Feiyue Ni, Hao-fan Wang, Mengying Hu, Xiaofeng Guo, and Lele Cheng. Lat: Latent translation with cycle-consistency for video-text retrieval. *arXiv preprint arXiv:2207.04858*, 2022.
- [Chen and Kae, 2019] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8415–8424, 2019.
- [Chen *et al.*, 2023a] Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13106–13117, 2023.
- [Chen *et al.*, 2023b] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.
- [Feng *et al.*, 2024] Aosong Feng, Weikang Qiu, Jinbin Bai, Kaicheng Zhou, Zhen Dong, Xiao Zhang, Rex Ying, and Leandros Tassioulas. An item is worth a prompt: Versatile image editing with disentangled control. *arXiv preprint arXiv:2403.04880*, 2024.
- [Gal *et al.*, 2022] 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. *arXiv preprint arXiv:2208.01618*, 2022.
- [Gatys *et al.*, 2015] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.
- [Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.
- [Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [Jiang *et al.*, 2023a] Jingxia Jiang, Jinbin Bai, Yun Liu, Junjie Yin, Sixiang Chen, Tian Ye, and Erkang Chen. Rsfdmnet: Real-time spatial and frequency domains modulation network for underwater image enhancement. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2560–2564. IEEE, 2023.
- [Jiang *et al.*, 2023b] Jingxia Jiang, Tian Ye, Jinbin Bai, Sixiang Chen, Wenhao Chai, Shi Jun, Yun Liu, and Erkang Chen. Five a + network: You only need 9k parameters for underwater image enhancement. *arXiv preprint arXiv:2305.08824*, 2023.
- [Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.
- [Kim *et al.*, 2023] Kangyeol Kim, Sunghyun Park, Junsoo Lee, and Jaegul Choo. Reference-based image composition with sketch via structure-aware diffusion model. *arXiv preprint arXiv:2304.09748*, 2023.
- [Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.
- [Kumari *et al.*, 2023] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.
- [Kumari *et al.*, 2024] Nupur Kumari, Grace Su, Richard Zhang, Taesung Park, Eli Shechtman, and Jun-Yan Zhu. Customizing text-to-image diffusion with camera viewpoint control. *arXiv preprint arXiv:2404.12333*, 2024.
- [Li *et al.*, 2023] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.
- [Lin *et al.*, 2018] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image



- compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9455–9464, 2018.
- [Liu *et al.*, 2023a] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [Liu *et al.*, 2023b] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023.
- [Liu *et al.*, 2023c] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023.
- [Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khali-dov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [Peng *et al.*, 2023] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.
- [Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [Rombach *et al.*, 2022a] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [Rombach *et al.*, 2022b] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
- [Seyfioglu *et al.*, 2023] Mehmet Saygin Seyfioglu, Karim Bouyarmane, Suren Kumar, Amir Tavanaei, and Ismail B Tutar. Dreampaint: Few-shot inpainting of e-commerce items for virtual try-on without 3d modeling. *arXiv preprint arXiv:2305.01257*, 2023.
- [Tsai *et al.*, 2017] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017.
- [Udandarao, 2022] Vishaal Udandarao. *Understanding and Fixing the Modality Gap in Vision-Language Models*. PhD thesis, Master’s thesis, University of Cambridge, 2022.
- [Xie *et al.*, 2023] Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*, 2023.
- [Yang *et al.*, 2023] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.
- [Ye *et al.*, 2023] Tian Ye, Sixiang Chen, Jinbin Bai, Jun Shi, Chenghao Xue, Jingxia Jiang, Junjie Yin, Erkang Chen, and Yun Liu. Adverse weather removal with codebook priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12653–12664, 2023.
- [Yi *et al.*, 2022] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *European Conference on Computer Vision*, pages 739–756. Springer, 2022.
- [Yi *et al.*, 2023] Xuanyu Yi, Jiajun Deng, Qianru Sun, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Invariant training 2d-3d joint hard samples for few-shot point cloud recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14463–14474, 2023.
- [Yi *et al.*, 2024] Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, and Hanwang Zhang. Diffusion time-step curriculum for one image to 3d generation. *arXiv preprint arXiv:2404.04562*, 2024.
- [Zhang *et al.*, 2023a] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023.
- [Zhang *et al.*, 2023b] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [Zhao *et al.*, 2023] Zhuoran Zhao, Jinbin Bai, DeLong Chen, Debang Wang, and Yubo Pan. Taming diffusion models for music-driven conducting motion generation. In *Proceedings of the AAAI Symposium Series*, volume 1, pages 40–44, 2023.