This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MOVIS: Enhancing Multi-Object Novel View Synthesis for Indoor Scenes

Ruijie Lu[1,2⋆] Yixin Chen[2⋆†] Junfeng Ni[2,3] Baoxiong Jia[2]
Yu Liu[2,3] Diwen Wan[1] Gang Zeng[1†] Siyuan Huang[2†]
⋆ Equal contribution † Corresponding author
[1] State Key Laboratory of General Artificial Intelligence, Peking University
[2] State Key Laboratory of General Artificial Intelligence, BIGAI [3] Tsinghua University

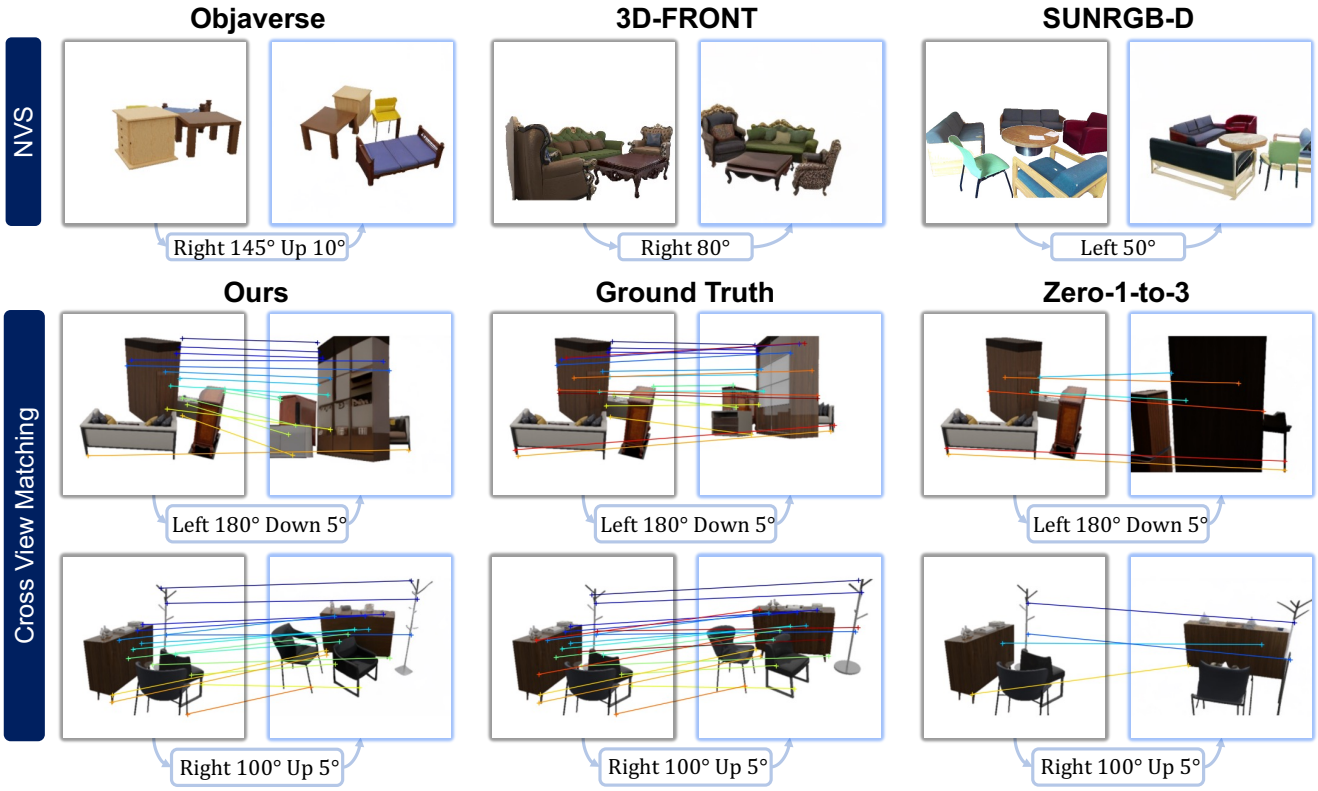


Figure 1. **Novel view synthesis and cross-view image matching**. The first row shows that MOVIS generalizes to different datasets on novel view synthesis (NVS). We also show visualizations of cross-view consistency compared with Zero-1-to-3 [43] and ground truth by applying image-matching. MOVIS can match a significantly greater number of points, closely aligned with the ground truth.

Abstract

Repurposing pre-trained diffusion models has been proven to be effective for NVS. However, these methods are mostly limited to a single object; directly applying such methods to compositional multi-object scenarios yields inferior results, especially incorrect object placement and inconsistent shape and appearance under novel views. How to enhance and systematically evaluate the cross-view consistency of such models remains under-explored. To address this issue, we propose MOVIS to enhance the structural awareness of the view-conditioned diffusion model for multi-object NVS in terms of model inputs, auxiliary tasks, and training strategy. First, we inject structure-aware features, including depth and object mask, into the denoising U-Net to enhance the model's comprehension of object instances and their spatial relationships. Second, we introduce an auxiliary task requiring the model to simultaneously predict novel view object masks, further improving the model's capability in differentiating and placing objects. Finally, we conduct an in-depth analysis of the diffusion sampling process and carefully devise a structure-guided timestep sampling

scheduler during training, which balances the learning of global object placement and fine-grained detail recovery. To systematically evaluate the plausibility of synthesized images, we propose to assess cross-view consistency and novel view object placement alongside existing image-level NVS metrics. Extensive experiments on challenging synthetic and realistic datasets demonstrate that our method exhibits strong generalization capabilities and produces consistent novel view synthesis, highlighting its potential to guide future 3D-aware multi-object NVS tasks. Our project page is available at https://jason-aplp.github.io/MOVIS/.

1. Introduction

Novel view synthesis (NVS) from a single image is imperative for various applications, including scene understanding [16, 19, 20, 25, 40] and interaction [9, 26, 27], interior designs [3, 18, 22, 77, 86], robotics [34–37, 46, 48, 62], *etc*. This is highly challenging as it requires understanding complex spatial structures from a single viewpoint while being able to extrapolate consistent and plausible content for unobserved areas. The substantial demands for comprehensive knowledge of the 3D world render it a difficult task, even for humans with rich priors of the 3D environments.

Recently, significant progress has been made in single-object image-to-3D generation [41, 42, 47, 65, 66, 70, 71, 82] empowered by powerful 2D diffusion models [17, 63]. Among them, one prominent line of research [5, 21, 38, 39, 41, 44, 60, 69, 80, 90] has achieved compelling results by building on insights from Zero-1-to-3 [43]: repurposing a pre-trained diffusion model as a novel view synthesizer by fine-tuning on large 3D object datasets can provide promising 3D-aware prior for image-to-3D tasks. However, these methods are mostly restricted to the single-object level. It remains unclear if this paradigm can be effectively extended to the multi-object level to facilitate more complex tasks like reconstructing an indoor scene. In Fig. 1, we visualize cross-view matching results of directly applying the novel view synthesizers [43] in multi-object scenarios, which showcases weak consistency with input views. Specifically, we believe that the lack of structural awareness is the primary reason for the disappearance, distortion, incorrect position and orientation of objects under novel views. While several works [64, 73] have explored training on mixed real-world scene datasets, the complexity introduced by multiple objects, such as spatial placement, per-instance geometry and appearance, and occlusion relationship, makes incorporating such awareness non-trivial.

Inspired by the discussion above, our paper seeks to address the question: *How to enhance the structural awareness of current diffusion-based novel view synthesizers?* We begin by identifying key challenges in extending single-object methods for multi-object NVS tasks. A multi-object image possesses more complicated structural information than a single-object one. The model must first grasp hierarchical structures within, which includes both high-level object placement, *e.g*., position and orientation, and low-level ones like per-object geometry and appearance. High-level structural information significantly reduces the ambiguity in object composition while low-level details are essential for capturing the characteristics of each object instance. Subsequently, the model needs to retain this hierarchical information captured from the input view while synthesizing novel-view images to ensure cross-view structural consistency. These capabilities are less critical in single-object NVS tasks due to the reduced ambiguity in one-to-one mapping but are crucial for effective multi-object NVS models.

Building on these insights, our technical designs are threefold. We first propose injecting structure-aware features, *i.e*., depth and object mask, from the input view as additional inputs to provide information on both high-level global placement and fine-grained local details. Secondly, we utilize the prediction of novel view object masks as an auxiliary task during training for the model to differentiate object instances, laying a solid foundation for fine-grained geometry and appearance recovery. Finally, through an in-depth analysis of the model's inference process, we highlight the importance of revising the noise timestep sampling schedule, which influences the learning focus in the training process. To be specific, larger timesteps emphasize global placement learning, while smaller timesteps focus on local fine-grained object geometry and appearance recovery. To endow the view-conditioned diffusion model with both capabilities, we propose a structure-guided timestep sampling scheduler that prioritizes larger timesteps in the initial stage, gradually decreasing over time to balance these two conflicting inductive biases. This design is fundamental to our proposed model's effectiveness in addressing the complexity of multi-object level NVS tasks.

To systematically assess the plausibility of synthesized novel view images, we additionally evaluate novel-view object mask and cross-view structural consistency apart from the existing NVS metrics. Specifically, we employ image-matching techniques [32, 78] to compare the input-view image with both the ground-truth and synthesized novel-view images. Cross-view structural consistency evaluates how closely the matching results align, providing a measure of the accuracy in recovering object placement, shape, and appearance. On the other hand, the object mask, as measured by Intersection over Union (IoU), assesses the precision of object placement. Extensive experiments demonstrate that our method excels at multi-object level NVS in indoor scenes, achieving consistent object placement, shape, and appearance. Notably, it exhibits strong generalization capabilities for generating novel views on unseen datasets, including both synthetic ones 3D-FRONT [14], Room-Texture [51] and Objaverse [11], as well as the real-

world SUNRGB-D [68].

In summary, our main contributions are:

1. We introduce structure-aware features as model inputs and incorporate novel view mask prediction as an auxiliary task during training. This enhances the model's understanding of hierarchical structures in multi-object scenarios, leading to improved NVS performance.
2. We present a novel noise timestep sampling scheduler to balance the learning of global object placement and fine-grained detail recovery, which is critical for addressing the increased complexity in multi-object scenarios.
3. We introduce additional metrics to systematically evaluate the novel view structural consistency. Through extensive experiments, our model demonstrates superiority in consistent object placement, geometry, and appearance recovery, showcasing strong generalization capability to unseen datasets.

2. Related Work

Single object NVS with generative models Synthesizing novel view images for single objects given a single-view image is an extremely ill-posed problem that requires strong priors. With great advances achieved in diffusion models [17, 63], research efforts [52, 70, 83] seek to distill priors [24, 59] learned from Text-to-Image (T2I) diffusion models via image captioning like [33]. However, this presents a huge gap between the image and semantics due to the ambiguity of the text, hindering the 3D consistency of these methods. On the other hand, view-conditioned diffusion models like Zero-1-to-3 [43] explore an Image-to-Image (I2I) generation paradigm that "teaches" the diffusion model to control viewpoints to synthesize plausible images under novel views, providing a more consistent 3D-aware prior. Subsequent work focuses on accelerating the generation speed [41, 69], enhancing the view consistency [5, 21, 39, 44, 80], or accelerating the training process [28]. However, all these methods deal with single and complete object novel view synthesis tasks since they usually fine-tune their model on Objaverse [10, 11], an extensive single-object level dataset, contrary to real images which normally consist of multiple or incomplete objects. The lack of specific model designs for compositional scenes also leads to significant inconsistencies when directly applying them to the multi-object scenarios, as shown in Fig. 1.

Multi-object 3D reconstruction with single object priors Following the advance in 3D-aware single object generative prior [43, 65], a line of research work [6, 7, 13, 54, 55, 77] focuses on extending their application to compositional multi-object scenarios. The core idea is to decompose object compositions into individual objects, thereby fully leveraging the powerful generative priors of single-object models. They first break down a multi-object composition into several components via segmentation models like SAM [31], and then complete every single object with amodal [49, 58, 84, 89] or inpainting [50, 63, 72] techniques. The object instances are lifted to 3D via image-to-3D models [43, 69, 75, 81] and finally composited into a whole utilizing spatial-aware optimization, 3D bounding box detection [1, 56] or carefully estimating the metric depth [29, 85]. However, this divide-and-conquer paradigm is limited by the user-specified spatial relations from language prompts [7] and relies heavily on the cascaded modules of detection [1, 29, 31], completion [50, 63] and 3D-aware object-level novel view synthesis (NVS) [41, 75] to provide priors for reconstruction. Unlike any of the above, our method aims to build an end-to-end image-conditioned novel view synthesis model that can directly cope with the increased complexity in multi-object compositions, especially in the multiple-object setting.

Scene-level NVS with sparse view input Early efforts [23, 45, 76, 88] attempted to directly perform scene-level NVS tasks by extracting image features from input-view images and inferring the underlying 3D representation [53]. With the development of Gaussian Splatting [30], recent works [2, 8] attempt to switch the underlying representation to Gaussian Splatting for efficiency. However, they mainly deal with synthesizing views near input ones with limited generative capabilities to the unseen region. Inspired by the great success of diffusion models [63] and the object-level 3D-aware novel-view synthesizer [43, 75], several recent works have also attempted to perform scene-level NVS tasks by directly conditioning the generative models on a single-view scene image or a monocular dynamic scene video [74]. ZeroNVS [64] proposes to train a view-conditioned diffusion model on a mixture of real-world datasets, MegaScenes [73] further scales up the training dataset with Internet-level data pairs for stronger generalization capabilities. However, all these works mainly deal with small view-change and simple scenarios in terms of object number, with few adaptations to tackle the multi-object complexity. In this work, we systematically examine the cross-view consistency of NVS by proposing new metrics, and explore the critical designs required to enhance the structural consistency of the view-conditioned diffusion models in the multi-object scenarios.

3. Method

In this section, we address the challenge of enhancing the structural awareness of diffusion-based novel view synthesizers for better cross-view consistency in multi-object scenarios. We begin with a brief introduction to diffusion models and view-conditioned diffusion models (Sec. 3.1). Next, we detail the key architectural designs of MOVIS,

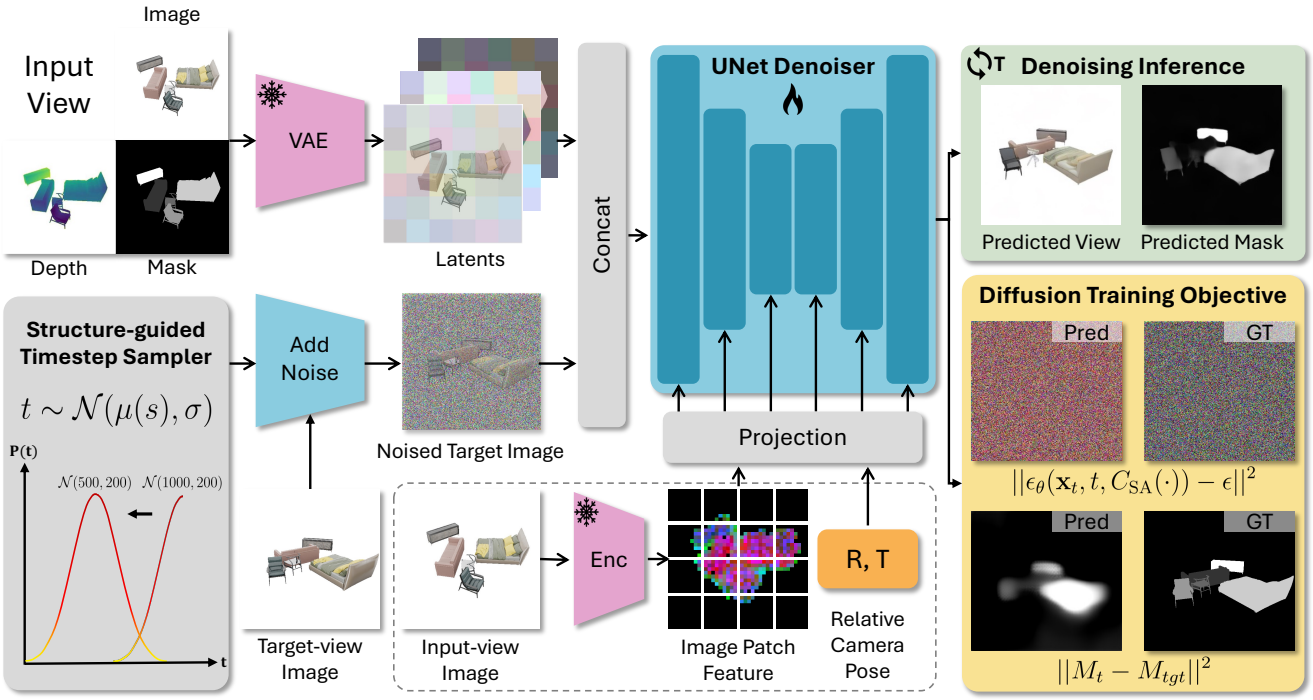


Figure 2. **Overview of MOVIS.** Our model performs NVS from the input image and relative camera change. We introduce structure-aware features as additional inputs and employ mask prediction as an auxiliary task (Sec. 3.2). The model is trained with a structure-guided timestep sampling scheduler (Sec. 3.3) to balance the learning of global object placement and local detail recovery.

including how we incorporate structural-aware features as input to improve the model's understanding of hierarchical structure information (Sec. 3.2) and how we introduce novel view mask prediction as an auxiliary task, instructing the model to differentiate the object instances with correct object placement (Sec. 3.2). Finally, we provide an in-depth analysis of the inference process and adopt a structure-guided timestep sampling scheduler (Sec. 3.3) to balance the learning of global object placement and local fine-grained object geometry and appearance recovery. We provide an overview of our model in Fig. 2.

3.1. Diffusion models as novel view synthesizers

Diffusion models have been recently repurposed as a novel view synthesizer. By training on posed image pairs $\{(\mathbf{x}_0, \hat{\mathbf{x}}_0)\}$ where $\hat{\mathbf{x}}_0 \in \mathbb{R}^{H\times W\times 3}$ denotes the input view image and $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times 3}$ denotes the target view, view-conditioned diffusion models [43, 79] use the input image $\hat{\mathbf{x}}_0$ and camera pose transformation as conditions to predict the target view image $\mathbf{x}_0$. Concretely, the learning objective of view-conditioned diffusion models is:

$$\mathbb{E}[||\epsilon_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon, t, C(\hat{\mathbf{x}}_0, R, T)) - \epsilon||_2^2], \qquad (1)$$

where R, T represent the relative camera pose transformation between the target view $\mathbf{x}_0$ and the input view $\hat{\mathbf{x}}_0$. $C(\hat{\mathbf{x}}_0, R, T)$ is the view-conditioned feature, combining the relative camera pose transformation with encoded image features to form a new 'pose-aware' feature map, taking the place of the origin CLIP [61] feature embedding. Moreover, input view image $\mathbf{x}_0$ will be concatenated with the noisy image as the input of the denoising U-Net. As discussed in Sec. 2, single-image-based NVS is extremely challenging, current methods inherit natural image priors from large-scale pre-training [63] and fine-tune diffusion models on large-scale 3D object datasets like Objaverse [11] to learn the transformation between objects in the input and novel views given the relative camera pose. Despite their ability to generalize to in-the-wild objects, these view-conditioned diffusion models struggle with multi-object scenarios like multi-furniture indoor scenes due to the scarcity of similar data and increased complexity arising from intricate object compositions. Our method builds on the insight of repurposing the diffusion model as a novel view synthesizer while emphasizing the inherent properties of multi-object scenarios in both model design and training strategy to facilitate multi-object NVS.

3.2. MOVIS

Our proposed method extends view-conditioned diffusion models to multi-object level, as illustrated in Fig. 2. The model leverages a pre-trained Stable Diffusion [63] and concatenates the 2D structural information from the input

view with a noisy target image as input. Additionally, it integrates a pre-trained image encoder [57] to capture semantic information, which is injected into the network through cross-attention alongside the relative camera pose. Moreover, it predicts novel view mask simultaneously as an auxiliary task to aid global object placement learning.

Structure-Aware Feature Amalgamation To synthesize plausible images under novel viewpoints, the model must first grasp the compositional structural information from the input view, laying a solid foundation for generation. To address the innate complexity in multi-object scenarios due to the intricate object relationship, we propose to leverage structure-aware features to facilitate model's comprehension. Specifically, we use depth maps and object masks as proxies for image-level structural information. Object masks provide a rough concept of object placement and shape as well as distinguishing distinct object instances, while depth maps encode the rough relative position and shape of the visible objects. Together with input-view images, these conditions provide both global structural information like object placement and local fine-grained details like object shape. Concretely speaking, we normalize the image rendered with object instance IDs of the input view to create a continuous object mask image $\widehat{\mathbf{M}}$. We then replicate the depth map $\widehat{\mathbf{D}}$ and object mask image $\widehat{\mathbf{M}}$ into three channels to simulate RGB images. These two structural-aware feature images, along with the input image $\hat{\mathbf{x}}_0$, are passed into a VAE to obtain latent features, which will be later concatenated with the noisy target view image $\mathbf{x}_t$ as input to the denoising U-Net. Note that both object mask and depth can be obtained with off-the-shelf detectors during the inference stage, such as SAM [31] and Marigold [29]. After introducing these additional conditions, the learning objective of MOVIS becomes:

$$\mathbb{E}[||\epsilon_\theta(\alpha_t\mathbf{x}_0+\sigma_t\epsilon, t, C_{\text{SA}}(\hat{\mathbf{x}}_0, R, T, \widehat{\mathbf{D}}, \widehat{\mathbf{M}}))-\epsilon||_2^2]. \quad (2)$$

We use $C_{\text{SA}}(\cdot)$ as a shorthand for the structure-aware view-conditioned feature throughout the paper.

Auxiliary Novel View Mask Prediction Task Input-view depth maps and mask images are intended to help the model indirectly understand the structure of multi-object compositions by incorporating additional structure-aware information into the input. To encourage the model to better grasp overall structure, particularly its ability to generate it, we propose leveraging structural information (*i.e.*, mask image) prediction under the target view as an auxiliary task, providing more direct supervision. Our approach draws inspiration from classifier guidance [12], where a classifier $p_\phi(y|x_t, t)$ guides the denoising process of image x_t to meet the criterion y via incorporating the gradient $\nabla_{x_t}\log p_\phi((y|x_t, t))$ during the inference process as an auxiliary guidance. Similarly, to improve the model's ability to learn compositional structure, particularly in synthesizing novel view plausible object placement (position and orientation), we introduce an auxiliary task during training: predicting object mask images $\mathbf{M}_t \sim p(\mathbf{M}_t|\mathbf{x}_t, t, C_{\text{SA}}(\cdot))$ under target view. This prediction is conditioned on the noisy target-view image $\mathbf{x}_t$, timestep t and input-view structure-aware feature $C_{\text{SA}}(\cdot)$, derived from the final layer of the denoising U-Net. We jointly train the mask predictor and denoising U-Net following:

$$\mathbb{E}[||\epsilon_\theta(\alpha_t\mathbf{x}_0+\sigma_t\epsilon, t, C_{\text{SA}}(\cdot))-\epsilon||_2^2+\gamma||\mathbf{M}_{tgt}-\mathbf{M}_t||_2^2], \quad (3)$$

where we use $\mathbf{M}_{tgt}$ to denote the ground-truth target-view image, and we use the weight $\gamma = 0.1$ to balance the diffusion loss and mask prediction loss.

3.3. Structure-Guided Sampling Scheduler

Inspired by previous works [4, 28] that identify the importance of different scheduling strategies, we provide an in-depth analysis of the inference process of multi-object novel view synthesis, where we adopt a DDIM [67] sampler:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{x}_0^{'} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\mathbf{F} + \sigma_t\epsilon_t. \quad (4)$$

where $\mathbf{x}_0^{'} = (\mathbf{x}_t - \sqrt{1-\alpha_t}\cdot\mathbf{F})/\sqrt{\alpha_t}$. We use $\mathbf{F}$ as a shorthand for $\epsilon_\theta(\mathbf{x}_t, t, C_{\text{SA}}(\cdot))$ and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We examine the predicted $\mathbf{x}_0^{'}$ (as in Eq. (4)) and the predicted mask image $\mathbf{M}_t$ at various timesteps during the inference process as they offer direct visualizations for analysis. These visualized results are presented in Fig. 3.

In Fig. 3, we observe that a blurry image, which indicates the approximate placement of each object, is quickly restored in the early stages (*i.e.*, larger t) of the inference process. This suggests that global structural information is prioritized for the model to learn during this stage. Accurate object placements are crucial for synthesizing reasonable novel view images, as incorrect placement predictions indicate a fundamental misunderstanding of the compositional structure. This underscores the importance of training the model with a larger t during the initial training periods, which is even more important for multi-object NVS scenarios considering the increased compositional complexity compared with a single object. Conversely, a mask with a clear boundary is not predicted until a later stage of the sampling process (*i.e.*, smaller t). This is because accurate mask prediction depends heavily on a relatively noiseless image. Therefore, to capture fine-grained geometry and appearance details of objects, it is essential to train the model with a smaller t during later training periods.

Recognizing the importance of timestep t in balancing the learning of global placement information and local fine-grained details, we propose to adjust the original timestep sampling process to:

$$t \sim \mathcal{U}(\mathbf{1}, \mathbf{1000}) \rightarrow t \sim \mathcal{N}(\mu(s), \sigma), \quad (5)$$

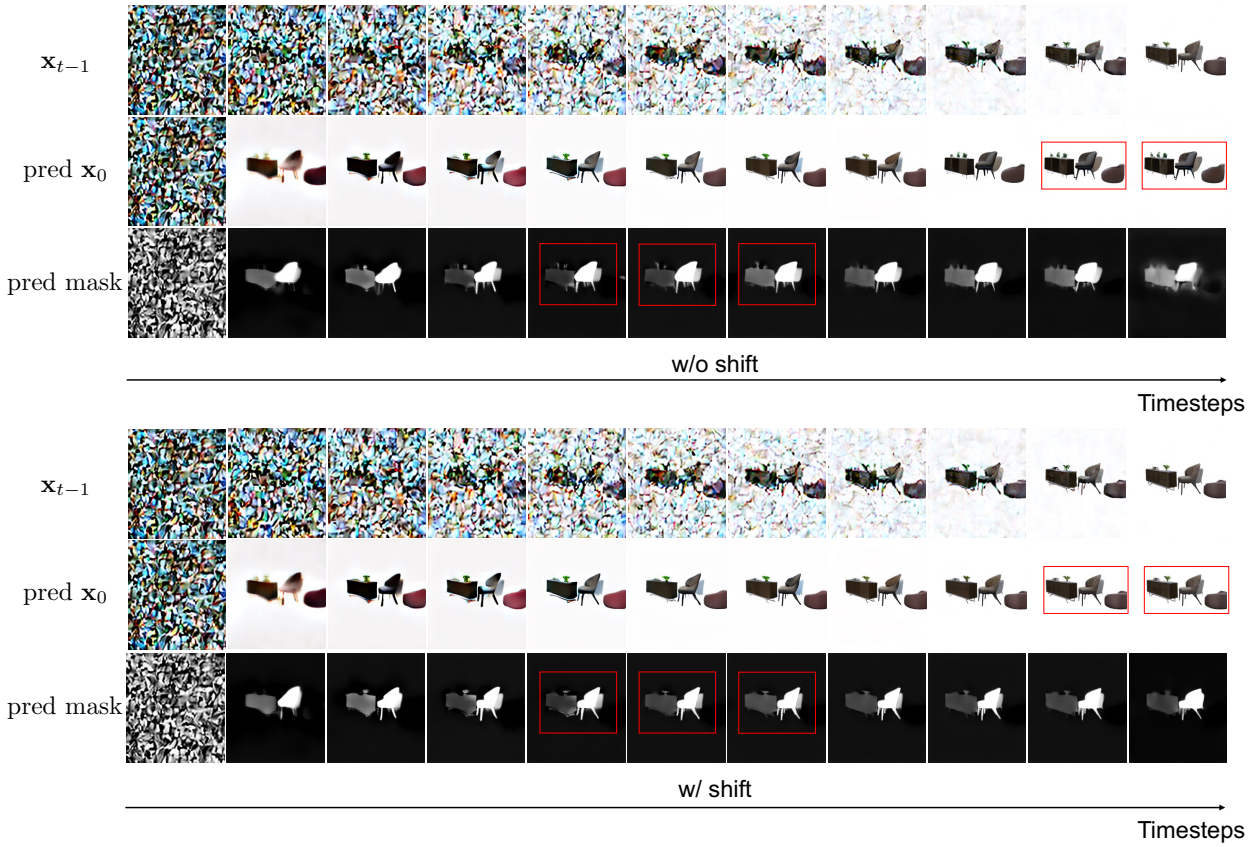


Figure 3. **Visualization of inference.** The early stage of the denoising process focuses on restoring global object placements, while the prediction of object masks requires a relatively noiseless image to recover fine-grained geometry. This motivates us to seek a balanced timestep sampling scheduler during training. The model trained *w/ shift* yields better mask prediction and thus recovers an image with more details and sharp object boundary. The *w/o shift* here refers to not shifting the μ value.

where $\mu(s) = \mu_{\text{local}} + (\mu_{\text{global}} - \mu_{\text{local}}) \cdot \frac{s}{T_s}$ and s denotes the model training iteration, T_s denotes the total number of training steps, $\sigma = 200$ is a constant variance. We sample the timestep t from a Gaussian distribution with mean $\mu(s)$ following a linear decay from a large value $\mu_{\text{global}} = 1000$ to a small value $\mu_{\text{local}} = 500$. This approach allows the model to initially learn correct global object placement information and gradually turn its focus to refining detailed object geometry in later training stages. In practice, we include a warmup period with 4000 training steps sampling t with a fixed $\mu(s) = \mu_{\text{global}}$. After the warmup, we use the linear decay schedule over 2000 steps, and then stabilize the learning for fine-grained details after 6000 steps where we use $\mu(s) = \mu_{\text{local}}$.

4. Experiments

4.1. Experiment Setup

We focus on multi-object composite NVS tasks in indoor scenes, with an emphasis on foreground objects, examining novel view structural plausibility regarding object placement, geometry, appearance, and cross-view consistency with input view. This choice stems from the recent advancements in object segmentation [31], while we leave the background modeling for future work.

Datasets. To facilitate the training and evaluation of our proposed method, we curate a scalable synthetic dataset Compositional 3D-FUTURE (C3DFS), comprising 100k composites for training and 5k for testing. Each composite is created by composing pre-filtered furniture items from 3D-FUTURE [15] using a heuristic strategy to avoid collision and penetration. Beyond C3DFS, we emphasize testing the generalization capability by benchmarking our method on Room-Texture [51] and Objaverse [11]. We also evaluate our model on diverse indoor scenes from both the synthetic dataset 3D-FRONT [14] and the real-world dataset SUNRGB-D [68]. Refer to *supplementary materials* for more details.

Baselines. We compare our method against three recent novel view synthesis methods including Zero-1-to-3 [43], ZeroNVS [64], and Free3D [90]. The original Zero-1-to-3 is trained on extensive object-level datasets. Therefore, we also re-train Zero-1-to-3 on our synthetic dataset C3DFS, denoted as Zero-1-to-3†, for a fair comparison. ZeroNVS is

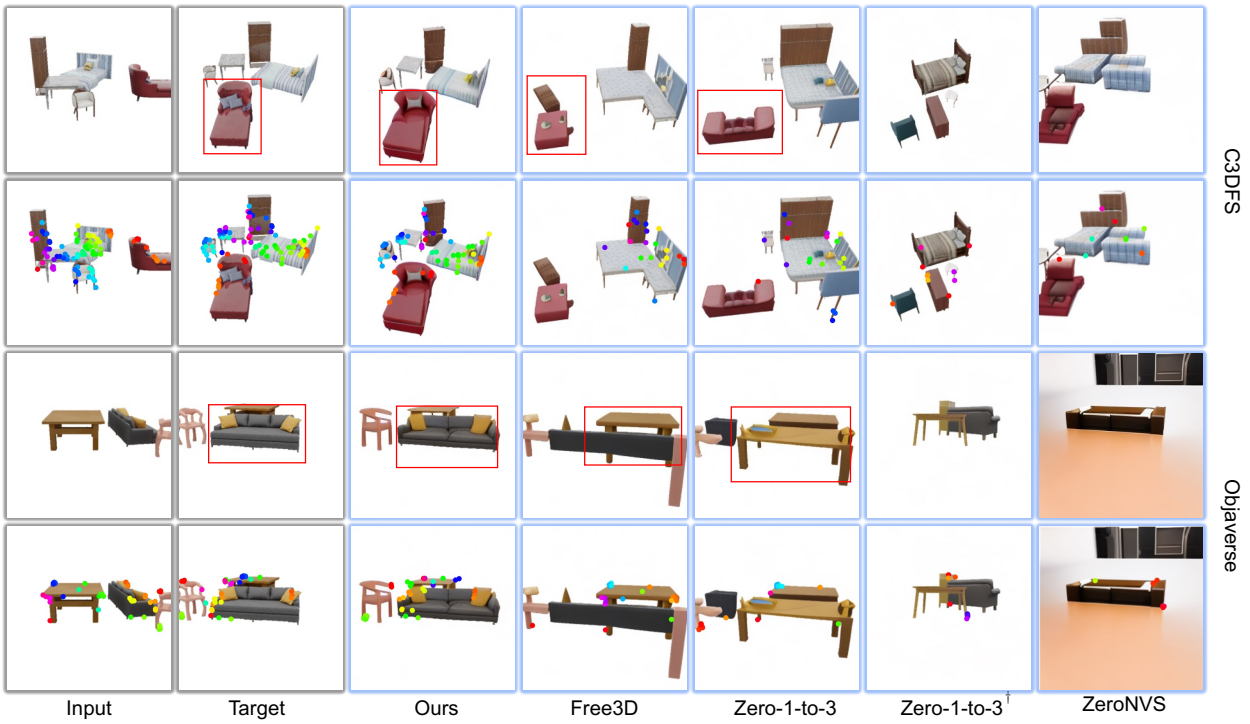


Figure 4. **Qualitative results of NVS and cross-view matching.** Our method generates plausible novel-view images across various datasets, surpassing baselines regarding object placement, shape, and appearance. In cross-view matching, points of the same color indicate correspondences between the input and target views. We achieve a higher number of matched points with more precise locations.

Table 1. **Quantitative results of multi-object NVS, Object Placement, and Cross-view Consistency**. We evaluate on C3DFS test set, along with *generalization experiments* on Room-Texture [51] and Objaverse [11]. † indicates re-training on C3DFS.

Dataset	Method	Novel View Synthesis			Placement	Cross-view Consistency	
		PSNR(↑)	SSIM(↑)	LPIPS(↓)	IoU(↑)	Hit Rate(↑)	Dist(↓)
C3DFS	ZeroNVS	10.704	0.533	0.481	21.6	1.4	135.2
	Zero-1-to-3	14.255	0.771	0.302	33.7	4.4	86.7
	Zero-1-to-3†	14.811	0.794	0.283	34.4	1.3	117.9
	Free3D	14.390	0.774	0.297	34.2	4.8	83.6
	Ours	**17.432**	**0.825**	**0.171**	**58.1**	**19.3**	**44.9**
Objaverse	ZeroNVS	10.557	0.513	0.486	17.3	1.3	135.8
	Zero-1-to-3	15.850	0.810	0.259	34.7	6.0	80.2
	Zero-1-to-3†	15.433	0.815	0.273	29.7	1.2	124.1
	Free3D	15.980	0.813	0.254	35.6	6.1	77.0
	Ours	**17.749**	**0.840**	**0.169**	**51.3**	**17.0**	**47.2**
Room-Texture	ZeroNVS	8.217	0.647	0.487	8.2	1.0	138.0
	Zero-1-to-3	9.860	0.712	0.406	13.9	2.3	104.2
	Zero-1-to-3†	8.342	0.657	0.452	13.5	0.4	156.1
	Free3D	9.623	0.705	0.415	19.7	2.1	104.6
	Ours	**10.014**	**0.718**	**0.366**	**24.2**	**4.4**	**78.1**

trained on a mixture of real-world images with background, so we use images with backgrounds as its input if possible for a fair comparison.

Metrics. We utilize PSNR, SSIM, and LPIPS as metrics for evaluating the quality of *Novel View Synthesis*. To assess global object *Placement*, we compute the foreground-background IoU with ground-truth masks. Finally, we propose metrics to evaluate *Cross-view Consis-*

tency with image-matching. More specifically, we first apply MASt3R [32] to acquire the image matching between the input-view image and target-view image for both ground truth and model predictions. With the ground-truth matching as references, we compute each method's *Hit Rate* and the nearest matching distance (*Dist.*). *Hit Rate* measures the proportion of predicted matches that align with the ground truth matches. *Dist.* quantifies the distance between the predicted matching and ground-truth matching in the target view. Please refer to *supplementary materials* for more details about the metrics.

4.2. Results and Discussions

Fig. 4 presents qualitative results of multi-object NVS and cross-view matching visualization on different datasets, with quantitative results in Tab. 1. We summarize the following key observations:

1. Our method realizes the highest PSNR and generates high-quality images under novel views, closely aligned with the ground truth images, especially regarding novel-view object placement (position and orientation), shape, and appearance. In contrast, the baseline models struggle to accurately capture the compositional structure under novel views. For example, in the first row, the red bed is incorrectly oriented in Zero-1-to-3 and is either missing or distorted in other baselines.
2. From the visualized cross-view matching results and the metrics in Tab. 1, it is evident that our method significantly outperforms the baseline approaches in *Cross-view Consistency*. It achieves a much higher IoU and *Hit Rate* while exhibiting a considerably lower matching distance. The visualized results are consistent with the metrics, further validating our method's accuracy in capturing cross-view structural consistency, which cannot be reflected by existing NVS metrics.
3. Our model exhibits strong generalization capabilities on unseen datasets, *e.g.*, Room-Texture and Objaverse. We demonstrate more qualitative results on in-the-wild datasets, including 3D-FRONT, SUNRGB-D, ScanNet++ [87], and RealEstate10K [91] in *supplementary materials*. We showcase potential applications, including object removal and reconstruction in *supplementary materials*. Further discussion about limitation and failure cases are presented in *supplementary materials*.

4.3. Ablation Study

To verify the efficacy of each component, we perform an ablation study on our key technical designs, including the depth input (*w/o depth*), mask prediction auxiliary task (*w/o mask*), and the scheduler (*w/o sch.* learns with a uniform sampler $t \sim \mathcal{U}(\mathbf{1}, \mathbf{1000})$). Results in Tab. 2 show that the auxiliary mask prediction task and the timestep sampler are the most critical components, significantly affecting all the metrics and the realistic object recovery as demonstrated by the misoriented brown cabinet in the example from Fig. 5. Without the scheduler, the model produces less accurate object positions, evident both qualitatively and quantitatively. Furthermore, removing depth or mask predictions weakens the model's understanding of spatial relationships and object existence. This also shows incorporating structure-aware features as inputs, though seemingly intuitive, offers the most straightforward approach to enhancing the model's structural awareness, particularly given recent advancements in monocular predictors [29, 31]. We present a more comprehensive discussion on the scheduler strategy and ablations in *supplementary materials*.

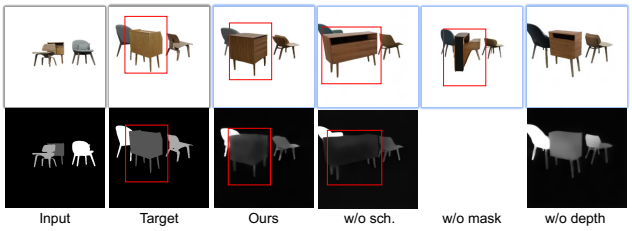


Figure 5. **Qualitative comparison for ablation study**. Excluding mask predictions or the scheduler reduces the model's ability to learn object placement, as shown by the brown cabinet example.

Table 2. **Ablation results on C3DFS**.

Method	Novel View Synthesis			Placement
	PSNR(↑)	SSIM(↑)	LPIPS(↓)	IoU(↑)
w/o depth	17.080	0.819	0.178	57.2
w/o mask	16.914	0.818	0.187	54.7
w/o sch.	16.166	0.808	0.212	49.1
Ours	**17.432**	**0.825**	**0.171**	**58.1**

5. Conclusion

We extend diffusion-based novel view synthesizers to handle multi-object compositions in indoor scenes. Our proposed model generalize well across diverse datasets with more accurate object placement, shape, and appearance, showing a stronger cross-view consistency with input view. The core of our approach lies in integrating structure-aware features as additional inputs, an auxiliary mask prediction task, and a structure-guided timestep sampling scheduler. These components enhance the model's awareness of compositional structure while balancing the learning of global object placement and fine-grained local shape and appearance. Given the prevalence of multi-object compositions in real-world scenes, we believe that our model designs and comprehensive evaluations can offer valuable insights for advancing NVS models in more complex environments.

Acknowledgment

This work is supported by the Sichuan Science and Technology Program (2023YFSY0008), China Tower-Peking University Joint Laboratory of Intelligent Society and Space Governance, National Natural Science Foundation of China (61632003, 61375022, 61403005), Grant SCITLAB20017 of Intelligent Terminal Key Laboratory of SiChuan Province, Beijing Advanced Innovation Center for Intelligent Robots and Systems (2018IRS11), and PEK-SenseTime Joint Laboratory of Machine Vision.

References

[1] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[3] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[4] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023. 5

[5] Yabo Chen, Jiemin Fang, Yuyang Huang, Taoran Yi, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Cascade-zero123: One image to highly consistent 3d with self-prompted nearby views. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[6] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *International Conference on 3D Vision (3DV)*, 2024. 3

[7] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[9] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[10] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 6, 7

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

[13] Andreea Dogaru, Mert Özer, and Bernhard Egger. Generalizable 3d scene reconstruction via divide and conquer from a single view. *arXiv preprint arXiv:2404.03421*, 2024. 3

[14] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 6

[15] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, 129:3313–3337, 2021. 6

[16] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. *arXiv preprint arXiv:2304.04321*, 2023. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[18] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[19] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *International Conference on Machine Learning (ICML)*, 2024. 2

[20] Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[21] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, et al. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[22] Zhipeng Huang, Wangbo Yu, Xinhua Cheng, ChengShu Zhao, Yunyang Ge, Mingyi Guo, Li Yuan, and Yonghong Tian. Roompainter: View-integrated diffusion for consistent

indoor scene texturing. *arXiv preprint arXiv:2412.16778*, 2024. 2

[23] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[24] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[25] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[26] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2

[27] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[28] Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3dim: Learning a generalizable single-image novel-view synthesizer in one day. *arXiv preprint arXiv:2310.03015*, 2023. 3, 5

[29] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 5, 8

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 3

[31] 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 *International Conference on Computer Vision (ICCV)*, 2023. 3, 5, 6, 8

[32] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 8

[33] 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 *International Conference on Machine Learning (ICML)*, 2023. 3

[34] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *International Conference on Robotics and Automation (ICRA)*, 2023. 2

[35] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In *International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[36] Yishu Li, Wen Hui Leng, Yiming Fang, Ben Eisner, and David Held. Flowbothd: History-aware diffuser handling ambiguities in articulated objects manipulation. *arXiv preprint arXiv:2410.07078*, 2024.

[37] Yuyang Li, Bo Liu, Yiran Geng, Puhao Li, Yaodong Yang, Yixin Zhu, Tengyu Liu, and Siyuan Huang. Grasp multiple objects with one hand. *IEEE Robotics and Automation Letters (RA-L)*, 2024. 2

[38] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[39] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 2, 3

[40] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[41] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[42] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[43] 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. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 4, 6

[44] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3

[45] Yu Liu, Baoxiong Jia, Yixin Chen, and Siyuan Huang. Slotlifter: Slot-guided feature lifting for learning object-centric radiance fields. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[46] Yu Liu, Baoxiong Jia, Ruijie Lu, Junfeng Ni, Song-Chun Zhu, and Siyuan Huang. Building interactable replicas of complex articulated objects via gaussian splatting. *arXiv preprint arXiv:2502.19459*, 2025. 2

[47] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[48] Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. Manigaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[49] Ruijie Lu, Yixin Chen, Yu Liu, Jiaxiang Tang, Junfeng Ni, Diwen Wan, Gang Zeng, and Siyuan Huang. Taco: Taming diffusion for in-the-wild video amodal completion. *arXiv preprint arXiv:2503.12049*, 2025. 3

[50] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[51] Rundong Luo, Hong-Xing Yu, and Jiajun Wu. Unsupervised discovery of object-centric neural fields. *arXiv preprint arXiv:2402.07376*, 2024. 2, 6, 7

[52] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[53] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[54] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[55] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[56] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[57] 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. *arXiv preprint arXiv:2304.07193*, 2023. 5

[58] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[59] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[60] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[61] 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 (ICML)*, 2021. 4

[62] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *Conference on Robot Learning (CoRL)*, 2023. 2

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4

[64] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[65] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2, 3

[66] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 2

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[68] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 6

[69] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2, 3

[70] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[71] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[72] Jiaxiang Tang, Ruijie Lu, Xiaokang Chen, Xiang Wen, Gang Zeng, and Ziwei Liu. Intex: Interactive text-to-texture synthesis via unified depth-aware inpainting. *arXiv preprint arXiv:2403.11878*, 2024. 3

[73] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah

Snavely. Megascenes: Scene-level view synthesis at scale. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[74] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[75] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3

[76] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[77] Qi Wang, Ruijie Lu, Xudong Xu, Jingbo Wang, Michael Yu Wang, Bo Dai, Gang Zeng, and Dan Xu. Roomtex: Texturing compositional indoor scenes via iterative inpainting. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[78] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[79] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 4

[80] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2, 3

[81] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[82] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2

[83] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[84] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9099–9109, 2024. 3

[85] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[86] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[87] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 8

[88] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[89] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[90] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6

[91] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 8