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

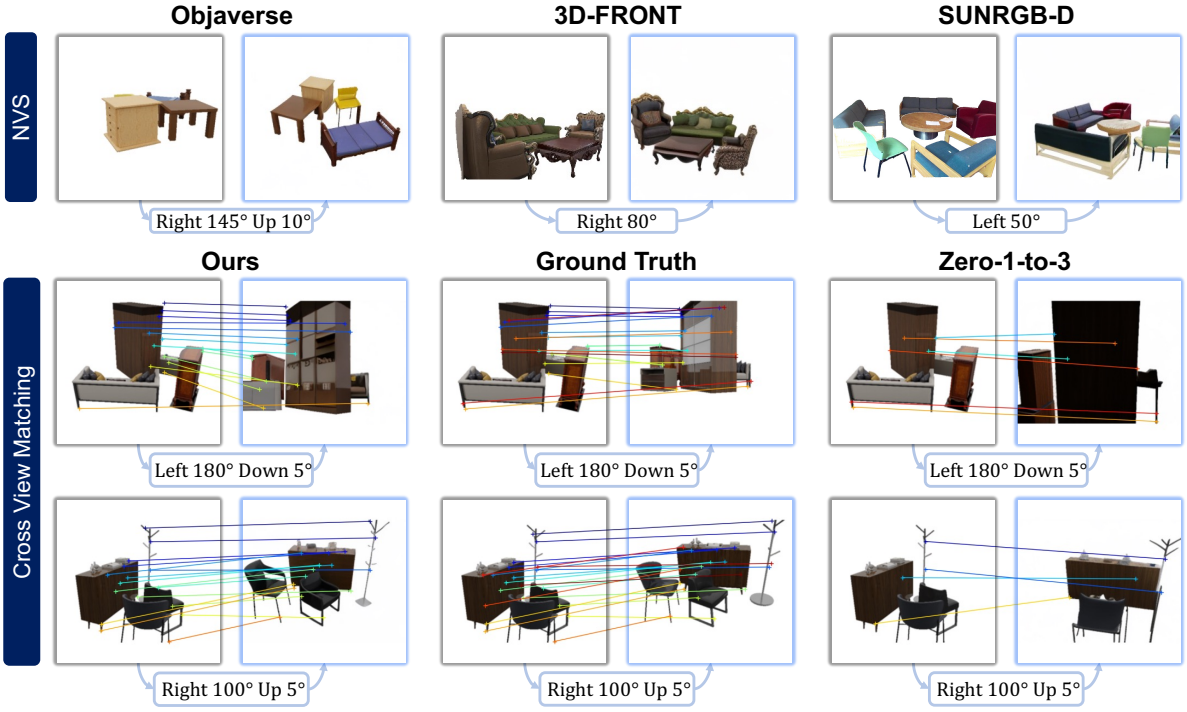

Figure 1. **Novel view synthesis and cross-view image matching**. The first row shows that MOVIS generalizes to different datasets on novel view synthesis (NVS). We also show visualizations of cross-view consistency compared with Zero-1-to-3 [8] and ground truth by applying image-matching. MOVIS can 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.

1. Introduction

Novel view synthesis (NVS) from a single image is challenging as it requires understanding complex spatial structures from a single viewpoint while being able to extrapolate consistent and plausible content for unobserved areas.

Recently, one prominent line of research [1, 3, 5–7, 9, 11, 15, 17, 18] has achieved compelling image-to-3D results by building on insights from Zero-1-to-3 [8]: 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, whether this paradigm can be effectively extended to the multi-object level to facilitate more complex tasks like reconstructing an indoor scene remains unclear. In Fig. 1, we visualize cross-view matching results of directly applying novel view synthesizers [8] in multi-object scenarios. 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.

Our paper seeks to address the question: *How to enhance the structural awareness of current diffusion-based novel view synthesizers?* We first propose injecting structure-aware features, *i.e.*, depth and object mask, from the input view as additional inputs. 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 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. 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.

We additionally evaluate novel-view object mask and cross-view structural consistency apart from the existing NVS metrics. Specifically, we employ image-matching techniques [4, 16] to compare the input-view image with both the ground-truth and synthesized novel-view images. 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.

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 [12] 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 [10] 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 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. 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. After introducing these additional conditions, the learning objective of MOVIS becomes:

$$\mathbb{E}[||\epsilon_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon, t, C_{\text{SA}}(\hat{\mathbf{x}}_0, R, T, \widehat{\mathbf{D}}, \widehat{\mathbf{M}})) - \epsilon||_2^2]. \quad (1)$$

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 To encourage the model to better grasp overall structure, we propose leveraging structural information (*i.e.*, mask image) prediction under the target view as an auxiliary task. Our approach draws inspiration from classifier guidance [2], 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. Similarly, to improve the model's ability to learn compositional structure, particularly in synthesizing novel view plausible object placement, 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 $C_{\text{SA}}(\cdot)$, using the final layer of the denoising U-Net. We jointly train the mask predictor and denoising U-Net following:

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

where we use $\mathbf{M}_{tgt}$ to denote the GT target-view image.

Structure-Guided Sampling Scheduler We provide an in-depth analysis of the inference process of multi-object novel view synthesis, where we adopt a DDIM [14] sam-

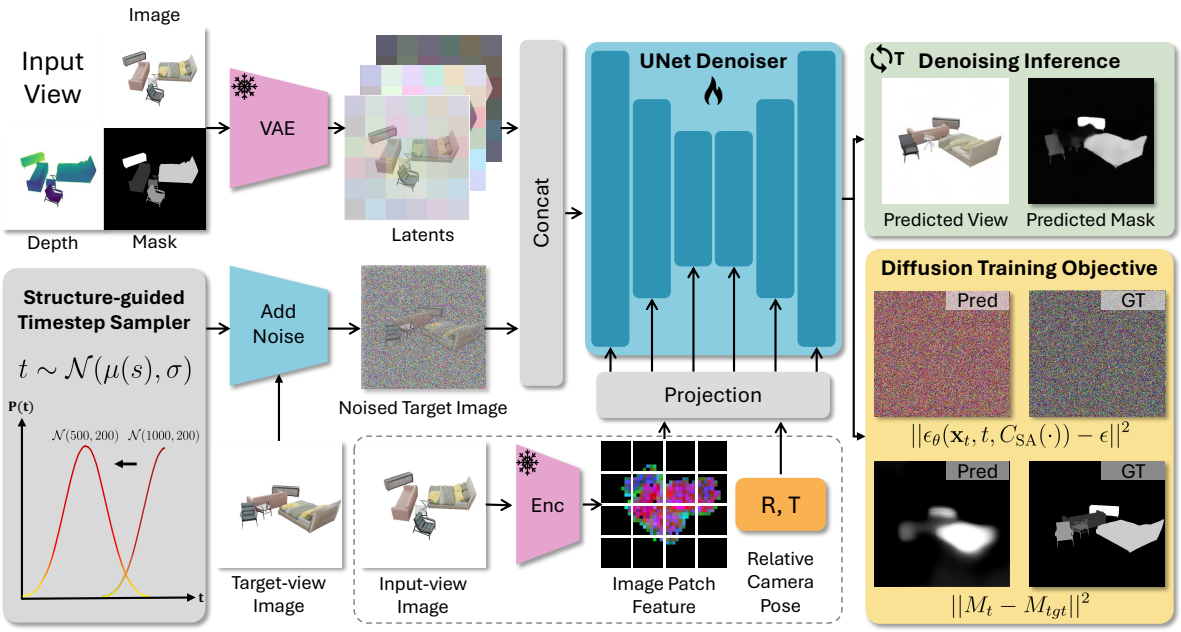

Figure 2. **Overview of MOVIS.** Our model performs NVS from the input image and relative camera change. We introduce structure-aware features as additional inputs and employ mask prediction as an auxiliary task (Sec. 2). The model is trained with a structure-guided timestep sampling scheduler (Sec.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

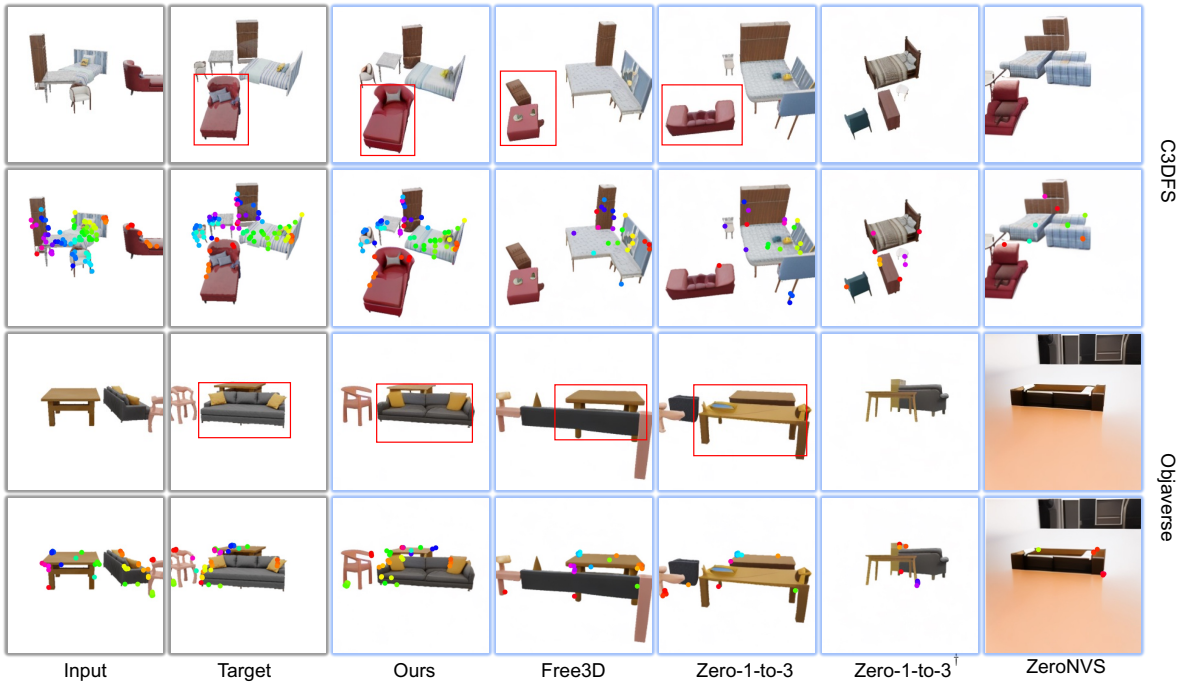

Figure 3. **Qualitative results of NVS and cross-view matching.** Our method generates plausible novel-view images across various datasets, surpassing baselines regarding object placement, shape, and appearance. In cross-view matching, points of the same color indicate correspondences between the input and target views. We achieve a higher number of matched points with more precise locations.

pler:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{x}_0^{'} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\mathbf{F} + \sigma_t\epsilon_t. \quad (3)$$

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

Table 1. **Quantitative results of multi-object NVS, Object Placement, and Cross-view Consistency**. We evaluate on C3DFS test set.

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**

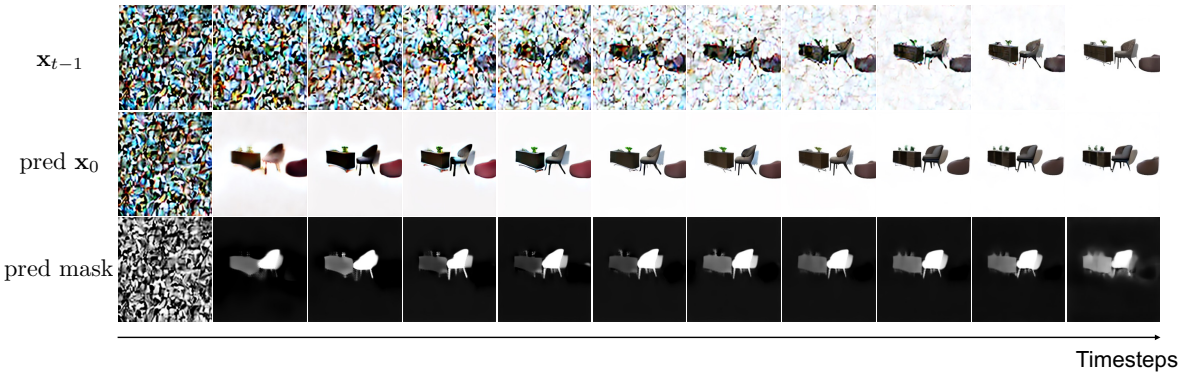

Figure 4. **Visualization of inference.** The early stage of the denoising process focuses on restoring global object placements, while the prediction of object masks requires a relatively noiseless image to recover fine-grained geometry. This motivates us to seek a balanced timestep sampling scheduler 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.

the predicted $\mathbf{x}_0^{'}$ (as in Eq. (3)) and the predicted mask image $\mathbf{M}_t$ at various timesteps during the inference process as they offer direct visualizations for analysis in Fig. 4.

In Fig. 4, 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. This underscores the importance of training the model with a larger t during the initial training periods. 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, it is essential to train the model with a smaller t during later training periods. We propose to adjust the original timestep sampling process to:

$$t \sim \mathcal{U}(\mathbf{1}, \mathbf{1000}) \rightarrow t \sim \mathcal{N}(\mu(s), \sigma), \tag{4}$$

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 $\mu_{\text{local}} = 500$.

3. Experiments

We benchmark our method against various object-level baselines including Zero-1-to-3 [8] and Free3D [18] as well as scene-level ZeroNVS [13], using NVS metrics and cross-view consistency metrics. Fig. 3 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.
3. Our model exhibits strong generalization capabilities on unseen datasets, *e.g.*, Objaverse.

References

[1] 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

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[3] 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

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

[5] 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

[6] 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.

[7] 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

[8] 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, 4

[9] 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

[10] 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. 2

[11] 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

[12] 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

[13] 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. 4

[14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[15] 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

[16] 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

[17] 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

[18] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4