World-consistent Video Diffusion with Explicit 3D Modeling

Qihang Zhang[1,2*] Shuangfei Zhai[1] Miguel Angel Bautista Martin[1] Kevin Miao[1]
Alexander Toshev[1] Josh Susskind[1] Jiatao Gu[1]
[1]Apple [2]The Chinese University of Hong Kong

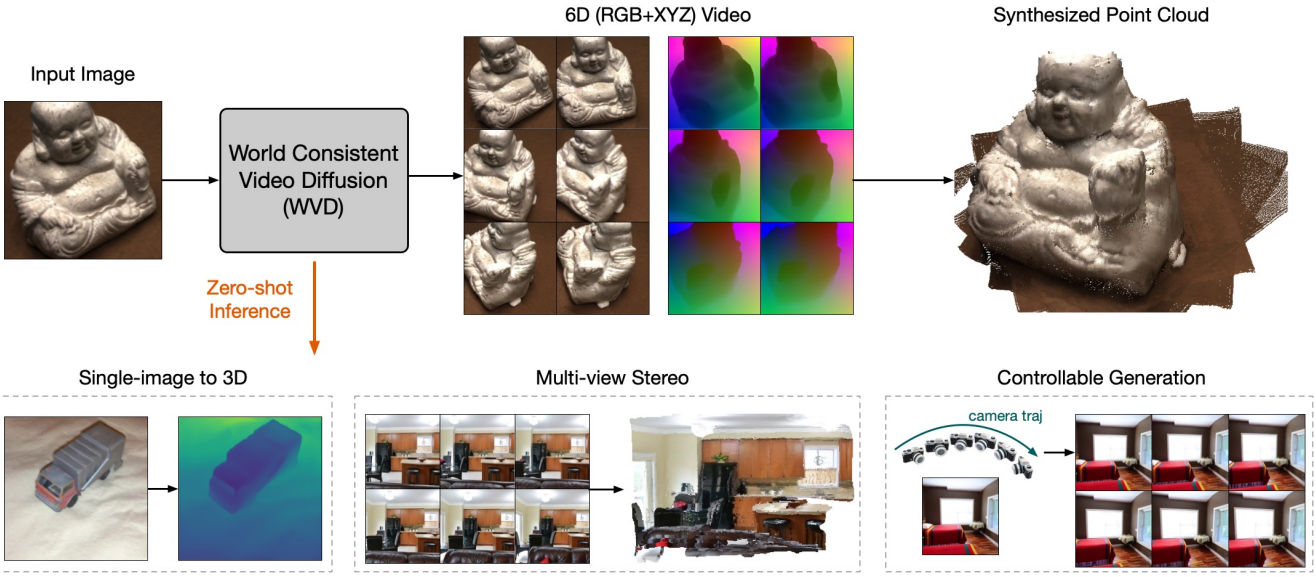

Figure 1. WVD predicts 6D videos from an image, unifying various 3D tasks with a single diffusion model.

Abstract

Recent advancements in diffusion models have set new benchmarks in image and video generation, enabling realistic visual synthesis across single- and multi-frame contexts. However, these models still struggle with efficiently and explicitly generating 3D-consistent content. To address this, we propose ***World-consistent Video Diffusion*** *(WVD), a novel framework that incorporates explicit 3D supervision using XYZ images, which encode global 3D coordinates for each image pixel. More specifically, we train a diffusion transformer to learn the joint distribution of RGB and XYZ frames. This approach supports multi-task adaptability via a flexible inpainting strategy. For example, WVD can estimate XYZ frames from ground-truth RGB or generate novel RGB frames using XYZ projections along a specified camera trajectory. In doing so, WVD unifies tasks like single-image-to-3D generation, multi-view stereo, and camera-controlled video generation. Our approach demonstrates competitive performance across multiple benchmarks, providing a scalable solution for 3D-consistent video and image generation with a single pretrained model.*

1. Introduction

Recent advancements in deep generative models have brought significant breakthroughs to the field of visual synthesis, with diffusion models [18, 47] emerging as the state-of-the-art approach for high-quality image generation [9, 38]. These models have demonstrated remarkable success in generating realistic images. By extending the input from single-frame image to multiple frames, diffusion models have also been applied to tasks such as video generation [1, 2, 4, 12, 15, 19, 20, 57] and multi-view image synthesis [11, 21, 26, 27, 43, 55, 63], where consistency across frames is crucial. In these applications, multi-frame consistency is typically learned in an implicit manner, e.g., through the attention mechanism that captures relationships across frames. Despite their success, multi-view (video) diffusion models face several limitations: they demand large amounts of data and significant computational resources for training, and they lack explicit guarantees for 3D consistency, often leading to 3D inconsistencies.

*Work done during internship at Apple MLR.

In contrast to these implicit methods, some approaches [5, 13, 30, 32, 62] seek to explicitly model 3D correspondences by embedding 3D inductive biases into the generative pipeline. These methods leverage techniques such as volume rendering [31], which can impose constraints that ensure 3D consistency in the generated images. However, the integration of 3D inductive biases tends to place heavy constraints on both the data and the architectural design, making it difficult to scale these methods to more complex datasets with diverse distributions.

To address these limitations, we propose a novel framework for multi-view and video generation that introduces explicit 3D supervision into diffusion models. Our method is designed to handle both RGB image generation and 3D geometry modeling within a unified framework. A major challenge in this integration arises from the inherent incompatibility between traditional 3D geometry representations and existing image architectures, such as the 2D Transformer-based models commonly used in diffusion models (e.g., DiT [34]). To resolve this, we propose to use XYZ images to represent 3D geometry, which are compatible with 2D Transformer architectures. Each pixel in an XYZ image records its corresponding global 3D coordinates. Unlike RGB images, which encode complex texture and lighting information, XYZ images are textureless and only capture geometric information, making them ideal for providing explicit 3D supervision during training.

Furthermore, because our model learns the joint distribution of RGB and XYZ images during the training phase, it can naturally perform conditional generation during inference using a flexible inpainting strategy [28, 29]. This enables the model to adapt to a wide range of tasks beyond image synthesis, including camera pose estimation, single-view and multi-view depth prediction from unposed images, and camera-conditioned novel view synthesis. This versatility allows our model to unify various generative and discriminative tasks under a single framework.

We refer to our proposed method as WVD. The major contributions of our work can be summarized as follows:

- We propose a novel approach to learn a multi-view diffusion model with explicit 3D supervision.
- Via a flexible inference strategy, WVD is capable of unifying various tasks within a single framework.
- WVD achieves competitive performance over different tasks, showcasing the potential to become a world-consistent 3D foundation model.

2. Related Work

Multi-view Diffusion Models. The advancement of multi-view diffusion models represents a significant step in generative modeling, combining the robust generation capabilities of diffusion frameworks with the complex requirement for cross-view consistency. Notable approaches like MVDream [43], ImageDream [55], Zero123++ [26], ConsistNet [63], SyncDreamer [27], and ViewDiff [21] adapt text-to-image diffusion models [38] to produce synchronized multi-view outputs. Video diffusion models [1, 2, 4, 12, 15, 19, 20, 57] learn multi-view consistency from extensive video datasets. Models like CameraCtrl [16], MotionCtrl [59], and Camco [61] enhance video diffusion models by introducing camera-specific conditions, which allow for controlled synthesis of novel views across different perspectives.

Estimating 3D from Multi-view Images. Estimating 3D structure from multi-view images remains a foundational challenge in 3D vision. Classical approaches, such as COLMAP [41], tackle this problem with a multi-stage pipeline involving keypoint detection and matching, RANSAC [10], Perspective-n-Point (PnP) solvers [10], and a final bundled adjustment step for refinement. While classical geometric methods are effective, they require extensive engineering and optimization, often making it challenging to achieve accurate solutions, especially with large or complex datasets. Modern approaches study end-to-end learning methods that simplify the 3D estimation pipeline while also learning 3D priors from data. For example, VGGsfm [54] introduces differentiability at every stage of the COLMAP pipeline, making the process more adaptable to gradient-based optimization. DUSt3R [56] takes this a step further by employing Vision Transformers to regress point clouds directly from unposed image pairs. Mast3R [25] builds on these methods, enhancing performance by predicting features that increase the accuracy of keypoint matching, resulting in more reliable 3D reconstructions. These recent end-to-end approaches reduce the need for complex engineering and iterative processes, offering an efficient alternative to traditional multi-view 3D reconstruction pipelines and paving the way for more robust and scalable 3D vision applications.

3. World-consistent Video Diffusion (WVD)

In this section, we present World-consistent Video Diffusion Models (WVD), which leverage diffusion models to jointly model the distribution of RGB and XYZ frames across different viewpoints. We begin by introducing foundational concepts of diffusion models and its application in modeling 3D content (Sec. 3.1), followed by an in-depth discussion of our architectural design (Sec. 3.2).

3.1. Preliminaries

Diffusion Models. Standard diffusion models [18] operate by iteratively transforming noise into structured data through a denoising process. More specifically, a data point $\mathbf{x}_0$ is progressively noised through a forward process, yielding a sequence $\{\mathbf{x}_t\}_{t=1}^{T}$ according to a variance-scheduled Gaussian distribution. The diffusion model aims to reverse

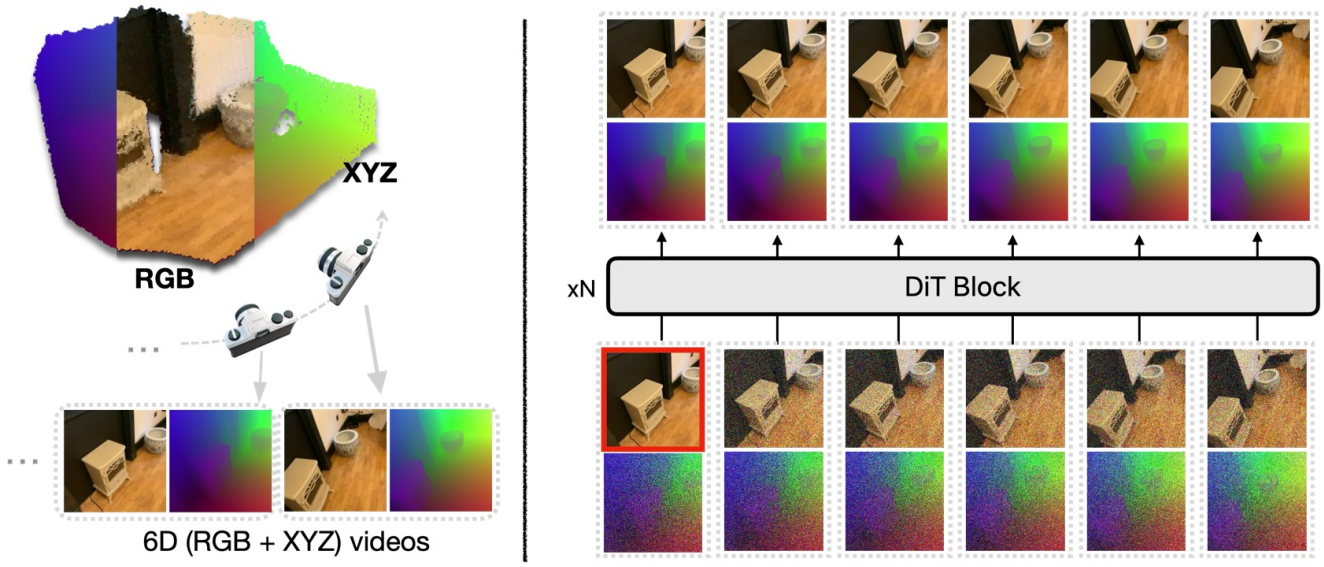

Figure 2. An illustration of **WVD** pipeline. The left part shows 6D videos formed by RGB and XYZ frames. On the right part, WVD iteratively denoises the 6D videos based on a specified RGB frame, which is highlighted with a red box.

this, parameterized as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, where θ denotes the model parameters. To reduce the computation cost for high-resolution inputs, LDM [38] improves by learning diffusion in the latent space $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ of a pretrained VAE [24]. Diffusion models are typically implemented using a UNet architecture [39]. Recently, however, Diffusion Transformers (DiT)[34] have emerged as a promising alternative. Leveraging the self-attention mechanism of Transformers to model intricate dependencies, DiT has demonstrated significant improvements in the fidelity of generated outputs and enhanced flexibility. This approach has shown potential across modalities, including images [6] and videos [3]. For instance, DiT can process videos by flattening and concatenating each frame into a single long sequence, allowing it to jointly denoise all frames.

Multi-view Diffusion Models. A common approach for diffusion models to learn 3D structure involves modeling the joint distribution of multi-view images [27, 43] and reconstructing 3D content in a second stage. This reconstruction is typically achieved either through optimization [31] or feed-forward prediction [22]. One can use DiT to process multi-view inputs similar to video diffusion, where DiT's attention layers operate across views. This implicitly captures 3D consistency, thus ensuring coherent image synthesis across perspectives. To make the diffusion process 3D controllable, approaches like CAT3D [11] condition the model on camera ray maps ($\mathbf{r}$) [46] using $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{r})$. This condition is crucial as it allows the trained model to generate novel views during inference.

However, this approach has two clear challenges: (1) It lacks explicit 3D guarantees, relying on the model to infer consistency purely from multi-view images. This often requires significant computational resources and high-quality data, yet it can still suffer from 3D inconsistency failures; (2) Furthermore, the dependence on camera ray inputs poses challenges for scaling to large datasets due to fundamental ambiguities in existing camera representations. These representations struggle to handle variations across datasets, necessitating non-trivial camera normalization [60], which further complicates the training process.

3.2. Approach

To tackle the primary challenges faced by traditional multi-view diffusion models, we propose **W**orld-consistent **V**ideo **D**iffusion (WVD), drawing inspiration from advancements in video diffusion models. Instead of incorporating additional camera control, our approach explicitly predicts 3D geometry by simultaneously diffusing over RGB frames and their corresponding point clouds. Specifically, the point clouds are projected into each frame as XYZ images.

XYZ Image Representation. Point clouds are a widely used representation of 3D geometry. However, their highly unstructured nature ($X \in \mathbb{R}^{N\times 3}$) poses significant challenges for learning with standard DiT architectures. To address this, we propose representing a 3D scene using multiple XYZ images, which provide a structured and learnable format. The transformation from a point cloud to an XYZ image is defined as:

$$\boldsymbol{x}^{\text{XYZ}} = \mathcal{R}(\mathcal{N}(X), X, C), \tag{1}$$

where $C = (P, K)$ represents the camera parameters, including the pose (P) and intrinsic (K) matrices. Here, $\mathcal{N}$ is a normalization function that centers and rescales the point clouds within the range [-1, 1], and $\mathcal{R}$ is a rasterizer that maps the normalized 3D point values onto an image plane, using the 3D positions and camera transformations. This representation ensures compatibility with existing architectures while preserving the geometric structure of the scene.

The XYZ image has the same shape as its RGB counterpart, with each pixel corresponding to a 3D point in the global coordinate system. By combining XYZ and RGB images into a unified 6D video representation, the model effectively captures a 3D region while maintaining compatibility with standard video diffusion architectures. Beyond its simplicity and learnability, representing 3D geometry using XYZ images offers several additional benefits:

- **Explicit Consistency Supervision:** XYZ images are texture-free and provide robust pixel alignment across views, unlike RGB images, which are influenced by variations in texture and lighting. When two pixels in different views share the same value in XYZ images, they correspond to the same location in the global 3D coordinate system. This property facilitates strong pixel correspondence across views, enabling direct 3D supervision during the generation of both XYZ and RGB images.
- **Elimination of Camera Control:** By encoding 3D geometry directly, XYZ images obviate the need for additional camera information to align multiple views, as required in existing methods [11, 16]. This approach reduces camera-related ambiguities, making it practical to scale up to larger and more complex datasets.

RGB-XYZ Diffusion. Following prior works [11], WVD learns a DiT-like model in the latent space, operating on a sequence of 6D video data $\{\boldsymbol{x}_n^{\text{RGB}}, \boldsymbol{x}_n^{\text{XYZ}}\}_{n=1}^{N}$. Since $\boldsymbol{x}_n^{\text{XYZ}}$ is pre-normalized, it can be directly processed using pretrained VAEs [38] without requiring additional fine-tuning. To improve computational efficiency, we concatenate the RGB and XYZ latents along the channel dimension before adding noise:

$$\boldsymbol{z}_n = [\mathcal{E}(\boldsymbol{x}_n^{\text{RGB}}); \mathcal{E}(\boldsymbol{x}_n^{\text{XYZ}})] \in \mathbb{R}^{L \times 2D}, \tag{2}$$

where L is the sequence length, D is the latent dimension. This design allows us to directly fine-tune pretrained image or video diffusion models, significantly enhancing training efficiency. The model can be trained using either text- or image-conditioned data, depending on the dataset. For instance, as illustrated in Fig. 2, in the case of image-conditioned generation, the added noise on the conditional image is simply removed at each iteration during training.

Post Optimization. As WVD directly predicts the point clouds in the global coordinates, we can easily perform a Perspective-n-Point (PnP) algorithm to recover the corresponding camera $C = (P, K)$ and depth maps $\boldsymbol{d}$ given the predicted XYZ images $\hat{\boldsymbol{x}}^{\text{XYZ}}$. In this paper, we directly perform gradient optimization over re-projection loss:

$$\min_{P,K,\boldsymbol{d}} \sum_{u,v} \|\tilde{\boldsymbol{x}}_{u,v}^{\text{XYZ}} - \hat{\boldsymbol{x}}_{u,v}^{\text{XYZ}}\|_2^2, \tag{3}$$

where $\tilde{\boldsymbol{x}}_{u,v}^{\text{XYZ}} = P^{-1}K^{-1}\boldsymbol{d}_{u,v}[u, v, 1]^{\text{T}}$, and u, v are the pixel coordinates. This post-optimization step is efficient and can be easily parallelized across views. Moreover, the optimized depth map $\boldsymbol{d}$ and camera parameters C provide a more accurate and physically consistent estimation $\tilde{\boldsymbol{x}}^{\text{XYZ}}$ of the original XYZ image, which can be highly beneficial for downstream tasks.

4. WVD as a 3D Foundation Model

WVD learns to generate RGB and XYZ frames together by modeling the joint probability $P(\text{RGB}, \text{XYZ})$, effectively capturing their interdependent structures and features. At inference time, this joint distribution can be leveraged to estimate conditional distributions, such as $P(\text{XYZ} \mid \text{RGB})$ or $P(\text{RGB} \mid \text{XYZ})$. This capability makes WVD a foundation for supporting a wide range of downstream tasks.

4.1. Single-image to 3D Tasks

Given its training methodology, WVD can be directly applied to various single-image tasks, including monocular depth estimation (as described in Eq. (3)), novel view synthesis, and 3D reconstruction. Notably, unlike traditional monocular depth estimation approaches that are typically supervised to infer depth from single-image inputs, our approach estimates depth through a generative process. By jointly sampling consistent surrounding views from the learnt data distribution, WVD produces depth predictions that are more 3D-grounded and consistent with the global scene geometry.

4.2. Multi-view Stereo Tasks

Since WVD learns the distribution of videos, it can also be applied to multi-view tasks with a collection of **unposed** RGB images provided. In this setup, the model predicts only the XYZ images through a diffusion process, following a procedure akin to "in-painting" [44]. At each diffusion step, the model's RGB predictions are replaced with the observed RGB values, ensuring consistency with the given inputs while generating the missing XYZ components. Consistent with the findings in [14], our early experiments revealed that incorporating additional Langevin correction steps [48] significantly enhances the quality and stability of the in-painting process. With the additional post-optimization steps (Eq. (3)), WVD not only reconstructs 3D geometry but also enables consistent multi-view video

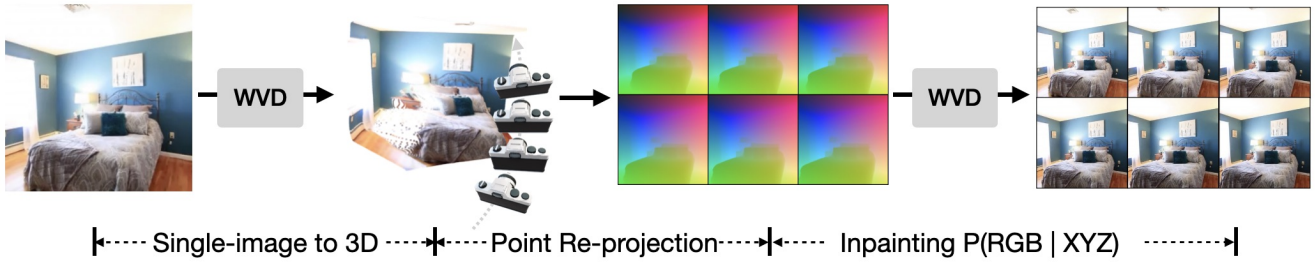

Figure 3. **Illustration of camera-controlled multi-view generation pipeline.** We first use WVD to infer the geometry from the input image, and then project it to obtain XYZ images for novel views. Next, we employ an inpainting strategy to sample RGB images.

depth estimation. This capability makes it highly valuable for applications that require accurate 3D scene interpretation from diverse viewing angles.

4.3. Controllable Generation Tasks

WVD also supports controllable video generation by leveraging the XYZ information in reverse mode. Similar to existing multi-view diffusion models, WVD is trained in a camera-agnostic manner, learning the underlying distribution of camera trajectories implicitly without requiring explicit conditional guidance. However, at inference time, the model can be adapted to enable video generation with camera control through point re-projection. As illustrated in Fig. 3, the pipeline involves the following steps:

1. **Single-image to 3D:** We first estimate the points of the input using standard WVD diffusion inference.
2. **Point Re-projection:** The synthesized point clouds are projected onto the target camera poses, producing partial XYZ images with corresponding projection masks, as described in Eq. (1);
3. **RGB & XYZ In-painting:** Finally, WVD regenerates the RGB images jointly with the projected XYZ images through an in-painting process.

Unlike the scenario in Sec. 4.2, the projected XYZ images in this case are typically incomplete, requiring the model to in-paint both the RGB and missing XYZ components during the diffusion process. In addition, the above point-guidance process enables us to maintain the union of synthesized points as a "spatial memory," where new video frames are guided by the projected points. This approach allows for the progressive generation of long video sequences while enforcing explicit consistency constraints, ensuring coherence across frames.

5. Experimental Results

5.1. Settings

Datasets. We train our model WVD on a mixture of datasets: RealEstate10K [71], ScanNet [7], MVImgNet [67], CO3D [37], and Habitat [40]. These datasets cover a broad range, from object-centric to scene-centric distributions. For RealEstate10K, MVImgNet, and CO3D, we use DUSt3R [56] to generate pseudo-ground-truth point clouds. ScanNet offers ground-truth depth maps that contain holes, which we fill using NeRF with depth regularization [8]. For Habitat, we directly utilize the rendered ground-truth point cloud. All images are center-cropped and resized to 256×256 resolution.

Table 1. Quantitative comparisons for single image to 3D.

Method	FID↓	KPM↑	FC↑
CameraCtrl [16]	**12.1**	88.6	94.0
MotionCtrl [59]	12.9	68.6	94.6
WVD	15.8	**95.8**	**95.4**
WVD w/o XYZ	18.3	72.3	95.0

Implementation details. Our Diffusion Transformer has 2 billion parameters and is implemented with rotary positional embedding [50] and RMSNorm [68]. A detailed model card is available in the Supplementary materials. As for training, we employ a learning rate of 3×10^{-4} using the AdamW optimizer, with the momentum parameter setting to $\beta = (0.99, 0.95)$. We train the model for 1 million steps, with an effective batch size of 128. The training takes approximately two weeks over 64 A100 GPUs.

5.2. Main Results

Single Image to 3D. Fig. 4 illustrates the synthesized RGB and XYZ frames conditioned on a single RGB frame. Our method effectively generates multi-view consistent frames with remarkable detail across a diverse range of visual distributions. Furthermore, we visualize the 3D scenes by projecting RGB pixels into 3D space using the corresponding coordinates from the XYZ frames. The resulting 3D point cloud exhibits realistic appearance and geometry, showcasing our ability to create 3D scenes from a single image.

For quantitative comparison with baselines, we choose

Figure 4. **Synthesized Multi-view RGB and XYZ Images by WVD, and associated reconstructed point clouds.** Input images are randomly sampled across the validation set.

the following metrics to measure the quality of synthesized frames: (1) Frechet Inception Distance (FID) [17] measures

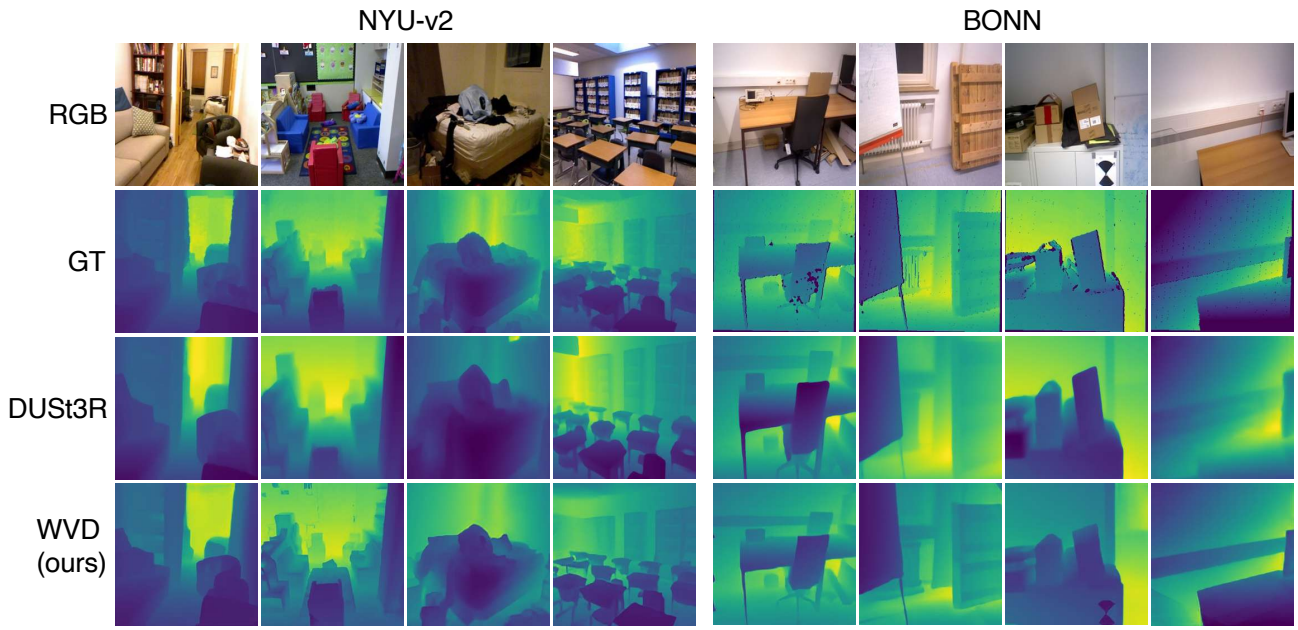

Figure 5. **Monocular depth estimation on NYU-v2 [45] and BONN [33] benchmarks.** We present RGB input images, ground-truth depth maps, and the predicted depth maps from DUSt3R (512 resolution) and WVD, respectively.

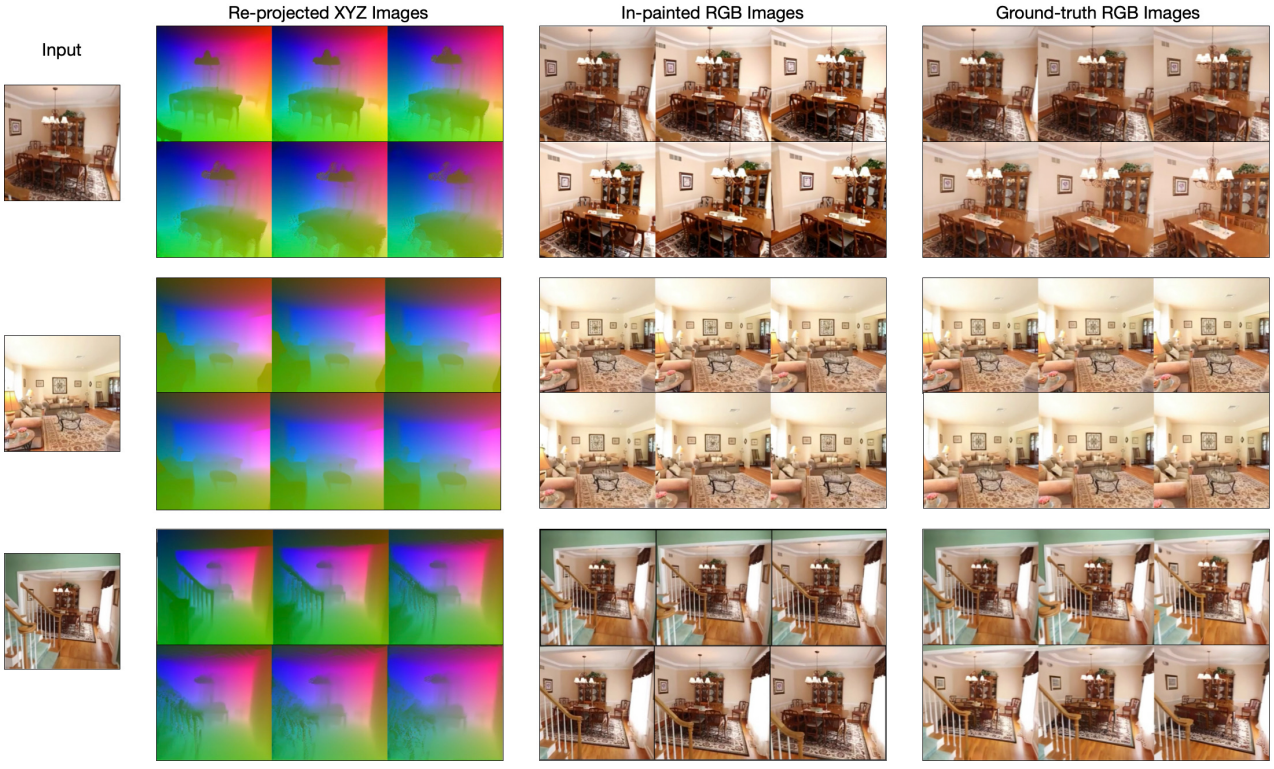

Figure 6. **Camera-controlled video generation.** By re-projecting XYZ images and using them as conditions, our method can control the camera movements in the synthesized videos, effectively replicating the trajectories of the real videos.

the per-frame appearance quality; (2) Key Points Matching (KPM) [58] assesses multi-view consistency by averaging

Table 2. Monocular depth estimation performance on NYU-v2 [45] and BONN [33]. *DUSt3R-512 was trained with higher resolution than ours.

Methods	NYU-v2 [45]		BONN [33]	
	Rel ↓	$\delta_{1.25}$ ↑	Rel ↓	$\delta_{1.25}$ ↑
RobustMIX [36]	11.8	90.5	-	-
SlowTv [49]	11.6	87.2	-	-
DUSt3R-224 [56]	10.3	88.9	11.1	89.1
DUSt3R-512 [56]*	<u>6.5</u>	<u>94.1</u>	8.1	93.6
WVD	**9.7**	**90.8**	**7.0**	**96.4**

the number of matching key points identified by a pretrained matching model [51]. We use the numbers obtained from the ground truth videos as a baseline and report the percentage of each method; (3) Frame Consistency (FC) [23] assesses a video based on the similarity of the CLIP image features among its frames. We choose CameraCtrl [16] and MotionCtrl [59] as baselines, and show the results in Fig. 4. Our method achieves comparable performance to CameraCtrl and MotionCtrl in terms of frame appearance. It consistently outperforms both baselines in multi-view consistency, as measured by KPM and FC, highlighting the advantages of jointly modeling XYZ images alongside RGB images.

Monocular depth estimation. As shown above, our method can generate multi-view RGB and XYZ images with single RGB frame. By converting the XYZ-encoded point clouds into dense depth maps, we enable monocular depth estimation. We evaluated our approach against other zero-shot methods on the NYU-v2 [45] and BONN [33] benchmarks for monocular depth estimation. We visualize the result in Fig. 5. Although our method is never trained over any depth prediction benchmark, our model can make precise prediction given monocular image as input. Quantitative comparison with baselines is presented in Tab. 2. On BONN, our method outperforms all baseline models. While DUSt3R trained at 512×512 achieves the best performance on NYU-v2, our model, despite being trained at a lower resolution (256×256), surpasses all other baseline methods.

Video depth estimation. As discussed in Sec. 4.2, our model can estimate the conditional distribution of P(XYZ | RGB) using an in-painting strategy. This allows us to adapt our model for estimating 3D geometry based on a set of unposed RGB images. We can sample point clouds from P(XYZ | RGB), and subsequently convert these point clouds into dense depth maps through post-optimization, effectively repurposing our method as a video depth estimator. We benchmark this capability and present the performance in Tab. 3. The results demonstrate that our method performs on par with state-of-the-art approaches..

Table 3. Video depth estimation performance on ScanNet++.

Method	ScanNet++	
	AbsRel ↓	$\delta_{1.03}$ ↑
COLMAP [41, 41]	14.6	34.2
Vis-MVSSNet [69]	8.9	33.5
MVS2D [64]	27.2	5.3
DeMon [53]	75.0	0.0
MVSNet [65]	65.2	28.5
Robust MVD [42]	7.4	38.4
DeepV2D [52]	**4.4**	54.8
DUSt3R-224 [56]	5.9	50.8
DUSt3R-512 [56]*	4.9	<u>60.2</u>
WVD	5.0	**57.2**

Camera-controlled Video Generation. As mentioned in Sec. 4.3, our model enables camera-controlled video generation. This is accomplished by first estimating 3D geometry from a single input image, which is then used to guide the generation of novel views.

We demonstrate this process in Fig. 6, showcasing the ground-truth videos, the corresponding projected XYZ images, and the videos generated by our method. The synthesized videos can mimic the camera motion observed in the real videos, highlighting the effectiveness of our approach for camera-controlled video synthesis.

Ablation of jointly predicting XYZ together with RGB frames. In Tab. 1, we assess the necessity of learning XYZ frames by training an RGB-only model. Without learning the XYZ frames, both image quality and multi-view consistency declines. This demonstrates that jointly learning the XYZ frames offers explicit 3D supervision, which enhances multi-view synthesis.

6. Discussions and Future Work

We introduce `WVD`, a DiT framework that jointly models the distribution of multi-view RGB and XYZ images, enabling direct 3D scene generation without the need for post-processing. Additionally, `WVD` can be adapted for various downstream tasks (e.g., monocular depth estimation, camera pose estimation) through a flexible inference strategy.

While `WVD` demonstrates the potential to serve as a 3D foundation model, the framework itself is not limited by modality. Future work could explore incorporating different modalities rather than 3D XYZ images (e.g., optical flow, splatter images) within our framework to support an even broader range of tasks.

Limitations. Our model currently has the following limitations: (1) We have only trained on static datasets, re-

stricting its application to static scenes. Extending this work to dynamic 4D datasets and jointly learning motion-related representations, such as optical flow, would be a valuable direction for future research. (2) Our model does not incorporate confidence maps, making it challenging to handle unbounded or outdoor scenes. Jointly modeling XYZ with confidence could improve performance in such scenarios.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1, 2

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 2

[3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 3

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. 1, 2

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. *arXiv preprint arXiv:2112.07945*, 2021. 2

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[11] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1, 3, 4

[12] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 1, 2

[13] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. *arXiv preprint arXiv:2302.10109*, 2023. 2

[14] Jiatao Gu, Qingzhe Gao, Shuangfei Zhai, Baoquan Chen, Lingjie Liu, and Josh Susskind. Control3diff: Learning controllable 3d diffusion models from single-view images. In *2024 International Conference on 3D Vision (3DV)*, pages 685–696. IEEE, 2024. 4

[15] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2

[16] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2, 4, 5, 8

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 2

[20] Jonathan Ho, Tim Salimans, Alexey A Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. 1, 2

[21] Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. Viewdiff: 3d-consistent image generation with text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5043–5052, 2024. 1, 2

[22] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao

Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 8

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[25] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 2

[26] 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, 2023. 1, 2

[27] 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. 1, 2, 3

[28] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[30] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 2

[31] 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. *arXiv preprint arXiv:2003.08934*, 2020. 2, 3

[32] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. *CVPR*, pages 11453–11464, 2021. 2

[33] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 7, 8

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 3

[35] 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. 12

[36] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 8

[37] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 5

[38] 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. 1, 2, 3, 4

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[40] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5

[41] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 8

[42] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *2022 International Conference on 3D Vision (3DV)*, pages 637–645. IEEE, 2022. 8

[43] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 3

[44] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3D Photography using Context-aware Layered Depth Inpainting. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4

[45] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. *European Conference on Computer Vision*, pages 746–760, 2012. 7, 8

[46] Vincent Sitzmann, Semon Rezchikov, William T Freeman, Joshua B. Tenenbaum, and Frédo Durand. Light Field Networks : Neural Scene Representations with Single-Evaluation Rendering. *arXiv:2106.02634*, 2021. 3

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[49] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the

world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15768–15779, 2023. 8

[50] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[51] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 8

[52] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 8

[53] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 8

[54] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 2

[55] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 1, 2

[56] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 5, 8

[57] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1, 2

[58] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 7

[59] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 5, 8

[60] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 3

[61] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 2

[62] Yinghao Xu, Sida Peng, Ceyuan Yang, Yujun Shen, and Bolei Zhou. 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18430–18439, 2022. 2

[63] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7079–7088, 2024. 1, 2

[64] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8574–8584, 2022. 8

[65] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 8

[66] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 12, 14

[67] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 5

[68] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[69] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214, 2023. 8

[70] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018. 12

[71] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 5

Appendix

A. Implementation Details

We begin by training a text-to-video Latent Diffusion Model (LDM) on web-scale datasets, which serves as the initialization for WVD. For this, we utilize the Variational Autoencoder (VAE) from SDXL [35]. We implement 3D self-attention to capture the spatial and temporal relationships between image patches. In addition to the timestep, we incorporate a binary mask to specify whether the frames are RGB or XYZ. This conditioning information is integrated through cross-attention. We remove the text embedding from the original model.

Classifier-free guidance. During training, we randomly select either a single RGB frame or a single XYZ frame for conditioning. With classifire-free-guidance (CFG) implemented, we randomly drop conditional images with a probability of 0.1 during training. The denoising score can be written as:

$$\boldsymbol{\epsilon} = (1+w)(k\boldsymbol{\epsilon}^{\text{RGB}} + (1-k)\boldsymbol{\epsilon}^{\text{XYZ}}) - w\boldsymbol{\epsilon}^{\text{Uncond}}, \quad \text{(A1)}$$

where $\boldsymbol{\epsilon}^{\text{RGB}}$, $\boldsymbol{\epsilon}^{\text{XYZ}}$, and $\boldsymbol{\epsilon}^{\text{Uncond}}$ represent the estimated scores with RGB conditioning, XYZ conditioning, and unconditional score, respectively. w is the guidance strength, and k balances the guidance between RGB and XYZ. In most cases, we select $k = 1$ to condition solely on RGB frames. For camera-controlled video generation, we use $k = 0.5$.

B. Camera-controlled Video Generation.

We present additional samples for camera-controlled video generation using the test set from RealEstate10K [70] in Fig. A1. As shown, the synthesized videos closely replicate the camera motion observed in the ground-truth videos, highlighting our model's camera-control capability. It's important to note that camera information was not utilized during training.

C. Multi-view Depth Estimation

We present samples over ScanNet++ [66] for multi-view depth estimation in Fig. A2. As demonstrated, our method can accurately estimate depth maps with video input.

D. Camera Estimation

As outlined in the main paper, our method can predict the corresponding XYZ frames through inpainting when provided with ground-truth RGB frames. In addition to video depth estimation, these XYZ frames can also be utilized for camera estimation tasks. Specifically, we can easily perform a Perspective-n-Point (PnP) algorithm to extract the camera poses from point clouds represented by the XYZ frames. We use gradient optimization over re-projection loss as specified in the main paper. Fig. A3 presents results from the test set of RealEstate10K [70]. Our method not only predicts accurate 3D geometry from unposed images but also estimates precise camera trajectories. The estimated camera poses are in close agreement with the ground truth.

E. In-the-wild Samples

We also evaluate our model on in-the-wild samples to assess its generalizability. As illustrated in Fig. A4, our model successfully generalizes to out-of-domain images, such as those generated by AIGC algorithms. It can produce novel view images, accurately estimate XYZ images, and reconstruct 3D scenes from a single image. This demonstrates that our method exhibits strong generalizability after training on a diverse mixture of datasets.

Figure A1. **Camera-controlled video generation.** For each sample, the first row shows the ground-truth video sequence, and the second row shows the synthesized frames which re-produce the camera trajectory. The conditioned frame is marked with a red box.

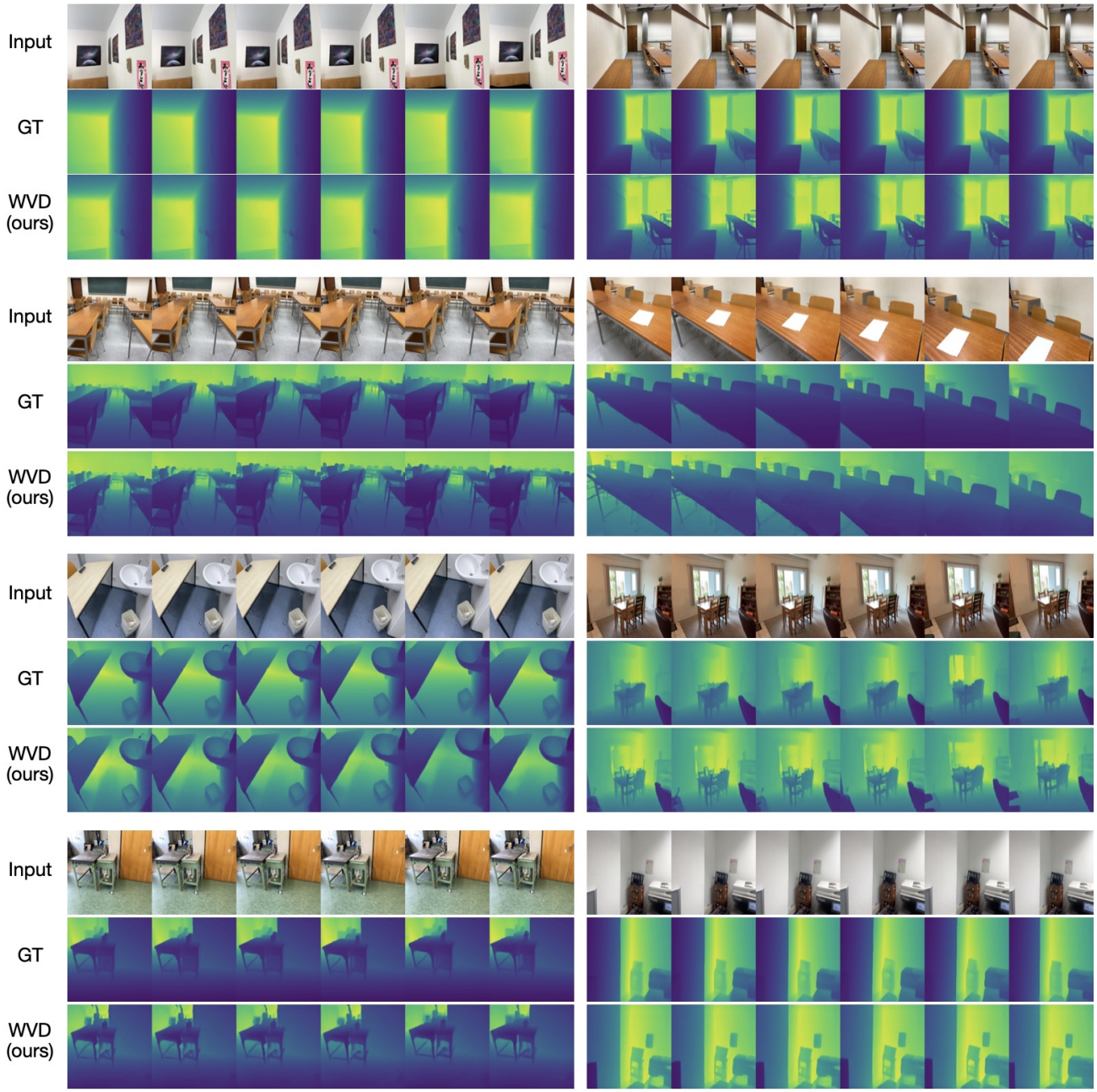

Figure A2. **Multi-view depth estimation on ScanNet++ [66].** For each sample, the first row presents the input video sequence, the second row shows the ground-truth depth maps. The third row shows the depth maps synthesized by our method.

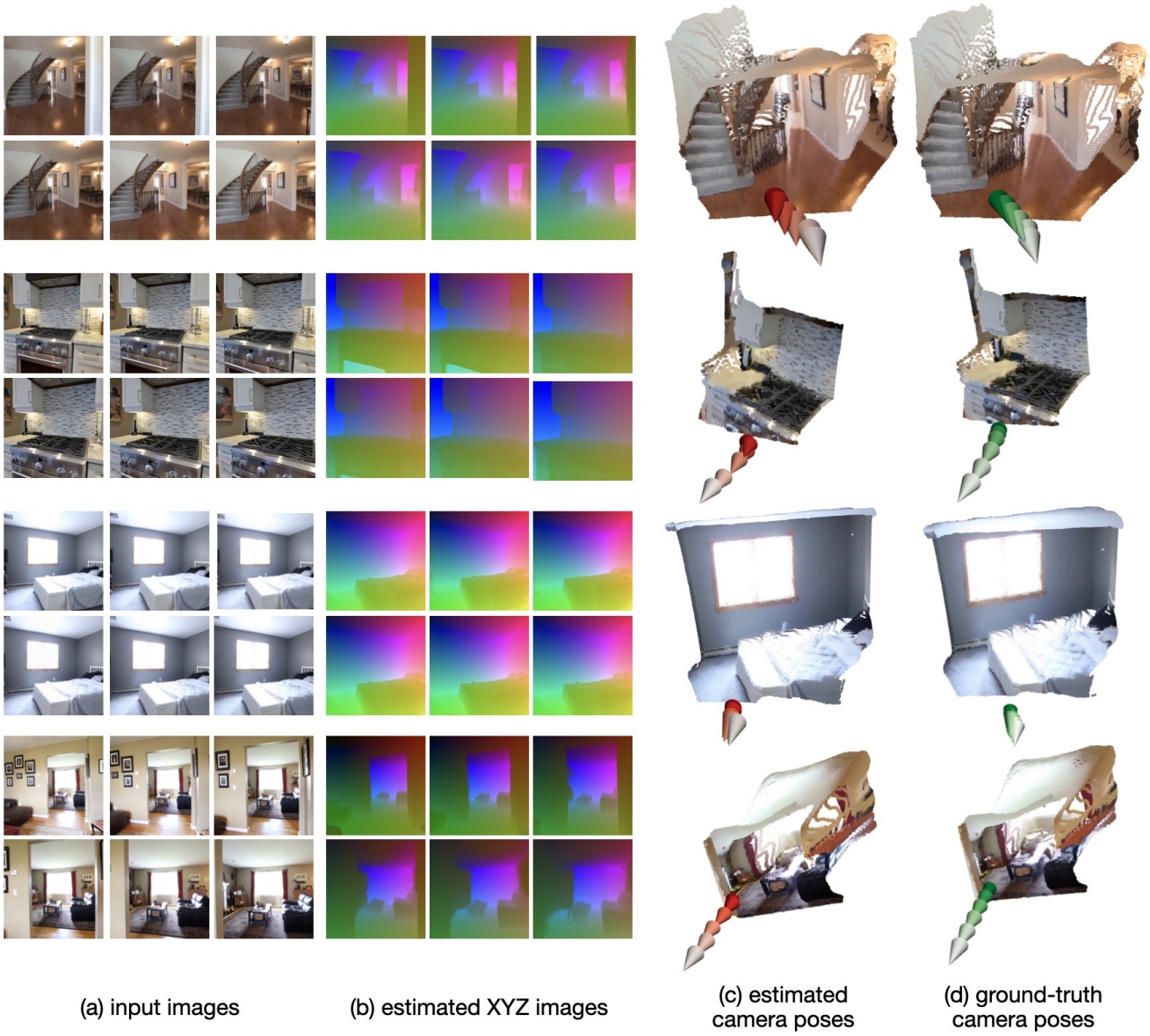

Figure A3. **Camera estimation.** Column (a) shows input unposed images. Column (b) shows estimated XYZ images by our method. Column (c) shows the estimated camera poses from the XYZ images, while column (d) provides the ground-truth camera poses.

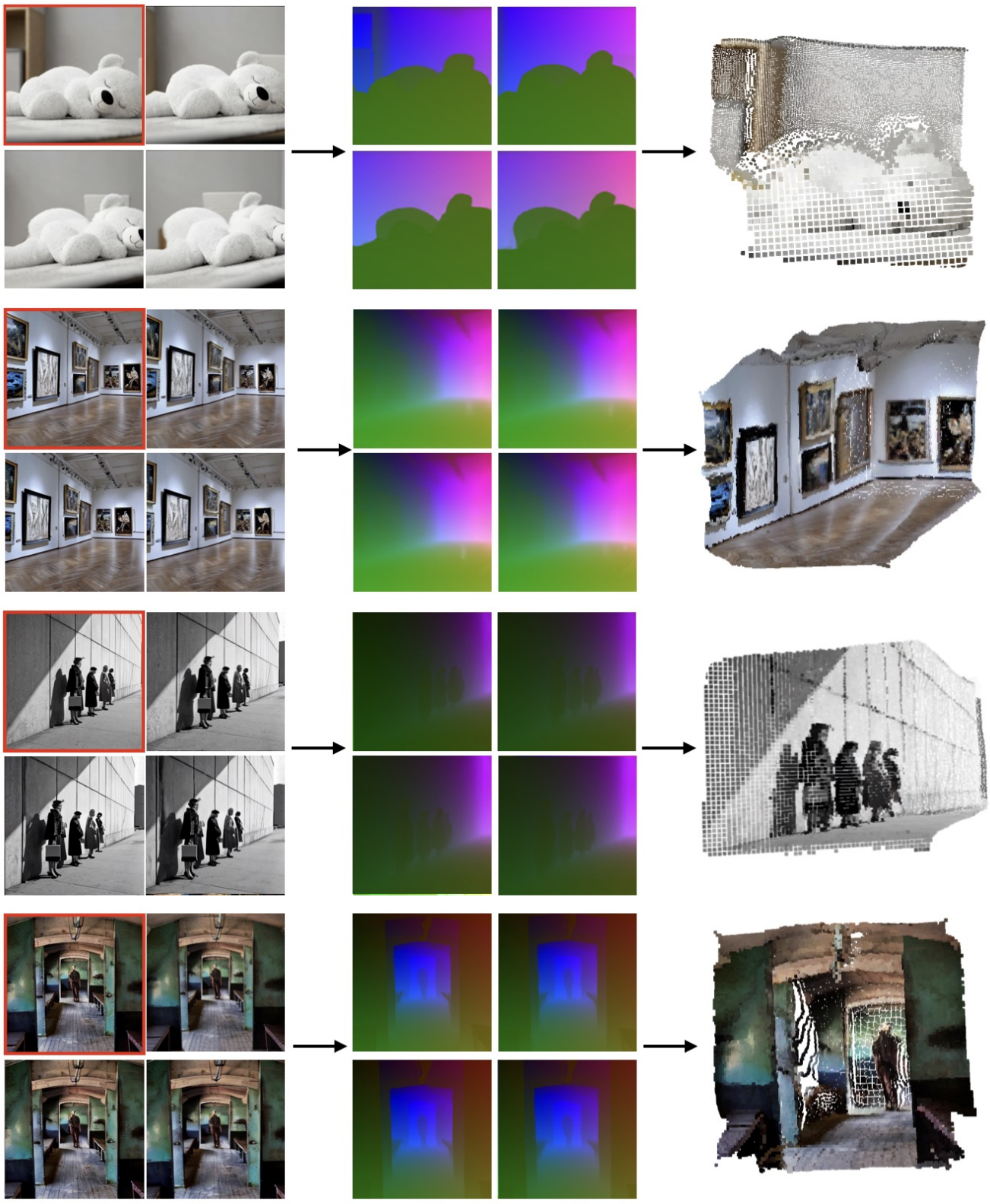

Figure A4. **In-the-wild samples.** We evaluate our model on in-the-wild samples to demonstrate its generalizability. The conditioned image is highlighted with a red box.