GeoWorld: Providing Full-frame Geometry Features to Facilitate 3D Scene Generation

Yuhao Wan[1,2], Lijuan Liu[2], Jingzhi Zhou[1], Zihan Zhou[3], Xuying Zhang[1], Dongbo Zhang[2], Shaohui Jiao[2], Qibin Hou[1,4], and Ming-Ming Cheng[4,1]✉

[1]VCIP & AAIS, Nankai University [2]ByteDance Inc.
[3]Renmin University of China [4]NKIARI, Shenzhen Futian

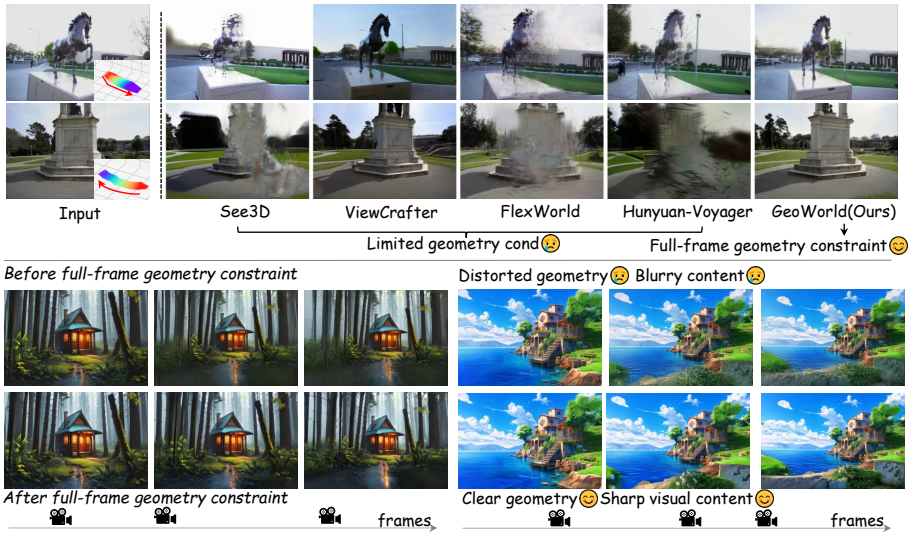


Fig. 1: Visual comparisons. Top: Comparison between our **GeoWorld** and previous methods. By incorporating full-frame geometry constraints, our approach achieves superior visual quality. Bottom left: Results before applying full-frame geometry constraints, which often suffer from geometric distortions and blurry content. Bottom right: Results after applying full-frame geometry constraints. By unlocking the potential of geometry models, our **GeoWorld** produces clear geometric structures and sharp visual details.

Abstract. Previous works that leverage video models for image-to-3D scene generation often suffer from geometric distortions and blurry content. Using video generation models to implicitly maintain geometric consistency according to a single-frame input is ineffective. In this paper, we present a two-stage method, named **GeoWorld**, that renovates the image-to-3D scene generation pipeline by providing full-frame geometry features. The first-stage video generation model, followed by a multi-view geometry model, produces **full-frame** geometry features, which are then used as a mental draft of geometric conditions to aid the second-stage video-generation model. A geometric loss is proposed to impose real-world geometric constraints, and a geometry adaptation module is introduced to ensure the effective utilization of geometry features. Thanks to full-frame geometric modeling, the two smaller video models in our two-stage

method can generate higher-fidelity 3D scenes than SOTA methods, while being even faster, *e.g.* 7.5× faster than Hunyuan-Voyager. Project page: https://peaes.github.io/GeoWorld.

Keywords: 3D scene generation · Video models · Diffusion models

1 Introduction

Generating a high-fidelity 3D scene from a single image has become a significant topic in recent years due to its high value in applications such as entertainment, interior and architectural design, and autonomous driving [10–12, 27, 51, 63, 73, 75, 78]. Leveraging deep learning methods for this task can significantly advance traditional 3D modeling pipelines. Given the limited information in a single image, a common approach is to use generative model priors to synthesize the scene content. Early methods [5, 6, 17, 29, 43, 65, 70, 76] often rely on 2D generative models [16, 44], which often lead to issues such as structural inconsistency and inconsistencies within the scene content.

Thanks to the advances in foundational 3D generative models, some recent works [4, 14, 20, 21, 28, 32, 36, 40, 45–47, 58, 64, 67, 68] use video models [18, 50, 62] and leverage their implicit 3D priors to alleviate the aforementioned issues. Such methods typically employ video models to synthesize a video under a specified camera trajectory from a single input image, and subsequently reconstruct a 3D scene from the generated video. However, generating high-fidelity videos from a single image remains challenging. As shown at the top of Fig. 1, these methods often suffer from geometric distortions and blurry content, which degrade the quality of the final 3D reconstruction. To address the above issues, a common approach is to provide additional geometric guidance for the model. As shown in Fig. 2(a), some previous works [4, 21, 67] have utilized estimated monocular depth maps as spatial priors or camera embeddings to assist in video generation. 'Optional' indicates that this step is not included in some methods.

Compared to tasks such as 3D reconstruction [34, 37, 52] or multi-view-to-3D scene generation [10, 68], image-to-3D scene generation is significantly more ill-posed. How to effectively provide geometric guidance for this task remains an open question. Previous works rely on limited geometric information extracted from a single input frame, which is insufficient to guide the generation of an entire video. Consequently, even larger models struggle to produce satisfactory results [4, 21]. However, providing corresponding geometric information for every frame is extremely challenging. Fortunately, we experimentally find that feeding rendered partial views into the video model produces coarse but content-complete condition views, as shown in Fig. 2(b). These views can be used to extract geometric information, thereby providing full-frame guidance for the subsequent generation process. Inspired by this, we propose a new two-stage pipeline paradigm for this task that incorporates full-frame geometric features into the video generation process. As shown in Fig. 2(b), in the first stage, we generate a video with complete content, and then leverage geometry models that can extract geometric

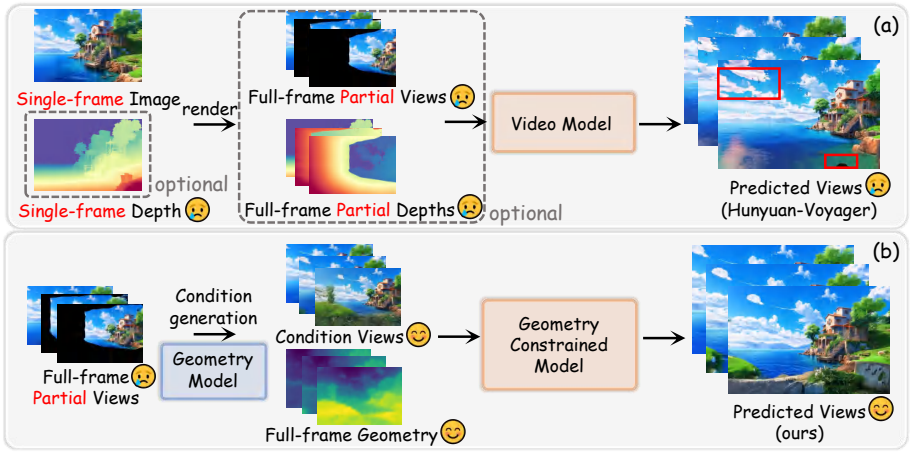


Fig. 2: Pipeline comparison. (a) Pipelines of previous methods [4, 21, 67]. Although details vary, their video models are conditioned only on single-frame information and limited geometric information. 'Optional' indicates that this step is not included in some methods. We can see that partial views or partial depths have very limited information. (b) Our GeoWorld leverages the geometric condition generation procedure and a geometry model to obtain full-frame geometry features for generation, rather than relying solely on geometry extracted from the input image.

information from multiple frames to provide geometry features. These features are then used to facilitate the second-stage video generation model.

To be specific, we first render the single-frame input using the given camera trajectory. The resulting rendered partial views are fed into a fine-tuned video model to generate a coarse but content-complete video. This video can then be used as input to a geometry model to extract full-frame geometry features, providing a mental draft of geometric conditions for the second-stage video generation model. Since the geometric information is derived from a model of limited accuracy, we further design a geometric alignment loss to compensate for this limitation. Instead of directly embedding the obtained geometry features as conditioning, our geometric alignment loss aligns the geometry features extracted from the predicted and ground-truth videos during training. This design imposes real-world geometric constraints on the model, and the number of frames is naturally consistent. Finally, we propose a geometry adaptation module to effectively exploit the extracted full-frame geometry features for improving video generation quality.

From Fig. 1, we can see that our GeoWorld is able to generate a high-fidelity 3D scene from a single image and a given camera trajectory, outperforming other state-of-the-art methods. Counterintuitively, although we introduce an additional procedure for obtaining full-frame geometry features, our two-stage method is more efficient, *e.g.* uses only 0.3× the model size and achieves 7.5×

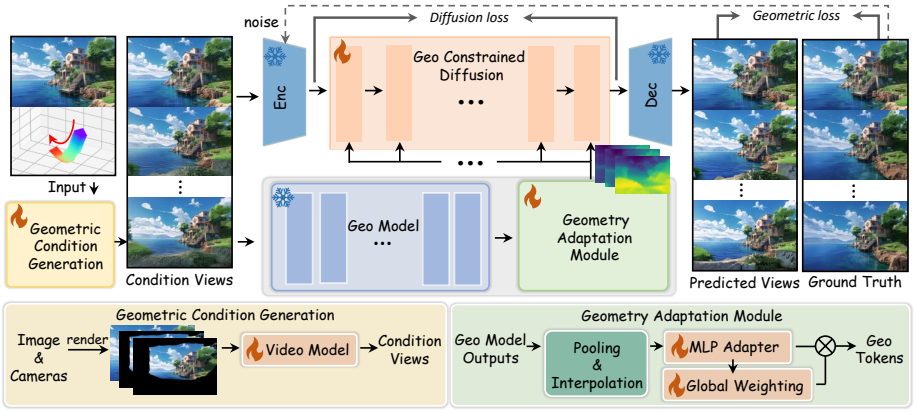


Fig. 3: Overview of our **GeoWorld**. During training, we perform the geometric condition generation procedure, feeding the obtained condition views into the geometry model to obtain full-frame geometry features, which are then processed by the geometry adaptation module. Together, the condition views and geometry features serve as input to the geometry-constrained diffusion model. The predicted views produced by this model are then used to reconstruct the 3DGS scene.

faster inference than Hunyuan-Voyager [21] (See Sec. 4.4). Our contributions can be summarized as follows:

- We propose GeoWorld, a novel two-stage pipeline paradigm for single-image-to-3D scene generation. We incorporate full-frame geometric features into the video generation process to alleviate the difficulty of generation caused by the limited input information inherent in this task.
- We explore how to leverage geometry models to assist the video generation process. In this exploration, we design a geometric condition generation procedure, a geometric alignment loss, and a geometry adaptation module to gradually unlock the potential of geometric information.
- Our GeoWorld outperforms previous methods qualitatively, achieves superior fidelity, yet is more computationally efficient.

2 Related work

2.1 3D Scene Generation

Existing scene generation methods can be broadly categorized into four types. The first category comprises view-by-view image inpainting approaches [5,25,30,39,61,65,66,76], which generate frames sequentially along a predefined camera trajectory. However, due to the lack of global semantic consistency, they struggle to produce semantically coherent full-scene results. The second category directly generates 3D scene representations [3,31,33,48,69] such as 3D Gaussian Splatting (3DGS) [23]. While these methods offer advantages in 3D consistency and reconstructability,

their performance is often constrained by the scarcity of high-quality 3D data, posing challenges for training robust models. The third category is compositional scene generation [1,7,22,55,63], where the key idea is to generate individual objects and place them plausibly within a scene layout. Although significant progress has been made in 3D object generation, these methods still face unresolved issues related to model stability, object placement, and occlusion handling. The final category is controllable video generation [4, 13, 36, 38, 46, 67, 74], which aims to synthesize a sequence of spatially consistent video frames given a camera trajectory. This is typically achieved by fine-tuning pre-trained video diffusion models, which can produce visually appealing videos. However, such methods often exhibit geometric inconsistencies and structural artifacts that degrade the quality of subsequent 3D reconstructions.

To overcome these limitations, our method leverages multi-frame geometric information and introduces constraints to enforce spatial consistency, leading to highly competitive reconstruction results.

2.2 Learning-Based 3D Reconstruction

Unlike traditional 3D reconstruction pipelines that require training separately for each individual scene, learning-based reconstruction methods leverage neural networks trained on large-scale scene datasets to encode strong scene priors, ultimately achieving impressive open-world generalization capabilities [9, 26, 34, 37, 41, 42, 52–54, 56, 59, 60, 72]. DUSt3R [54] directly regresses a 3D point cloud from a pair of RGB images. It extracts features from the two views using a transformer architecture enhanced with cross-attention mechanisms, and then feeds the fused features into a regression head to predict the point cloud along with a confidence map. Its successor, MASt3R [26], maintains the two-view regression paradigm but introduces a confidence-weighted loss to improve prediction reliability. Recent methods have generalized the two-view alignment-based architecture to handle multi-view scenarios, allowing for the joint processing of long frame sequences—up to 100 frames or more—as demonstrated by models such as Fast3R [60] and VGGT [52]. VGGT scales up the core idea of DUSt3R into a 1.2B-parameter transformer that jointly predicts camera intrinsics and extrinsics, dense depth maps, 3D point clouds, and 2D tracking features. The features extracted by such architectures exhibit strong 3D consistency and can support a wide range of downstream 3D tasks. Since video generation models often lack explicit 3D consistency control, introducing geometry features to guide the video generation process emerges as a natural and promising direction.

3 Methodology

As mentioned in Sec. 1, our goal is to leverage geometry models to provide reliable full-frame conditional signals for the video generation process. In practice, we use VGGT [52] as the geometry model. The overall architecture of our GeoWorld is shown in Fig. 3. Our design consists of three components: a geometric

condition generation procedure to obtain full-frame geometry features, a geometric alignment loss to introduce real-world geometric constraints, and a geometry adaptation module to utilize the geometry features effectively. For the video generation process, during training, we first fine-tune a video model. We then let this model perform inference on the entire training set, and the newly generated data is used to train the geometry-constrained diffusion model. Finally, we use the predicted views to reconstruct the 3DGS scene.

3.1 Geometric Condition Generation

The primary challenge of utilizing geometry models to help the video generation process lies in obtaining suitable geometry features. With only a single input frame, one can extract geometry for that frame alone, which provides limited guidance for subsequent frames in the video. We solve this by introducing a geometric condition generation procedure to attain appropriate full-frame geometry features. As shown in Fig. 3, this procedure employs a fine-tuned video model to generate conditional views from the single-frame input and the given camera trajectory. In the full GeoWorld pipeline, these conditional views are then processed by the geometry model to obtain full-frame geometry features. Specifically, the entire process consists of two components: rendering and completion.

Rendering. The goal of this component is to obtain a video, in which each frame is the rendering of the input single-frame image under the given camera trajectory. This video serves as the input to the video model. The rendering procedure differs between the training and inference phases to ensure higher-quality inputs during training. During training, we reconstruct the 3DGS scene using all available frames from the dataset, and then start from a random frame, extract its depth from the 3DGS, and perform back-projection and pairing process [4]. During inference, we directly estimate the point cloud of the single-frame input under the given camera trajectory using MAST3R [26], and then perform back-projection.

Completion. The goal of this component is to use the fine-tuned video model to complete the rendered video into a content-complete one. Specifically, we first collect a batch of training data to fine-tune the video model directly. Then, the fine-tuned video model performs inference on the entire training dataset, and the newly generated data is used to train the geometry-constrained diffusion model shown in Fig. 3. In terms of model design, since only single-frame geometry features are available at this stage, we simply use them as additional conditions and embed them into the model through cross-attention.

Through this process, we can obtain full-frame geometry features with the help of the geometry model. These outcomes also reflect, to some extent, the limitations of directly fine-tuning video models for geometrically consistent video generation. In Sec. 4.5, we further discuss the quality of the conditional views (Fig. 11) and validate the role of the geometry features in the subsequent optimization process through comparisons (Fig. 9).

3.2 Geometric Alignment Loss

Although the geometric condition generation process provides relatively complete geometric information, it is derived from a model of limited accuracy. To compensate for this limitation, we introduce real-world geometric information to guide the video generation process toward geometrically 3D scene synthesis. Specifically, we incorporate a geometric alignment loss into the diffusion objective, which compares the geometric features extracted from the generated and ground-truth videos. The loss is computed as the mean squared error between the two corresponding features obtained from the geometry model. In GeoWorld, this geometry model corresponds to the aggregator module of VGGT [52], with the decoding stage omitted to preserve complete geometric representations.

Specifically, given the ground-truth data I, a randomly sampled timestep t, and the corresponding Gaussian noise ϵ, we first encode I using the pre-trained VAE encoder and apply the forward process to obtain the input $\mathbf{x}$ for the geometry-constrained diffusion model. Geometry tokens c_{geo} are obtained as described in Sec. 3.3. The diffusion loss is then defined as:

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}, c_{geo}, t)\|_2^2\right]. \tag{1}$$

Given the predicted views I_{pred} and the geometry model G, the geometric loss is defined as:

$$\mathcal{L}_{geo} = \|G(I) - G(I_{pred})\|_2^2. \tag{2}$$

By leveraging the priors of the geometry model, this design could implicitly provide the model with real-world geometric information, ensuring that the optimization direction of the geometry-constrained model is geometrically consistent with the ground truth. The geometric alignment loss is defined as:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda\mathcal{L}_{geo}, \tag{3}$$

where λ is used as the weight for the geometric loss. As shown in Fig. 10, we observe that the outputs exhibit fewer geometric distortions and artifacts after incorporating the geometric alignment loss.

3.3 Geometry Adaptation Module

The goal of this subsection is to enable the model to effectively utilize the geometry features. In the latent space, since the output g of the geometry model differs from the input $\mathbf{x}$ of the video model along the frame, height, and width dimensions, we first perform pooling along the frame dimension and interpolation along the height and width dimensions to obtain g_{resize}, aligning its size with $\mathbf{x}$. We then train an MLP-based adapter to process the resized features and align them with the latent space of the video model, resulting in g_{ada}.

Since the generation of condition views lacks geometric guidance, some regions inevitably exhibit ambiguous geometric structures. To prevent these ambiguous structures from misleading the model, we perform a global weighting on g_{ada}

after processing it with the MLP adapter. We use an MLP-based predictor to integrate global information, similar to the Squeeze-and-Excitation block proposed in [19], and output a global weight for each token. By visualizing the global weights, we observe that low-weight tokens contain limited useful geometric information (Sec. 4.5). As a result, we multiply the global weights with g_{ada} and empirically discard the bottom 50% of tokens with lower weights to achieve a better performance, resulting in the final geometry tokens c_{geo}.

Finally, we fuse the obtained c_{geo} with $\mathbf{x}$ using a single-frame cross-attention in each layer. Given the query Q from $\mathbf{x}$, key K from c_{geo}, and value V from c_{geo}, the formulation can be written as follows:

$$\mathrm{CA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V}, \tag{4}$$

where B is an aligned relative position embedding and $\sqrt{d_k}$ is a scaling factor [8].

4 Experiments

4.1 Experimental Settings

Model and training details. We use Wan2.1-1.3B [50] as our video model. During training, we set the batch size, learning rate, input image resolution, video frame length, and the weight for the geometric loss to 16, 5e-5, 192×336, 17, and 0.2, respectively. The video model in the geometric condition generation procedure and the geometry-constrained model are trained for 7000 and 2000 iterations, respectively. Our geometry adaptation module and the geometry-constrained diffusion model are trained simultaneously. The training process is conducted on 8 NVIDIA A100 GPUs.

Training Dataset. We use DL3DV [35] as our training dataset and construct training pairs following the dataset construction method proposed in FlexWorld [4] (See Sec. 3.1). Specifically, we sample two epochs from DL3DV and select the top 25% of cases with the smallest average camera translation and rotation to ensure the quality of the partial views, resulting in approximately 5000 video pairs. We found that this simple filtering strategy effectively removes low-quality and facilitates training.

Testing and evaluation. We use the RealEstate10K (RE10K) [77] and the Tanks and Temples (Tanks) [24] as our test datasets, following the same construction method as FlexWorld. We randomly select 100 video clips from RE10K and 100 video clips from Tanks, each of which consists of 49 frames. We adopt PSNR, SSIM [57], LPIPS [71], FID [15], and FVD [49] as our evaluation metrics.

4.2 Comparisons with State-of-the-Art Methods

Quantitative comparisons. We show the quantitative comparisons between our GeoWorld and recent state-of-the-art methods (See3D [38], ViewCrafter [67],

Table 1: Quantitative comparison of our GeoWorld with recent state-of-the-art methods on **novel view synthesis**. The best performances are in **bold** and the second performances are underlined.

Datasets	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
RealEstate10K	**See3D** [38]	14.60	0.5307	0.4402	38.25	378.4
	ViewCrafter [67]	14.37	0.4854	0.4670	32.35	445.1
	FlexWorld [4]	14.28	0.5223	0.4418	**30.56**	**270.4**
	Hunyuan-Voyager [21]	14.85	0.5430	0.4357	52.32	569.6
	GeoWorld (ours)	**17.28**	**0.6193**	**0.3297**	31.00	311.7
Tanks and Temples	**See3D** [38]	13.00	0.3977	0.5400	53.36	571.9
	ViewCrafter [67]	12.53	0.3651	0.5558	41.33	716.3
	FlexWorld [4]	12.99	0.3938	0.5298	**38.69**	**422.6**
	Hunyuan-Voyager [21]	12.60	0.3855	0.5769	68.26	969.3
	GeoWorld (ours)	**14.99**	**0.4625**	**0.4556**	39.39	507.9

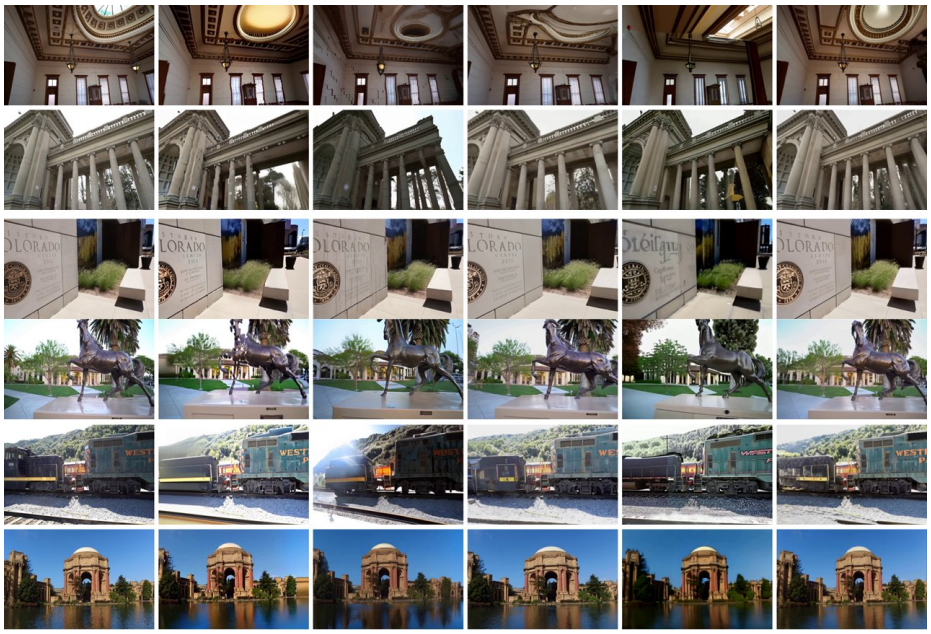
Table 2: Quantitative comparison of our GeoWorld with recent state-of-the-art methods on **3D scene generation**. The best performances are in **bold** and the second performances are underlined.

Method	RealEstate10K			Tanks and Temples		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
See3D [38]	14.67	0.5315	0.4413	13.14	0.3979	0.5420
ViewCrafter [67]	13.34	0.4847	0.4754	12.02	0.3649	0.5717
FlexWorld [4]	13.71	0.4775	0.5174	12.46	0.3543	0.5785
Hunyuan-Voyager [21]	13.52	0.4780	0.5189	12.27	0.3528	0.5950
GeoWorld (ours)	**16.64**	**0.5970**	**0.4284**	**15.00**	**0.4642**	**0.5058**

FlexWorld [4], and Hunyuan-Voyager [21]) in Tab. 1 and Tab. 2. Tab. 1 reports the quantitative comparisons in novel view synthesis (i.e., the video generation process). We obtain the input of the video model following the method described in Sec. 3.1. As shown, our GeoWorld outperforms previous methods in terms of fidelity (PSNR, SSIM) and achieves competitive perceptual quality (LPIPS, FID, and FVD). In particular, the LPIPS score is lower than those of all previous methods, while the FID and FVD scores are only slightly higher than FlexWorld [4].

Tab. 2 presents the quantitative comparisons in 3D scene generation. We reconstruct the generated videos into 3DGS and render images from corresponding camera poses to compute the evaluation metrics. As shown, our GeoWorld surpasses previous methods across all metrics, including fidelity (PSNR, SSIM) and perceptual quality (LPIPS). All these results demonstrate the effectiveness of our method.

Qualitative comparisons. Fig. 4 and Fig. 5 show the qualitative comparison results. Fig. 4 presents the qualitative comparisons in novel view synthesis (i.e., the video generation process). As shown, the videos generated by our GeoWorld contain fewer artifacts, richer details, and exhibit stable camera control. Fig. 5 shows the qualitative comparisons in 3D scene generation. As illustrated, our



(a)Ground Truth (b)See3d (c)ViewCrafter (d)FlexWorld (e)Voyager (f)GeoWorld(ours)

Fig. 4: Qualitative comparison of GeoWorld with state-of-the-art methods on **novel view synthesis**.

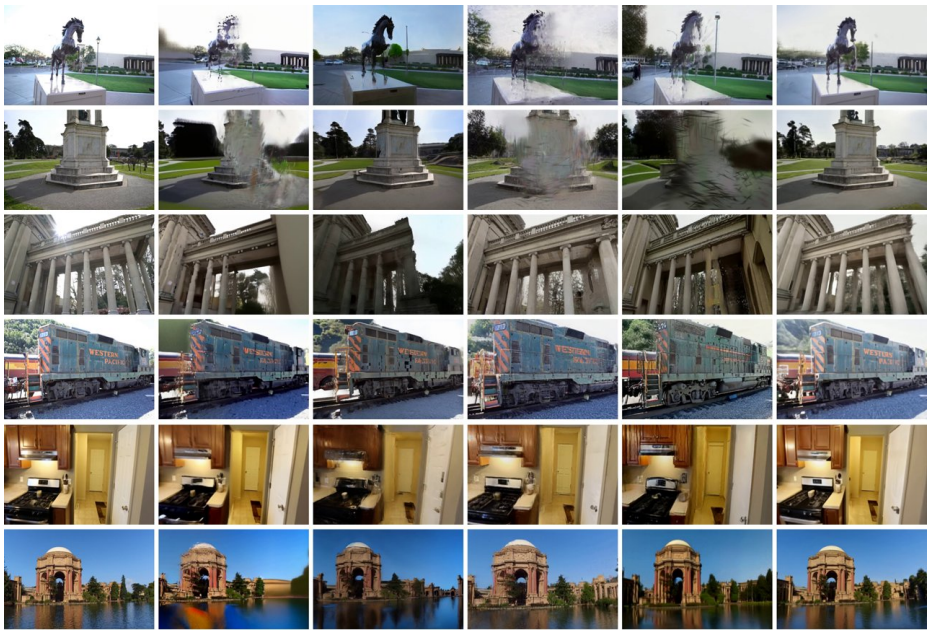
GeoWorld can maintain high-quality geometric structures, while producing fewer artifacts compared to previous methods.

4.3 3D Evaluation

To further evaluate the capabilities of the image-to-3D scene model, we conduct multi-view consistency and image matching tests using MEt3R [2] and MASt3R [26]. We reconstruct the generated videos into 3DGS and render images from corresponding camera poses to serve as the inputs for testing.

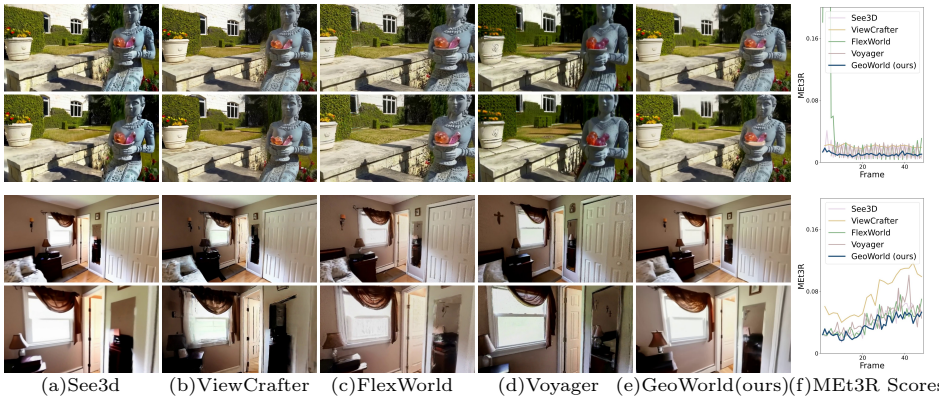
Multi-view consistency. Fig. 6 shows the MEt3R [2] results. We calculate the MEt3R score for each adjacent frame in a scene (49 frames), with lower scores indicating better multi-view consistency. As shown, our GeoWorld has the lowest MEt3R scores. Furthermore, since the videos used in the test are rendered by 3DGS, the lower MET3R score also indicates the stability of the 3DGS structure and that 3DGS contains fewer geometric conflict areas.

Image matching. Fig. 7 shows the image matching results. We input two frames from one scene, then use MASt3R [26] for image matching test and visualize the matched points. As seen, GeoWorld achieves the best matching results, and the output frames adhere to the geometric logic of 3D space at the pixel level. These test results demonstrate the superiority of our GeoWorld and validate the rationale behind our use of geometric conditions.



(a)Ground Truth (b)See3d (c)ViewCrafter (d)FlexWorld (e)Voyager (f)GeoWorld(ours)

Fig. 5: Qualitative comparison of our GeoWorld with recent stage-of-the-art methods on **3D scene generation**.

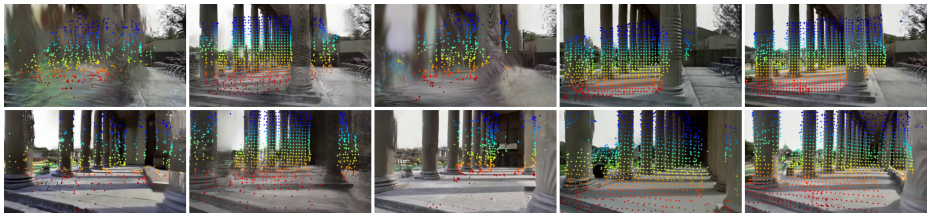


(a)See3d (b)ViewCrafter (c)FlexWorld (d)Voyager (e)GeoWorld(ours)(f)MEt3R Scores

Fig. 6: MEt3R results [2]. Lower scores indicate better multi-view consistency.

4.4 Cost Analysis

We report parameters (the sum of the two-stage models for GeoWorld), training time (train on 8 NVIDIA A100 80 GB GPUs, See3D [38] uses 114 NVIDIA A100 40 GB GPUs), inference time, and inference GPU memory cost in the Tab. 3 (inference on 1 NVIDIA A100 80 GB GPU). Train time and inference time of GeoWorld represent the total duration of our two-stage pipeline. As shown, our method has lower numbers of parameters and compute cost among video



Matching: 481 (a) See3d [38] | Matching: 890 (b) ViewCrafter [67] | Matching: 434 (c) FlexWorld [4] | Matching: 946 (d) Voyager [21] | Matching: **1250** (e) GeoWorld (ours)

Fig. 7: Image matching results [26]. 'Matching' denotes the number of matched points between two views. A higher count of matches indicates superior multi-view consistency of the views.

Table 3: Cost analysis. Our GeoWorld has lower numbers of parameters and compute cost among video diffusion model–based methods (ViewCrafter [67], FlexWorld [4], and Hunyuan-Voyager [21]). See3D [38] has lower GPU memory cost since it is based on a 2D diffusion model.

Model	PSNR(RE10K)	Params.	Train. Time	Infer. Size	Infer. Time	Infer. Mem.
See3d [38]	14.60	1.6B	25days	512*512*25	73s	9G
ViewCrafter [67]	14.37	2.6B	-	576*1024*25	120s	24G
FlexWorld [4]	14.28	5.0B	7days	576*1024*49	201s	28G
Hunyuan-Voyager [21]	14.85	12.8B	-	512*768*49	1110s	48G
GeoWorld (ours)	**17.28**	3.9B	4days	576*1024*49	148s	31G

diffusion model–based methods (ViewCrafter [67], FlexWorld [4], and Hunyuan-Voyager [21]), with an acceptable increase in GPU memory used to load the VGGT model during inference (See3D [38] has lower GPU memory cost since it is based on a 2D diffusion model).

4.5 Ablation Study

Direct analysis of gains from the geometric conditions. We perform two experiments: (i) The removal of geometric conditions throughout the entire process(Fig. 8(c)). The purpose of (i) is to provide a straightforward 'base model' to demonstrate the gains from our method. (ii) The model conditioned on depth maps(Fig. 8(d)). The purpose of (ii) is to demonstrate that the embedding of geometric features is superior to simple geometric conditions. As shown in Fig. 8, GeoWorld could provide clearer geometry and sharper visual content. This indicates that providing no geometric conditions or providing simple geometric conditions for the video model are both worse than GeoWorld's method of providing high-quality geometric conditions.

Effectiveness of the geometry-constrained diffusion model. We then evaluate the effectiveness of geometry-constrained diffusion model by comparing the quality of the condition views produced by the geometric condition generation procedure with predicted views produced by the geometry-constrained diffusion model. Fig. 9 shows the visual comparisons. It can be clearly observed that

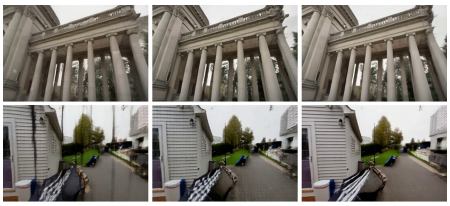
(a)Input (b)Ground Truth (c)w/o geo. cond. (d)depth only (e)GeoWorld

Fig. 8: Gains from the geometric conditions. As shown, providing no geometric conditions or providing simple geometric conditions for the video model are both worse than GeoWorld's method of providing high-quality geometric conditions.

(a) Before geo-constrained model

(b) After geo-constrained model

Fig. 9: Visual comparisons of the geometry constraints.



(a)Emb. directly (b)+GAL (c)+GAM

Fig. 10: Visual comparisons of the design of the geometry-constrained diffusion model. 'GAL': geometric alignment loss. 'GAM': geometry adaptation module.

Table 4: Design of geometry-constrained diffusion model.

Model	PSNR↑	LPIPS↓
Embed directly	16.77 (+0.00)	0.3381 (+0.0000)
+ Geometric Alignment Loss	16.96 (+0.19)	0.3284 (-0.0097)
+ Geometry Adaptation Module		
+ Resize	17.17 (+0.40)	0.3286 (-0.0095)
+ Global Weighting	17.28 (+0.51)	0.3292 (-0.0089)

the refined videos contain almost no blurry regions, the generated objects are sharper, more detailed, and exhibit clearer geometric structures, demonstrating the effectiveness of geometric constraints.

Design of the geometry-constrained diffusion model. We then verify the rationality of our model design. The design of the geometry-constrained diffusion model consists of two main components: the geometric alignment loss described in Sec. 3.2 and the geometry adaptation module described in Sec. 3.3. The quantitative comparison results on the RE10K [77] test set are presented in Tab. 4. 'Embed directly' refers to directly embedding the geometry features into the model via cross-attention. As shown, each component of our model contributes to improvements in the PSNR metric. In particular, introducing the geometric

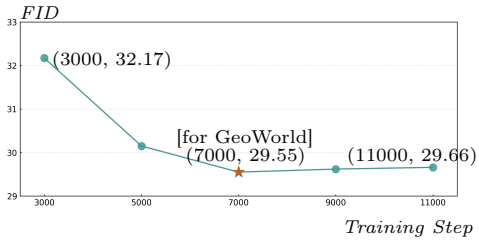


Fig. 11: Quantitative comparisons of the geometric condition generation procedure.

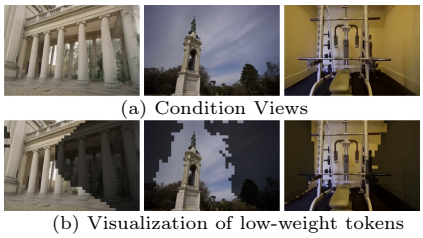
(a)7000 Steps (b)11000 Steps (c)After geo. constraint

Fig. 12: Visual comparisons of the geometric condition generation procedure.

alignment loss leads to an improvement in the LPIPS metric, indicating enhanced perceptual quality, which is consistent with our objective of incorporating real-world geometric information. We also provide visual comparisons in Fig. 10. It can be observed that after introducing the geometric alignment loss, the distorted geometry and artifacts in the outputs are notably reduced, while further incorporating the geometry adaptation module leads to clearer geometric structures, reduces grid-like artifacts, and produces cleaner visual results.

Discussion about the geometric condition generation procedure. The geometric condition generation procedure in our overall pipeline requires additional discussion, mainly concerning the performance of the video model fine-tuned during this process. If the fine-tuned video model is undertrained, it may fail to provide sufficient geometric information and could even introduce incorrect guidance to the geometry-constrained model. If the fine-tuned video model is trained for a sufficiently long time and can already generate highly detailed videos, the subsequent refinement process would become unnecessary. The results presented in Fig. 11 and Fig. 12 might address these concerns. As shown in Fig. 11, the model trained for 7000 steps achieves the lowest FID, and further increasing the training steps does not lead to additional improvements. The visual comparisons in Fig. 12 show that the fine-tuned video model (7000 steps) in the geometric condition generation procedure still suffers from geometric distortions and blurry content even after an additional 4000 training steps (11000 steps). In contrast, the geometry-constrained model (with an additional 2000 steps) produces videos with clearer geometry and sharper visual content.

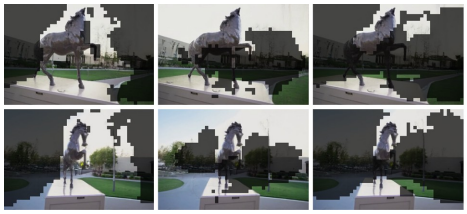
Analysis of global weighting process. We analyze the global weighting process proposed in Sec. 3.3. Fig. 13 presents the visualization results. The first row shows the condition views provided by the geometric condition generation procedure, while the darker regions in the second row visualize the bottom 50% low-weight tokens identified by the global weighting process. It can be observed that our global weighting process effectively filters out low-quality tokens. As illustrated in the first column, this process is able to detect blurry regions within the condition views. In the second and third columns, when the conditional views contain few blurry regions, the process tends to retain regions containing richer geometric structures (e.g., the sky and plain-colored walls contain limited geometric information).



(a) Condition Views

(b) Visualization of low-weight tokens

Fig. 13: Visualization of the global weighting process. The darker regions in the second row visualize the low-weight tokens identified by the process.



(a) 50% (b) 30% (c) 70%

Fig. 14: Visual comparisons of different discard ratios in the global weighting process. The darker regions visualize the discarded tokens.

Table 5: Quantitative results of different discard ratios in the global weighting process.

Discard ratio	PSNR↑	SSIM↑
50% (For GeoWorld)	**17.28** (+0.00)	**0.6193** (+0.0000)
30%	17.17 (-0.11)	0.6173 (-0.0020)
70%	17.22 (-0.06)	0.6188 (-0.0005)

Ablation on the discard ratio in the global weighting process. Fig. 14 presents the visualization of the discarded tokens. As shown, when the discard ratio is set to 30%, some low-quality regions(e.g., blurry content or regions that contain limited geometric information) are still retained. When the ratio increases to 70%, although most low-quality regions are effectively removed, a considerable portion of the geometrically informative main regions is also discarded. A discard ratio of 50% strikes a favorable balance between the two. Furthermore, Tab. 5 presents the quantitative results of different discard ratios. As can be seen, a discard ratio of 50% achieves the best PSNR and SSIM performance.

5 Conclusions

In this paper, we propose GeoWorld, a novel pipeline paradigm that can provide full-frame geometric features to the video generation process to alleviate the high generation difficulty caused by the limited input information inherent in this task. Furthermore, we explore how to leverage geometry models to obtain full-frame features and assist the video generation process. Extensive experiments demonstrate the effectiveness of our method, which outperforms previous methods qualitatively while achieving superior fidelity and competitive perceptual quality quantitatively. We expect that our GeoWorld can inspire future research and provide a new perspective for designing image-to-3D scene generation models.

Acknowledgments. This work was funded by the National Natural Science Foundation of China under 62522607, 62495061, and 625B2093, and the Fundamental Research Funds for the Central Universities (Nankai University).

References

1. Ardelean, A., Özer, M., Egger, B.: Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In: 2025 International Conference on 3D Vision (3DV). pp. 616–626. IEEE (2025)
2. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6034–6044 (2025)
3. Cai, Y., Zhang, H., Zhang, K., Liang, Y., Ren, M., Luan, F., Liu, Q., Kim, S.Y., Zhang, J., Zhang, Z., et al.: Baking gaussian splatting into diffusion denoiser for fast and scalable single-stage image-to-3d generation and reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25062–25072 (2025)
4. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. arXiv preprint arXiv:2503.13265 (2025)
5. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
6. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R., Cohen-Or, D.: Set-the-scene: Global-local training for generating controllable nerf scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2920–2929 (2023)
7. Dong, W., Yang, B., Yang, Z., Li, Y., Hu, T., Bao, H., Ma, Y., Cui, Z.: Hiscene: creating hierarchical 3d scenes with isometric view generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9783–9792 (2025)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fang, X., Gao, J., Wang, Z., Chen, Z., Ren, X., Lyu, J., Ren, Q., Yang, Z., Yang, X., Yan, Y., et al.: Dens3r: A foundation model for 3d geometry prediction. arXiv preprint arXiv:2507.16290 (2025)
10. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024)
11. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: International Conference on Learning Representations. vol. 2024, pp. 22841–22860 (2024)
12. Guo, J., Ding, Y., Chen, X., Chen, S., Li, B., Zou, Y., Lyu, X., Tan, F., Qi, X., Li, Z., et al.: Dist-4d: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27231–27241 (2025)
13. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
14. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023)
18. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Huang, J., Yang, Y., Yang, B., Ma, L., Ma, Y., Liao, Y.: Gen3r: 3d scene generation meets feed-forward reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25358–25369 (2026)
21. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
22. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23646–23657 (2025)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
24. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
25. Lei, J., Tang, J., Jia, K.: Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8422–8434 (2023)
26. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
27. Li, B., Guo, J., Liu, H., Zou, Y., Ding, Y., Chen, X., Zhu, H., Tan, F., Zhang, C., Wang, T., et al.: Uniscene: Unified occupancy-centric driving scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11971–11981 (2025)
28. Li, G., Zheng, S., Xu, S., Chen, J., Li, B., Hu, X., Zhao, L., Jiang, P.T.: Magicworld: Interactive geometry-driven video world exploration. arXiv preprint arXiv:2511.18886 (2025)
29. Li, H., Shi, H., Zhang, W., Wu, W., Liao, Y., Wang, L., Lee, L.h., Zhou, P.Y.: Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In: European Conference on Computer Vision. pp. 214–230. Springer (2024)
30. Li, M.F., Ku, Y.F., Yen, H.X., Liu, C., Liu, Y.L., Chen, A.Y., Kuo, C.H., Sun, M.: Genrc: Generative 3d room completion from sparse image collections. In: European Conference on Computer Vision. pp. 146–163. Springer (2024)
31. Li, X., Lai, Z., Xu, L., Qu, Y., Cao, L., Zhang, S., Dai, B., Ji, R.: Director3d: Real-world camera trajectory and 3d scene generation from text. Advances in neural information processing systems **37**, 75125–75151 (2024)
32. Li, X., Wang, T., Gu, Z., Zhang, S., Guo, C., Cao, L.: Flashworld: High-quality 3d scene generation within seconds. arXiv preprint arXiv:2510.13678 (2025)

33. Liang, H., Cao, J., Goel, V., Qian, G., Korolev, S., Terzopoulos, D., Plataniotis, K.N., Tulyakov, S., Ren, J.: Wonderland: Navigating 3d scenes from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 798–810 (2025)
34. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
35. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
36. Liu, L., Li, W., Zhang, D., Wang, S., Jiao, S.: Idcnet: Guided video diffusion for metric-consistent rgbd scene generation with precise camera control. arXiv preprint arXiv:2508.04147 (2025)
37. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: Worldmirror: Universal 3d world reconstruction with any-prior prompting. arXiv preprint arXiv:2510.10726 (2025)
38. Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., Wang, X.: You see it, you got it: Learning 3d creation on pose-free videos at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2016–2029 (2025)
39. Ma, Y., Zhan, D., Jin, Z.: Fastscene: Text-driven fast 3d indoor scene generation via panoramic gaussian splatting. arXiv preprint arXiv:2405.05768 (2024)
40. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., Zhang, K.: Yume1. 5: A text-controlled interactive world generation model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7752–7761 (2026)
41. Meng, M., Zhu, Y., Zhao, Y., Li, Z., Zhu, Z.: 3d indoor scene geometry estimation from a single omnidirectional image: A comprehensive survey. Computational Visual Media (2025)
42. Peng, Z., Zhou, K., Shao, T.: Gaussian-plus-sdf slam: High-fidelity 3d reconstruction at 150+ fps. Computational Visual Media (2025)
43. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
45. Shin, J., Li, Z., Zhang, R., Zhu, J.Y., Park, J., Shechtman, E., Huang, X.: Motionstream: Real-time video generation with interactive motion controls. arXiv preprint arXiv:2511.01266 (2025)
46. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024)
47. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. arXiv preprint arXiv:2512.14614 (2025)
48. Szymanowicz, S., Insafutdinov, E., Zheng, C., Campbell, D., Henriques, J.F., Rupprecht, C., Vedaldi, A.: Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. In: 2025 International Conference on 3D Vision (3DV). pp. 670–681. IEEE (2025)
49. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. In: ICLR Workshop (2019)

50. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
51. Wang, C., Peng, H.Y., Liu, Y.T., Gu, J., Hu, S.M.: Diffusion models for 3d generation: A survey. Computational Visual Media **11**(1), 1–28 (2025)
52. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
53. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10510–10522 (2025)
54. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
55. Wang, Y., Qiu, X., Liu, J., Chen, Z., Cai, J., Wang, Y., Wang, T.H.J., Xian, Z., Gan, C.: Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting. Advances in Neural Information Processing Systems **37**, 67575–67603 (2024)
56. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
57. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
58. Wu, R., He, X., Cheng, M., Yang, T., Zhang, Y., Kang, Z., Cai, X., Wei, X., Guo, C., Li, C., et al.: Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. arXiv preprint arXiv:2602.02393 (2026)
59. Wu, T., Yuan, Y.J., Zhang, L.X., Yang, J., Cao, Y.P., Yan, L.Q., Gao, L.: Recent advances in 3d gaussian splatting. Computational Visual Media **10**(4), 613–642 (2024)
60. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21924–21935 (2025)
61. Yang, S., Tan, J., Zhang, M., Wu, T., Wetzstein, G., Liu, Z., Lin, D.: Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025)
62. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations. vol. 2025, pp. 83048–83077 (2025)
63. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions on Graphics (TOG) **44**(4), 1–19 (2025)
64. Ye, D., Zhou, F., Lv, J., Ma, J., Zhang, J., Lv, J., Li, J., Deng, M., Yang, M., Fu, Q., et al.: Yan: Foundational interactive video generation. arXiv preprint arXiv:2508.08601 (2025)
65. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5916–5926 (2025)

66. Yu, H.X., Duan, H., Hur, J., Sargent, K., Rubinstein, M., Freeman, W.T., Cole, F., Sun, D., Snavely, N., Wu, J., et al.: Wonderjourney: Going from anywhere to everywhere. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6658–6667 (2024)
67. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
68. Zhai, S., Ye, Z., Liu, J., Xie, W., Hu, J., Peng, Z., Xue, H., Chen, D., Wang, X., Yang, L., et al.: Stargen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26822–26833 (2025)
69. Zhang, K., Bi, S., Tan, H., Xiangli, Y., Zhao, N., Sunkavalli, K., Xu, Z.: Gs-lrm: Large reconstruction model for 3d gaussian splatting. In: European Conference on Computer Vision. pp. 1–19. Springer (2024)
70. Zhang, Q., Wang, C., Siarohin, A., Zhuang, P., Xu, Y., Yang, C., Lin, D., Zhou, B., Tulyakov, S., Lee, H.Y.: Towards text-guided 3d scene composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6829–6838 (2024)
71. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
72. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21936–21947 (2025)
73. Zhang, X., Zhou, Y., Wang, K., Wang, Y., Li, Z., Jiao, S., Zhou, D., Hou, Q., Cheng, M.M.: Ar-1-to-3: Single image to consistent 3d object via next-view prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26273–26283 (2025)
74. Zhao, Y., Lin, C.C., Lin, K., Yan, Z., Li, L., Yang, Z., Wang, J., Lee, G.H., Wang, L.: Genxd: Generating any 3d and 4d scenes. In: International Conference on Learning Representations. vol. 2025, pp. 95479–95499 (2025)
75. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12405–12414 (2025)
76. Zhou, S., Fan, Z., Xu, D., Chang, H., Chari, P., Bharadwaj, T., You, S., Wang, Z., Kadambi, A.: Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In: European Conference on Computer Vision. pp. 324–342. Springer (2024)
77. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
78. Zhou, Y., Li, Z., Ouyang, Z., Chen, Y., Du, R., Zhou, D., Fu, B., Liu, Y., Gao, P., Cheng, M.M., et al.: Onevae: Joint discrete and continuous optimization helps discrete video vae train better. arXiv preprint arXiv:2508.09857 (2025)