Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

Haoyu Wu[1*], **Diankun Wu**[2*], **Tianyu He**[1†], **Junliang Guo**[1], **Yang Ye**[1], **Yueqi Duan**[2], **Jiang Bian**[1]

[1]Microsoft Research [2]Tsinghua University

Abstract

Videos inherently represent 2D projections of a dynamic 3D world. However, our analysis suggests that video diffusion models trained solely on raw video data often fail to capture meaningful geometric-aware structure in their learned representations. To bridge this gap between video diffusion models and the underlying 3D nature of the physical world, we propose Geometry Forcing, a simple yet effective method that encourages video diffusion models to internalize latent 3D representations. Our key insight is to guide the model's intermediate representations toward geometry-aware structure by aligning them with features from a pretrained geometric foundation model. To this end, we introduce two complementary alignment objectives: Angular Alignment, which enforces directional consistency via cosine similarity, and Scale Alignment, which preserves scale-related information by regressing unnormalized geometric features from normalized diffusion representation. We evaluate Geometry Forcing on both camera view–conditioned and action-conditioned video generation tasks. Experimental results demonstrate that our method substantially improves visual quality and 3D consistency over the baseline methods. Project page: https://GeometryForcing.github.io.

1 Introduction

Learning to simulate the physical world and predict future states is a cornerstone of intelligent systems [27]. Recent advances in generative modeling [31, 60, 53, 9], coupled with the availability of large-scale video datasets, have led to significant progress in generating realistic visual environments conditioned on text descriptions [50, 81, 56, 25] or agent actions [32, 26, 17, 8]. However, these approaches typically aim to model pixel distributions across video frames, overlooking a fundamental principle: *videos are 2D projections of a dynamic 3D world* [24]. By focusing solely on image-space generation, such models often struggle to maintain geometric coherence and long-term consistency, particularly in autoregressive settings where small errors can accumulate over time [11, 15, 34].

Building on this motivation, a growing line of research has explored explicitly modeling the dynamic 3D structure of the physical world [49, 94, 1, 88, 47, 35], as opposed to implicitly learning distributions in 2D pixel space. For example, WVD [88] proposes transforming 3D coordinates into images and jointly modeling the RGB and geometric information using diffusion models. While effective to some extent, representing 3D information in a tractable form remains challenging, and the reliance on additional annotations imposes limitations on scalability.

In this work, we aim to bridge the gap between video diffusion models and the underlying dynamic 3D structure of the physical world. We begin with a fundamental question: *Can video diffusion*

*Equal contribution.

†Project lead.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NextVid.

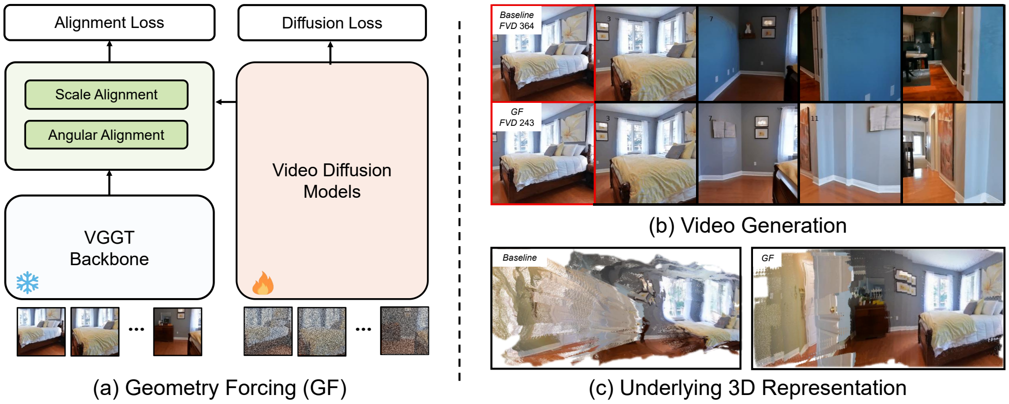

Figure 1: **Geometry Forcing equips video diffusion models with 3D awareness. (a)** We propose Geometry Forcing (GF), a simple yet effective paradigm to internalize geometric-aware structure into video diffusion models by aligning with features from a pretrained geometric foundation model, *i.e.*, VGGT [69]. **(b)** Compared to the baseline method [63], our method produces more consistent generations both temporally and geometrically. **(c)** Features learned by the baseline model fail to reconstruct meaningful 3D geometry, whereas our method internalize 3D representation, enabling accurate 3D reconstruction from the intermediate features.

models implicitly learn 3D information through training on raw video data, without explicit 3D supervision? To investigate this, we analyze a pretrained autoregressive video diffusion model [63] by introducing a DPT [59] head that maps its intermediate features to corresponding depth maps [69]. As illustrated in Fig. 1(c), we observe that features learned solely from raw video data fail to yield meaningful geometric representations, highlighting a potential gap in the geometric understanding of video diffusion models trained without additional guidance.

To address this limitation, we propose *Geometry Forcing (GF)*, a simple yet effective approach that encourages video diffusion models to *internalize* 3D representations during training. Inspired by recent advances in semantic REPresentation Alignment (REPA) for image diffusion models [85], we align intermediate features of video diffusion models with the *geometric representations* extracted from a pretrained 3D foundation model [69]. To align these two representations, our method introduces two complementary alignment objectives: Angular Alignment and Scale Alignment. Angular Alignment enforces directional consistency between the diffusion model's intermediate features and geometric representations by maximizing their cosine similarity. Scale Alignment, in contrast, preserves the scale information of the geometric representations by predicting unnormalized geometric features from normalized diffusion features. The decoupled formulation of Angular and Scale Alignment allows the model to capture both directional and scale-related aspects of geometry, while improves stability during training and expressiveness in the learned representations.

We evaluate the effectiveness of GF on two widely adopted benchmarks: camera view-conditioned video generation on RealEstate10K [93] and action-conditioned video generation on Minecraft environment [6]. Experimental results demonstrate that our method delivers substantial gains in geometric consistency and visual quality over the baseline methods. For example, GF reduces the FVD from 364 to 243 on RealEstate10K benchmark. Moreover, the ability to reconstruct explicit geometry during inference opens up opportunities for integrating structured memory into long-term world modeling.

2 Related Work

2.1 Interactive World Modeling

A world simulator seeks to model the underlying dynamics of the physical world by predicting future states conditioned on current observations and conditions [50, 8, 10, 52, 26, 3, 2]. We review prior works through the lenses of interactive video generation, 4D generation, and consistent world modeling.

Interactive Video Generation. Recent advancements in generative models [31, 60, 53, 42], fueled by the availability of large-scale video datasets, have positioned video generation as a promising approach to world modeling. Beyond text-to-video synthesis [12, 13, 38, 68, 40, 43, 82], interactive video generation [84] that emphasizes responding interactive control signals evolves rapidly. Existing models incorporate different signals like camera controls [28, 86, 63] and action controls [17, 26, 22, 61]. Building on this progress, our work introduces a novel training pipeline that enhances 3D consistency in video generation, enabling more coherent and realistic simulation of spatial scenes.

Interactive 4D Generation. In contrast to data-driven video simulators, 4D-based simulators [16, 5, 75, 83, 39] explicitly model dynamic 3D structures [36, 48, 77]. Building upon static 3D content generation [58], these methods evolve from object-centric 4D modeling [79, 4], to more complex dynamic scenes [49, 94]. Recent works further integrate video priors to improve the realism and temporal coherence of 4D [1, 35, 47, 14], and explore leveraging video priors for robust 4D world modeling. For example, TesserAct [92] predicts RGB, depth, and surface normals to reconstruct temporally consistent 4D scenes. While our work shares the goal of unifying 3D and video generation, it differs by injecting 3D geometric priors into the video representation to improve both temporal and spatial coherence.

Consistent World Modeling. A key challenge in world modeling lies in maintaining consistency over long video sequences. To address this, prior works have explored different forms of memory and contextual guidance. Frame-level context mechanisms [11, 63, 23, 55, 76] introduce frame-level context guidance by adding noise to context frames during training. Alternatively, several methods leverage 3D information to enforce spatial coherence. For example, WorldMem [78] maintain a memory bank indexed by field-of-view overlap to retrieve relevant historical frames. WVD [88] propose jointly modeling RGB frames and point maps to main consistency. In contrast to these approaches, we propose a unified method that internalizes 3D representations directly into the video diffusion model, enabling stronger and more stable geometric consistency across time.

2.2 3D Foundation Models

3D foundation models (3DFMs) [69, 41, 54, 80, 90, 62, 70, 71] have recently shown remarkable progress, offering end-to-end learning with fast and robust inference. These models are capable of predicting a wide range of 3D properties, such as camera poses [90], depth maps [54], and dense point clouds [69], directly from diverse visual inputs.

Due to their accuracy, efficiency, and robustness, 3DFMs are becoming essential for enabling in downstream tasks like spatial reasoning [74, 33, 20], autonomous driving [21], SLAM [45, 46], and beyond. Inspired by their strong 3D capabilities, we explore incorporating 3D representations into video diffusion models to enhance temporal and spatial consistency for world modeling.

3 Preliminaries

Our approach builds upon autoregressive video diffusion models [11, 63, 15] and incorporates a 3D foundation model [69] into the training process to guide geometric learning. In this section, we provide a brief overview of both components to establish the foundation for our method.

3.1 Autoregressive Video Diffusion Models

Training. We formulate our training pipeline based on Flow Matching [42, 44] with Transformer backbone [67, 7], aiming for both simplicity and scalability. Let $\mathbf{x} = \{x_1, \ldots, x_I\}$ denote a video sequence sampled from the data distribution, we assign an independent timestep for each frame $\mathbf{t} = \{t_1, \ldots, t_I\}$ and corrupt frames via interpolation:

$$x_i^{t_i} = (1 - t_i) \cdot x_i^0 + t_i \cdot \epsilon_i, \quad \text{where} \quad \epsilon_i \sim \mathcal{N}(0, I).$$

The target velocity field is defined as the difference between noise and clean input. We train a neural network v_θ to minimize the Flow Matching loss:

$$\mathcal{L}_{\text{FM}} = \left\| v_\theta(\mathbf{x}^{\mathbf{t}}, \mathbf{t}) - (\boldsymbol{\epsilon} - \mathbf{x}) \right\|^2 .$$

Sampling. At inference time, the sampling follows a simple probability flow ODE:

$$\mathrm{d}\mathbf{x} = v_\theta(\mathbf{x}^\mathbf{t}, \mathbf{t}) \cdot \mathrm{d}\mathbf{t}.$$

In practice, we iteratively apply the standard Euler solver [19] to sample data from noise. For autoregressive generation, we initialize the inputs with a clean context and generate subsequent frames sequentially, conditioning each prediction on the previously generated frames.

3.2 Visual Geometry Grounded Transformer

Visual Geometry Grounded Transformer (VGGT) [69] is a feed-forward model that directly outputs various 3D attributes of a scene, including camera parameters, point maps, depth maps, and 3D point tracks, from one, a few of its projected 2D views.

VGGT is composed of a Transformer backbone and multiple prediction heads. To make the Transformer focus within each frame and globally in an alternate way, the model employ Alternating-Attention mechanism that interleaves frame-wise self-attention (intra-frame structure) and global self-attention (inter-frame context). For each frame, local and global features are integrated into a unified latent representation, which is subsequently processed by a set of task-specific heads to produce corresponding 3D attributes. In our work, we leverage the features from the Transformer backbone of VGGT to provide geometric priors for video diffusion models.

4 Geometry Forcing

4.1 Method Overview

Motivation. Recent advances in video diffusion models have enabled the simulation of the world directly from large-scale video datasets. However, these models often overlook a fundamental property of visual data: videos are 2D projections of an dynamic 3D world. To address this, we seek to narrow the gap between video diffusion models and the dynamic 3D structure of the world.

Observation. We begin by examining whether video diffusion models are capable of implicitly learning 3D information when trained solely on raw video data, without explicit 3D supervision. To probe the geometric content of their learned representations, we adopt a strategy inspired by linear probing [29]: we freeze the parameters of a pretrained video diffusion model [63] and train a DPT [59] head to map intermediate features to corresponding depth map [69]. This allows us to assess the extent to which geometric information is encoded in the model's feature space. The results, presented in Fig. 1(c), indicate that features learned solely from raw video data do not produce meaningful geometric representations, suggesting a limited capacity of the model to encode dynamic 3D structure without explicit geometric guidance.

Challenge. Bridging the gap between video diffusion models and the dynamic 3D structure of the world presents significant challenges, primarily due to the limited annotated 3D data. A straightforward approach is to jointly model RGB and geometric information within an end-to-end architecture. However, relying heavily on 3D annotations can hinder the scalability and generalization ability of the models, particularly when applied to large and diverse real-world video datasets.

In this work, inspired by recent advances in REPA [85], we propose *Geometry Forcing (GF)* that aligns the features of video diffusion models with geometric representations, encouraging the model to internalize geometric information. Our approach builds upon video diffusion models described in Sec. 3.1. In Sec. 4.2, we introduce two regularization objectives designed to facilitate representation alignment between the diffusion model and geometric foundation model. The overall training objective, along with additional functional extensions, is summarized in Sec. 4.3.

4.2 Geometric Representation Alignment

To improve the geometric consistency of the learned representations, we introduce two complementary alignment objectives: *Angular Alignment* and *Scale Alignment*. These objectives are designed to align the latent features of the diffusion model with intermediate representations from a pretrained geometric foundation model [69], ensuring both directional consistency and scale preservation of geometric features within the feature space.

Angular Alignment. Angular Alignment enforces directional correspondence between the hidden states of the diffusion model, denoted by h, and specified target features, denoted by y. We select intermediate features from the Transformer backbone of VGGT [69] as y, as these features preserve both local and global information within each frame and can be further used to reconstruct various explicit geometric representations. In practice, the target features $y \in \mathbb{R}^{L\times N\times P\times D}$, where L denotes the number of layers, N denotes the number of input images, P denotes the patch count, and D denotes the feature dimension. To achieve Angular Alignment, we first use a lightweight projector f_ϕ to map the diffusion latents $h \in \mathbb{R}^{N\times P'\times D'}$ to y's shape. The Angular Alignment loss is then defined as:

$$\mathcal{L}_{\text{Angular}} = -\frac{1}{LNP}\sum_{\ell=1}^{L}\sum_{n=1}^{N}\sum_{p=1}^{P}\cos\left(y_{\ell,n,p},\ f_\phi(h_{n,p})\right),$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. This loss aligns hidden states independently at both the frame and patch levels. Since the VGGT backbone already incorporates cross-frame attention, we do not explicitly enforce global alignment across frames in the loss.

Scale Alignment. While Angular Alignment ensures directional consistency, it disregards feature scale that could also encode geometric information. Although direct mean squared error (MSE) loss could supervise magnitudes, it often leads to optimization instability and model collapse due to inherent scale difference across models. To address this issue, we introduce Scale Alignment, which preserves scale information through predicting the scale of target features given normalized diffusion hidden states. Specifically, we first normalize $f_\phi(h)$ to unit length. Then we use a another lightweight prediction head g_φ to predict the full target features from normalized inputs:

$$\hat{h}_{\ell,n,p} = \frac{f_\phi(h_{n,p})}{\|f_\phi(h_{n,p})\|_2}, \quad \tilde{y}_{\ell,n,p} = g_\varphi(\hat{h}_{\ell,n,p}).$$

The Scale Alignment loss is defined as:

$$\mathcal{L}_{\text{Scale}} = \frac{1}{LNP}\sum_{\ell=1}^{L}\sum_{n=1}^{N}\sum_{p=1}^{P}\|\tilde{y}_{\ell,n,p} - y_{\ell,n,p}\|_2^2.$$

This decomposition stabilizes training while capturing both directional and scale attributes of geometric representations.

4.3 3D-aware Autoregressive Video Diffusion Models

Building on the autoregressive video diffusion framework and the proposed alignment objectives, we now present the overall training objective:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{Angular}}\cdot\mathcal{L}_{\text{Angular}} + \lambda_{\text{Scale}}\cdot\mathcal{L}_{\text{Scale}}.$$

Given that the intermediate features of our model are well-aligned with geometric representations, an appealing consequence is the model's ability to predict explicit 3D geometry during inference. This enables unified generation of both video and 4D, effectively bridging the gap between videos and the underlying dynamic 3D structure of the physical world, as illustrated in Fig. 1. Moreover, the ability to reconstruct explicit geometry during inference provides a structured and interpretable form of memory, which can be further utilized to support long-term world modeling and reasoning. We leave the exploration of such geometry-based memory mechanisms as a promising direction for future work.

Discussion. Teacher Forcing [73] is a widely adopted training paradigm for autoregressive models [57, 9, 37]. To combine autoregressive nature with diffusion models, Diffusion Forcing [11] is introduced, which trains video diffusion models using independently sampled noise levels for each frame. More recently, Self Forcing [34] is proposed to addressing exposure bias in autoregressive video diffusion models. Orthogonal to these methods, Geometry Forcing focuses on improving the spatial structure of the learned representations by aligning the intermediate representation of autoregressive video diffusion models with geometry-aware signals from a pretrained 3D foundation model. Our approach provides structural supervision at the representational level, encouraging the model to internalize 3D consistency throughout training.

Figure 2: **Qualitative comparison of camera view-conditioned video generation under full-circle rotation.** Videos are generated from a single input frame and corresponding per-frame camera poses simulating a full 360° rotation. Our method (GF) is compared with DFoT [63], VideoREPA [91], and REPA [91]. The results demonstrate that the baseline methods fail to maintain temporal consistency, while our proposed GF consistently revisit the starting viewpoint.

5 Experiments

In this section, we evaluate Geometry Forcing (GF) on camera view-conditioned video generation on RealEstate10K [93] dataset and action-conditioned video generation on Minecraft environment [6]. We also provide more illustration and visualization in Appendix.

Implementation Details. For camera view-conditioned video generation [93], we apply GF to the Diffusion Forcing Transformer [63]. Training uses 16-frame videos at 256×256 resolution for 2,500 iterations with a learning rate of 8×10^{-6} and batch size 8. During inference, we condition the model on the first frame and generate 256 frames. For action-conditioned video generation, we apply GF to Next-Frame Diffusion [15], training on 32-frame videos at 384×224 resolution for 2,000 steps with a learning rate of 6×10^{-5} and batch size 32. By default, we set $\lambda_{\text{Angular}} = 0.5$ and $\lambda_{\text{Scale}} = 0.05$ to balance the contribution of each loss component. All experiments are conducted on 8 NVIDIA A100 GPUs.

Evaluation Metrics. We evaluate visual quality using standard video generation metrics, including FVD (Fréchet Video Distance) [66], PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index) [72], and LPIPS (Learned Perceptual Image Patch Similarity) [89].

To further evaluate geometric consistency, we introduce two metrics: Reprojection Error (RPE) [18] and Revisit Error (RVE) [78]. Reprojection Error (RPE) quantitatively measures multi-view geometric consistency by calculating the average reprojection discrepancy between projected and observed pixel locations across multiple views. Revisit Error (RVE) assesses long-range temporal consistency by examining discrepancies between initial and revisited frames under complete camera rotation. We provide more details of these metrics in the Appendix (Sec. B.4).

5.1 Main Results

This section presents the main experimental results, comparing our method against state-of-the-art approaches across different tasks. The evaluation results demonstrate the effectiveness and generalization ability of our method in both short- and long-term video generation.

Table 1: Quantitative comparison on the RealEstate10K dataset for both short-term (16-Frame) and long-term (256-Frame) video generation. Our method (Geometry Forcing) achieves the best performance across all metrics. **bold** values denote the best, and Underlined values indicate the second best. * indicates the method is conditioned on the first frame only.

Method	Frames	FVD↓	LPIPS↓	SSIM↑	PSNR↑	RPE↓	RVE↓
DFoT [63]	16	252	0.40	0.50	14.40	–	–
REPA [85]	16	221	0.37	0.54	**15.20**	–	–
VideoREPA [91]	16	210	0.37	0.54	**15.20**	–	–
Geometry Forcing (ours)	16	<u>193</u>	**0.32**	**0.58**	<u>14.70</u>	–	–
Geometry Forcing (ours) + REPA	16	**179**	<u>0.34</u>	<u>0.54</u>	<u>15.00</u>	–	–
Cosmos* [2]	256	934	0.68	0.20	10.25	–	–
DFoT [63]	256	364	0.55	0.36	11.40	0.3575	297
REPA [85]	256	297	0.54	0.36	11.51	<u>0.3337</u>	315
VideoREPA [91]	256	455	0.56	0.35	11.50	0.3823	**190**
Geometry Forcing (ours)	256	<u>243</u>	**0.51**	**0.38**	<u>11.87</u>	<u>0.3337</u>	272
Geometry Forcing (ours) + REPA	256	**237**	**0.51**	<u>0.37</u>	**12.10**	**0.3264**	<u>236</u>

Table 2: **Ablation study on target representation**. We compare the effect of aligning the diffusion model with different target representations: DINOv2 (semantic), VGGT (geometric), and their combination. The joint use of both representation achieves the best FVD.

Target Representation	FVD-256
Baseline	364
DINOv2 Only	297
VGGT Only	243
VGGT + DINOv2	**237**

Table 3: **Ablation study on alignment loss.** Angular and Scale Alignment losses are evaluated for long-term video generation, with MSE as a naive baseline of aligning both angular and scale information. The combination of Angular and Scale Alignment yields the best results.

Alignment Loss	FVD-256
Baseline	364.0
Angular	253.0
Angular + Scale	**243.0**
MSE	1648.0

Camera view-conditioned Video Generation. We conduct comprehensive evaluation of GF on the RealEstate10K [93] dataset, comparing against state-of-the-art baselines. We report results for both short-term (16-Frame) and long-term (256-Frame) video generation in Tab. 1.
As shown in Tab. 1, our method consistently outperforms all baselines across multiple evaluation metrics, including FVD, LPIPS, SSIM, and PSNR, in both the short-term and long-term generation settings. These results highlight the effectiveness of GF in enhancing visual fidelity, temporal stability, and 3D spatial consistency, thereby enabling more realistic and coherent world modeling.

Action-conditioned Video Generation. To demonstrate the generality of our method, we apply GF to Next-Frame Diffusion [15] model. As shown in Tab. 5, the model achieves a lower FVD score which indicates GF can be seamlessly integrated into video diffusion models and leads to measurable gains. Note that, there exists a large data distribution gap between real world and Minecraft. This results demonstrate that GF generalize well on out-of-domain distribution.

5.2 Qualitative Results

Fig. 2 presents qualitative comparisons on the RealEstate10K dataset. Each video is generated from a single input frame along with camera poses simulating 360° rotation. We compare GF against baselines: DFoT [63], REPA [85], and VideoREPA [91]. As shown in Fig. 2, our method reconstructs the initial frame when the camera completes rotation, while producing reasonable and realistic intermediate views. In contrast, the baseline methods fail to maintain temporal and scene consistency, resulting in implausible results and unable to revisit the starting viewpoint. These results highlight the superior long-term 3D consistency and scene understanding of our approach.

5.3 Ablation Studies

We provide a series of ablation studies to validate the design of GF.

Table 4: **Ablation study on method to integrate geometry information.** We compare external condition (via ControlNet) with internal alignment (via Geometry Forcing).

Method	FVD-256↓
Baseline	364
External Condition	275
Geometry Forcing (ours)	**243**

Table 5: **Evaluation on action-conditioned video generation in Minecraft.** FVD results of NFD before and after applying Geometry Forcing (GF) on 16-Frame generation show clear improvement.

Method	FVD-16↓
NFD	216
NFD + GF	**205**

Which Representation Should be Aligned? To validate the effectiveness of geometric representation, we compare two target representations in GF: VGGT [69], trained on 3D datasets with strong geometric priors, and DINOv2 [51], trained on 2D images focusing on semantic features. As shown in Tab. 2, aligning with VGGT consistently outperforms DINOv2 on both long-term and short-term generation tasks, highlighting the advantage of geometric alignment over semantic supervision.

To further explore their complementarity, we combine VGGT and DINOv2 features as joint supervision targets. Results in Tab. 2 show that integrating geometric and semantic signals leads to additional gains, suggesting that the two types of representations are orthogonal and can enhance each other when used together. However, as we mainly focus on bridging the gap between the video diffusion model and the dynamic 3D structure of the real world, we only use VGGT features in further experiments.

Alignment Loss. GF consists of two alignment objectives: Angular Alignment and Scale Alignment. To validate their effectiveness, we compare three alignment loss types: (1) Angular Alignment alone (Sec. 4.2), (2) Angular Alignment with Scale Alignment (Sec. 4.2), and (3) MSE loss between VGGT and diffusion features. As shown in Tab. 3, the combination of Angular Alignment and Scale Alignment achieves best performance, indicating the benefit of aligning both angular and scale-related information. Although direct mean squared error (MSE) also supervises magnitudes, the change of feature scale of the diffusion model may cause collapse in the following layers. These results highlight that neither Angular Alignment nor Scale Alignment alone is sufficient.

How Can Geometry Information be Integrated into Video Diffusion Models? To validate the effectiveness of internalizing geometric representation in the video diffusion model, we compare two strategies to incorporate geometric representation: internal alignment through GF and external guidance via an additional ControlNet [87] (Geometry ControlNet). In the external guidance experiment, we obtain intermediate features from the transformer backbone of VGGT (identical to the one used in GF). Then we feed the intermediate features into a ControlNet attached to DFoT. This approach introduces geometry information as external conditions. In contrast, GF encourages the model to internalize geometric features.

As shown in Tab. 4, while the external guidance produces improvements over the baseline DFoT model, it still underperforms compared to GF. This suggests that integrating geometric priors into the model is more effective than supplying them as external conditions. By aligning internal features with geometric representations, GF enables deeper geometric understanding and yields better performance in terms of perceptual quality and structural consistency.

Which Layer Should be Aligned? As shown in Fig. 3, we also explore applying alignment at different layers of the video diffusion model [63], which uses a 7-layer U-ViT [7] backbone (3 downsampling layers, 1 bottleneck layer, 3 upsampling layers). Aligning at layer 3 yields the best FVD-256 score while preserving FVD-16 performance.

Mitigating Exposure Bias in Autoregressive Video Diffusion Model via Geometry Forcing. Exposure bias is a long-standing challenge in autoregressive video generation [11, 63, 64, 15, 34]. While previous methods attempted to address it through memory mechanisms or context guidance, GF offers an orthogonal solution. As shown in Fig. 4, GF mitigates long-term drift and reduces the accumulation of error during generation significantly by aligning 3D geometric representation. These results validate integrating 3D representation enables more reliable and coherent long-term video synthesis.

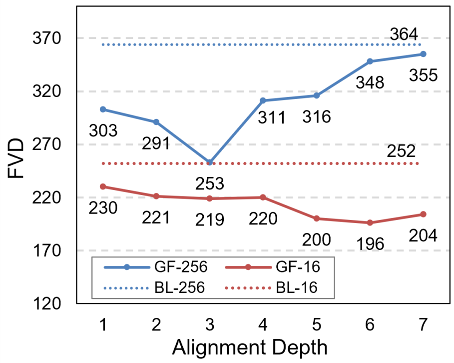

Figure 3: **Ablation study on alignment depth.** We present FVD-256 and FVD-16 results for different alignment layers of diffusion model which suggest mid-level feature is most effective to improve video quality.

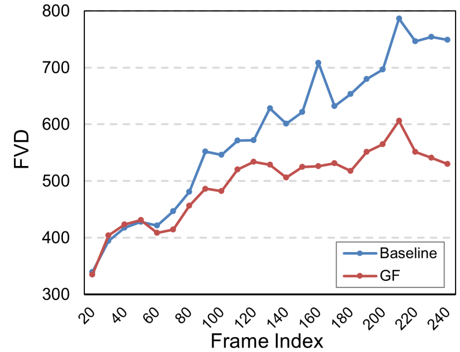

Figure 4: **Exposure bias analysis.** This figure shows the trend of FVD scores during long-term video generation. Compared to the baseline, GF results in significantly lower FVD after 100 frames.

Table 6: **User study.** Average scores (1–5) on Camera Following, Object Consistency, and Scene Continuity. Each user was shown one case at a time and asked to rate each dimension on a scale of 1 to 5. Higher values indicate better quality.

Method	Camera Following	Object Consistency	Scene Continuity
DFoT	3.56	2.73	2.74
REPA	3.82	3.55	3.66
VideoREPA	3.31	3.05	2.82
Geometry Forcing	**4.40**	**4.44**	**4.52**

5.4 User Study

While Reprojection Error (RPE) and Revisit Error (RVE) provide useful signals for measuring 3D consistency, they only capture specific geometric aspects and may miss perceptual artifacts or unrealistic dynamics that humans can easily notice. Additionally, we conduct a user study focusing on three aspects of 3D consistency. 1) **Camera Following**: Whether the camera in the video moves smoothly and accurately follows the given pose trajectory. 2) **Object Consistency**: Whether objects remain consistent in shape, appearance, and position across frames. 3) **Scene Continuity**: Whether the generated parts of the scene beyond the context frames remain coherent and reasonable.

As shown in Tab. 6, GF consistently outperforms all baselines across the three aspects of 3D consistency, demonstrating its effectiveness in producing geometrically coherent videos.

6 Conclusion

This paper introduces Geometry Forcing (GF), a simple yet effective framework that enhances the geometric consistency of autoregressive video diffusion models by aligning their internal representations with geometry-aware features. Motivated by the observation that video diffusion models trained on raw pixel data often fail to capture meaningful 3D structure, our method proposes two alignment objectives (Angular Alignment and Scale Alignment) guide the latent feature align with 3D-aware freature from geometric foundation model. Empirical results on both camera-conditioned and action-conditioned video generation benchmarks demonstrate that GF significantly improves visual quality and 3D consistency, yielding lower FVD scores and more stable scene dynamics.

Limitations. The primary limitation of this work lies in its scale. While GF consistently improves geometric consistency and visual quality, its full potential remains unexplored under large-scale training. In particular, we have not yet investigated its effectiveness when applied to larger models and more extensive video datasets, which may further amplify its benefits.

Future Work. Future directions include scaling GF on larger datasets to build 3D-consistent world simulators, and applications for long video generation by treating 3D representation as memory.

Acknowledgment

This work was conducted during an internship at Microsoft Research Asia and was supported by the company. We would like to acknowledge Kiwhan Song and Boyuan Chen for their valuable advice and assistance in helping us reproduce the results of Diffusion Forcing Transformer.

References

[1] Aether, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025.

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[3] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[4] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024.

[5] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024.

[6] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

[7] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

[8] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

[10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[11] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

[12] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[13] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[14] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025.

[15] Xinle Cheng, Tianyu He, Jiayi Xu, Junliang Guo, Di He, and Jiang Bian. Playing with transformer at 30+ fps via next-frame diffusion. *arXiv preprint arXiv:2506.01380*, 2025.

[16] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

[17] Decart, Quevedo Julian, McIntyre Quinn, Campbell Spruce, Chen Xinlei, and Wachen Robert. Oasis: A universe in a transformer. 2024.

[18] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

[19] Leonhard Euler. *Institutionum calculi integralis*, volume 4. impensis Academiae imperialis scientiarum, 1845.

[20] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

[21] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *ArXiv*, abs/2412.06777, 2024.

[22] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

[23] Michael Fuest, Vincent Tao Hu, and Björn Ommer. Maskflow: Discrete flows for flexible and efficient long video generation. *arXiv preprint arXiv:2502.11234*, 2025.

[24] Andrew S Glassner. *An introduction to ray tracing*. Morgan Kaufmann, 1989.

[25] Google. Veo 3. `https://deepmind.google/models/veo/`, 2025.

[26] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

[27] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[28] 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.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[30] 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.

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

[32] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[33] Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. Mllms need 3d-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025.

[34] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[35] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[37] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, pages 25105–25124. PMLR, 2024.

[38] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[39] Yao-Chih Lee, Yi-Ting Chen, Andrew Wang, Ting-Hsuan Liao, Brandon Y Feng, and Jia-Bin Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024.

[40] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.

[41] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[43] Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8009–8019, 2025.

[44] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[45] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16651–16662, 2025.

[46] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl(4) manifold. *ArXiv*, abs/2505.12549, 2025.

[47] Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025.

[48] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[49] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11453–11464, 2021.

[50] OpenAI. Sora. `https://openai.com/index/sora/`, 2024.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[52] J Parker-Holder, P Ball, J Bruce, V Dasagi, K Holsheimer, C Kaplanis, A Moufarek, G Scully, J Shar, J Shi, et al. Genie 2: A large-scale foundation world model. *URL: https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model*, 2024.

[53] 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.

[54] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.

[55] Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *arXiv preprint arXiv:2505.20171*, 2025.

[56] A Polyak, A Zohar, A Brown, A Tjandra, A Sinha, A Lee, A Vyas, B Shi, CY Ma, CY Chuang, et al. Movie gen: A cast of media foundation models. 2024a. *arXiv preprint arXiv:2410.13720*, 2024.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[58] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2349–2359, 2023.

[59] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[61] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024.

[62] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

[63] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

[64] Mingzhen Sun, Weining Wang, Gen Li, Jiawei Liu, Jiahui Sun, Wanquan Feng, Shanshan Lao, SiYu Zhou, Qian He, and Jing Liu. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7364–7373, 2025.

[65] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[66] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[68] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[69] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[70] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[71] 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.

[72] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[73] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[74] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

[75] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, Yue Qian, Xiaohang Zhan, and Yueqi Duan. 4d-fly: Fast 4d reconstruction from a single monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 16663–16673, June 2025.

[76] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

[77] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025.

[78] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

[79] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024.

[80] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *arXiv preprint arXiv:2501.13928*, 2025.

[81] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[82] Yang Ye, Junliang Guo, Haoyu Wu, Tianyu He, Tim Pearce, Tabish Rashid, Katja Hofmann, and Jiang Bian. Fast autoregressive video generation with diagonal decoding. *arXiv preprint arXiv:2503.14070*, 2025.

[83] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025.

[84] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video. *arXiv preprint arXiv:2504.21853*, 2025.

[85] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[86] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[87] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[88] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21685–21695, 2025.

[89] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[90] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025.

[91] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025.

[92] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[93] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[94] Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. Compositional 3d-aware video generation with llm director. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper is foundational research and not tied to particular applications, let alone deployments.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Appendix for *Geometry Forcing*:Marrying Video Diffusion and 3D Representation for Consistent World Modeling

A Limitations

Our method's reliance on VGGT (trained mainly on static scenes) constrains performance in dynamic environments. Generalization to significant motion scenarios requires further research.

B Implementation Details

B.1 Dataset

RealEstate10K [93]. This dataset contains camera poses for 10 million video frames, suitable for evaluating 3D consistency and camera navigation in generated videos. We use a resolution of 256×256 pixels.

Minecraft [6]. This game dataset includes action annotations, enabling evaluation of video generation in dynamic environments with camera motion.

Alignment Projection To maximize geometric information retention, we aggregate features from all transformer blocks of the VGGT backbone as alignment targets. For computational efficiency, we apply bilinear interpolation to reduce the spatial dimensions from the original resolution to a manageable size of 512×512 tokens.

The alignment is performed using a Conv3D-based projector that operates on the latent dimensions. To accommodate multi-layer and multi-target alignment scenarios, we initialize independent projectors for each feature layer and target representation. This design ensures effective dimensional compatibility between the U-ViT feature space and the target geometric representations while maintaining computational efficiency.

B.2 Training

Model Architecture. We adopt a U-ViT backbone for video generation, with geometric feature alignment integrated at the third transformer block.

Training Data. The model is trained on 10,000 video clips sampled from the RealEstate10K training dataset, each comprising 16 consecutive frames.

Training Protocol. Training proceeds for 2 epochs using a learning rate of 8×10^{-6} and a global batch size of 40. The geometric alignment loss is combined with the standard diffusion training objective.

B.3 Inference

A key advantage of Geometry Forcing is its inference-time efficiency which introduces no computational overhead during sampling. We demonstrate results using a DDIM sampler with 50 steps, though the approach is compatible with any standard diffusion sampling algorithm.

B.4 Metrics

In this section,we introduce the detailed implementation of Reprojection Error (RPE) and Revisit Error (RVE).

Reprojection Error. Reprojection error (RPE) is a widely used metric in visual SLAM to evaluate multi-view geometric consistency. Following **(author?)** [18], we utilize DROID-SLAM [65] to reconstruct scene. Specifically, DROID-SLAM first extracts corresponding features across frames and then refines camera poses (G_t) and per-pixel depth estimates (d_t) through its differentiable Dense Bundle Adjustment (DBA) optimization, enforcing optical flow constraints and achieving robust structure-from-motion. The reprojection error is then computed by measuring the average Euclidean distance between the projected and observed pixel locations of co-visible 3D points across multiple frames. Formally, RPE is defined as:

$$RE = \frac{1}{|\mathcal{V}|} \sum_{(i,j)\in\mathcal{V}} \left\| \mathbf{p}_{ij}^{*} - \Pi(\mathbf{P}_{ij}) \right\|_2 , \tag{1}$$

where $\mathcal{V}$ denotes the set of valid feature correspondences, $\mathbf{p}ij$ is the observed pixel location in generated video frames, $\mathbf{P}ij$ represents the corresponding reconstructed 3D point derived from refined depths and camera poses, and Π denotes the camera projection function. Lower RPE values indicate better 3D alignment, reduced spatial artifacts, and enhanced spatio-temporal stability, thereby effectively reflecting the overall geometric coherence and consistency of the generated videos.

Revisit Error. Revisit Error evaluates long-range temporal consistency under full camera rotation, inspired by the setup proposed in WorldMem [78]. For each of 100 randomly sampled RealEstate10K video clips, we extract the first frame and initial camera pose. A camera trajectory of 256 frames is then constructed by rotating the initial camera pose around the Y-axis. We assess revisit consistency by comparing the first and final frame using reconstruction FID (rFID) [30]. Larger discrepancies indicate greater geometric or appearance drift, suggesting weaker long-term 3D consistency.

B.5 3D Reconstruction from Diffusion Features

In this section, we provide a detailed overview of the 3D reconstruction process illustrated in Fig. 1(c).

Reconstruction using Geometry Forcing Features. We extract features from the Geometry Forcing (GF) model and pass them through the depth prediction head of VGGT to obtain the predicted depth map.

Reconstruction using Diffusion Features. Motivated by our linear probing experiments, we investigate the 3D reconstruction capability of intermediate features extracted from DFoT [63]. Specifically, we freeze the pretrained DFoT backbone and train a DPT head [59] to regress depth maps from its intermediate representations. The target depth maps are provided by the VGGT model [69], serving as ground-truth supervision. The DPT head adopts the same architecture as the depth prediction module used in VGGT but is trained from scratch. We optimize the DPT head for 2500 steps using a learning rate of 1×10^{-4} and a batch size of 4.

C Discussion

C.1 Computational Efficiency

The geometric alignment loss increases per-step computation by 50%, but accelerates convergence to reduce total training time. For fine-tuning, our method requires only several thousand steps and completes within hours, offering substantial efficiency gains over full pre-training.

C.2 Analysis of Geometric and Semantic Representations

We analyze the roles of geometric and semantic representation alignment in video generation. First, these representations exhibit considerable overlap rather than orthogonality. Semantic representations like DINOv2 [51] demonstrate zero-shot depth estimation capabilities (see Section 7.5 and Figure 7 in the original paper), indicating inherent geometric understanding. Conversely, geometric representations such as VGGT utilize DINOv2 features as inputs, thereby encoding semantic information.

Second, experimental results in Table 1 and Table 2 show that VGGT alignment primarily enhances 3D consistency, while DINOv2 alignment improves visual quality. The combination of both representations achieves superior performance compared to either individual approach.

Finally, the distinct contributions of each representation can be characterized as follows: semantic alignment enhances object realism and visual details, whereas geometric alignment ensures structural consistency and shape coherence throughout the generated video sequences.

C.3 3D Consistency and Exposure Bias Mitigation

As shown in Figure 4, the FVD metric increases at a slower rate when Geometry Forcing is employed, indicating effective mitigation of exposure bias in long-term video generation. The underlying mechanism can be understood through the inherent stability of 3D scenes: while the number of generated frames increases, the underlying scene geometry remains same. Geometry Forcing enables the model to internalize this geometric consistency, thereby reducing error accumulation when regenerating frames from previously encountered viewpoints.

C.4 Supplementary Visualizations

In order to better understand the geometry influences, we provide comprehensive visual results.

Figure 5: **Qualitative comparisons on camera-conditioned video generation.** All the videos are generated given first frame and per-frame camera pose. We comprehensively compare GF (ours) with DFoT [63], VideoREPA [91], REPA [91]. The results demostrate consistency in long-term video generation both inside (left) and outside (right) scenes.

Fig. 5 presents qualitative comparisons on the RealEstate10K dataset. Given the same first frame and per-frame camera trajectory as input, we compare our proposed GF method with three strong baselines: DFoT [63], REPA [91], and VideoREPA [91].

As shown in Fig. 5, our method generates visually coherent and geometrically consistent videos over long time horizons even context is limited. In particular, GF better preserves object shapes and scene layouts that is visible in context, while generating reasonable scenes not seen in the context. In contrast, baseline models often exhibit drift, shape distortion, or abrupt transitions. These results highlight the effectiveness of internalizing geometric priors to enhance spatial and temporal consistency in video generation.