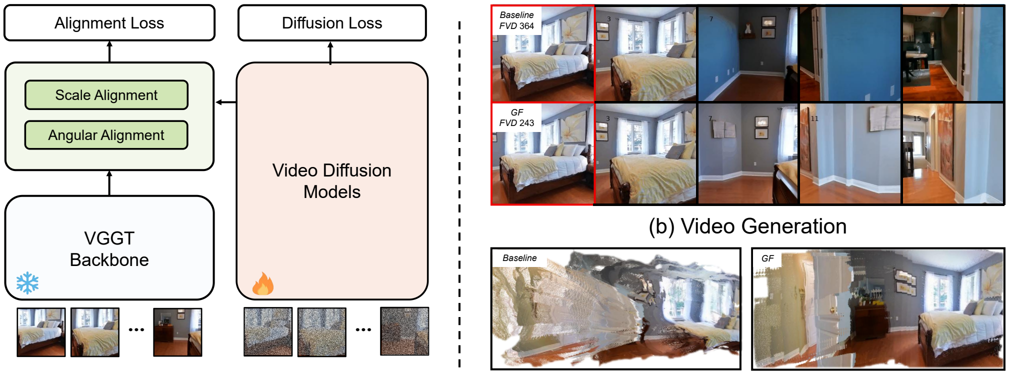
Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

Anonymous authors
Paper under double-blind review

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 the 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 3D representations. Our key insight is to guide the model's intermediate representations toward geometry-aware structure by aligning them with features from a 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 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.



(a) Geometry Forcing (GF)

(b) Video Generation

(c) Underlying 3D Representation

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 geometric foundation model, *i.e.*, VGGT (Wang et al., 2025b). **(b)** Compared to the baseline method (Song et al., 2025), 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 internalizes 3D representation, enabling accurate 3D reconstruction from the intermediate features.

1 Introduction

Learning to simulate the physical world and predict future states is a cornerstone of intelligent systems (Ha & Schmidhuber, 2018). Recent advances in generative modeling (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Brown et al., 2020), coupled with the availability of large-scale video datasets, have led to significant progress in generating realistic visual environments conditioned on text descriptions (OpenAI, 2024; Yang et al., 2024; Polyak et al., 2024; Google, 2025) or agent actions (Hu et al., 2023; Guo et al., 2025; Bar et al., 2025). 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* (Glassner, 1989). 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 (Chen et al., 2024a; Cheng et al., 2025; Huang et al., 2025b).

Building on this motivation, a growing line of research has explored explicitly modeling the dynamic 3D structure of the physical world (Zhu et al., 2024; Aether et al., 2025; Jiang et al., 2025), as opposed to implicitly learning distributions in 2D pixel space. For example, Zhang et al. (2025a) proposes transforming 3D coordinates into point maps and jointly modeling the RGB and 3D information. 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 models implicitly learn 3D information through training on raw video data, without explicit 3D supervision*? To investigate this, we analyze a pretrained video diffusion model (Song et al., 2025) by introducing a DPT (Ranftl et al., 2021) head that maps its intermediate features to corresponding depth maps (Wang et al., 2025b). 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 (Yu et al., 2024a), we align features of video diffusion models with the *geometric representations* from a pretrained 3D foundation model (Wang et al., 2025b). 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 features and geometric representations by maximizing their cosine similarity. Scale Alignment, in contrast, preserves the scale information of the geometric representations by predicting 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 improving 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 (Zhou et al., 2018) and action-conditioned video generation on Minecraft environment (Baker et al., 2022). 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. 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 (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2023; Bruce et al., 2024; Parker-Holder et al., 2024; Alonso et al., 2024; Valevski et al., 2024), have positioned video generation as a promising approach to world modeling. Beyond text-to-video synthesis (Chen et al., 2023; 2024b; Kong et al., 2024; Wan et al., 2025; Li et al., 2024; Liu et al., 2025a; Ye et al., 2025), interactive video generation (Yu et al., 2025b) that emphasizes responding interactive control signals evolves rapidly. Existing models incorporate different signals like camera controls (He et al., 2024; Yu et al., 2024b; Song et al., 2025) and action controls (Decart et al., 2024; Guo et al., 2025; Feng et al., 2024; Shin et al., 2024). 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 (Chung et al., 2023; Bahmani et al., 2024b; Wu et al., 2025b; Yu et al., 2025a; Lee et al., 2024) explicitly model dynamic 3D structures (Kerbl et al., 2023; Mildenhall et al., 2021; Xiang et al., 2025). Building upon 3D content generation (Raj et al., 2023), these methods evolve from dynamic objects (Xu et al., 2024; Bahmani et al., 2024a) to complex dynamic scenes (Niemeyer & Geiger, 2021; Zhu et al., 2024). Recent works integrate video priors to improve the realism and temporal coherence of 4D (Aether et al., 2025; Jiang et al., 2025; Mai et al., 2025; Chen et al., 2025). For example, TesserAct (Zhen et al., 2025) 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 video representation to improve 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 solutions. Frame-level context mechanisms (Chen et al., 2024a; Fuest et al., 2025; Po et al., 2025; Wu et al., 2025c) improve consistency by training with noisy context frames. Meanwhile, other methods leverage 3D information. For example, Xiao et al. (2025) maintain a memory bank indexed by field-of-view overlap to retrieve relevant historical frames. Zhang et al. (2025a) propose jointly modeling RGB frames and point maps to maintain 3D consistency. In contrast, we propose to directly internalize 3D representations into video diffusion models, enabling more stable geometric consistency.

2.2 3D Foundation Models

3D foundation models (3DFMs) (Li et al., 2025; Yang et al., 2025; Smart et al., 2024; Wang* et al., 2025; Wang et al., 2024) have recently shown remarkable progress, applying end-to-end framework with fast and robust inference. These models are capable of predicting different 3D properties including camera poses (Zhang et al., 2025b), depth maps (Piccinelli et al., 2024), and dense point clouds (Wang et al., 2025b), directly from visual inputs.

Due to their accuracy, efficiency, and robustness, 3DFMs are becoming essential for enabling downstream tasks like spatial reasoning (Wu et al., 2025a; Huang et al., 2025a; Fan et al., 2025), autonomous driving (Fei et al., 2024), SLAM (Liu et al., 2025b; Maggio et al., 2025), 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 (Chen et al., 2024a; Song et al., 2025; Cheng et al., 2025) and incorporates a 3D foundation model (Wang et al., 2025b) 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 training pipeline based on Flow Matching (Lipman et al., 2023; Liu et al., 2023) with Transformer backbone (Vaswani et al., 2017; Bao et al., 2023), aiming for simplicity and scalability. Let $\mathbf{x} = \{x_1, \ldots, x_I\}$ denote a video sequence sampled from the data distribution, we assign 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:

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

In practice, we iteratively apply the standard Euler solver (Euler, 1845) 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) (Wang et al., 2025b) is a feed-forward model that directly outputs 3D attributes of a scene, including camera parameters, point maps, depth maps.

VGGT has a Transformer backbone and multiple prediction heads. The model employ Alternating-Attention mechanism that interleaves frame-wise self-attention and global self-attention to extract local and global information. For each frame, local and global features are integrated into a unified representation, which is subsequently processed by task-specific heads to produce 3D attributes. We leverage the features from the Transformer backbone of VGGT to extract geometric representation.

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 (He et al., 2020): we freeze the parameters of a pretrained video diffusion model (Song et al., 2025) and train a DPT (Ranftl et al., 2021) head to map intermediate features to corresponding depth map (Wang et al., 2025b). 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 (Yu et al., 2024a), 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 (Wang et al., 2025b), 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 (Wang et al., 2025b) 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 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:

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

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 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. We leave the exploration of such geometry-based memory mechanisms as a promising direction for future work.

Discussion. Teacher Forcing (Williams & Zipser, 1989) is a widely adopted training paradigm for autoregressive models (Radford et al., 2019; Brown et al., 2020; Kondratyuk et al., 2024). To combine autoregressive nature with diffusion models, Diffusion Forcing (Chen et al., 2024a) propose to train video diffusion models with independent noise levels for each frame. More recently, Self Forcing (Huang et al., 2025b) is proposed to address exposure bias in autoregressive video diffusion models. Orthogonal to these methods, Geometry Forcing focuses on improving the structure of the learned representations by aligning the intermediate representation of video diffusion models with geometry-aware signals from 3D foundation model. Our approach provides structural supervision at representational level, encouraging the model to internalize 3D consistency throughout training.

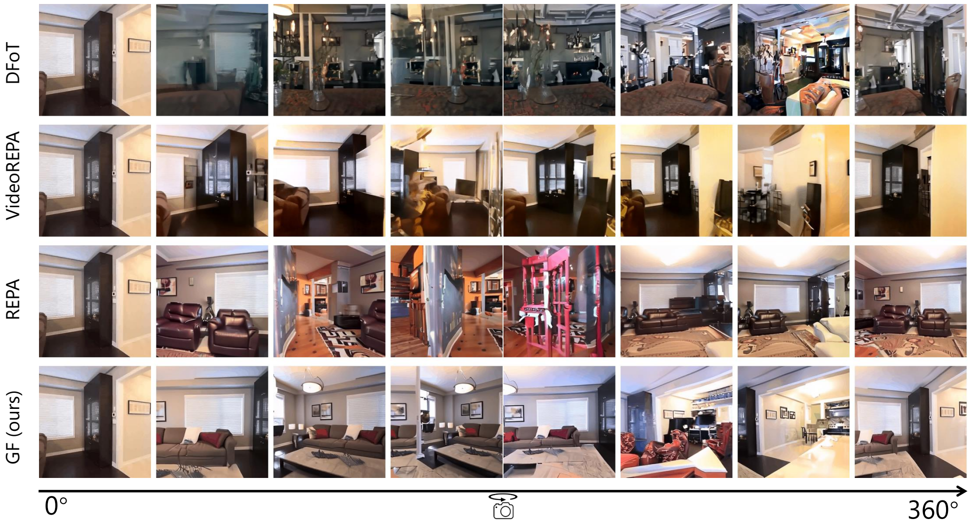


Figure 2: **Qualitative comparison of camera view-conditioned video generation under full-circle rotation.** Videos are generated from a single frame and per-frame camera poses simulating a full 360° rotation. Our method (GF) is compared with DFoT (Song et al., 2025), VideoREPA (Zhang et al., 2025c), and REPA (Yu et al., 2024a). The results demonstrate that the baseline methods fail to maintain temporal consistency, while our proposed GF consistently revisits the starting viewpoint.

5 Experiments

In this section, we evaluate Geometry Forcing (GF) on camera view-conditioned video generation on RealEstate10K (Zhou et al., 2018) dataset and action-conditioned video generation on Minecraft environment (Baker et al., 2022). We also provide more illustration and visualization in Appendix.

Implementation Details. For camera view-conditioned video generation, we apply GF on Diffusion Forcing Transformer (Song et al., 2025), training on 16-frame 256×256 videos for 2,500 steps with a learning rate of 8×10^{-6} and batch size 8. Inference used first frame and per-frame camera poses. For action-conditioned video generation, we apply GF to Next-Frame Diffusion (Cheng et al., 2025), training on 32-frame 384×224 videos for 2,000 steps with a learning rate of 6×10^{-5} and batch size 32. We set $\lambda_{\text{Angular}} = 0.5$ and $\lambda_{\text{Scale}} = 0.05$ to balance each loss component. All experiments are conducted on 8 NVIDIA A100 GPUs.

Evaluation Metrics. We evaluate visual quality using FVD (Fréchet Video Distance) (Unterthiner et al., 2018), PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index) (Wang et al., 2004), and LPIPS (Learned Perceptual Image Patch Similarity) (Zhang et al., 2018).

To further evaluate geometric consistency, we introduce Reprojection Error (RPE) (Duan et al., 2025) and Revisit Error (RVE) (Xiao et al., 2025). Reprojection Error (RPE) quantitatively measures multi-view geometric consistency by calculating the average reprojection discrepancy between projected and observed pixel locations across frames. 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. C.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.

Camera view-conditioned Video Generation. We conduct comprehensive evaluation of GF on the RealEstate10K (Zhou et al., 2018) 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.

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 (Song et al., 2025)	16	252	0.40	0.50	14.40	–	–
REPA (Yu et al., 2024a)	16	221	0.37	0.54	**15.20**	–	–
VideoREPA (Zhang et al., 2025c)	16	210	0.37	0.54	**15.20**	–	–
Geometry Forcing (ours)	16	193	**0.32**	**0.58**	14.70	–	–
Geometry Forcing (ours) + REPA	16	**179**	0.34	0.54	15.00	–	–
Cosmos* (Agarwal et al., 2025)	256	934	0.68	0.20	10.25	–	–
DFoT (Song et al., 2025)	256	364	0.55	0.36	11.40	0.3575	297
REPA (Yu et al., 2024a)	256	297	0.54	0.36	11.51	0.3337	315
VideoREPA (Zhang et al., 2025c)	256	455	0.56	0.35	11.50	0.3823	**190**
Geometry Forcing (ours)	256	243	**0.51**	**0.38**	11.87	0.3337	272
Geometry Forcing (ours) + REPA	256	**237**	**0.51**	0.37	**12.10**	**0.3264**	236

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

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 (Cheng et al., 2025) 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 per-frame camera poses simulating a full 360° rotation. We compare GF against three strong baselines: DFoT (Song et al., 2025), REPA (Yu et al., 2024a), and VideoREPA (Zhang et al., 2025c). 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 coherence and scene consistency, resulting in implausible intermediate frames 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.

Which Representation Should be Aligned? To validate the effectiveness of geometric representation, we compare two target representations in GF: VGGT (Wang et al., 2025b), trained on 3D

Table 4: **Ablation study on explicit and implicit geometry information.** We compare explicit geometry condition with internal alignment (ours).

Method	FVD-256↓
Baseline	364
Rendered Image Injection	280
Latent Feature Injection	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**

datasets with strong geometric priors, and DINOv2 (Oquab et al., 2023), 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.

Explicitly or Implicitly Integrate Geometry Information into Video Diffusion Models? To assess the benefit of internalizing geometric representations, we compare two ways of incorporating geometry into the video diffusion model: internal alignment via GF and external guidance via a ControlNet (Zhang et al., 2023). For external guidance, we test two settings: (1) injecting VGGT intermediate features into a ControlNet attached to DFoT, and (2) reconstructing the 3D scene, rendering it into 2D images, and injecting the rendered images as geometric conditions. As shown in Tab. 4, using the same VGGT features, GF outperforms explicit feature injection, and rendered-image conditioning also lags behind GF. These results show that while explicit geometric cues are helpful, internal alignment through GF provides consistently stronger supervision. By aligning internal features with geometric representations, GF enables deeper geometric understanding and yields better performance in terms of perceptual quality and structural consistency. Full evaluation results is listed here Tab. 8.

Which Layer Should be Aligned? As shown in Fig. 3, we also explore applying alignment at different layers of the video diffusion model (Song et al., 2025), which uses a 7-layer U-ViT (Bao et al., 2023) 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 (Chen et al., 2024a; Song et al., 2025; Sun et al., 2025; Cheng et al., 2025; Huang et al., 2025b). 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.

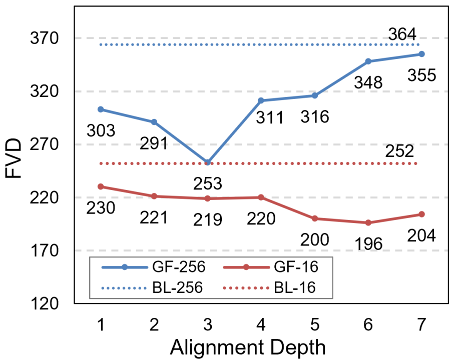


Figure 3: **Ablation study on alignment depth.** We present FVD-256 and FVD-16 results for different alignment layers of diffusion model which suggest mid-level feature is most effective to improve video quality.

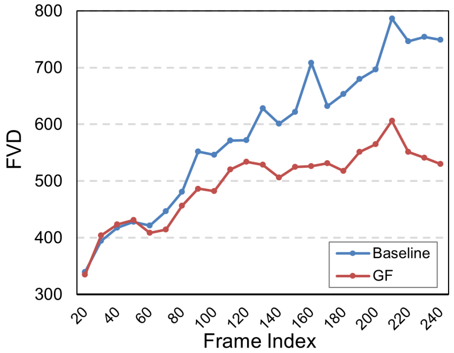


Figure 4: **Exposure bias analysis.** This figure shows the trend of FVD scores during long-term video generation. Compared to the baseline, GF results in significantly lower FVD after 100 frames.

Table 6: **User study.** Average scores on Camera Following, Object Consistency, and Scene Continuity. Each user has 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.

We compare GF with DFoT (Song et al., 2025), REPA (Yu et al., 2024a), and VideoREPA (Zhang et al., 2025c). 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.

Reproducibility

We provide comprehensive implementation details, including model architectures, training configurations, and data preprocessing procedures, in Appendix C to ensure reproducibility.

References

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.

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.

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.

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*, pp. 53–72. Springer, 2024a.

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*, pp. 7996–8006, 2024b.

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.

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*, pp. 22669–22679, 2023.

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

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.

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.

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, 2024a.

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.

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*, 2024b.

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.

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.

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.

Decart, Quevedo Julian, McIntyre Quinn, Campbell Spruce, Chen Xinlei, and Wachen Robert. Oasis: A universe in a transformer. 2024. URL `https://oasis-model.github.io/`.

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.

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

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.

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. URL `https://api.semanticscholar.org/CorpusID:274610426`.

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.

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

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

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

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.

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

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.

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*, pp. 9729–9738, 2020.

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.

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

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.

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

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*, 2025b.

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.

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.

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*, pp. 25105–25124. PMLR, 2024.

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.

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.

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.

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.

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.

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*, pp. 8009–8019, 2025a.

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.

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*, pp. 16651–16662, 2025b.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl(4) manifold. *ArXiv*, abs/2505.12549, 2025. URL `https://api.semanticscholar.org/CorpusID:278739766`.

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.

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.

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*, pp. 11453–11464, 2021.

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

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.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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*, pp. 10106–10116, 2024.

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.

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.

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.

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*, pp. 2349–2359, 2023.

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

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)*, pp. 10684–10695, 2022.

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*, pp. 2070–2080, 2024.

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.

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

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*, pp. 7364–7373, 2025.

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.

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.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

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.

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.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. Spatialvid: A large-scale video dataset with spatial annotations, 2025a. URL `https://arxiv.org/abs/2509.09676`.

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*, 2025b.

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.

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*, pp. 20697–20709, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. URL `https://arxiv.org/abs/2507.13347`.

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.

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

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*, 2025a.

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)*, pp. 16663–16673, June 2025b.

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*, 2025c.

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*, pp. 21469–21480, 2025.

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.

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.

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.

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.

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.

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*, pp. 5916–5926, 2025a.

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*, 2025b.

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*, 2024a.

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*, 2024b.

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*, pp. 3836–3847, 2023.

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*, pp. 21685–21695, 2025a.

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*, pp. 586–595, 2018.

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*, pp. 21936–21947, 2025b.

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*, 2025c.

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.

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.

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.

APPENDIX FOR ICLR26 SUBMISSION *Geometry Forcing*:MARRYING VIDEO DIFFUSION AND 3D REPRESENTATION FOR CONSISTENT WORLD MODELING

A DECLARATION OF LLM USAGE

We used large language models (LLMs) to aid or polish writing. Details are described in the paper. The use is limited to language editing (grammar, spelling, and word choice), code formatting (e.g., adding comments to the code). All scientific ideas, analysis, and conclusions were conceived, validated, and interpreted independently by the authors. We gratefully acknowledge the assistance of large language models in our work.

B 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.

C Implementation Details

C.1 Dataset

RealEstate10K (Zhou et al., 2018). 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 (Baker et al., 2022). 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.

C.2 Training

Model Architecture. We adopt a U-ViT Bao et al. (2023) 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.

C.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.

C.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 Duan et al. (2025), we utilize DROID-SLAM (Teed & Deng, 2021) 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

Table 7: **Ablation study on teacher model** Our method (Geometry Forcing) is compatible with different teachmodel including VGGT and Pi3**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 (Song et al., 2025)	256	364	0.55	0.36	11.40	0.3575	297
Pi3 Wang et al. (2025)	256	309	0.53	0.38	11.53	0.x	x
Geometry Forcing (VGGT)	256	243	**0.51**	**0.38**	11.87	0.3337	272

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 (Xiao et al., 2025). 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) (Heusel et al., 2017). Larger discrepancies indicate greater geometric or appearance drift, suggesting weaker long-term 3D consistency.

C.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 (Song et al., 2025). Specifically, we freeze the pretrained DFoT backbone and train a DPT head (Ranftl et al., 2021) to regress depth maps from its intermediate representations. The target depth maps are provided by the VGGT model (Wang et al., 2025b), 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.

D Supplementary Experiments

D.1 Ablation on Teacher Model

Geometry Forcing don't depend on specific 3D foundation model but still requires the 3D foundation to be feed-forward and support multiple images input which is required by online training. We conduct Geometry Forcing algorithm on Pi^3 model and also achieves significant improvement on video generation as shown in Tab. 7.

D.2 Explicit Geometry Control

We provide full evaluation comparison between explicit control and ours Geometry forcing in Tab. 8.

D.3 Alignment Context Length

Our Geometry Forcing input 16 frames into VGGT model to extract the latent representation and then align first 16 frame during training. We provide the ablation results on different alignment

Table 8: **Ablation study on explicit and implicit geometry information.** Our method (Geometry Forcing) achieves the best performance across all metrics on the RealEstate10K dataset for long-term (256-Frame) video generation. **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 (Song et al., 2025)	256	364	0.55	0.36	11.40	0.3575	297
Render Image Injection	256	280	0.52	0.37	11.99	x	x
Latent Feature Injection	256	364	0.55	x	x	0.x	x
Geometry Forcing (ours)	256	243	**0.51**	**0.38**	11.87	0.3337	272

Table 9: **Ablation study on GF alignment context length.** Geomtry Forcing-n indicates n frames is used to extract VGGT feature during training. The results is evaluated on the RealEstate10K dataset for long-term (256-Frame) video generation. **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 (Song et al., 2025)	256	364	0.55	0.36	11.40	0.3575	297
Geometry Forcing-4	256	261	0.51	0.38	12.21	x	x
Geometry Forcing-8	256	257	0.50	x	x	0.38	12.27
Geometry Forcing-16 (default)	256	243	**0.51**	**0.38**	11.87	0.3337	272

context length in Tab. 9. The results indicates that the when the alignment context length is longer, the 3D information is more complete, thus lead to better results.

D.4 Multiple Layer Alignment

Since the large number of potential combination of layers selected for alignment, we provide the result on align with last 3 layers of our diffusion model in Tab. 10. However, increasing number of layers to align doesn't lead to better performance.

D.5 Text-to-video generation

We extend our Geometry Forcing method to general text-to-video generation tasks. Our model is trained on 2K videos from the Wang et al. (2025a), which provides videos with detailed scene and camera descriptions. Experimental results demonstrate that our approach achieves improvements across multiple evaluation dimensions, including visual aesthetics, motion smoothness, and motion quality, as detailed in Table 11. These results indicate that Geometry Forcing can extend effectively to dynamic text-to-video training, even though VGGT itself is trained on static scenes.

E Discussion

E.1 Computational Efficiency

We perform a detailed profiling of our method on a NVIDIA A800 GPU and report both the execution time and floating-point operations (FLOPs) for different components of our model during the training stage in Table 12. The VGGT Feature Alignment contributes an additional 52.5% in execution time and 60.4% in total FLOPs. Although this alignment process increases the per-step computation compared to the base diffusion model, it significantly accelerates convergence, thereby reducing the overall training duration. For fine-tuning, our method requires only a few thousand steps and completes within hours, yielding substantial efficiency gains over full pre-training. Additionally, during inference, our method does not introduce any additional computational cost compared to other methods that use explicit or implicit memory.

We also provided a feature extraction time of VGGT model in Fig 5. The result shows that the extraction time increases from 0.1s to 0.8s when the input view from 1 to 12.

Table 10: **Alignment on Multiple Layers.** Comparison of aligning VGGT features at the middle layer vs. the last three layers of the diffusion model using Geometry Forcing.

Method	Frames	FVD↓	LPIPS↓	SSIM↑	PSNR↑	RPE↓	RVE↓
DFoT (Song et al., 2025)	256	364	0.55	0.36	11.40	0.3575	297
Geometry Forcing (Last 3 Layers)	256	309	0.53	0.38	11.53	0.xxx	xxx
Geometry Forcing (Mid)	256	243	**0.51**	**0.38**	11.87	0.3337	272

Table 11: **Evaluation on text-conditioned video generation.** FVD results of Wan2.1 1.3B before and after applying Geometry Forcing (GF) on 81-frame generation show clear improvement.

Method	Aesthetic Quality↑	Imaging Quality↑	Motion Smoothness↑
Wan2.1	0.58	0.56	0.98
Wan2.1 + GF	**0.59**	**0.59**	**0.99**

E.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 (Oquab et al., 2023) 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.

E.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.

E.4 Failure Case Analysis

Although our methods significantly improve visual quality and geometric consistency in video generation, they still struggle in certain complex scenarios. As shown in Fig. 6, the transparent, reflective glass table intermittently disappears and reappears across frames, indicating that the model still has difficulty handling reflective materials.

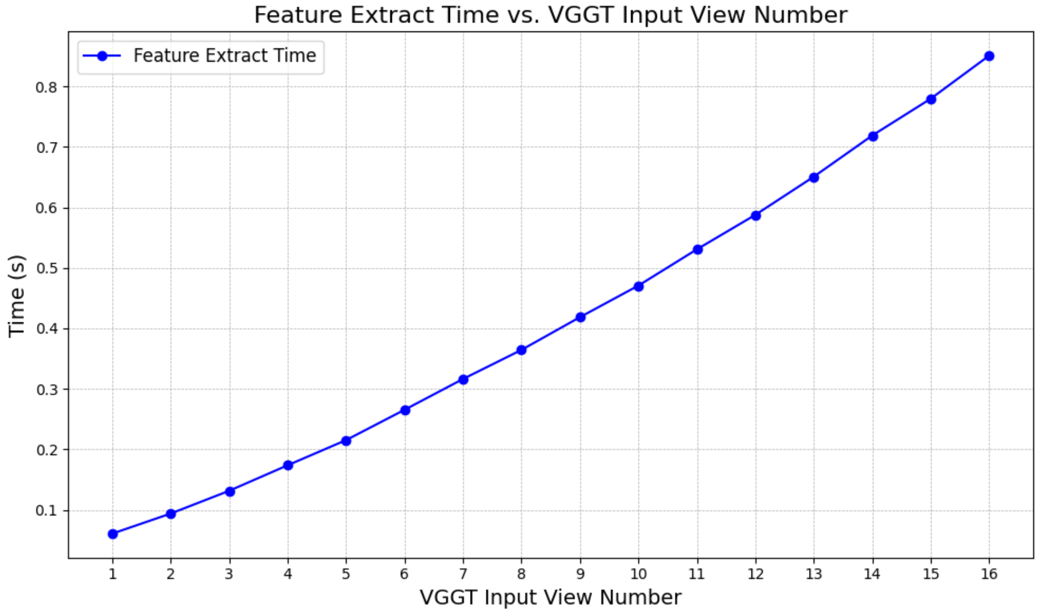


Figure 5: **VGGT Feature Extraction Time.** The feature extraction time of VGGT model increases along the number of inout view.

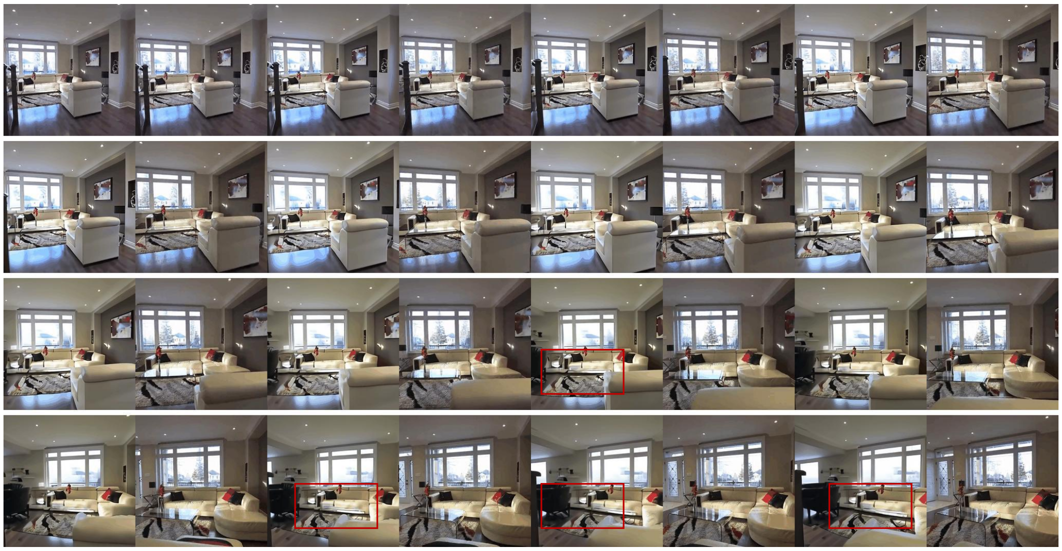


Figure 6: **Failure Case Analysis.** The transparent, reflective glass table intermittently disappears and reappears across frames, indicating that the model still has difficulty handling reflective materials. The red box indicates when the table disappears.

Table 12: **Training Stage Profiling.** We report the execution time and floating-point operations (FLOPs) for different components during a training step.

Pipeline Stage	Time		FLOPs	
	Value (s)	Percentage (%)	Value (T)	Percentage (%)
Forward (Frozen)				
VGGT Encoding	0.853	53.4%	93.3	60.4%
Forward (Learnable)				
Projector	0.017	1.1%	0.1	0.1%
Diffusion Backbone	0.220	13.8 %	17.7	11.5%
Backward (Learnable)				
Projector + Diffusion Backbone	0.506	31.7%	43.5	28.1%
Total per Step	**1.597**	**100.0%**	**154.6**	**100.0%**

F Supplementary Visualizations

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

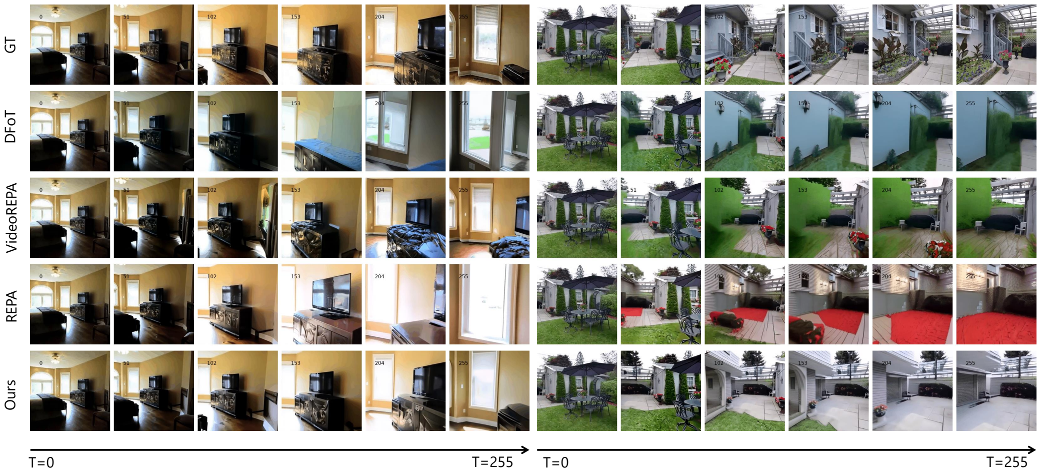


Figure 7: **Qualitative comparisons on camera-conditioned video generation.** All the videos are generated given first frame and per-frame camera pose. We comprehensively compare GF (ours) with DFoT (Song et al., 2025), VideoREPA (Zhang et al., 2025c), REPA (Yu et al., 2024a). The results demonstrate consistency in long-term video generation both inside (left) and outside (right) scenes.

Fig. 7 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 (Song et al., 2025), REPA (Yu et al., 2024a), and VideoREPA (Zhang et al., 2025c).

As shown in Fig. 7, our method generates visually coherent and geometrically consistent videos over long time horizons even when 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.

Qualitative Ablation on Alignment loss. To further assess the impact of the proposed scale alignment loss, we conduct qualitative comparisons between models trained with and without this component (Fig. 8). While angular alignment alone helps maintain basic geometric coherence, the lack of scale supervision often leads to inconsistent camera motion, manifesting as unstable perspective changes or unnatural object scaling. By introducing the scale alignment loss, our method produces noticeably smoother viewpoint transitions and more reliable camera-following behavior, demonstrating its effectiveness in stabilizing multi-frame geometry.

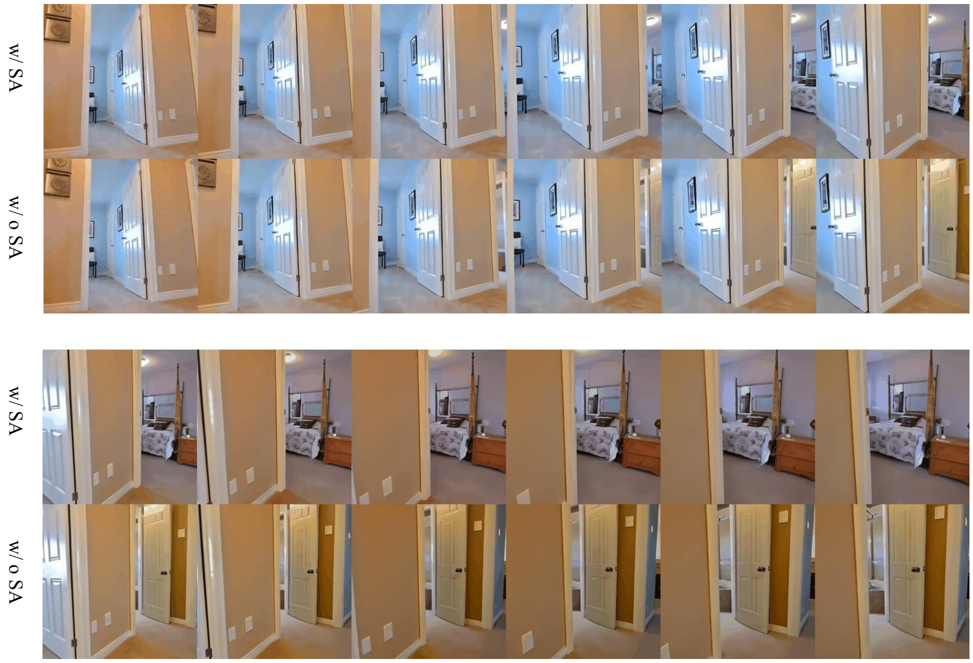


Figure 8: **Qualitative comparison of the Alignment Loss.** “w/ SA” denotes models trained with both angular alignment and scale alignment losses, while “w/o SA” refers to models trained using only angular alignment. Incorporating scale alignment enables the model to generate videos with more stable and realistic camera-following behavior.