FANTASYWORLD: GEOMETRY-CONSISTENT WORLD MODELING VIA UNIFIED VIDEO AND 3D PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality 3D world models are pivotal for embodied intelligence and Artificial General Intelligence (AGI), underpinning applications such as AR/VR content creation and robotic navigation. Despite the established strong imaginative priors, current video foundation models lack explicit 3D grounding capabilities, thus being limited in both spatial consistency and their utility for downstream 3D reasoning tasks. In this work, we present FANTASYWORLD, a geometry-enhanced framework that augments frozen video foundation models with a trainable geometric branch, enabling joint modeling of video latents and an implicit 3D field in a single forward pass. Our approach introduces cross-branch supervision, where geometry cues guide video generation and video priors regularize 3D prediction, thus yielding consistent and generalizable 3D-aware video representations. Notably, the resulting latents from the geometric branch can potentially serve as versatile representations for downstream 3D tasks such as novel view synthesis and navigation, without requiring per-scene optimization or fine-tuning. Extensive experiments show that FANTASYWORLD effectively bridges video imagination and 3D perception, outperforming recent geometry-consistent baselines in multi-view coherence and style consistency. Ablation studies further confirm that these gains stem from the unified backbone and cross-branch information exchange.

Figure 1: FANTASYWORLD overview. Given multimodal inputs (image, text, and camera trajectory), the model generates photorealistic videos along the specified views while constructing an implicit 3D representation for consistent geometry.

1 INTRODUCTION

Building world models has long been viewed as a key step toward Artificial General Intelligence (AGI). By modeling environments, objects, causal relations, and temporal dynamics, world models enable agents to predict, plan, and generate, endowing them with human-like understanding and creative abilities. At the core of this vision lies the construction of high-quality, diverse 3D environments, which support a wide spectrum of applications spanning AR/VR content creation, robotic navigation, and embodied AI at large. Plenty of approaches have been explored towards the goal of generating coherent 3D scenes, involving leveraging 2D generative priors to enable 3D generation

(Chen et al., 2023; Bahmani et al., 2024) via SDS (Poole et al., 2022), building entire 3D scenes by multi-view images with NeRF (Mildenhall et al., 2021) and 3DGS (Kerbl et al., 2023).

Recently, camera-guided video diffusion models have gained popularity, where image frames implicitly encode the scene's 3D structure with multiple images from different viewpoints, as videos are natural 2D projections of dynamic 3D worlds. Such methods harness the spatiotemporal capacity of video diffusion to improve 3D consistency and multi-view modeling, such as ReconX (Liu et al., 2024a), Gen3C (Ren et al., 2025), DimensionX (Sun et al., 2024), and ViewCrafter (Yu et al., 2024b), to name a few. These models excel at producing geometrically plausible local shapes and mitigating view inconsistency through temporal priors.

However, despite being trained on vast Internet video corpora and exhibiting strong imaginative priors, generative video models lack explicit 3D supervision, making it difficult to preserve real-world structure and spatial consistency. This reveals a central challenge in world model construction: *how to inject reliable geometric grounding into video generation without sacrificing creative capacity*.

Lately, the emergence of 3D foundation models, such as DUSt3R (Wang et al., 2024a), MASt3R (Duisterhof et al., 2025), Fast3R (Yang et al., 2025a), and VGGT (Wang et al., 2025a), demonstrates that robust geometry can be predicted in a single forward pass without additional 3D reconstruction, providing a scalable solution to geometry-consistent generation. Consequently, several works attempt to couple 2D video generation with 3D reasoning. Voyager (Huang et al., 2025a) trains an end-to-end model that jointly predicts RGB and depth while maintaining world consistency through a cache and geometry-injected frames, and then reconstructs explorable 3D scenes. Geometry Forcing (Wu et al., 2025) further marries video diffusion with an explicit 3D representation to strengthen geometric consistency during training. Matrix-3D (Yang et al., 2025c) shows two practical routes from omni-directional video to a navigable world, including an optimization path and a large panorama reconstruction model that directly infers 3D Gaussians from video latents. WonderWorld (Yu et al., 2025a) focuses on interactive single-image scene authoring with layered Gaussian surfels and guided depth diffusion for fast, connected scene creation. Vidu4D (Wang et al., 2024c) shows that even a single generated video can support high-quality 4D reconstruction when paired with dynamic Gaussian surface elements. Complementary to these trends in content creation, GaussianWorld (Zuo et al., 2025) frames perception as streaming 4D occupancy forecasting, underscoring the value of a 3D world representation that can generalize across embodied tasks.

Despite significant progress in video-to-3D modeling, several limitations remain. First, most video generative models operate purely within the video domain, yielding features that cannot directly support 3D reasoning. When explicit 3D reconstruction is desired, they often resort to additional scene-specific optimization with NeRF (Mildenhall et al., 2021) or 3DGS (Kerbl et al., 2023), which introduces computation overhead. Second, video imagination and 3D perception remain weakly coupled at inference, preventing mutual reinforcement, which effectiveness is evidenced by recent works in 3D scene understanding (Huang et al., 2025b). For instance, although Voyager (Huang et al., 2025a) predicts RGB and depth jointly to maintain world consistency, and WorldExplorer (Schneider et al., 2025) leverages video-based imagination for scene exploration, in both cases the two processes operate largely independently, highlighting the persistent limitation of weak coupling between video generation and 3D perception. Third, many approaches, such as Geometry Forcing (Wu et al., 2025), integrate 3D priors by fine-tuning video foundation models (VFMs) while keeping a 3D model like VGGT frozen, which incurs substantial computational cost and risks compromising the VFM's general generative capacity.

To address these limitations, we introduce FANTASYWORLD, a geometry-enhanced framework that efficiently produces reusable 3D-consistent features by augmenting frozen VFMs with an additional trainable branch for geometric inference, tightly coupling video imagination and 3D perception without expensive per-scene optimization or fine-tuning. Instead of predicting depth or point clouds from RGB images, we directly infer camera parameters and 3D signals from video latents. This is inspired by VGGT but achieves tighter integration between generative and geometric modeling. Concretely, we split the backbone of video foundation models (i.e., Wan2.1 in our case) into Preconditioning Blocks (PCB) that inject video priors and stabilize latents, and Integrated Reconstruction and Generation Blocks (IRG) that fuse spatiotemporal tokens with a geometry co-encoder to predict a geometry-aware implicit 3D field. As a result, our model generates camera-conditioned video features alongside an explicit 3D representation in a single forward pass, without relying on additional

3D reconstruction (e.g., NeRF or 3DGS) or iterative memory refinement as in Voyager (Huang et al. (2025a)).

To this end, our contributions are as follows: (i) Unified video-3D modeling: We propose FANTASYWORLD, a geometry-enhanced framework that jointly predicts video latents and an implicit 3D field through a single backbone, preserving imaginative priors while exposing explicit geometry. (ii) 2D/3D cross-branch supervision: We introduce constraints that let geometry supervise video features and video priors regularize 3D prediction, ensuring 3D-consistent frames inference. (iii) Potential for generalizable 3D features: We expect that the resulting video-3D representations serve as versatile features for downstream tasks, such as novel view synthesis and navigation, without per-task adaptation, which has been evidenced by recent works, such as AnySplat (Jiang et al., 2025).

2 RELATED WORK

2.1 FEED-FORWARD RECONSTRUCTION

Feed-forward reconstruction methods have achieved promising results in recovering 3D properties of a scene from a set of images in a single pass (Wang et al., 2024b; 2025b; Wang & Agapito, 2024; Duisterhof et al., 2025; Tang et al., 2025; Yang et al., 2025b; Zhang et al., 2024; 2025b; Wang et al., 2025a;c). DUSt3R (Wang et al., 2024b) and MASt3R (Duisterhof et al., 2025) directly estimate point clouds from images, which is suitable for challenging scenarios such as low-texture regions, but can only input two images at a time. Fast3R (Yang et al., 2025b) can process thousands of frames at once, surpassing the aforementioned restrictions. VGGT (Wang et al., 2025a) introduces a more generalized framework capable of supporting input from one frame to multiple frames and producingg multiple 3D attributes. Furthermore, its feature backbone serves as a versatile feature extractor for various downstream tasks. Recently, many methods have integrated diffusion knowledge into reconstruction techniques, thereby enriching the 3D modeling process with the ability to imagine occluded or unobserved areas (Fu et al., 2024; Hu et al., 2025; Ke et al., 2024; Lu et al., 2025; Yang et al., 2024; Zhu et al., 2024; 2023; Liang et al., 2025). In this work, we adopt an architecture similar to VGGT as our 3D implicit feature extractor. Specifically, we extract implicit 3D features from WanDiT block and integrate them within a multi-task learning paradigm.

2.2 GEOMETRY-AWARE VIDEO GENERATION

The consistency of generated videos with the real physical world is a crucial challenge in simulation. Many previous works have attempted to incorporate geometric constraints during the video generation process, primarily categorized into explicit guidance and implicit guidance. For explicit guidance, many prior works explicitly incorporate 3D signals (such as point clouds, mesh, etc.) obtained from the first frame into the model, thereby demonstrably improving generation consistency (Yu et al., 2024a; Huang et al., 2025a; Yang et al., 2025d; Cao et al., 2025). These methods often suffer from insufficient point cloud accuracy and limited scope when subjected to significant viewpoint changes. For implicit guidance, many prior works investigate methods to enable models to perceive 3D structural information within the diffusion process (Zhang et al., 2025a; Team et al., 2025; Wu et al., 2025). Geometry-Forcing (Wu et al., 2025) aligns the model's intermediate representations with the features of a pre-trained geometric foundational model, risks undermining the creativity of large-scale video generation models trained on massive video data.

2.3 JOINT 3D AND VIDEO GENERATION

Integrating both video and 3D structural information enables the generation of comprehensive 3D scenes for applications in embodied AI and simultaneously facilitates a better representation of the underlying physical world. AETHER (Team et al., 2025) unifies RGB–depth modeling and couples reconstruction with video generation. DeepVerse (Chen et al., 2025a) achieves interactive joint generation via an autoregressive paradigm and control via text-specified control signals. Voyager (Huang et al., 2025a) incorporates point cloud projection within its joint generation framework to achieve camera control.

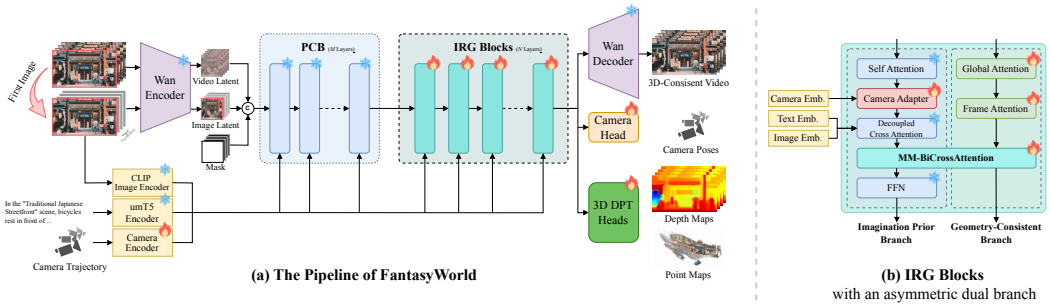

Figure 2: Overview of FANTASYWORLD. Inputs (image, text, camera) are processed by PCBs and stacked IRG blocks, where an asymmetric dual-branch design couples video synthesis with 3D reasoning. The model outputs geometry-consistent video frames and task-agnostic 3D features.

In this paper, we introduce an implicit 3D feature encoding branch designed to decode various 3D attributes. Through mutual enhancement achieved during the diffusion process, it simultaneously yields video features and 3D structural features.

3 METHODOLOGY

3.1 OVERVIEW

FANTASYWORLD is a unified feed-forward model for joint video and 3D scene generation. Given a reference image, an optional text prompt, and a target camera trajectory, the model produces a video aligned with the specified views while simultaneously constructing an implicit 3D representation. Inputs are encoded by pretrained backbones: CLIP (Radford et al., 2021) for images, umT5 (Chung et al., 2023) for text, and a learned camera encoder, following Wan's Plücker-ray design (Wan et al., 2025), for camera poses. These signals jointly condition both the video and geometry branches during training and inference.

As shown in Fig. 2 (a), the front end employs *Preconditioning Blocks (PCBs)* that reuse the frozen WanDiT denoiser to supply partially denoised latents, ensuring the geometry pathway operates on meaningful features rather than pure noise. The backbone then consists of stacked *Integrated Reconstruction and Generation (IRG) Blocks*, which iteratively refine video latents and geometry features under multimodal conditioning. Each IRG block contains an asymmetric dual-branch structure (Fig. 2 (b)): an *Imagination Prior Branch* for appearance synthesis and a *Geometry-Consistent Branch* for explicit 3D reasoning, coupled through lightweight adapters and cross attention.

The outputs are twofold. The imagination branch generates geometry-consistent video frames along the trajectory, while the geometry branch produces task-agnostic 3D features decoded by DPT heads into depth maps, point maps, and camera poses. This unified design supports downstream tasks such as novel view synthesis, pose estimation, and depth prediction, all without per-scene optimization.

In summary, FANTASYWORLD bridges generative video priors with structured geometric reasoning in a single forward pass, producing outputs that are both photorealistic and geometrically consistent.

3.2 PRECONDITIONING BLOCKS

Denoising diffusion models progressively remove noise across timesteps, revealing structure and details in the signal. Recent theory (Han et al., 2025) shows that the denoising objective balances learning signal and noise, enabling structural information to emerge gradually—consistent with the empirical observation that features become more informative as denoising unfolds. We further observe a similar effect along network depth: even at a fixed timestep, deeper WanDiT layers produce clearer spatial structure (Fig. 3), suggesting that denoising progresses across both time and depth.

Motivated by this, we introduce *Preconditioning Blocks (PCBs)* at the front end of our framework. We reuse the first 16 frozen layers of Wan2.1, following its camera conditioning design, to partially denoise video latents. This ensures that inputs to the geometry branch contain geometry-relevant

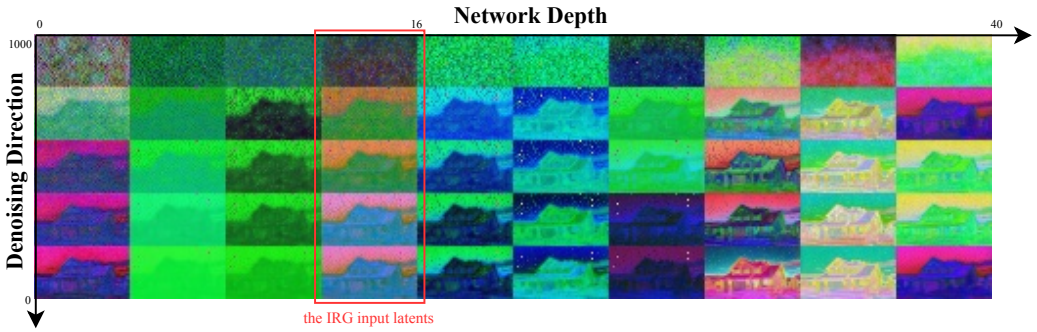

Figure 3: PCA over timestep–block pairs: rows vary timesteps top → bottom, columns vary blocks left → right; the red rectangle marks the IRG input latents.

cues rather than pure noise, reducing gradient variance and avoiding early training dominated by high-noise latents. PCBs thus bridge noisy initialization and geometry-aware processing, providing stable supervision from the start and allowing the geometry branch to focus on refining structure.

3.3 Integrated Reconstruction and Generation Blocks

At the core of FantasyWorld are the *Integrated Reconstruction and Generation (IRG) Blocks*, stacked as the fundamental units of the model. Each IRG block adopts an asymmetric dual-branch design (Fig. 2 (b)): the *Imagination Prior Branch* reuses the pretrained Wan2.1 backbone to propagate appearance-rich spatiotemporal features, while the *Geometry-Consistent Branch* projects them into a geometry-aligned latent space. Unlike VGGT, which relied on DINO features, we bridge the geometry pathway to Wan latents, ensuring geometry is inferred in the same domain as video synthesis and avoiding feature mismatch.

IRG blocks unify the two branches through bidirectional cross-attention (MM-BiCrossAttn (Liu et al., 2024b; 2025)). Here, geometric cues regularize video features for multi-view coherence, while video priors provide imaginative signals that complete occluded regions and refine geometry. As cooperative units, stacked IRG blocks progressively enhance both video latents and geometry features, forming the core mechanism where imagination and structure converge.

The geometry branch outputs an implicit representation decoded by a custom 3D DPT head. This head performs temporal decoding aligned with the WanVAE video frames. Motivated by the depth-wise emergence of structure observed in Sec. 3.2, the head inverts the conventional reassemble strategy: instead of drawing fine features from early encoder layers dominated by noise, it sources them from late diffusion blocks where semantics are stronger and denoising is mature. Anchoring predictions in these stable features improves depth accuracy, stabilizes pose estimation, and enhances the consistency of the implicit 3D field.

3.4 Bridge to Unify Training

Two-stage framework. Training in FantasyWorld follows two stages: first *bridging* the geometry branch to the Wan feature space, then *unifying* geometry and video through bidirectional cross-attention. Stage 1 freezes the Wan2.1 backbone and adapts the geometry branch to consume hidden features; Stage 2 introduces cross-branch adapters for joint optimization, while keeping the backbone frozen. Please refer to A.2, A.3 A.4 for more implementation and training details.

Stage 1: Latent Bridging. We select hidden features from block 16 of Wan2.1 and feed them to the geometry branch through a lightweight transformer adapter that maps to a geometry-aligned latent space. This latent is encoded and decoded into camera, depth, and point map predictions, while only the geometry branch is trained. Supervision is applied via:

$$\mathcal{L}_{\text{geo}} = \alpha\,\mathcal{L}_{\text{depth}} + \beta\,\mathcal{L}_{\text{pmap}} + \gamma\,\mathcal{L}_{\text{camera}}$$

where $\mathcal{L}_{\text{depth}}$ follows Video Depth Anything (Chen et al., 2025b), $\mathcal{L}_{\text{pmap}}$ follows VGGT (Wang et al., 2025a), and $\mathcal{L}_{\text{camera}}$ is a Huber penalty. This stage ensures the geometry branch operates stably on Wan latents instead of raw noise.

Stage 2: Unified Co-Optimization. From block 16 onward, we insert one bidirectional cross-attention adapter after each of 24 transformer blocks, aligned with the geometry branch. For video tokens X_v and geometry tokens X_g, with projections (Q_v, K_v, V_v) and (Q_g, K_g, V_g), attention is:

$$A = \text{softmax}\left(\frac{Q_v K_g^\top}{\sqrt{d_k}}\right)$$

and updates are:

$$X_v^+ = X_v + \gamma_v A V_g, \qquad X_g^+ = X_g + \gamma_g A^\top V_v$$

with learnable gates γ_v, γ_g. Geometry-to-video updates enforce 3D consistency, while video-to-geometry updates inject generative priors. Camera parameters are embedded with a pose encoder (Wan et al., 2025); we modify the pose adapter to predict only the shift β_i, injected additively:

$$f_i = f_{i-1} + \beta_i,$$

applied to the first 24 of 40 blocks.

Training objective. The final objective is:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{z_0,\epsilon,t,c}\Big[\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2\Big] + \lambda \mathcal{L}_{\text{geo}}$$

combining the standard diffusion loss with geometry supervision which aggregates depth, point map, and camera supervision. The weight λ balances video generation and geometry learning, enforcing multi-view coherence and enabling cross-branch co-adaptation.

4 EXPERIMENT

We conduct comprehensive experiments to evaluate FANTASYWORLD across large-scale benchmarks and diverse scenarios. Sec. 4.1 outlines datasets, training protocols, and evaluation settings. Sec. 4.2 assesses world generation, including comparisons with state-of-the-art baselines, analysis under varying camera motions, and ablations on the geometry branch. Sec. 4.3 examines geometric fidelity through quantitative reconstruction metrics and qualitative visualizations, further validating the role of explicit geometry modeling.

4.1 IMPLEMENTATION DETAILS

Datasets. Our training corpus consists of about 180k video clips collected from a diverse mix of real-world and simulated sources, with geometric supervision obtained through multiple strategies. For the RealEstate10K (Zhou et al., 2018) and ACID datasets (Liu et al., 2021), we generate multi-view consistent depth maps using a reconstruction-based pipeline. For additional datasets, which include the real-world DL3DV (Ling et al., 2024), WildRGB (Xia et al., 2024), and ScanNet (Dai et al., 2017) together with the simulated TartanAir (Wang et al., 2020), we apply the Cut3R (Wang et al., 2025b) to extract geometric labels.

Training Details. We train our model with the AdamW optimizer using a learning rate of 10^{-5}. The process consists of two stages. In Stage 1 (latent bridging), the geometry branch is trained for 20,000 steps with a global batch size of 64. During this stage, only the geometry branch parameters are updated, adapting it to the feature space of the frozen video backbone. In Stage 2 (unified co-optimization), training is performed with 81-frame clips at resolutions of 592×336 and 336×592. In this stage, the lightweight interaction modules (bidirectional cross-attention and camera control adapter) are fine-tuned for 10,000 steps with a global batch size of 112, while the core backbones remain frozen. We train Stage 1 with 64 H20 GPUs for 36 hours and Stage 2 with 112 H20 GPUs for 144 hours.

Evaluation. We conduct evaluation on 1,000 samples drawn from the photorealistic subset of the WorldScore static benchmark (Duan et al., 2025). The static split of WorldScore is divided into photorealistic and stylized subsets. Since our objective is to model real-world environments for embodied intelligence, we focus on photorealistic video synthesis. In contrast, evaluating 3D consistency

in stylized videos is ill-defined and less representative of performance in realistic scenarios, and is therefore excluded from our study. We assess our model along two complementary dimensions:

- **World Generation**: For a holistic evaluation of controllable world generation, we report performance on the WorldScore benchmark, which measures performance across multiple axes, including camera and object control, content alignment, 3D consistency, and perceptual quality.
- **Geometric Fidelity**: To quantify geometric fidelity, we reconstruct each scene with 3DGS, and report PSNR, SSIM, and LPIPS.

Together, these metrics capture both the controllability of video-based world generation and the structural accuracy of 3D geometry. For more evaluation details, please refer to A.5.

4.2 World Generation

We evaluate the video generation capabilities of FantasyWorld by comparing it against WonderWorld (Yu et al., 2025a), Uni3C Cao et al. (2025), Voyager (Huang et al., 2025a), and AETHER (Team et al., 2025), which represent the most relevant state-of-the-art baselines for 3D world generation.

Quantitative Comparison. We evaluate under two camera-motion settings. The *Small* configuration follows the WorldScore static benchmark, using 1,000 photorealistic samples to ensure consistency with prior work. We additionally introduce a *Large* setting with 100 curated cases featuring wide orbital or panning trajectories (up to 90°), to test robustness under challenging viewpoints. As shown in Table 1, FantasyWorld achieves the highest scores on all consistency-related metrics (*3D Consist.*, *Photo Consist.*, and *Style Consist.*). Moreover, it yields the lowest standard deviation across samples, suggesting that our method is not only more accurate on average but also more stable across diverse scenes. In the Large setting, the gains are particularly pronounced, demonstrating strong geometric stability even under substantial viewpoint shifts. Compared to baselines that prioritize instruction following and camera manipulation (Yu et al., 2025a), our framework focuses instead on embedding geometric awareness into video features. Although this design does not optimize for camera or content alignment benchmarks, it enables stable and reusable 3D representations (our main focus), which the results demonstrate to be both effective and robust.

Qualitative Analysis. Under large camera motion, distinct failures emerge: WonderWorld shows tearing and holes indicative of missing geometry; in Uni3C and Voyager, first-frame point-cloud priors quickly fall out of view, causing style drift in Uni3C and multi-view misalignment in Voyager; AETHER, despite generating RGB-D point clouds, often produces incoherent, low-detail content (see Fig. 4). In contrast, FantasyWorld predicts an implicit 3D representation that evolves with the video, ensuring stable geometry and appearance across time and yielding coherent, 3D-consistent results without the failures of static priors.

Ablation on the Geometry Branch. To assess the role of explicit geometry modeling, we compare our full model with a variant where the geometry branch and bidirectional cross-attention are removed. As shown in Table 1, removing the geometry branch leads to declines in *Photo Consist.* and *Style Consist.*, and an especially severe drop in *3D Consist.*, underscoring its critical role in ensuring multi-view coherence and high-fidelity generation.

4.3 Geometric Fidelity

Quantitative Comparison. We evaluate video–geometry consistency using 3DGS (Kerbl et al., 2023) on 100 *RealEstate10K* samples, comparing three settings (Tab. 2): removing the geometry branch with VGGT initialization, our full model with VGGT initialization, and our full model with feed-forward point-cloud initialization. Results show that, under the same VGGT initialization, adding the geometry branch consistently improves PSNR/SSIM and reduces LPIPS, confirming its role in enforcing 3D consistency. Direct initialization from our predicted point clouds yields slightly lower scores than VGGT initialization, but still provides competitive results, indicating that our geometry branch produces meaningful 3D structure without relying on external supervision.

Qualitative Analysis. We compare reconstructed 3D scenes from Voyager, AETHER, Uni3C, and our method on indoor and outdoor cases (Fig. 5). All methods use the same pipeline: VGGT predicts

Table 1: WorldScore with *Small* vs. *Large* camera motion.

Method	Motion	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.
WonderWorld	Small	82.85±19.69	67.86±23.56	55.79±34.89	**92.32**	47.63	**79.09**	**69.03**
AETHER	Small	79.84±14.68	58.68±38.59	72.09±32.62	57.44	52.26	28.06	41.11
Uni3C	Small	78.59±21.08	85.48±20.98	88.32±18.47	62.94	45.83	47.40	57.00
Voyager	Small	56.00±26.32	80.68±16.32	72.89±29.78	45.92	**57.69**	48.36	44.74
Ours w/o 3D	Small	79.77±16.06	83.86±8.73	92.54±12.90	57.94	37.33	43.31	55.85
Ours w/ 3D	Small	**83.31**±14.24	**86.11**±7.97	**94.22**±9.11	57.05	34.46	38.45	57.40
WonderWorld	Large	63.70±24.37	3.22±8.47	35.95±33.47	**96.28**	38.61	**97.10**	**72.46**
AETHER	Large	63.97±17.39	33.11±23.99	61.99±32.24	4.43	34.78	33.69	35.09
Uni3C	Large	73.95±17.55	46.78±32.64	71.43±29.38	8.69	34.28	77.88	51.12
Voyager	Large	13.82±19.96	9.52±17.17	61.34±35.29	0.00	**49.23**	64.10	39.21
Ours w/o 3D	Large	72.06±20.14	56.98±23.60	81.59±22.23	9.32	34.44	75.85	46.96
Ours w/ 3D	Large	**74.83**±16.31	**60.61**±21.39	**82.02**±19.56	11.24	31.96	77.20	50.46

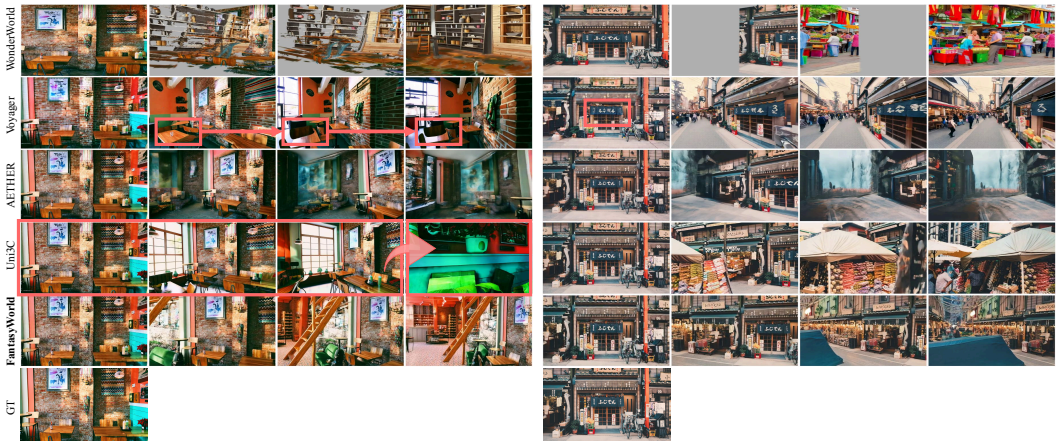

Figure 4: Qualitative comparison of world generation. WonderWorld shows missing regions, Voyager suffers from temporal incoherence and degraded first-frame fidelity, AETHER produces low-detail outputs, and Uni3C exhibits abrupt stylistic shifts. In contrast, FANTASYWORLD maintains stronger 3D consistency and coherent style across views.

Table 2: 3DGS reconstruction on *RealEstate10K*. Post-reconstruction (Post Rec) indicates the 3DGS initialization source: either from the VGGT point cloud or our own feed-forward point cloud.

Method	Post Rec.	PSNR ↑	SSIM ↑	LPIPS ↓
Ours w/o 3D	VGGT	26.89	0.84	0.17
Ours w/ 3D	VGGT	**28.24**	**0.86**	**0.14**
Ours w/ 3D	-	26.54	0.85	0.19

point maps from generated frames, then reconstructs point clouds. Baselines show structural artifacts and duplication (e.g., blurred signage, layered walls, distorted layouts), while our method yields cleaner, more complete geometry. Indoors, walls and corners stay rectilinear and closed; outdoors, text and facades are sharper, indicating stronger 3D consistency.

Ablation on the Geometry Branch. As shown in Table 2, the same geometry-ablated variant described in Sec. 4.2 produces weaker reconstruction quality, while the full FANTASYWORLD achieves clearer geometry and thus confirms the geometry branch is crucial for robust 3D representation.

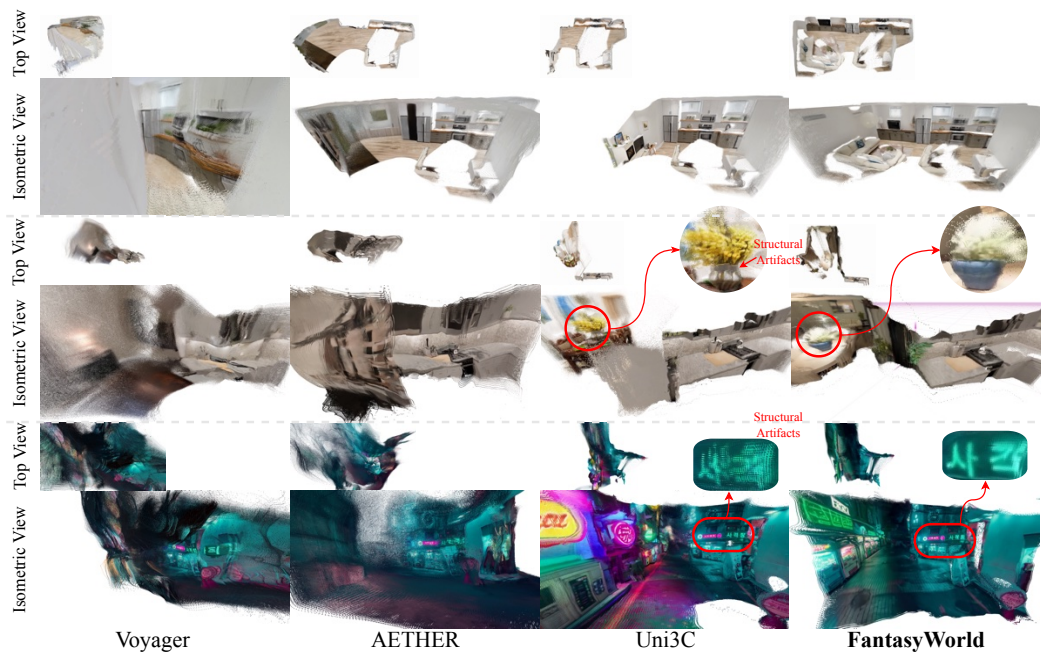

Figure 5: Qualitative Comparison of Geometry Fidelity.

5 Conclusion and Limitations

In this work, we presented FantasyWorld, a unified feed-forward model designed to generate 3D-consistent, explorable virtual worlds in a single pass. Our core contribution is a novel architecture that bridges the imaginative power of a pre-trained video diffusion backbone with the geometric rigor of a VGGT-style 3D geometry branch. To achieve this, we separate the prediction of 3D geometry from video appearance, introducing a dedicated geometry branch that operates on the video model's internal features while preserving its powerful generative capabilities. This is facilitated by a bidirectional cross-attention mechanism that allows video and geometry to mutually reinforce one another during generation. Our experiments show that the generated videos not only maintain strong visual realism but also achieve higher multi-view coherence and improved geometric fidelity compared to existing methods, offering a practical and efficient path toward creating structured, reusable world models for embodied AI.

Our model is currently designed for fixed-length clip generation, and extending it to continuous, long-range synthesis remains an important next step. Achieving this will require developing effective caching or streaming mechanisms for our implicit 3D representation to maintain state over time. Recent work such as Context-as-Memory (Yu et al., 2025b) has already begun to address this challenge in the domain of long video generation by introducing memory retrieval strategies that preserve scene consistency across extended trajectories. In contrast, our primary focus has been on validating the potential of large video diffusion backbones to serve as world models when paired with geometric supervision. This design choice allowed us to clearly demonstrate our central hypothesis: that an implicit 3D representation, learned directly from a video model's hidden features, is a powerful and effective tool for enforcing 3D consistency.

Reproducibility Statement

We ensure reproducibility by detailing FantasyWorld's architecture (PCB, IRG, bidirectional cross-attention) in Sec. 3, dataset composition and preprocessing in Sec. 4.1 and A.5, training protocols in Sec. 4.1 and A.4, and evaluation metrics, splits, results, and ablation studies in Sec. 4.1, 4.2, and 4.3. Baseline code and pretrained weights are documented in A.5, and an anonymous link to our implementation and datasets is provided in A.6.

REFERENCES

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, 2024.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. *arXiv preprint arXiv:2504.14899*, 2025.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025a.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22246–22256, 2023.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22831–22840, 2025b.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations*, 2023.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *CoRR*, 2025.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pp. 1–10. IEEE, 2025.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2005–2015, 2025.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation, 2025a. URL `http://arxiv.org/abs/2506.04225`.

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

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. AnySplat: Feed-forward 3d gaussian splatting from unconstrained views, 2025. URL `http://arxiv.org/abs/2505.23716`.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 798–810, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024a.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024b.

Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11250–11263, 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.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5283–5293, 2025.

Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling, 2025. URL http://arxiv.org/abs/2503.18945.

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.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

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, 2024a.

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, 2024b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

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. *arXiv preprint arXiv:2507.13347*, 2025c.

Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels, 2024c. URL http://arxiv.org/abs/2405.16822.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025. URL http://arxiv.org/abs/2507.07982.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22378–22389, 2024.

Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, et al. Unipad: A universal pre-training paradigm for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15238–15250, 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21924–21935, 2025a.

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, 2025b. URL http://arxiv.org/abs/2501.13928.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, Eric Li, Yang Liu, Yikai Wang, Hao-Xiang Guo, and Yahui Zhou. Matrix-3d: Omnidirectional explorable 3d world generation, 2025c. URL http://arxiv.org/abs/2508.08086.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025d.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6658–6667, 2024a.

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, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025b.

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.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

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.

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.

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.

Haoyi Zhu, Honghui Yang, Xiaoyang Wu, Di Huang, Sha Zhang, Xianglong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, Tong He, et al. Ponderv2: Pave the way for 3d foundation model with a universal pre-training paradigm. *arXiv preprint arXiv:2310.08586*, 2023.

Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. Spa: 3d spatial-awareness enables effective embodied representation. *arXiv preprint arXiv:2410.08208*, 2024.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6772–6781, 2025.

A Appendix

A.1 Declaration on LLM Usage

We used a large language model (LLM) to aid in the preparation of this manuscript. Specifically, the LLM was employed to polish the writing, improve readability, and refine grammar and style. All technical content, including the model design, experimental methodology, and results, was conceived, implemented, and validated by the authors. The LLM did not contribute to the development

of algorithms, the execution of experiments, or the analysis of data. Its role was limited to assisting with language clarity and presentation.

A.2 Camera Motion Control

For camera motion control, we largely follow the methodology of Wan et al. (2025), which consists of a camera pose encoder and a pose adapter. Our camera pose encoder mirrors their design, transforming camera parameters into multi-level feature embeddings via Plücker coordinates and a series of convolutional operations.

Our primary modification lies in the camera pose adapter. While the original work utilizes a full Adaptive Layer Normalization (AdaLN) to predict both scaling (γ_i) and shifting (β_i) parameters, we propose a streamlined variant that exclusively generates the shifting parameters. The features are integrated into the video latent f_{i-1} at each layer i through a simple additive projection:

$$f_i = f_{i-1} + \beta_i,$$

where β_i is derived from the camera embeddings. Also the camera control module is adapt only the first 24 block among the 40 transformer blocks.

A.3 3D DPT Head Architecture

Our model employs a custom 3D DPT head, which extends the 2D version from VGGT (Wang et al., 2025a) to process video sequences via two key innovations, as shown in Fig 6.

First, we invert the spatial reassembly logic to align with the nature of diffusion backbones. Conventional DPTs, designed for standard encoders, assume that shallow layers contain high-frequency spatial details and therefore upsample them most aggressively. Diffusion models operate differently: they progressively denoise the input at each block. This means deeper blocks produce features that are not only semantically richer but also less noisy and more structurally reliable. Consequently, we invert the DPT logic. Our Geometry-Consistent Branch uses features from blocks $\{8,12,18,24\}$ and upsamples features from the deepest layers the most, while downsampling those from the shallower ones. This anchors the fusion process in the highest-quality information the backbone provides.

Second, we temporally upsample each feature stream using two sequential Temporal Blocks after the reassemble block. The design of the upsample block is inspired by the WanVAE decoder (Wan et al., 2025). Each block first doubles the temporal resolution and then applies a causal 3D convolution. This process results in a total 4x temporal upsampling factor, transforming an input of t frames into a smooth output sequence of $T = 4(t-1)+1$ frames.

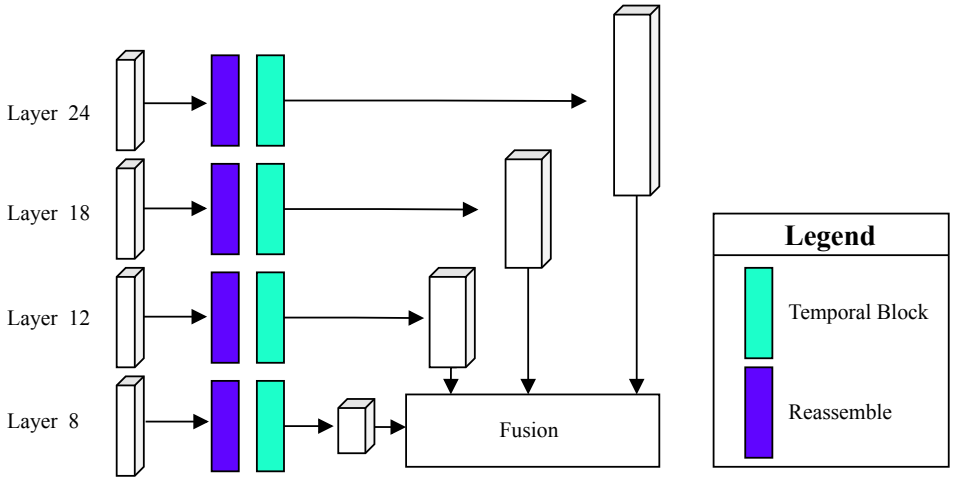

Figure 6: Our 3D DPT head with temporal blocks for temporal upsampling.

A.4 Training Details for the Geometry Branch

Camera Pose Prediction. Following VGGT (Wang et al., 2025a), we concatenate one learned camera token and four register tokens to the video token sequence. Because camera trajectories are temporally smooth in videos, we apply a lightweight 1D convolution to the camera token stream to perform temporal upsampling. For each frame f_i, the geometry head predicts a 9D camera parameter vector $\mathbf{g}_i \in \mathbb{R}^9$ (rotation, translation, and field-of-view).

Depth and Point Map prediction We use hidden features from the Geometry-Consistent Branch's layers $\{8, 12, 18, 24\}$ with shapes $\mathbb{R}^{t\times(h\times w)\times c}$. The features are fed a Depth 3D DPT head and a Point-Map 3D DPT head. The depth head produces $D \in \mathbb{R}^{T\times H\times W}$, and the point head produces a 3D point map $P \in \mathbb{R}^{T\times H\times W\times 3}$ together with a confidence (uncertainty) map $\Sigma^P \in \mathbb{R}^{T\times H\times W}$, where $h = \frac{H}{16}$, $w = \frac{W}{16}$, and $T = (t-1)\times 4 + 1$.

Camera loss. We supervise camera parameters with a robust Huber loss, where $\hat{\mathbf{g}}_i$ is the ground truth:

$$\mathcal{L}_{\text{camera}} \;=\; \sum_{i=1}^{N} \big\|\hat{\mathbf{g}}_i - \mathbf{g}_i\big\|_{\epsilon}.$$

Depth loss. We adapt Video Depth Anything (Chen et al., 2025b) and combine a temporal gradient matching term with a per-frame scale-sensitive spatial loss term:

$$\mathcal{L}_{\text{depth}} \;=\; \alpha\,\mathcal{L}_{\text{TGM}} \;+\; \beta\,\mathcal{L}_{\text{frame}},$$

where $\mathcal{L}_{\text{TGM}}$ enforces temporal consistency of depth gradients, and $\mathcal{L}_{\text{frame}}$ measures per-frame depth error without scale/shift normalization.

Point-map loss. Following VGGT (Wang et al. (2025a)), we penalize both point positions and local gradients, weighted by the predicted uncertainty:

$$\mathcal{L}_{\text{pmap}} \;=\; \sum_{i=1}^{N} \Big\|\Sigma_i^P \odot \big(\hat{P}_i - P_i\big)\Big\| \;+\; \Big\|\Sigma_i^P \odot \big(\nabla\hat{P}_i - \nabla P_i\big)\Big\| \;-\; \gamma\,\log\Sigma_i^P.$$

Total objective. The geometry branch is trained with

$$\mathcal{L}_{\text{geo}} \;=\; \mathcal{L}_{\text{depth}} \;+\; \mathcal{L}_{\text{pmap}} \;+\; 3\,\mathcal{L}_{\text{camera}}.$$

A.5 Evaluation Details

To ensure the reproducibility of our evaluation, we provide the implementation details of all baselines considered in our benchmark. For each method, we list the official GitHub repository, the specific commit used in our experiments, and the resolution and number of frames generated per sample.

Repositories and Commits.

- **Voyager:** `https://github.com/Tencent-Hunyuan/HunyuanWorld-Voyager` (commit `54a658b`)
- **Uni3C:** `https://github.com/alibaba-damo-academy/Uni3C` (commit `75ed6e2`)
- **AETHER:** `https://github.com/InternRobotics/Aether` (commit `f2221b8`)
- **WonderWorld:** `https://github.com/KovenYu/WonderWorld` (commit `5cf1146`)

Resolution and Frame Settings. We followed the official default inference settings of each baseline:

- **FantasyWorld (ours):** 336×592, 81 frames
- **Voyager:** 512×768, 49 frames
- **Uni3C:** 480×768, 81 frames
- **AETHER:** 480×720, 41 frames

- **WonderWorld:** 512×512, 50 frames

All baselines were run with the official inference scripts and default hyperparameters from the respective repositories.

A.6 CODE AVAILABILITY.

All our code, model design, training scripts, are available and fully reproducible at: `https://anonymous.4open.science/r/submit-F3040042/`