GeoWorld: A Geometry-First World Model for Reconstruction and Imagination

Baicheng Li[1,2,†] **Jialin Liu**[2,†] **Dong Wu**[1,†] **Weijian Xie**[2] **Zhichao Ye**[2]
Donghui Shen[2] **Nan Wang**[2] **Guofeng Zhang**[2,4,*] **Haomin Liu**[2] **Hongbin Zha**[1,3,*]

[1] State Key Laboratory of GAI, School of IST, Peking University, Beijing, China
[2] InSpatio
[3] School of Artificial Intelligence and Computer Science,
Anqing Normal University, Anqing, China
[4] State Key Lab of CAD&CG, Zhejiang University

† Equal contribution. * Corresponding authors.

Abstract

Large visual world models have become increasingly capable at synthesizing plausible views and camera trajectories, largely because they inherit strong priors from image and video diffusion. Yet in robotics, spatial computing, AR anchoring, and digital twins, a world is useful only if it can be measured and acted upon: appearance plausibility cannot substitute for geometrically coherent structure. This motivates a geometry-native alternative, where the generative state is the feature space of a geometry foundation model rather than RGB pixels or an appearance latent. Existing geometry-space world models validate this direction, but they still occupy a narrow operating range. When target views lie within observed support, full multi-view interaction limits trajectory length; when observations become sparse, geometry-only training lacks the broad visual priors needed to infer structure beyond observed support. We present **GeoWorld**, a geometry-latent world model that expands this range without returning generation to RGB space. GeoWorld builds a compact geometry latent, conditions it with camera ray geometry, and uses camera-aware tiered attention with relative pose bias to make long-view denoising practical. For sparse-observation extrapolation, a Bridge module translates hidden states from a frozen camera-conditioned video diffusion model into geometry-latent conditioning, borrowing video priors without making the video model the scene generator. Experiments show that GeoWorld scales geometry-space diffusion to long camera trajectories while improving geometric consistency in prior-assisted extrapolation.

1 Introduction

What should a 3D world model generate? Much recent progress begins from an appearance-centric answer: given a few observations and a camera path, generate views that look plausible and temporally coherent. This answer is natural for visual simulation and content creation, and it benefits enormously from the scale of image and video diffusion models. But many 3D applications ask for a stricter object. In digital twins, robot planning, AR anchoring, spatial measurement, and embodied simulation, a world model is useful only if the generated scene remains geometrically consistent and spatially measurable. A view can look correct while the surfaces behind it drift, bend, or disagree across cameras.

This distinction turns the generation space into a first-order design choice. Pixel-first systems use powerful visual priors and recover geometry after appearance has been generated. Geometry-aware RGB systems add depth, 3D heads, or geometric latents to a visual generator, improving spatial

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

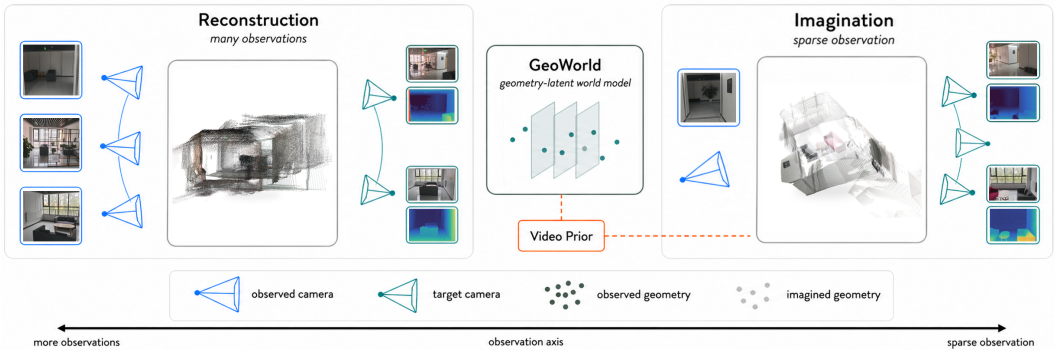

Figure 1: GeoWorld spans an observation axis. Posed observations support reconstruction; video-prior features help extrapolate unseen geometry under sparse observations.

consistency while still keeping appearance as the organizing space. A geometry-native route makes a different commitment: the denoising state itself lives in the feature space of a geometry foundation model. Its appeal is not merely that it can output depth or point maps, but that the quantity being generated is already geometric before any rendering or decoding step.

This route is promising, but not yet sufficient as a general world model. Existing geometry-space diffusion models tend to be strongest when target views remain within observed support. In that regime, they can preserve multi-view structure better than appearance-first pipelines; however, real scans and walkthroughs often contain dozens of target views, where full multi-view attention becomes expensive. At the other end of the observation spectrum, one or two views may leave large parts of the target camera path unsupported. In that regime the model needs a world prior, but geometry-side training data is far smaller than the video corpora that train modern generative models. Current systems therefore face an uncomfortable split: geometry preserves the right state, while video supplies the stronger imagination prior.

GeoWorld is designed around the idea that this split should be reduced rather than accepted. We keep generation in a compact geometry-foundation latent and expand its usable range along an observation axis. Near the within-support end, the central challenge is scalable aggregation: target views must exchange information without all-to-all attention. GeoWorld therefore builds a camera-aware communication graph, injects ray-based camera embeddings, and adds a relative pose bias so cross-view attention can distinguish a neighboring camera from a distant or oppositely facing one. Near the sparse-observation end, the challenge is prior: the model must infer structure that the conditioning views do not determine. GeoWorld reads hidden states from a frozen camera-conditioned video diffusion model and translates them into geometry-latent conditioning. The video model is used as feature memory, not as the generator of the final scene.

Figure 1 summarizes this axis from supported reconstruction to sparse-observation imagination.

The language of reconstruction and imagination follows from this axis. Reconstruction refers to the within-support side, where target views are largely explained by available evidence and the main issue is consistent trajectory completion. Imagination refers to the sparse-observation side, where target cameras leave the support of the observations and the model must infer plausible unseen structure. In experiments, we use the measurable term *beyond-cond extrapolation* for this latter regime.

Our contributions are:

- We present GeoWorld, a geometry-latent world model that uses geometry-foundation features as the generative state and spans both long-view trajectory completion and beyond-cond extrapolation.
- We make geometry-space diffusion scalable to dense camera trajectories through a compact latent state and camera-structured cross-view denoising, avoiding the cost of full multi-view interaction while retaining geometric communication.
- We connect frozen video priors to geometry-native generation by translating pre-decoding hidden states into geometry-latent conditioning and training the denoiser to use this prior selectively under sparse observations.

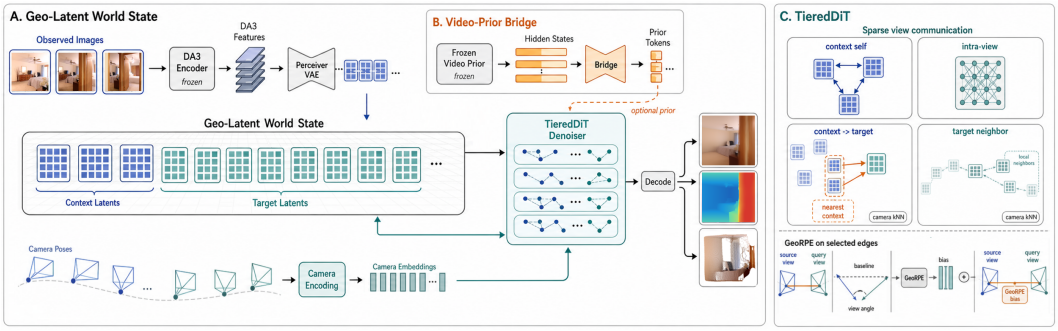

Figure 2: Method anatomy. GeoWorld denoises compact geometry latents with observed views, cameras, and optional Bridge features; TieredDiT and GeoRPE provide sparse camera-aware communication.

2 Related Work

Visual geometry models. Feed-forward visual geometry models have changed the substrate on which 3D generation can be built. DUSt3R [28] reframes stereo and multi-view reconstruction as pointmap regression; VGGT [26] broadens this idea into a unified model for cameras, depth, point maps, and tracks; and Depth Anything 3 [12] provides strong any-view geometry features at scale. Follow-up work extends this paradigm toward dense matching, sparse-view reconstruction, long image collections, persistent state, real-time video reconstruction, and dynamic scenes [11, 23, 25, 27, 34, 16, 38]. GeoWorld treats such geometry features as the generative state, asking how they can be compressed, denoised over long trajectories, and augmented with priors when observations become sparse.

Geometry-space world models. A smaller but increasingly relevant line turns geometry features from predictors into world states. VGGT-World [22] autoregressively forecasts frozen VGGT tokens, while GLD [8] repurposes geometric foundation features as the latent space for multi-view diffusion. Gen3R [7] sits near the sparse-observation side by combining generative video priors with feed-forward 3D reconstruction, and concurrent systems inject geometric supervision or 3D-aligned representations into video diffusion backbones [6, 30]. These methods expose the tradeoff behind our design: geometry-native states preserve explicit spatial structure, but existing systems are either expensive for long view sets or weak when the scene is poorly observed; video-centered models bring stronger imagination priors, but generation remains organized around appearance. GeoWorld keeps the denoising state geometry-native while treating video diffusion hidden states as optional prior memory.

View synthesis and sparse reconstruction. Classical novel-view synthesis and reconstruction optimize a scene representation from posed images. NeRF-style radiance fields and 3D Gaussian Splatting produce high-quality renderings when observations constrain the target views [18, 9], while feed-forward methods such as pixelSplat, MVSplat, DepthSplat, and Matrix3D learn sparse-view reconstruction without per-scene optimization [3, 4, 33, 17]. These methods are strongest when desired views are explainable from the input images; beyond observed support, they rely on interpolation, learned completion biases, or unstable/empty geometry rather than an explicit world prior. Camera-controlled generation expands this setting with pose-free splatting, diffusion regularization, geometric warping, and video priors [35, 32, 19, 20, 37, 31, 10, 21, 1, 2]. GeoWorld instead asks for a geometry-latent formulation spanning dense trajectory completion, where geometric consistency and scalability dominate, and beyond-cond extrapolation, where external priors must help without replacing the geometric world state.

3 Method

GeoWorld is a conditional diffusion model whose state is a compact geometry latent rather than pixels. The input is a set of observed posed images and a sequence of target cameras; the output is a sequence of target-view geometry latents that are decoded into geometry-aware scene predictions. The same interface covers both regimes studied in this paper: when target views lie within observed support, the model should aggregate view evidence without losing camera geometry; when observations are

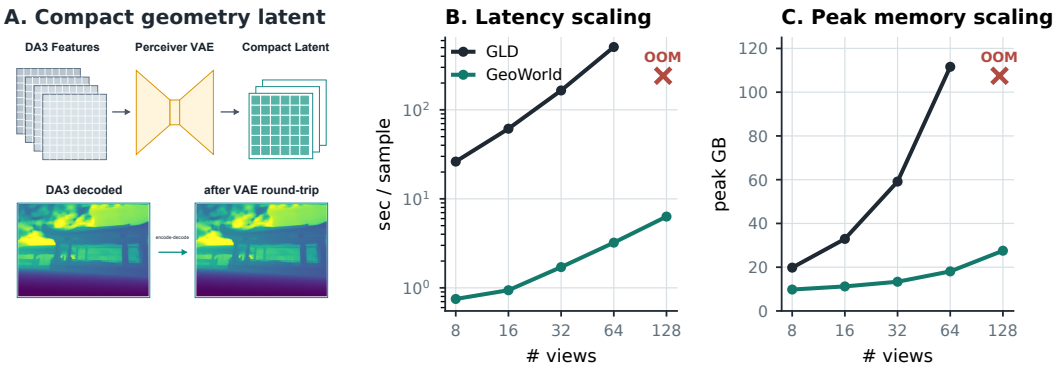

Figure 3: Compact representation and scalability. Left: geometry-feature compression. Right: latency and memory for full versus tiered communication.

sparse, it should be able to borrow a broader world prior without handing generation back to an RGB video model. Figure 2 summarizes the system.

3.1 Geometry-Latent World State

Let a trajectory contain observed views and target cameras. A frozen geometry foundation model, DA3 [12], provides the feature space for both conditioning and supervision: observed images are encoded into geometry features, and during training the target-view images supervise generated features in the same space. Unlike RGB diffusion or RGB-VAE latent diffusion, GeoWorld treats the generated object as a geometry feature before any rendering or RGB decoding step. This choice makes geometry consistency part of the denoising state itself: depth, point structure, and cross-view agreement are not post-hoc products of a generated appearance sequence.

The raw geometry features are too large for long trajectory diffusion, so GeoWorld introduces a learned bottleneck between the geometry foundation model and the denoiser. A Perceiver-style geometry VAE reads multi-level geometry tokens and maps each view into a compact spatial latent grid; a lightweight decoder maps generated latents back to the geometry feature space. The bottleneck is trained for high-fidelity feature reconstruction rather than for independent image generation, so distribution modeling remains the role of the downstream diffusion model. Its role is therefore representational: it preserves the geometry foundation model's usable information while reducing the per-view state enough for multi-view denoising.

Diffusion is trained with flow matching [14]. During training, target geometry latents are corrupted along a continuous noise path, and the denoiser predicts the velocity field used to integrate from noise to data. Its inputs are the noisy target latents, diffusion time, observed-view latents, target camera conditioning, and optional Bridge features. This gives reconstruction and imagination a common denoising interface along the observation axis. What changes is not the output space, but the evidence available to condition it: supported observations provide direct geometric evidence, while sparse observations require an additional prior.

3.2 Camera-Aware Denoising

Long-view geometry diffusion cannot rely on full view-to-view attention. An all-to-all view graph treats every camera as equally relevant to every other camera, even though the trajectory already tells us which views overlap, which views are nearby, and which directions are unlikely to exchange useful evidence. GeoWorld therefore turns camera geometry into the organizing principle of the denoiser.

First, every view is conditioned on its own camera through an absolute ray field derived from intrinsics and extrinsics. The ray field is embedded at the geometry-latent resolution and added to view tokens, so the denoiser sees each target as a posed geometric query rather than as an index in a sequence. Second, camera centers and viewing directions define a sparse communication graph. Observed views form a reference set; each target view selects nearby observed views for evidence and nearby target views for local trajectory consistency, using camera distance and viewing-direction agreement as the organizing cues. This gives a view graph that follows scene support rather than token order alone.

Our tiered denoiser implements this graph with four communication paths: condition self-attention, within-view self-attention, condition-to-target attention, and target-neighbor attention. For fixed neighbor budgets, cross-view communication grows approximately linearly with the number of views instead of quadratically. In complexity terms, replacing full view interaction with fixed-size condition and target neighborhoods changes the dominant multi-view term from $\mathcal{O}(V^2 N_t^2)$ to $\mathcal{O}(V N_t^2)$, where V is the number of views and N_t is the number of compressed latent tokens per view. Figure 3 visualizes the practical consequence in latency and memory.

Sparse connectivity decides *which* views exchange information, but it does not tell attention *how* the selected cameras are geometrically related. GeoWorld therefore adds a Geometric Relative Pose Encoding (GeoRPE) to the selected cross-view edges. GeoRPE maps a frame-pair pose descriptor into a per-head attention bias and applies it to condition-to-target and target-neighbor attention. This is intentionally a frame-level relation rather than an explicit epipolar router: after geometry-feature compression, reliable pixel-level correspondence is not guaranteed, but camera-level relative pose remains well defined. The result is a denoiser that is sparse for scalability yet still camera-aware in its cross-view reasoning.

3.3 Video-Prior Bridge

Sparse observations create a different failure mode. If target cameras leave the support of the input view, geometry evidence alone underdetermines the hidden scene. Large video diffusion models contain broad priors over layouts, surfaces, and plausible continuations, but decoding a video and then reconstructing geometry would route the world model through an RGB bottleneck. GeoWorld instead uses the video model as feature memory.

Bridge reads hidden states from a frozen camera-conditioned video diffusion model, WAN [24], and translates them into the common geometry-latent interface used by observed-view conditioning. The camera-conditioned prior follows the prescribed trajectory, while Bridge makes its hidden scene cues compatible with the denoiser without making the video model the scene generator. The distinction is important: the final output is still produced by geometry-latent diffusion, while the video prior supplies missing structure only as conditioning. This design avoids a decoded RGB detour in which video samples are first committed to pixels and only then reconstructed into geometry.

Bridge also needs a reliability mechanism. In sparse extrapolation, the video prior may be useful but imperfect: hidden states can contain plausible layout cues while still disagreeing with the commanded cameras or the observed geometry. Prior-Reliance Training exposes the denoiser to varying prior quality by degrading or removing prior features while keeping observed geometry clean. This asymmetric training encourages the geometry-latent generator to treat Bridge features as optional evidence, using them mainly where the observations do not determine the target geometry. Additional implementation and evaluation details are given in Appendix A.

4 Experiments

The experiments evaluate GeoWorld along the observation axis introduced above. We first test dense trajectory completion under a unified 64-view protocol, where target poses are known and geometric accuracy is the primary constraint. We then move to beyond-cond extrapolation, where a single image provides only partial support and the model must synthesize plausible newly exposed structure. Finally, we ablate the sparse camera-aware denoiser and the Bridge pathway to separate improvements from view communication, relative-pose bias, hidden-state video priors, and prior-reliance regularization.

4.1 Protocol

Trajectory completion. The main benchmark uses 64-view trajectories. Unless otherwise stated, each method receives the same $K{=}2$ endpoint conditioning views and predicts the remaining 62 target views at known target poses. Although K is small, target cameras mostly lie inside the span of the two observations; the task is therefore reconstruction/interpolation-like, completing a supported trajectory rather than inventing content outside observed coverage. We report RGB fidelity with PSNR, SSIM, and LPIPS, and geometry fidelity with depth AbsRel, RMSE, and δ_1 when depth or geometry is available. When an output depth is defined up to a global scale, we calibrate that

Table 1: Main trajectory-completion benchmark. Matched 64-view protocol on DL3DV-Evaluation [13] and RE10K [39]. Best/second-best are bolded/underlined per dataset.

Data	Method	Type	PSNR↑	SSIM↑	LPIPS↓	AbsRel↓	RMSE↓	δ_1 ↑
DL3DV	DepthSplat [33]	sparse recon.	12.39	0.409	0.478	0.214	2.838	0.741
	MVSplat [4]	sparse recon.	13.35	0.343	0.490	0.367	3.581	0.555
	Matrix3D [17]	feed-forward 3D	11.26	0.317	0.622	0.443	4.359	0.437
	MVGenMaster [2]	RGB-primary NVS	16.12	**0.457**	0.469	0.207	2.430	0.770
	GLD [8]	geometry-space diff.	**16.34**	0.456	0.482	0.204	2.391	0.773
	GeoWorld	geometry-space diff.	16.21	0.451	**0.465**	**0.193**	**2.274**	**0.784**
RE10K	DepthSplat [33]	sparse recon.	14.30	0.654	0.239	0.114	0.826	0.879
	MVSplat [4]	sparse recon.	21.52	0.724	**0.142**	0.108	0.812	0.907
	Matrix3D [17]	feed-forward 3D	20.66	0.714	0.206	0.102	0.724	0.900
	MVGenMaster [2]	RGB-primary NVS	21.42	0.715	0.147	0.105	0.831	0.909
	GLD [8]	geometry-space diff.	22.61	**0.763**	0.192	0.093	0.690	0.914
	GeoWorld	geometry-space diff.	**22.64**	0.760	0.183	**0.086**	**0.653**	**0.922**

scale once from the conditioning view(s) and keep it fixed over the target trajectory. Appendix B summarizes the full protocol and baseline adaptations.

Beyond-cond extrapolation. For imagination, each method receives a single conditioning image and follows the same five prescribed trajectories, which deliberately leave the input view's support. Since many unseen continuations can be plausible, we do not score against one ground-truth RGB future. Instead, we evaluate input consistency, adjacent-view reprojection consistency, and newly exposed regions. Depth errors measure local geometry under the same trajectory-level convention, Inlier measures the fraction of valid reprojections below an error threshold, Coverage reports how much adjacent-view support is available, and novel-region appearance scores expose collapse into low-frequency content. Appendix B.1 gives the exact definitions.

Baselines. The direct geometry-space baseline is GLD [8]. We also compare to sparse-view reconstruction methods such as MVSplat and DepthSplat [4, 33], a camera-controlled feed-forward variant based on Matrix3D [17], and MVGenMaster [2] as an RGB-primary fixed-pose reference. Matrix3D is evaluated with an 8-view sliding-window extension to cover the 64-view trajectory. Gen3R [7] is reserved for the beyond-cond setting because its native protocol starts from a single image and assumes a continuous camera trajectory. For the beyond-cond study, we additionally evaluate a camera-conditioned WAN [24] video model under the same commanded trajectories and GeoWorld without Bridge, which isolates the contribution of hidden-state prior conditioning.

4.2 Dense Reconstruction

Table 1 tests the reconstruction side on both object/scene-centric DL3DV trajectories and real-estate RE10K trajectories. The two endpoint observations provide direct support for most target views, so the comparison emphasizes long trajectory completion under known cameras rather than one-image extrapolation. This evaluation separates two requirements: retaining the geometric accuracy associated with geometry-space diffusion, and operating at trajectory length rather than the small view counts favored by full multi-view attention. GLD, MVGenMaster, and MVSplat remain competitive on particular appearance metrics under their native biases. GeoWorld reports the strongest geometry scores in this table while staying close on RGB fidelity, and uses the scalable denoising path analyzed below.

Figure 4 complements the table by visualizing long-range appearance and geometry consistency. The qualitative examples show the same pattern as the metrics: sparse reconstruction and RGB-primary references can preserve local texture, but their geometry becomes less stable as the requested trajectory extends. GeoWorld keeps surfaces more coherent across distant targets while avoiding the washed-out appearance often produced by conservative completion.

Table 2 separates practical inference scaling from reconstruction quality. It compares the full multi-view communication used by GLD against the camera-aware sparse communication pattern used by GeoWorld as the number of views grows. Together with Figure 3, it measures whether geometry-space diffusion remains practical at trajectory lengths where quadratic view attention becomes costly.

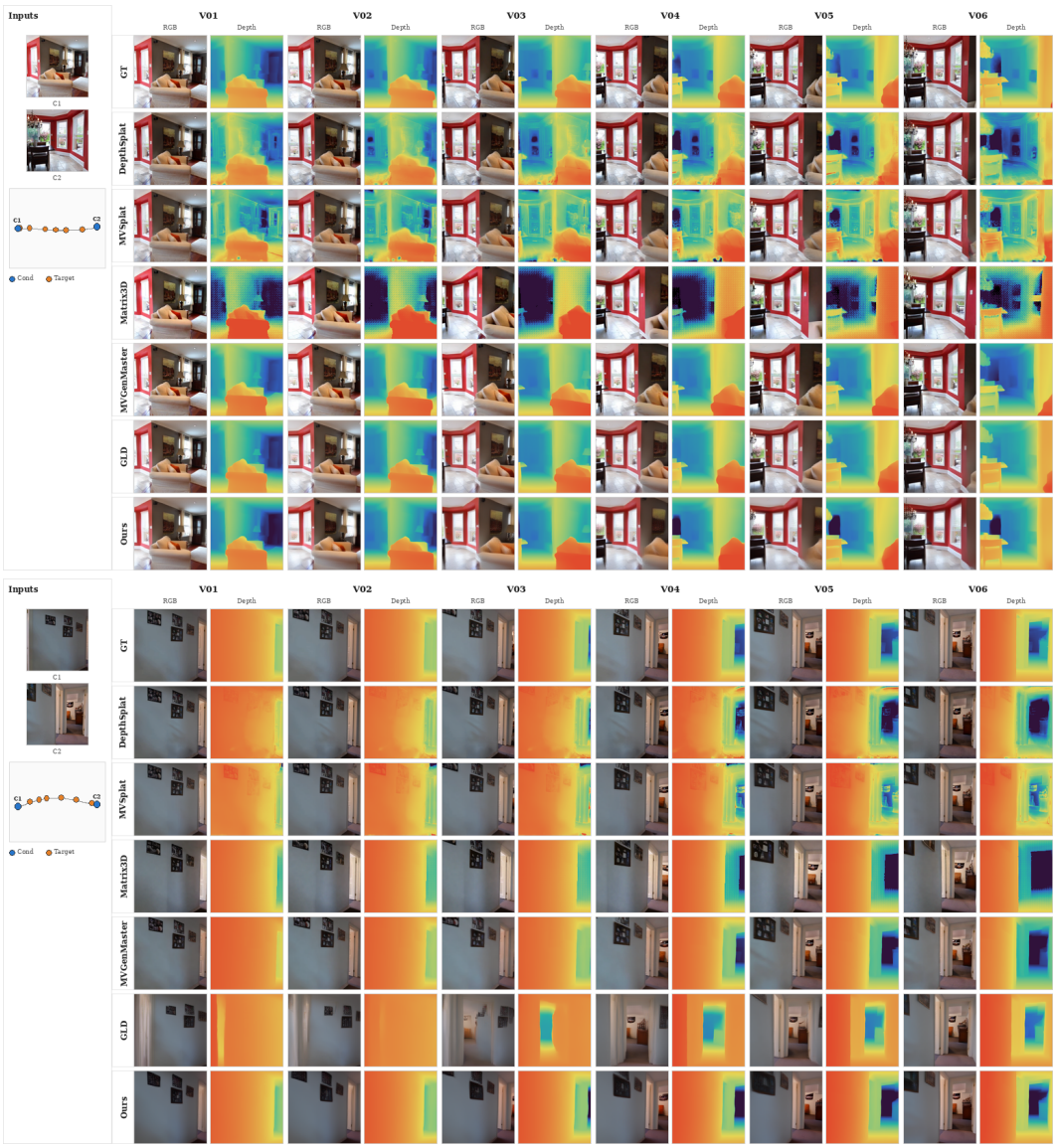

Figure 4: Dense 64-view reconstruction qualitative comparison. All methods use the same endpoint views and target cameras.

Table 2: Scaling and efficiency. Time (s) and peak memory (GB) on one 140GB H200 at 504×378 resolution. View counts include two endpoint conditions.

Method	8 views		16 views		32 views		64 views		128 views	
	Time↓	Mem.↓	Time↓	Mem.↓	Time↓	Mem.↓	Time↓	Mem.↓	Time↓	Mem.↓
GLD	26.22	19.80	61.37	32.93	165.10	59.14	508.12	111.60	OOM	OOM
GeoWorld	**0.75**	**9.79**	**0.94**	**11.21**	**1.71**	**13.35**	**3.22**	**18.08**	**6.32**	**27.53**

4.3 Beyond-Cond Extrapolation

The second regime asks whether the model can infer plausible structure when target cameras leave observed support. We compare GLD, Gen3R, decoded camera-conditioned WAN video followed by the same geometry evaluator, GeoWorld without Bridge, and GeoWorld with Bridge. This grouping separates geometry-only extrapolation, RGB-primary generation, and our hidden-state Bridge that imports video priors while keeping the generator in geometry space.

Table 3 shows that Bridge improves the extrapolation tradeoff where geometry-only generation is under-constrained. Without Bridge, GeoWorld behaves similarly to GLD: it remains consistent

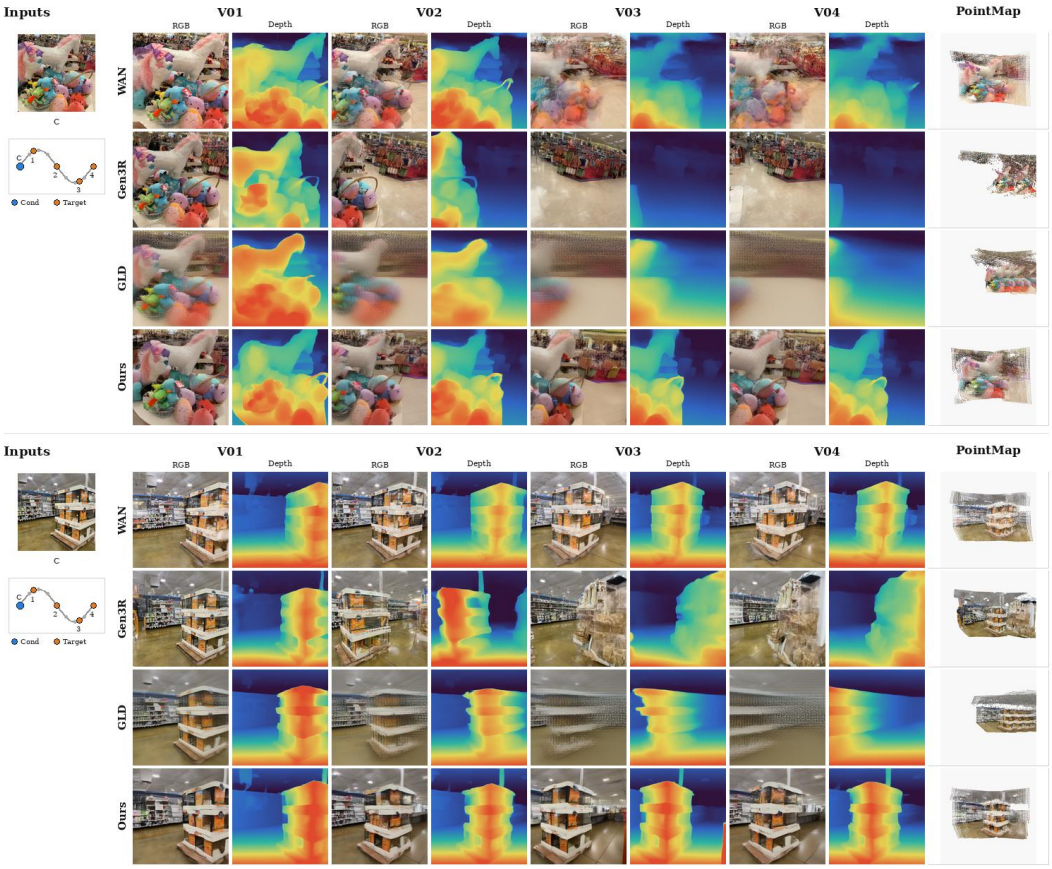

Figure 5: Beyond-cond qualitative comparison. Rows compare geometry-space baselines, appearance-primary references, and GeoWorld with Bridge across target distance.

Table 3: Beyond-cond extrapolation. Metrics cover adjacent geometry, reprojection support, and seen/novel appearance over 900 cases. Best/second-best directional metrics are bolded/underlined; Coverage is diagnostic.

	Geometry			Appearance		
Method	AbsRel↓	Inlier↑	Coverage	RGB-L1 (Seen)↓	RGB-L1 (Novel)↓	RMS Contrast (Novel)↑
GLD [8]	0.1545	0.8166	0.6400	0.2172	0.1536	0.5726
Gen3R [7]	<u>0.1323</u>	<u>0.8550</u>	0.7134	0.2217	0.1546	**0.8306**
Camera-cond. WAN [24]	0.1396	0.8428	0.6982	**0.1964**	**0.1187**	<u>0.8124</u>
GeoWorld w/o Bridge	0.1512	0.8194	0.6356	0.2208	0.1518	0.5903
GeoWorld	**0.1268**	**0.8667**	0.7285	<u>0.2039</u>	<u>0.1275</u>	0.8046

near observed support, but its lower reprojection support and muted novel-region contrast relative to prior-assisted variants indicate conservative extrapolation. With Bridge, GeoWorld improves adjacent-view geometry and support while also producing richer novel regions, suggesting that the WAN prior adds usable structure rather than only sharper texture.

The camera-conditioned WAN baseline confirms that a strong video prior is already useful in this regime: it gives the best RGB-L1 scores and high novel-region contrast. However, decoded video is still less camera-faithful than the geometry-latent generator, and its geometry scores trail both Gen3R and GeoWorld with Bridge. GeoWorld therefore does not simply beat WAN on color; instead, it uses WAN hidden states to improve imagination while retaining stronger geometric consistency. Gen3R remains a strong generative reconstruction reference with the highest novel-region contrast, but GeoWorld obtains the best geometry scores while staying close to WAN on appearance.

4.4 Ablations and Analysis

The ablations separate the two sides of GeoWorld rather than forcing them into one mixed metric. On the reconstruction side, we keep the 64-view, $K=2$ protocol fixed and compare sparse communication

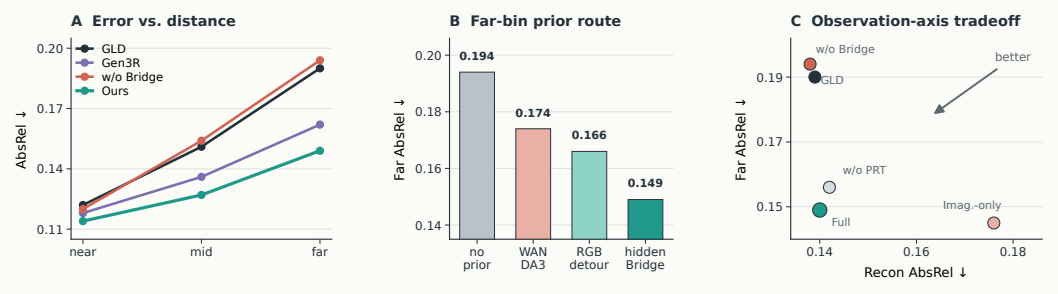

Figure 6: Analysis of extrapolation and design choices. Left: distance-binned geometry error. Middle: far-bin prior route. Right: a diagnostic reconstruction-imagination plane illustrating the operating bias of each variant.

Table 4: Ablation studies. Reconstruction uses the 64-view protocol; imagination uses far-bin beyond-cond metrics. Best directional values are bolded.

(a) Reconstruction

Variant	PSNR↑	LPIPS↓	AbsRel↓	δ_1 ↑
Full GeoWorld	**19.43**	**0.324**	**0.140**	**0.853**
Random sparse graph	18.62	0.361	0.176	0.811
Trajectory-local sparse graph	19.30	0.329	0.145	0.847
w/o GeoRPE	19.14	0.336	0.151	0.838

(b) Imagination

Variant	Far AbsRel↓	Far Inlier↑	Cov.	Seen L1↓	Novel C.↑
Full GeoWorld	**0.149**	**0.842**	0.689	0.214	0.756
w/o Bridge	0.194	0.774	0.574	0.231	0.501
WAN→DA3 baseline	0.174	0.801	0.625	0.210	0.709
RGB-detour prior	0.166	0.817	0.651	**0.207**	0.724
w/o PRT	0.156	0.835	0.676	0.216	**0.771**

patterns with the same view-neighbor budget. This asks whether the proposed camera-aware graph and GeoRPE are necessary, or whether random or trajectory-local sparse attention is sufficient. On the imagination side, we keep the single-image beyond-cond protocol fixed and isolate how video priors enter the geometry denoiser. In particular, we distinguish a decoded WAN→DA3 baseline, which never uses the GeoWorld denoiser, from an RGB-detour prior, which still uses GeoWorld but feeds it DA3–VAE latents re-encoded from decoded WAN frames. Figure 6 includes an imagination-only configuration biased toward sparse-observation extrapolation as an analysis endpoint on this observation axis, showing why a design that can absorb external priors without drifting away from reconstruction support is needed.

Table 4a ablates the reconstruction-side communication design. The random and trajectory-local sparse graphs control for the possibility that a reduced attention budget can rely on generic sparsification or sequence-order neighborhoods alone; GeoRPE then tests whether camera-aware edges also benefit from relative-pose modulation. The full design remains best across metrics, indicating that sparse edge selection and relative-pose bias contribute complementary gains.

Table 4b focuses on how the video prior enters the geometry denoiser. The WAN→DA3 row measures decoded video as an external prior baseline. The RGB-detour row is stricter: it gives the same GeoWorld denoiser access to DA3–VAE latents re-encoded from decoded WAN frames. The comparison to hidden-state Bridge tests whether the prior is more useful before it is committed to pixels, while the PRT ablation checks whether degrading Bridge features during training improves far-view reliability. RGB-detour and w/o PRT can improve individual appearance diagnostics, but they give up some far-view inlier ratio and coverage; GeoWorld with Bridge therefore favors usable geometric support over maximizing one texture statistic.

5 Conclusion

GeoWorld addresses a representational tension in 3D world modeling: the strongest generative priors often live in video models, while geometry-critical applications require the generated scene to remain native to geometry. The proposed system keeps denoising in a compact geometry latent, conditions it with camera ray geometry and relative pose bias, scales it to dense trajectories with camera-structured denoising, and uses Bridge to import frozen video-diffusion hidden states only as auxiliary conditioning for sparse-observation extrapolation. This separates the role of the video model from the role of the generator. Rather than replacing geometry-space generation with appearance generation, GeoWorld shows how the operating range of a geometry-native model can be expanded toward both long-view reconstruction and prior-assisted imagination.

Acknowledgments and Disclosure of Funding

We thank the reviewers and area chairs for their constructive feedback.

This work is partially supported by NSFC (U22A2061) and 230601GP0004.

Disclosure: Jialin Liu, Weijian Xie, Zhichao Ye, Donghui Shen, Nan Wang, Guofeng Zhang, and Haomin Liu are affiliated with InSpatio.

References

[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

[2] Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. MVGen-Master: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *CVPR*, pages 6045–6056, 2025.

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pages 19457–19467, 2024.

[4] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3d gaussian splatting from sparse multi-view images. In *ECCV*, pages 370–386, 2024.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

[6] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. FantasyWorld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

[7] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3R: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026.

[8] Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. *arXiv preprint arXiv:2603.22275*, 2026.

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

[10] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *arXiv preprint arXiv:2405.17414*, 2024.

[11] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *ECCV*, 2024.

[12] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[13] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024.

[14] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[15] 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 *ICCV*, 2021.

[16] 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. *arXiv preprint arXiv:2412.09401*, 2024.

[17] 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 *CVPR*, 2025. arXiv:2502.07685.

[18] 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. In *ECCV*, 2020.

[19] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. arXiv:2310.17994.

[20] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. GenWarp: Single image to novel views with semantic-preserving generative warping. In *NeurIPS*, 2024. arXiv:2405.17251.

[21] 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.

[22] Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. VGGT-World: Transforming VGGT into an autoregressive geometry world model. *arXiv preprint arXiv:2603.12655*, 2026.

[23] 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. *arXiv preprint arXiv:2412.06974*, 2024.

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

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

[26] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025.

[27] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025.

[28] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, pages 20697–20709, 2024.

[29] 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 *IROS*, 2020.

[30] 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. *arXiv preprint arXiv:2507.07982*, 2025.

[31] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024.

[32] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. ReconFusion: 3d reconstruction with diffusion priors. In *CVPR*, 2024. arXiv:2312.02981.

[33] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024.

[34] 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.

[35] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025. Oral.

[36] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

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

[38] 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. In *ICLR*, 2025.

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

A Additional Method Details

This supplement provides implementation and evaluation details that do not fit in the main paper. The goal is to make the long-view protocol and baseline adaptation unambiguous while keeping the main text focused on the core design.

Training recipe. GeoWorld is trained in stages with the DA3 encoder and WAN video model kept frozen. We first train the geometry VAE for 100K optimization steps on multi-view data from ScanNet [5], ScanNet++ [36], DL3DV [13], RE10K [39], TartanAir [29], and ACID/ACID-Large [15], using reconstruction losses on decoded DA3 feature levels. The geometry-latent denoiser is then trained with flow matching for 300K steps using AdamW, a base learning rate of 2×10^{-5}, weight decay 0.01, gradient clipping at 1.0, and EMA decay 0.9995. Bridge is initialized from WAN-hidden-to-geometry-latent pretraining and jointly tuned with the denoiser under Prior-Reliance Training using a lower learning rate of 5×10^{-6}. The geometry VAE uses AdamW with learning rate 1×10^{-4} and cosine decay. All reported results use EMA weights.

Geometry-latent state. GeoWorld uses a frozen DA3 [12] encoder as the source geometry representation. For each view, DA3 features from four levels are concatenated into 1296 spatial tokens with 6144 channels and compressed by a Perceiver-style geometry VAE into 256 tokens of width 256, arranged as a 16×16 latent grid for the denoiser. The VAE is trained as a high-fidelity feature autoencoder; the reconstruction loss is applied to the decoded DA3 feature levels, with larger weight on the finest level used by the renderer. Distribution modeling is left to the downstream flow-matching denoiser. A common latent format is used for target generation and reference conditioning, which lets observed frames, decoded priors, and hidden-state Bridge features enter the same geometry-latent interface.

This joint multi-level bottleneck differs from GLD's feature-level strategy [8], which reduces the diffusion target by operating on selected DA3 feature levels and relying on frozen DA3 components with cascaded decoding to recover the remaining representation. GeoWorld instead autoencodes the four DA3 feature levels into one compact latent state and decodes that state back to the full feature set. This keeps generation in a single denoising interface while avoiding a level-wise generation schedule at inference.

Table S1: Geometry-VAE latent budget. Compression is the deterministic representation-size ratio against the pre-VAE DA3 feature tensor described above, where larger values mean smaller latent states. VAE recovery measures DA3 feature round-trip fidelity under the same training schedule for all budgets.

Latent grid	Tokens	Latent scalars/view	Compression vs. raw DA3	VAE recovery↑
8×8	64	16,384	486.0×	90.7%
8×16	128	32,768	243.0×	94.9%
16×16	256	65,536	121.5×	97.0%
16×32	512	131,072	60.8×	97.2%

Table S1 reports the latent-budget sweep used to choose the default 16×16 representation. All variants are trained with the same VAE training schedule, so the comparison isolates the effect of latent capacity. Recovery denotes normalized DA3 feature round-trip fidelity. It improves substantially from 64 to 128 and from 128 to 256 tokens, but saturates from 256 to 512 tokens while the per-view diffusion state doubles. We therefore use the 256-token setting as the default balance between round-trip fidelity and multi-view denoising cost.

Flow-matching parameterization. Let z_{data} denote a clean geometry latent and $z_{\mathrm{noise}} \sim \mathcal{N}(0, I)$ denote the initial noise. We use a linear interpolation path between data and noise and train the denoiser to predict the corresponding flow-matching velocity. During sampling, we integrate the probability-flow ODE with the standard deterministic solver used in our reported experiments. In reconstruction mode, condition-view latents are provided through the clean reference channel while target-view latents are denoised; in Bridge mode, the same reference channel can be filled by prior features for unobserved target slots.

Camera-aware sparse communication. For every view, intrinsics and extrinsics are converted into a Plucker ray field at the image plane and embedded at the latent-grid resolution. Camera centers and principal directions are extracted from the ray field and used to construct a sparse view graph. In the reported model, each target view selects up to eight nearest condition views and eight nearest target-neighbor views under a scale-normalized camera distance that averages translation distance and viewing-direction angle. Condition views attend densely to one another, each target view attends to the selected condition views, and target views exchange information only through this local target-neighbor graph. For fixed neighbor counts, this changes the dominant cross-view communication cost from quadratic in the number of views to approximately linear in the trajectory length, while preserving dense spatial attention within each view.

GeoRPE. Sparse view selection determines which frames can communicate, but it does not encode how the selected cameras are geometrically related. GeoRPE adds a learned frame-pair bias to condition-to-target and target-neighbor attention. For an ordered query-source frame pair, the descriptor includes normalized baseline magnitude, viewing-direction angle, direction dot product, baseline projections onto the query and source viewing directions, and two signed normalized translation components. A small MLP maps this descriptor to per-head additive attention biases, initialized at zero so that enabling GeoRPE starts from the original sparse denoiser. This keeps relative pose visible without introducing pixel-level epipolar routing, which is ill-defined after Perceiver compression.

Bridge and Prior-Reliance Training. Bridge uses a frozen camera-conditioned WAN video diffusion model as feature memory rather than as a direct RGB generator. The WAN branch receives the same input image and commanded target cameras used by the beyond-cond protocol, and it does not receive target RGB frames; its hidden states are therefore tied to the prescribed trajectory rather than sampled as an unconstrained video. We read hidden states from three WAN transformer layers (5, 20, and 35), concatenate them, and feed them to a Perceiver Bridge with 256 learned output queries. The Bridge output follows the same channel-first latent-map format as the geometry VAE output, with 256 channels on a 16×16 grid, and can be inserted into the reference-conditioning channel of the denoiser. The Bridge is timestep-conditioned by the WAN noise level; in the main prior-assisted evaluations we use a fixed prior state at noise level 0.5. Before joint diffusion training, Bridge is aligned to the DA3–VAE latent space so that its outputs are readable by the geometry denoiser.

Prior-Reliance Training prevents the model from treating the external prior as ground truth in sparse extrapolation. During joint training, WAN hidden states are extracted at randomly sampled noise levels, translated by Bridge, and inserted only into target/reference slots; observed geometry remains clean. With probability 0.5, target prior features are corrupted independently across target slots. The corruption samples among additive Gaussian feature noise, spatial smoothing, channel dropout, full zeroing, and an unchanged branch. Training also mixes the number of observed views, so the same denoiser sees single-image extrapolation and more supported reconstruction episodes. This exposes the model to a continuum from reliable to absent prior information, allowing sparse extrapolation to use scene-level video priors while remaining anchored to observed geometry.

B Evaluation Protocol and Baseline Adaptation

The dense reconstruction benchmark uses held-out DL3DV [13] and RE10K [39] evaluation episodes. Each episode contains a 64-view trajectory, two conditioning views, and 62 target views with known target cameras. All methods receive the same conditioning images and the same target camera sequence. RGB metrics are computed on target views after resizing to the common evaluation resolution. For geometry-latent outputs, we follow GLD [8] and use its MAE-style color decoder for RGB evaluation. Geometry metrics use each method's comparable depth or geometry output when available, with a shared reconstruction fallback used only for adapted RGB-only pipelines. When an output depth is defined up to a global scale, we estimate that scale from the conditioning view(s) using the evaluation depth and apply it unchanged to all target views in the episode before computing AbsRel, RMSE, and δ_1. Figure S1 provides additional qualitative examples under this matched dense protocol.

The matched protocol adapts baselines to this common long-trajectory, known-camera setting rather than claiming to reproduce each method's native evaluation. GLD is run as the direct geometry-space

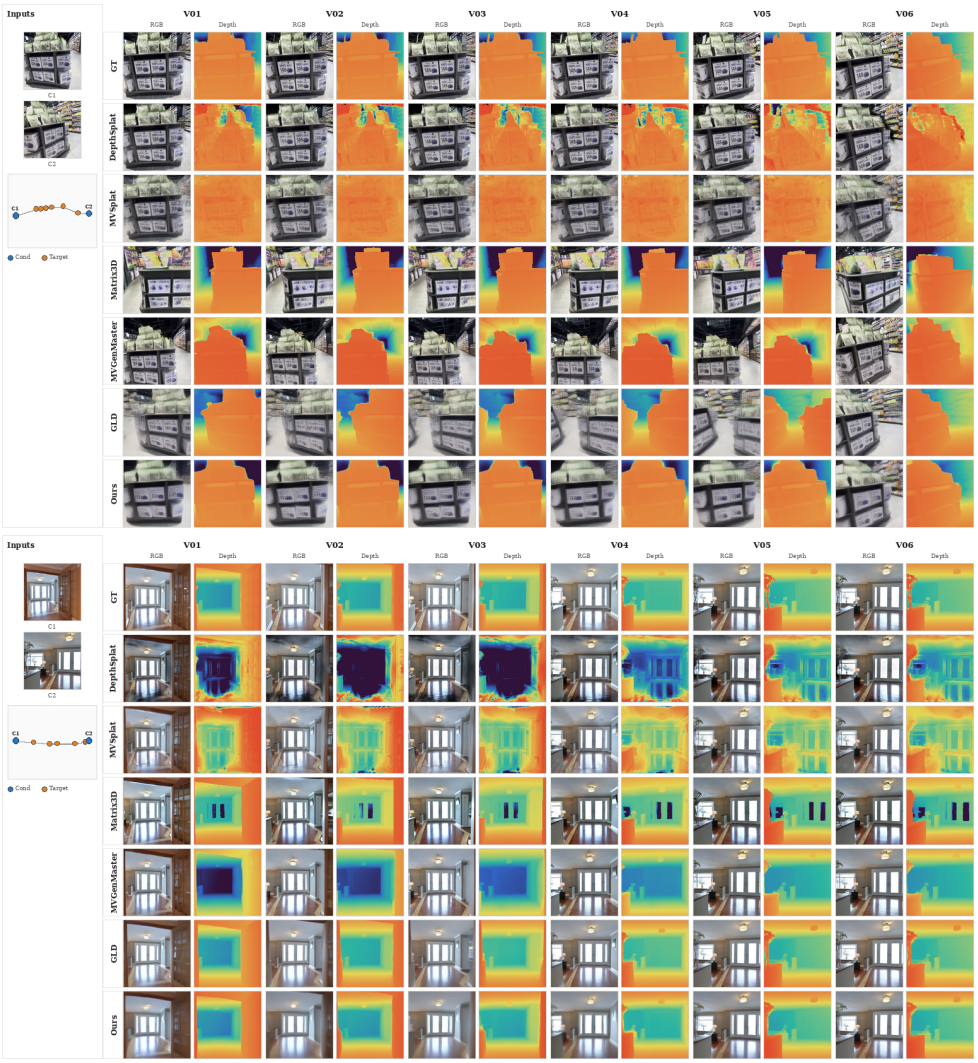

Figure S1: Additional dense reconstruction examples. All methods receive the same two conditioning views and the same 64-view target camera trajectory.

diffusion baseline with the same endpoint conditions and target cameras. DepthSplat and MVSplat are sparse-view reconstruction systems, so they receive the two conditioning views and render the requested target cameras from their predicted scene representation. Matrix3D is a feed-forward 3D predictor over a limited view set; we therefore use an 8-view sliding-window extension to cover the 64-view trajectory. MVGenMaster is treated as an RGB-primary fixed-pose novel-view reference. Gen3R is used only for beyond-cond extrapolation, where its single-image trajectory-generation setting matches the protocol more naturally than the dense $K{=}2$ benchmark. WAN is evaluated as a camera-conditioned video prior/reference in the beyond-cond study rather than as a native geometry generator.

Beyond-cond evaluation. The beyond-cond setting starts from one input image and follows prescribed camera trajectories that move beyond the observed support. Because a single ground-truth continuation is not the only plausible future, the evaluation focuses on camera-conditioned consistency rather than direct future-frame PSNR. Adjacent generated views are checked by depth-based reprojection under the commanded cameras. Seen and novel regions are separated according to whether pixels project back to the input view. This protocol penalizes videos that are self-consistent but ignore the input condition or commanded camera path.

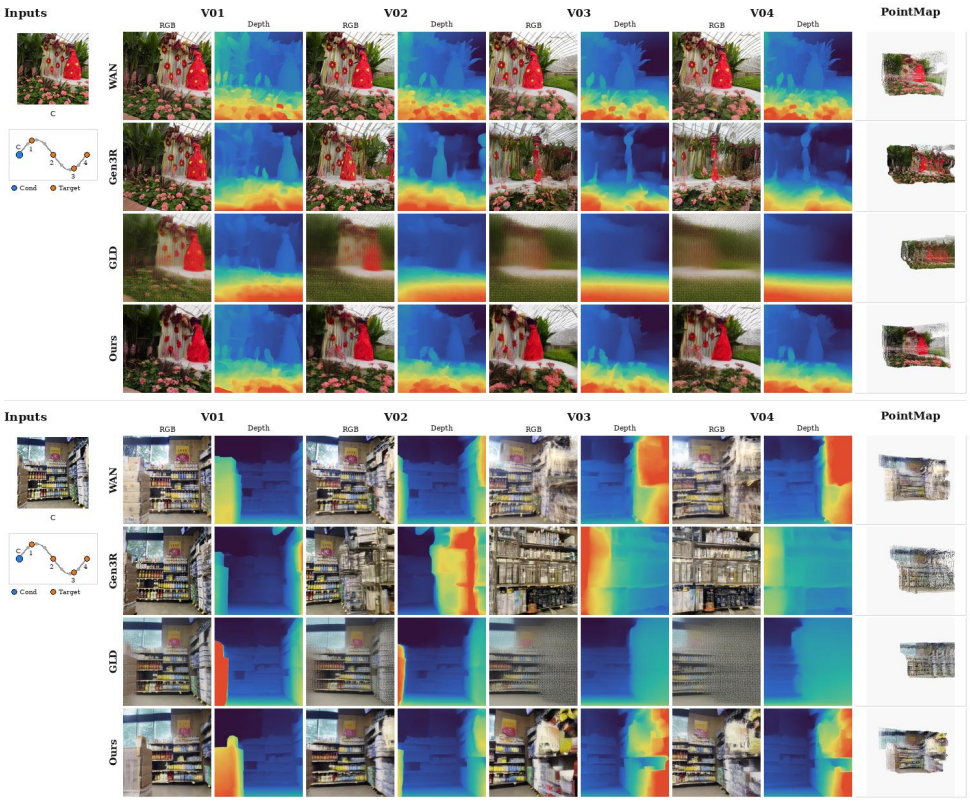

Figure S2: Additional beyond-cond qualitative examples. The layout follows Figure 5 with one conditioning image, representative commanded views, and fused point maps.

B.1 Beyond-Cond Metrics

For adjacent generated views, we sample a regular pixel grid with stride 12, back-project sampled pixels with the predicted depth in view i, transform them with the known relative camera, and reproject them into view $i{+}1$. *AbsRel* is the relative depth discrepancy on valid adjacent reprojections after the depth protocol described above. *Inlier* is the fraction of valid reprojections whose geometry error is below 0.25 under the same convention. *Coverage* is the fraction of sampled source points that yield a valid adjacent-view reprojection with valid target depth; it is a support diagnostic, not a monotonic quality score by itself. When reporting near, mid, and far analysis bins, we partition target views by their normalized camera-path distance from the conditioning view and use the same bin boundaries for every method.

For appearance, *RGB-L1 (Seen)* is computed on target pixels whose projected 3D points fall inside the conditioning image, and therefore checks input consistency. *RGB-L1 (Novel)* is computed on sampled target pixels whose projection falls outside the conditioning image and measures whether newly exposed content remains locally compatible across adjacent generated views. *RMS Contrast (Novel)* reports the root-mean-square contrast in novel regions, which helps identify methods that achieve low L1 by producing overly smooth extrapolations. These metrics should be read jointly: high contrast without geometry consistency can indicate hallucination, while low error with low coverage can indicate conservative generation. Figure S2 shows additional beyond-cond examples for this protocol.

C Supplementary Experiment Tables

Tables S2–S4 provide supplementary checks for conditioning robustness, long-sequence behavior, and the prior route used by Bridge. The reconstruction diagnostics in Tables S2 and S3 follow the same DL3DV 64-view protocol used in the main benchmark and are intended to analyze trends under controlled changes rather than to introduce a separate leaderboard. Table S4 isolates whether WAN is

Table S2: Conditioning-view sensitivity. The same DL3DV 64-view protocol is evaluated while varying the number of observed views; depth metrics follow the main benchmark protocol.

K condition views	PSNR↑	SSIM↑	LPIPS↓	AbsRel↓	RMSE↓	δ_1 ↑
1	14.36	0.421	0.552	0.281	2.754	0.653
2	16.21	0.451	0.465	0.193	2.274	0.784
4	17.12	0.474	0.421	0.158	1.821	0.834
8	17.34	0.480	0.408	0.151	1.746	0.842

Table S3: Long-sequence generalization. The DL3DV diagnostic protocol is evaluated while increasing the total trajectory length under the same $K{=}2$ endpoint-conditioning setup. Time and peak reserved memory are measured on one 140GB H200 at $504{\times}378$ resolution with GeoWorld's standard sampler; quality metrics follow the main benchmark protocol.

Total views	Time↓	Mem.↓	PSNR↑	SSIM↑	LPIPS↓	AbsRel↓	RMSE↓	δ_1 ↑
32	1.71	13.35	17.31	0.472	0.417	0.151	1.766	0.837
64	3.22	18.08	16.21	0.451	0.465	0.193	2.274	0.784
96	4.76	22.79	15.98	0.446	0.482	0.209	2.326	0.765
128	6.32	27.53	15.72	0.431	0.514	0.222	2.401	0.742

Table S4: Prior route ablation. Metrics use the full beyond-cond aggregation. The RGB-detour variant decodes WAN into RGB frames, re-encodes those frames through the DA3–VAE path, and inserts the resulting latents into the reference-conditioning slot.

Prior route	AbsRel↓	Inlier↑	Coverage	RGB-L1 Seen↓	RGB-L1 Novel↓	Novel Contrast↑
No prior	0.1512	0.8194	0.6356	0.2208	0.1518	0.5903
WAN RGB → DA3 → VAE	0.1349	0.8541	0.7102	0.1988	0.1229	0.7817
Hidden-state Bridge	0.1268	0.8667	0.7285	0.2039	0.1275	0.8046

most useful as decoded RGB that is re-encoded into the geometry-latent space, or as hidden-state features consumed directly by Bridge.

The RGB-detour comparison is intended to separate two ways of using the video prior. If decoded WAN frames are sufficient, then re-encoding them through the same DA3–VAE path used for image observations should provide a competitive reference prior. The detour is indeed strong on RGB-L1, but its lower geometry consistency and novel-region contrast relative to hidden-state Bridge support the interpretation that direct hidden-state conditioning provides prior features that are better matched to the geometry-latent denoising process. The main ablation table reports far-bin diagnostics, so its values should be read as the harder extrapolation subset rather than as duplicates of this full-trajectory aggregation.

D Limitations

GeoWorld assumes that target cameras are specified. This is natural for posed reconstruction, novel-view synthesis, and prescribed camera-path simulation, but the current system does not choose camera trajectories or optimize actions by itself. Extending geometry-latent generation toward camera planning or interactive control is a promising direction.

The present denoiser also operates on a requested set of views with bidirectional communication. This is appropriate for offline trajectory completion and extrapolation analysis, where the camera path is known in advance. For online applications, a streaming or causal geometry-latent variant that updates the world state as new observations arrive would be useful.

Finally, beyond-cond extrapolation is inherently multi-modal. A generated continuation can be plausible even when it differs from a held-out capture, so we evaluate input consistency, reprojection consistency, and novel-region support instead of relying only on direct RGB matching. Bridge improves this regime by importing video priors into the geometry-latent denoiser, but it can still

choose one plausible continuation among many when the input image does not determine the hidden scene.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The main claims are supported by the method design and empirical evaluations.

2. **Limitations**

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

   Answer: [Yes].

   Justification: Limitations and scope are discussed in the appendix.

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

   Justification: This paper focuses on methodology design and empirical evaluation and does not introduce new theoretical proofs.

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The method, experimental protocol, implementation details, and evaluation metrics are described in the main paper and appendix.

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

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

   Answer: [No].

   Justification: Code is not included with this submission. The paper describes the method and evaluation protocols in the main text and appendix.

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: Training and evaluation settings are specified in the main paper and appendix.

7. **Experiment statistical significance**

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

   Answer: [No].

   Justification: The paper reports aggregate evaluation metrics, but does not report error bars, confidence intervals, or formal statistical significance tests.

8. **Experiments compute resources**

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

   Answer: [No].

Justification: Hardware, memory, and runtime are reported for the scalability experiments, and the appendix gives training schedules. Full training compute costs and resource details for every experiment are not reported.

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The work uses existing research datasets and does not involve human-subject experiments or deployed decision systems.

10. **Broader impacts**

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

    Answer: [No].

    Justification: The paper motivates applications in robotics, spatial computing, and digital twins, but does not include a dedicated discussion of potential negative societal impacts.

11. **Safeguards**

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

    Answer: [N/A].

    Justification: This paper does not release any models or datasets that possess a high risk for misuse.

12. **Licenses for existing assets**

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

    Answer: [No].

    Justification: Existing datasets, baselines, and pretrained models are cited, but the paper does not enumerate the license and terms of use of every asset.

13. **New assets**

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

    Answer: [N/A].

    Justification: This paper does not introduce or release new datasets or standalone assets in the current submission.

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

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

    Answer: [N/A].

    Justification: This research relies on existing datasets and does not involve crowdsourcing or human subjects.

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

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

    Answer: [N/A].

Justification: This paper does not involve crowdsourcing or human subjects, so IRB approval is not applicable.

16. **Declaration of LLM usage**

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

    Answer: [N/A].

    Justification: LLMs are not used as an important, original, or non-standard component of the core method.