WORLD-GROUNDED CAMERA TRAJECTORY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Driven by advancements in 3D world reconstruction, the demand for world-aware video generation and navigation is rapidly growing, where generating controllable camera trajectories is the prerequisite for ensuring 3D consistency. While existing methods show promise, they heavily rely on character-driven motions or single RGB-D images, leaving text-driven trajectory generation in unconstrained 3D point clouds largely unexplored. In this paper, we present **WorldDirector**, a novel framework that generates world-grounded camera trajectories directly from 3D point clouds and text prompts. To achieve precise spatial grounding, we propose camera-aware world tokens that explicitly encode the observer's initial perspective by concatenating point-wise geometry features with 3D coordinates transformed into the starting camera's local coordinate system. This representation allows a Diffusion Transformer (DiT) to effectively map semantic intent onto geometric constraints through cross-attention layers, ensuring that the generated trajectories are both physically plausible and strictly consistent with the 3D scene. To support this task, we introduce WorldTraj, an extensive dataset consisting of roughly 136K triplets of 3D worlds, camera trajectories, and text prompts spanning 90K diverse environments. Extensive experiments demonstrate that WorldDirector successfully generates camera trajectories that are highly consistent with both the 3D world geometry and the semantic instructions in the text prompts. Code and models will be publicly released.

1 INTRODUCTION

The paradigm shift in large-scale foundation models has driven remarkable progress in both 3D reconstruction Kerbl et al. (2023); Yu et al. (2024); Wang et al. (2025b;c); Chen et al. (2026) and realistic video generation Wan et al. (2025); Blattmann et al. (2023), fueling growing demand for world-aware video synthesis and autonomous robotic navigation Huang et al. (2026); Chen et al. (2025); Wang et al. (2025a); Mao et al. (2025); Li et al. (2025a); Brooks et al. (2024). Since such tasks fundamentally rely on how a viewpoint traverses an environment, generating spatially coherent camera trajectories is a prerequisite for 3D geometric consistency. Without trajectories grounded in 3D space, generated videos often suffer from severe structural artifacts, such as inconsistencies upon scene revisit and geometrically implausible transitions.

While existing camera trajectory generation methods show promise, they exhibit critical limitations when applied to unconstrained 3D environments. Most prior works predominantly rely on character-driven motions Courant et al. (2024); Li et al. (2026); Courant et al. (2026), where the camera passively follows a human character Loper et al. (2015); Pavlakos et al. (2019), or derive trajectories from single RGB-D images Zhang et al. (2025b), or a text prompt Jiang et al. (2024); Li et al. (2024). Lacking a holistic understanding of global 3D geometry and semantics, these approaches fail to generate proactive, text-driven trajectories that physically interact with diverse 3D worlds. Consequently, generating camera trajectories directly grounded in unconstrained 3D point clouds based on natural language instructions remains largely unexplored.

In this paper, we present **WorldDirector**, a novel framework that produces world-grounded camera trajectories directly from 3D point clouds and text prompts, usable for downstream applications such as world-consistent video generation (Fig. 1). WorldDirector employs a unified Diffusion Transformer (DiT) that denoises camera trajectory tokens, learning complex kinematic motions while

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

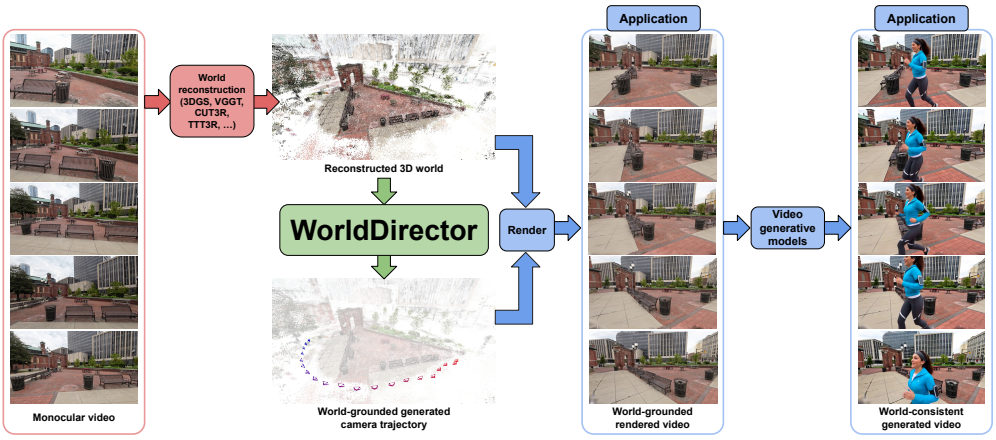

Figure 1: Given a reconstructed 3D world, WorldDirector generates a world-consistent camera trajectory that strictly adheres to text prompts. As a downstream application, the rendered views from these trajectories can serve as explicit structural guidance for video-to-video models Pika Labs (2023); Runway AI (2023); Wan et al. (2025), enabling the synthesis of final videos that maintain rigorous consistency with the underlying 3D world.

strictly adhering to the spatial layout of the environment, so that generated trajectories are both physically plausible and contextually consistent with the 3D world.

Central to our framework is the introduction of camera-aware world tokens, which explicitly encode the observer's initial spatial context – context typically lost in standard point cloud processing. Unlike conventional methods that use world features in a global coordinate system Zhang et al. (2025c), we transform the 3D point cloud into a local coordinate system centered on the starting camera's position and orientation, then concatenate these coordinates with high-level geometry features from a pre-trained point encoder Zhang et al. (2025c) to form the tokens. This lets the DiT's cross-attention layers directly query the 3D environment from the observer's own perspective, aligning physical consistency and semantic intent with the input world.

A major bottleneck in world-aware trajectory generation is the absence of suitable training data: existing datasets focus on isolated camera motions and lack comprehensive 3D world context. To address this, we construct WorldTraj, a large-scale dataset of approximately 136K triplets of 3D worlds, camera trajectories, and text prompts spanning 90K diverse worlds, providing the structural and semantic alignments needed to train and evaluate world-grounded camera trajectory generation in unconstrained environments.

Extensive evaluations across diverse 3D environments show that WorldDirector significantly outperforms existing baselines. Our method further enables new possibilities for downstream video generation by providing 3D-aware trajectories as spatio-temporal control signals, mitigating structural inconsistencies and geometric distortions to produce world-grounded, consistent videos.

Our main contributions are as follows:

- We propose WorldDirector, a unified DiT-based framework that generates world-grounded camera trajectories from 3D point clouds and text, integrating multimodal contexts so that generated motions are physically plausible and consistent with the 3D environment.
- We introduce camera-aware world tokens to explicitly encode the observer's initial spatial perspective. By transforming 3D coordinates into the starting camera's local coordinate system, our model precisely aligns semantic instructions with the geometric layout.
- We construct WorldTraj, a large-scale dataset featuring 136K triplets of 3D worlds, camera trajectories, and text across 90K environments, providing the structural and semantic alignments needed to train and evaluate world-grounded models.

2 RELATED WORKS

Camera trajectory generation. The generation of camera trajectories has been extensively studied under various conditioning signals. A major line of research focuses on character-centric camera control Courant et al. (2024); Li et al. (2026); Courant et al. (2026), where trajectories are generated to actively track human motions Loper et al. (2015); Pavlakos et al. (2019). While effective for cinematic tracking shots, these methods render the camera as a reactive observer constrained by local character movements, rather than an active explorer of the environment. To navigate without relying on human subjects, alternative frameworks estimate future camera paths from single RGB-D images Zhang et al. (2025b). However, operating heavily on 2.5D partial observations inherently restricts their ability to perceive global spatial structures, often leading to spatially disjointed trajectories. More recently beyond field of view (FoV) of the single image, text-guided trajectory generation has emerged Jiang et al. (2024), enabling intuitive semantic control over camera kinematics. Director3D Li et al. (2024) performs joint generation of 3D representations and camera trajectory. However, its camera generation is not conditionally grounded in the text, as the prompts do not explicitly specify camera movements. Despite this semantic flexibility, these models typically generate motions in an isolated latent space or co-generate them with synthetic scenes, lacking the capability to adapt to pre-existing, complex 3D environments. Unlike these approaches, the proposed WorldDirector directly conditions trajectory generation on unconstrained 3D point clouds, ensuring that the generated camera motions are strictly grounded in the given global spatial layout.

3D world reconstruction. The synthesis of large-scale 3D environments has seen remarkable progress, particularly in the domain of 3D scene reconstruction. Reconstruction methods Kerbl et al. (2023); Yu et al. (2024); Wang et al. (2025b;c); Chen et al. (2026) typically optimize implicit or explicit 3D representations from multi-view image collections, yielding high-fidelity digital replicas of real-world scenes. While these approaches successfully construct expansive static 3D spaces, they do not inherently provide mechanisms for dynamic spatial exploration. Consequently, when these environments are utilized for downstream tasks such as world-aware video generation, the absence of spatially coherent camera control often leads to geometric distortions and structural inconsistencies upon revisiting a scene. To fully harness the potential of these reconstructed worlds, our WorldDirector acts as a crucial complementary module, supplying the controllable, world-grounded camera trajectories required to render structurally consistent dynamic observations from static 3D spaces.

World-consistent video generation. Recent advancements have introduced 3D camera motion as a core condition for video generation. Methods such as GDC Van Hoorick et al. (2024), Uni3C Cao et al. (2025), GEN3C Ren et al. (2025), TrajectoryCrafter YU et al. (2025), ReCamMaster Bai et al. (2025), ReCapture Zhang et al. (2025a), and GeometryForcing Wu et al. (2026) perform camera-controlled video generation from single images or short clips, effectively acting as dynamic novel-view synthesis. However, their spatial exploration is fundamentally restricted by the narrow field of view of the 2D source. To encapsulate broader spatial contexts, recent frameworks like Wu et al. Wu et al. (2025a) and NeoVerse Yang et al. (2026) synthesize videos directly from explicit 3D worlds. Despite these architectural shifts, a critical bottleneck remains: all these generative models strictly rely on manually pre-defined camera trajectories. Our WorldDirector explicitly addresses this gap. By autonomously generating text-driven, world-grounded camera trajectories, WorldDirector serves as an ideal spatial-temporal control signal to seamlessly drive these downstream world-consistent video generation pipelines.

3 WORLDDIRECTOR

Fig. 2 illustrates the overall architecture of WorldDirector. Given a 3D point cloud, a directorial text prompt, and the extrinsic and intrinsic parameters of an initial camera, WorldDirector is designed to generate a spatially coherent camera trajectory. We represent the 3D world as a set of point coordinates, a format that ensures compatibility across various 3D modalities. To resolve scale ambiguity, both the 3D points and trajectories are normalized following the scale-consistent protocol of VGGT Wang et al. (2025b).

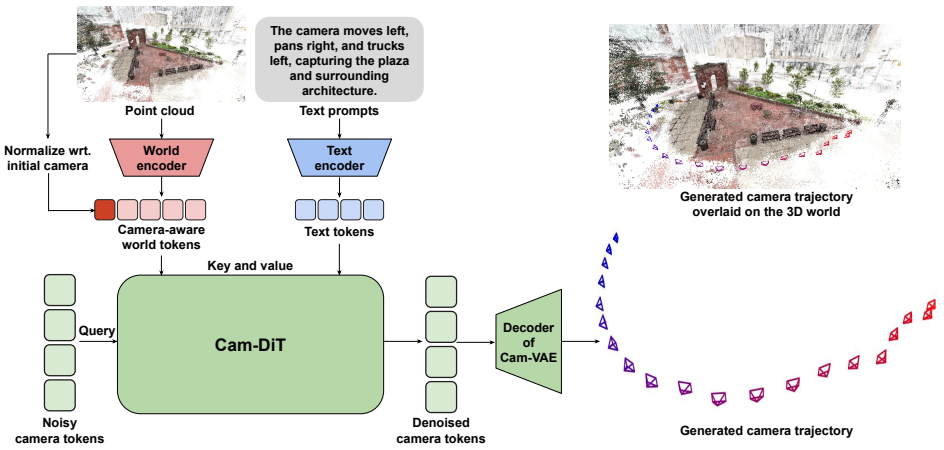

Figure 2: The overall architecture of WorldDirector, which generates spatially coherent camera trajectories conditioned on camera-aware world tokens and text tokens. The right-top visualization shows the overlaid generated camera trajectory on the 3D world.

3.1 Relative Camera Latent Space

In our framework, the camera parameters of the first frame are provided as a condition. Unlike image pixels (e.g., $[0, 255]$), raw camera trajectories consist of translation and rotation parameters that exhibit unbounded scales. Direct diffusion on this raw space often leads to severe instability. To mitigate this, we introduce a Camera Variational Autoencoder (Cam-VAE) to canonicalize the trajectories. Specifically, the Cam-VAE represents camera motions relative to the first frame: the initial extrinsic matrix is mapped to the identity, and its intrinsic focal length is normalized to 1, with subsequent frames predicted as relative transformations and focal length multipliers. Prior to diffusion training, we train the Cam-VAE with standard reconstruction and KL divergence loss functions. The Cam-VAE compresses these relative, normalized trajectories into a bounded latent representation, allowing our pipeline to operate efficiently as a latent diffusion model Rombach et al. (2022).

3.2 Camera-aware World Tokens and Text Tokens

We project the 3D world and text prompts into a unified context space for cross-attention to provide precise spatial and semantic guidance. The world representation is designed to be *camera-aware* by explicitly encoding the spatial relationship between the environment and the observer. First, we extract base world tokens using Concerto Zhang et al. (2025c), a point encoder pre-trained via joint self-supervised learning between 2D images and 3D point clouds Oquab et al. (2024); Wu et al. (2025b). While Concerto normalizes the point cloud by centering it at its mean for feature extraction, this process inevitably discards the absolute pose of the initial camera.

To preserve this critical geometric context, we transform the 3D points into the camera coordinate system, setting the initial camera position as the origin and its orientation as the axis alignment. These camera-relative coordinates are then concatenated with the Concerto features, allowing the diffusion model to remain fully aware of the initial camera's pose and perceive the scene layout from its specific perspective. Simultaneously, the text prompt is encoded into semantic tokens using UMT5 Chung et al. (2025). Finally, these camera-aware world tokens and text tokens serve as key and value pairs within the cross-attention layers of the diffusion model to guide the trajectory denoising process.

3.3 Trajectory Generation via Cam-DiT

We employ a Diffusion Transformer (DiT) Peebles & Xie (2023), termed Cam-DiT, to generate latent camera trajectories conditioned on the 3D world and text prompts. During the reverse diffusion process, the noisy latent camera tokens serve as queries within the DiT blocks. By alternately cross-attending to the text tokens and the camera-aware world tokens across successive layers, Cam-DiT predicts the noise added to the latent trajectory at each timestep. The model is trained using a

Table 1: Comparison of existing datasets with camera trajectory annotations. Notably, our WorldTraj is the only dataset that provides trajectories spanning over 90K diverse 3D environments.

Datasets	World	Cam. traj.	Text	# of worlds	# of cam. traj.	Average duration (s)	# of vocab.
CCD Jiang et al. (2024)	✗	✓	✓	0	25K	7.2	48
ET Courant et al. (2024)	✗	✓	✓	0	115K	3.75	1,790
DataDoP Zhang et al. (2025b)	✗	✓	✓	0	29K	**14.4**	8,698
PulpMotion Courant et al. (2026)	✗	✓	✓	0	**193K**	5.85	4,599
WorldTraj (Ours)	✓	✓	✓	**90K**	136K	11.7	**15,480**

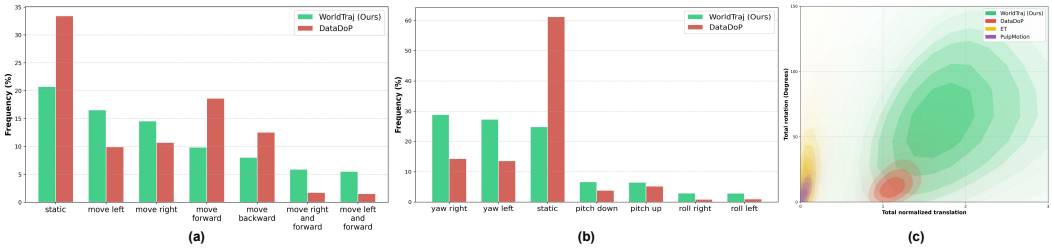

Figure 3: Comparison of camera trajectory annotations across various datasets. (a) Translation tagging distribution (top 7) of the camera trajectories. (b) Rotation tagging distribution (top 7) of the camera trajectories. (c) Joint density distribution of total normalized translation and total rotation. Following DataDoP Zhang et al. (2025b), translations are normalized to account for the scale ambiguity inherent in monocular videos.

standard noise prediction objective, which minimizes the error between the added noise and the predicted noise. This allows Cam-DiT to iteratively recover the denoised latent tokens, which are then passed through the Cam-VAE decoder to reconstruct the relative camera trajectory. Finally, to obtain the absolute trajectory in the 3D world space, we multiply these decoded relative parameters by the provided initial frame's actual extrinsic and intrinsic matrices.

4 WorldTraj Dataset

As shown in Tab. 1, while several datasets provide camera trajectory annotations, existing benchmarks fall short in supporting world-grounded camera trajectory generation due to a lack of comprehensive 3D-text-motion alignment. To bridge this critical gap, we introduce WorldTraj, a novel dataset comprising 136K triplets of 3D worlds, camera trajectories, and text descriptions spanning 90K diverse scenes (Fig. 5). Furthermore, as illustrated in Fig. 3, WorldTraj provides significantly more dynamic camera trajectories compared to existing datasets Courant et al. (2024); Zhang et al. (2025b); Courant et al. (2026).

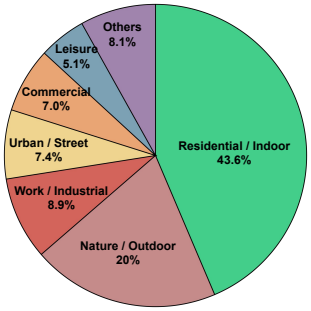

Figure 5: Diversity and distribution of the 3D worlds in our WorldTraj dataset.

The overall construction pipeline is illustrated in Fig. 4. Specifically, we build our dataset by leveraging complementary assets from two public sources: DL3DV-10K Ling et al. (2024) and DynamicVerse Wen et al. (2025). The DL3DV-10K dataset offers captured videos of static environments, from which optimized 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) representations and accurate camera trajectories are extracted. Conversely, the DynamicVerse dataset supplies captured videos of dynamic scenes, providing essential dynamic object masks and dense point maps.

4.1 World and Camera Trajectory Annotations

For the 3DGS scenes included in the DL3DV-10K dataset, we convert the Gaussians into a 3D point cloud. Specifically, we sample a set of discrete points from each 3D Gaussian based on its underlying probability distribution. The number of points extracted is proportional to the Gaussian's opacity. The spatial coordinates of these points are drawn from a multivariate normal distribution defined by the Gaussian's mean position and its covariance matrix, which is derived from the scale and

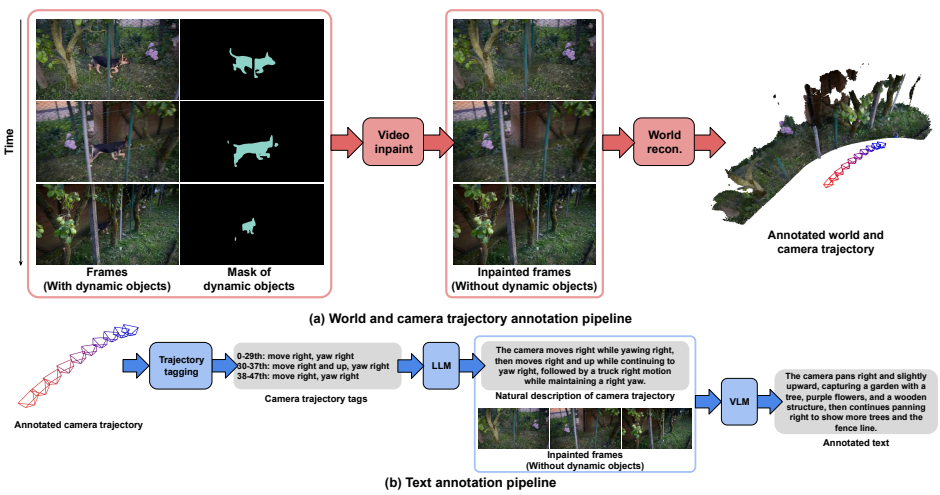

Figure 4: The construction pipeline of our WorldTraj dataset. Given in-the-wild monocular videos, we reconstruct the 3D scenes, extract camera trajectories, and annotate the corresponding text descriptions.

Table 2: World and camera trajectory annotation accuracy on the Sintel dataset Butler et al. (2012).

Methods	Dynamic objects	World accuracy		Camera accuracy		
		AbsRel↓	δ1.25↑	ATE↓	RPE_t↓	RPE_r↓
DynamicVerse Wen et al. (2025)	Present	0.224	0.674	**0.179**	0.084	0.297
DA3 Lin et al. (2025)	Present	**0.169**	**0.749**	0.213	0.112	**0.244**
STTN Zeng et al. (2020) + DA3 (Ours)	Inpainted	0.247	0.626	0.210	**0.075**	0.255

Table 3: Text annotation accuracy of subset of our WorldTraj dataset.

Inputs to VLM Team (2025)	Score↑
Inpainted frames	0.83
Inpainted frames + LLM text (Ours)	**3.7**

rotation parameters. We use camera trajectory annotations of DL3DV-10K dataset as it is. For the DynamicVerse dataset, we process the videos to obtain static 3D environments. We first remove moving objects using the provided dynamic object masks and apply STTN Zeng et al. (2020) to inpaint the occluded regions, resulting in static videos. Then, we utilize Depth Anything V3 Lin et al. (2025) to lift these inpainted static videos into 3D point clouds and obtain camera trajectory annotations. We show how our trajectory generation is robust to the inpainting in Tab. 8. Overall, DL3DV-10K and DynamicVerse provide 7K and 83K worlds—each derived from an equal number of videos—amounting to 129.8 and 163.7 total hours, respectively. Tab. 2 shows that our annotation pipeline achieves world reconstruction accuracy comparable to baselines while delivering superior camera trajectory accuracy. These results confirm our effectiveness in generating high-fidelity paths aligned with the 3D world. Detailed evaluation procedures are provided in Sec. A.

4.2 Text Annotations

To generate comprehensive text descriptions that capture both camera motion and visual content, we process the extracted camera trajectories from both DL3DV-10K and DynamicVerse through a multi-stage annotation pipeline. First, we apply a camera trajectory tagging method Zhang et al. (2025b) to segment each trajectory and assign discrete motion labels (e.g., "moving forward") based on predefined rules. These discrete tags are then translated into fluent, natural language descriptions using an LLM Team (2025). Finally, we input the inpainted video, generated via STTN as described above, along with the LLM-generated motion descriptions into a Vision-Language Model (VLM) Team (2025) to incorporate scene-specific visual details. We empirically found that explicitly providing this trajectory tagging information significantly improves the VLM's output quality, as shown in Tab. 3. We provide detailed evaluation procedures of the table in Sec. A. This process yields the final text annotations, which simultaneously describe the temporal camera dynamics and the spatial contents of the environment. Furthermore, we provide text captions for the original videos featuring dynamic objects, enabling the trajectories generated by our WorldDirector to be utilized for dynamic video generation. While our text annotations are predominantly kinematic, we show in Sec. A.5 that WorldDirector can be readily fine-tuned to follow semantic-level instructions as well.

Table 4: Ablation study on the effectiveness of world grounding, the Cam-VAE, and the camera-aware world tokens. The last row is ours.

Settings			Trajectory quality			Text-camera consistency		World-camera consistency	
World	Cam-VAE	Camera-aware world tokens	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
✗	✓	✗	20.8731	0.9091	0.7974	41.5922	0.1892	0.5866	0.7643
✓	✓	✗	17.5624	0.9409	0.8219	44.5243	0.219	0.6081	0.4431
✓	✗	✗	627.6212	0.7753	0.2112	22.8918	0.1558	**0.658**	**0.0533**
✓	✓	✓	**16.5335**	**0.9577**	**0.8392**	**45.1494**	**0.2411**	0.6464	0.3101

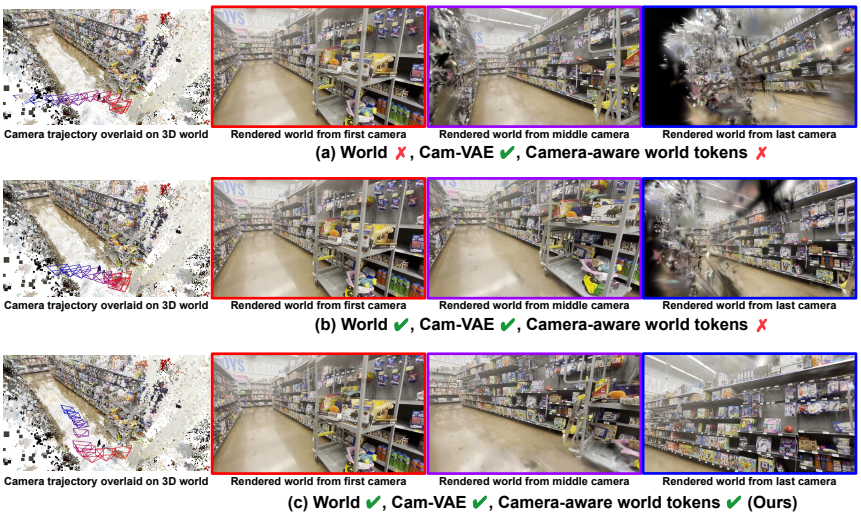

Figure 6: Qualitative comparison of generated camera trajectories across different ablation settings. The rightmost three columns show rendered world via 3DGS along the generated camera trajectories. The input text prompt is "The camera moves left and forward, tilting up and panning to reveal the toy aisle."

5 Experiments

5.1 Protocols

Datasets. Since existing datasets lack 3D worlds, camera trajectories, and text, we use our WorldTraj dataset for training and evaluation. Following Zhang et al. (2025b), we use a 9:1 train-test split. To assess out-of-domain generalization, we further evaluate on TartanAir Wang et al. (2020), a synthetic dataset with diverse motion patterns, and ScanNet Dai et al. (2017)/ARKitScene Baruch et al. (2021), which features real-world indoor reconstructions. They are not used during the training and only used for the evaluation purpose. These tests demonstrate WorldDirector's robustness across unseen synthetic and real environments. ScanNet and ARKitScene's results are in the Tab. 11 and 12, respectively. We used subset of WorldTraj dataset for ablation studies.

Evaluation metrics. We evaluate our model using metrics established in GenDoP Zhang et al. (2025b), while introducing Visible Point IoU (VP-IoU) to assess world-grounding accuracy and collision rate to assess physical plausibility. Trajectory Quality is measured by CLaTr-FID (equivalent to FID Heusel et al. (2017) in the CLaTr embedding space), Density, and Coverage. Text-camera Consistency is quantified via CLaTr-CLIP (text-trajectory similarity) and F1-score (motion tag classification). Our proposed VP-IoU evaluates World-camera Consistency by measuring the intersection over union of 3D point indices visible from the ground-truth (GT) and generated camera paths, verifying precise spatial alignment with the 3D world. Finally, collision rate (%) measures the fraction of trajectory segments with points within a fixed, scene-scale-normalized radius of the camera position, verifying that generated trajectories remain physically plausible and collision-free within the 3D world. We describe analysis of the new metric (VP-IoU) in Tab. 10 and details of metrics in Sec. B.1.

Table 5: Comparison to state-of-the-art camera trajectory generation methods on our WorldTraj dataset. GenDoP* is trained on our dataset.

Methods	Trajectory quality			Text-camera consistency		World-camera consistency	
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
Director3D Li et al. (2024)	528.3769	0.7438	0.2976	5.2554	0.0385	0.524	1.795
GenDoP Zhang et al. (2025b)	121.5615	0.7975	0.4663	40.4976	0.175	0.5649	2.009
GenDoP Zhang et al. (2025b)*	77.3137	0.8266	0.5894	44.5611	0.2013	0.5781	1.373
WorldDirector (Ours)	**6.9994**	**0.9864**	**0.8431**	**51.2048**	**0.298**	**0.6747**	**0.557**

Table 6: Comparison to state-of-the-art camera trajectory generation methods on TartanAir dataset Wang et al. (2020). GenDoP* is trained on our dataset.

Methods	Trajectory quality			Text-camera consistency		World-camera consistency	
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
Director3D Li et al. (2024)	493.1621	0.1369	0.1601	9.7389	0.0159	0.361	1.061
GenDoP Zhang et al. (2025b)	214.4058	0.2147	0.2764	44.9786	0.0684	0.3681	3.230
GenDoP Zhang et al. (2025b)*	104.2535	0.3655	0.4731	48.7682	0.0739	0.3788	1.531
WorldDirector (Ours)	**77.5955**	**0.5109**	**0.5948**	**49.804**	**0.1204**	**0.4813**	**0.136**

5.2 Ablation studies

World grounding with camera-aware world tokens. As demonstrated in Tab. 4, the transition from the first row to the second row shows that incorporating world inputs leads to a consistent improvement across all metrics, most notably nearly halving the collision rate (0.7643 to 0.4431). This confirms that 3D geometric context is fundamental for generating trajectories that are grounded in the physical environment and physically plausible. Interestingly, while adding basic world features (e.g., Concerto Zhang et al. (2025c) features) enhances general trajectory quality and physical plausibility, the world-consistency metric, VP-IoU, exhibits its most significant gain only when our camera-aware world tokens are introduced in the last row.

The primary reason for this sharp increase in VP-IoU of the last row is the explicit spatial conditioning provided by our encoding. By concatenating the Concerto features with 3D coordinates normalized to the initial camera's local coordinate system, the model gains a precise, ego-centric geometric prior of the scene. This allows the diffusion model to recognize the exact spatial relationship between the starting viewpoint and surrounding structures. As illustrated in Fig. 6, while basic world inputs prevent the model from "blindly" following text, it is the camera-aware tokens that ensure the trajectory is strictly anchored to the scene's geometry, effectively maximizing the overlap of visible points throughout the path. This confirms that explicit camera-world alignment is the key driver for achieving high-fidelity, world-consistent camera movements.

Stable camera trajectory generation with Cam-VAE. The second and third rows of Tab. 4 demonstrate the critical role of the Cam-VAE in stabilizing the generation process. While the model without the Cam-VAE paradoxically yields a better VP-IoU and collision rate, neither reflects superior world-grounding or physical plausibility. As evidenced by its exceptionally poor CLaTr-FID score, both are merely artifacts of severe, high-frequency oscillations in the generated camera pose: these unnatural vibrations cause the camera's field of view to wildly sweep across the scene, artificially inflating the visible point overlap, while the erratic, jittery motion incidentally avoids sustained contact with scene geometry, at the cost of catastrophic trajectory degradation.

This underlying instability occurs because raw camera parameters consist of highly heterogeneous components; specifically, translation vectors are inherently unbounded, while rotations operate on a fundamentally different, bounded manifold. Applying diffusion directly to this mixed and unconstrained raw space leads to severe training instability. Our Cam-VAE resolves this by canonicalizing and mapping these disparate signals into a unified, bounded latent representation. Operating within this well-regularized latent space stabilizes the diffusion process and naturally acts as a low-pass filter against high-frequency noise, yielding smooth, realistic, and physically plausible camera motions.

5.3 Comparison to state-of-the-art methods

As reported in Tab. 5 and 6, WorldDirector consistently outperforms state-of-the-art camera trajectory generation methods across nearly all metrics, most notably achieving the highest world-camera consistency (VP-IoU) and lowest collision rate on both benchmarks – demonstrating its unrivaled

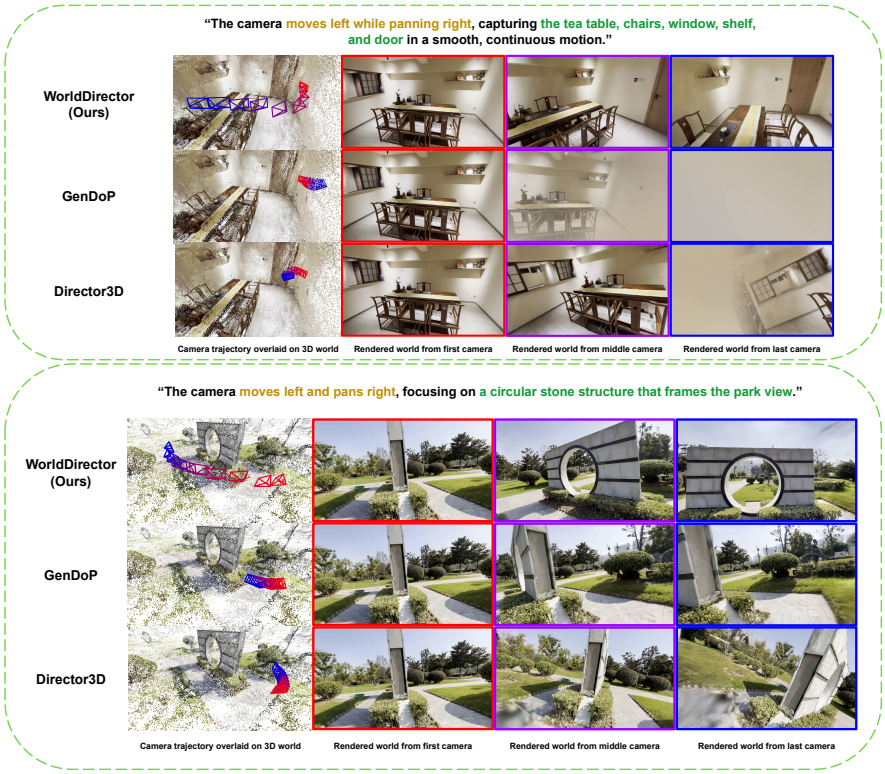

Figure 7: Qualitative comparison between our WorldDirector and state-of-the-art camera trajectory generation methods Li et al. (2024); Zhang et al. (2025b). The rightmost three columns show rendered world via 3DGS along the generated camera trajectories.

ability to ground generated motions within complex 3D environments while remaining physically plausible. This gap reflects the inherent limitations of existing approaches: GenDoP Zhang et al. (2025b), relying only on a single RGB-D view and text, remains unaware of the 3D layout beyond its limited FoV, while Director3D Li et al. (2024) relies only on text. Qualitative comparisons in Fig. 7 reinforce these results, with WorldDirector achieving comprehensive scene coverage while strictly adhering to text prompts, accurately executing complex panning and navigation tasks where baselines fail to maintain spatial or semantic alignment.

Our experiments also highlight the significant utility of WorldTraj. To ensure a rigorous comparison, we retrained GenDoP on WorldTraj using its official codebase, and it consistently outperformed its original version across all test scenarios – underscoring the effectiveness of WorldTraj's larger and more diverse collection of world-camera pairs (Sec. 4). Even for GenDoP, which lacks explicit 3D grounding due to its single RGB-D input, WorldTraj's rich variety of scenes and trajectories serves as a superior training signal. Character-centric models Courant et al. (2024); Li et al. (2026); Courant et al. (2026) are excluded, as their reliance on explicit character trajectories falls outside our text-driven, world-grounded setting. We used official implementations of Director3D Li et al. (2024) and GenDoP Zhang et al. (2025b) for our experiments. We compare long-horizon generation capability of state-of-the-art methods in Tab. 13.

6 Conclusion

We presented WorldDirector, a framework that generates world-grounded camera trajectories from 3D point clouds and text prompts. Its camera-aware world tokens explicitly preserve the observer's initial spatial context, aligning generated trajectories with the 3D world. To support this task, we also constructed WorldTraj, an extensive dataset of 136K triplets spanning 90K diverse worlds. Our evaluations validate that WorldDirector produces accurate, contextually consistent, and physically plausible trajectories, providing a robust foundation for downstream applications such as world-consistent video generation.

AI USE STATEMENT

In this work, we used generative AI tools only to edit the writing of the paper to improve readability and clarity (a recommended-disclosure task). None of the required-disclosure tasks listed in the ICLR AI Policy for Authors apply to this work. All AI-suggested edits were reviewed and verified by the authors before being incorporated. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *NeurIPS*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 2024.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

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. *ACM TOG*, 2025.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. In *NeurIPS*, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3D reconstruction as test-time training. In *ICLR*, 2026.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *ICLR*, 2025.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the Exceptional Trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024.

Robin Courant, Xi Wang, David Loiseaux, Marc Christie, and Vicky Kalogeiton. Pulp Motion: Framing-aware multimodal camera and human motion generation. In *ICLR*, 2026.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. In *ICLR*, 2026.

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. In *Computer Graphics Forum*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023.

Sikuang Li, Chen Yang, Jiemin Fang, Taoran Yi, Jia Lu, Jiazhong Cen, Lingxi Xie, Wei Shen, and Qi Tian. WorldGrow: Generating infinite 3D world. *arXiv preprint arXiv:2510.21682*, 2025a.

Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3D: Real-world camera trajectory and 3D scene generation from text. In *NeurIPS*, 2024.

Yu Li, Menghan Xia, Gongye Liu, Jianhong Bai, Xintao Wang, Conglang Zhang, Yuxuan Lin, Ruihang Chu, Pengfei Wan, and Yujiu Yang. AdaViewPlanner: Adapting video diffusion models for viewpoint planning in 4D scenes. In *ICLR*, 2026.

Zhuoling Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. LongDiff: Training-free long video generation in one go. In *CVPR*, 2025b.

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.

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 *CVPR*, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Pika Labs. Pika: An evolutionary video-to-video and text-to-video generation platform, 2023. URL `https://pika.art/`.

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 *CVPR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Runway AI. Gen-2: The next step forward for generative ai, 2023. URL `https://runwayml.com/`.

Qwen Team. Qwen3 technical report, 2025.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative Camera Dolly: Extreme monocular dynamic novel view synthesis. In *ECCV*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Dilin Wang, Hyunyoung Jung, Tom Monnier, Kihyuk Sohn, Chuhang Zou, Xiaoyu Xiang, Yu-Ying Yeh, Di Liu, Zixuan Huang, Thu Nguyen-Phuoc, et al. WorldGen: From text to traversable and interactive 3D worlds. *arXiv preprint arXiv:2511.16825*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *CVPR*, 2025c.

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.

Kairun Wen, Yuzhi Huang, Runyu Chen, Hui Zheng, Yunlong Lin, Panwang Pan, Chenxin Li, Wenyan Cong, Jian Zhang, Junbin Lu, et al. DynamicVerse: A physically-aware multimodal framework for 4D world modeling. In *NeurIPS*, 2025.

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. In *ICLR*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *NeurIPS*, 2025a.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *CVPR*, 2025b.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4d world model with in-the-wild monocular videos. In *CVPR*, 2026.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3D gaussian splatting. In *CVPR*, 2024.

Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *CVPR*, 2025a.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. GenDoP: Auto-regressive camera trajectory generation as a director of photography. *arXiv preprint arXiv:2504.07083*, 2025b.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2D-3D self-supervised learning emerges spatial representations. In *NeurIPS*, 2025c.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. RIFLEx: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 2023.

In this supplementary material, we provide more experiments, discussions, and other details that could not be included in the main text due to the lack of pages. The contents are summarized below:

- Sec. A: WorldTraj dataset
- Sec. B: Comparison to state-of-the-art methods
- Sec. C: Ablation studies
- Sec. D: Additional visualizations
- Sec. E: Architecture details
- Sec. F: Implementation details
- Sec. G: Limitations

A WorldTraj dataset

A.1 Examples of WorldTraj dataset

World and camera trajectory annotations. Fig. 8 and 9 present additional examples produced by our dataset construction pipeline. As described in Sec. 4 of the main paper, our dataset construction pipeline reconstructs the 3D world (scene) and estimates the corresponding camera trajectories from real-world videos. The examples highlight the robustness and generalizability of the pipeline in diverse real-world videos.

Text annotations. Examples in Fig. 10 illustrate the effectiveness of our text annotation pipeline. Details of the pipeline can be found in Sec. 4 of the main paper. This design enables the captions to jointly describe camera motion and scene context precisely.

A.2 Computational cost for dataset construction

Tab. 7 summarizes the average processing time per video, total runtime, and peak VRAM consumption for each stage of our dataset construction pipeline. The dataset construction is conducted on one NVIDIA RTX Pro 6000 GPU taking approximately 572 hours.

Table 7: Processing time and hardware usage for each stage of our dataset annotation pipeline.

Tasks	Avg. time /video (s)	Total time (h)	Peak VRAM (GB)
1. Inpainting	1.3	34	25
2. Point cloud scene and camera annotation	6.9	173	70
3. Camera trajectory tagging	0.2	5	80
4. Camera and scene captioning	14.4	360	80
Total	22.8	572	80

A.3 Details of world and camera trajectory annotation evaluation

Tab. 2 shows that in our pipeline, camera trajectory estimation is improved compared to the baseline (second row). This improvement stems from the fact that inpainting dynamic objects eliminates distracting foreground movements, allowing the estimation model to focus exclusively on the camera's egomotion. While our pipeline exhibits a marginal degradation in world accuracy—likely due to artifacts introduced during the inpainting process with STTN Zeng et al. (2020)—this degradation remains limited. The trade-off is justified by the ability to reconstruct complete and hole-free 3D point clouds, highlighting a practical balance between strict depth accuracy and obtaining a consistent world representation. Following the evaluation protocol of DynamicVerse Wen et al. (2025), we report AbsRel and $\delta < 1.25$ for depth, and ATE, RPEt, and RPEr for camera metrics. Depth evaluation is computed only on background regions, and all methods are aligned using least-squares in disparity space.

To assess robustness to inpainting-induced world reconstruction artifacts, we use static TartanAir/ScanNet scenes with available GT scene/camera, synthetically composite dynamic objects into the video, remove them via our inpainting pipeline, and reconstruct the world at increasing severity (measured by the average inpainted mask ratio per scene). Tab. 8 shows that while world recon-

Table 8: Robustness of our trajectory generation quality under worlds reconstructed with different inpainting mask ratios on TartanAir.

Dynamic/inpainted mask ratio	Trajectory quality			Text-camera consistency		World-camera consistency	
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
0%	65.1512	0.5446	0.6052	**45.6173**	<u>0.1082</u>	<u>0.3797</u>	**0.1815**
7.67%	<u>63.47</u>	**0.5754**	**0.6239**	<u>45.5656</u>	0.106	**0.3802**	0.1908
23.34%	**63.38**	<u>0.5567</u>	<u>0.6104</u>	45.4905	**0.1118**	0.3793	<u>0.1905</u>

struction accuracy degrades with increasing inpainting mask ratio, downstream trajectory generation quality is only marginally affected, often matching or even exceeding the quality obtained from a clean (0% mask) input. This demonstrates that WorldDirector's generation quality is robust to inpainting-induced world reconstruction artifacts. Note: the 0%-mask values here differ slightly from those in Tab. 6, as here we align the reconstructed camera trajectory to the GT scale and use this reconstructed, scale-aligned world as input.

A.4 Details of text annotation evaluation

To evaluate the quality of the text annotations in Tab. 3, we employ an LLM-as-a-judge framework Zheng et al. (2023) using Qwen3 Team (2025). Specifically, the LLM is prompted to assess the semantic consistency between the rule-based camera tags and the generated final annotations on a Likert scale ranging from 0 to 5. The results confirm that providing LLM-generated text as an additional semantic prior allows the VLM to produce descriptions that are more precisely aligned with the intended camera movement. The detailed prompt used for this evaluation, including the scoring rubric and contextual instructions, is provided in List. 1.

```
You are a camera-motion caption evaluator.

Compare a caption with ordered GT camera-motion tags.
Use only these motion labels.

Translation:
move right, move left, move forward, move backward, move up, move down,
    static

Rotation:
yaw right, yaw left, pitch up, pitch down, roll right, roll left, static

Phrase mapping:
pan left/right = yaw left/right
tilt up/down = pitch up/down
truck/slide left/right = move left/right
dolly/push in = move forward
pull back/dolly back = move backward

Verdict labels:
FULL_MATCH: same motion directions and temporal order
TEMPORAL_COMPRESSION: correct motions but multiple GT segments are merged
PARTIAL_MATCH: some correct motion, but an important component is missing
WRONG_DIRECTION: explicit wrong direction
WRONG_ORDER: correct motions but wrong temporal order
NO_MATCH: mostly unrelated or contradictory

Return ONLY valid compact JSON.
No markdown. No explanation outside JSON.
The reason should be short and factual.

Schema:
{
"verdict": "...",
"pred": ["translation + rotation", "..."],
"missing": ["..."],
"extra": ["..."],
"reason": "..."
```

Table 9: Evaluation on WorldTraj with original (kinematic) and semantic-level prompts.

Testing prompt type	Model	Trajectory quality			Text-camera consistency		World-camera consistency
		CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑
Kinematic	Original	6.9994	0.9864	0.8431	51.2048	0.298	0.6747
	Semantic-finetuned	**6.3332**	**1.02**	**0.8606**	**51.9758**	**0.2988**	**0.6781**
Semantic	Original	39.2955	0.7832	0.6644	27.5924	0.1614	0.6082
	Semantic-finetuned	**9.4341**	**0.8972**	**0.8083**	**32.724**	**0.1912**	**0.6303**

Table 10: Mean human preference score by VP-IoU range.

VP-IoU range	Mean preference↑	Chosen best (%)↑	Chosen worst (%)↓
0.0 - 0.4	2.241	20.1	60.4
0.4 - 0.6	2.452	14.5	40.6
0.6 - 0.8	3.735	49.0	**13.2**
0.8 - 1.0	**4.151**	**76.8**	13.7

```
}

Evaluate:
```

Listing 1: Prompt for LLM-as-a-judge to evaluate our text annotation pipeline.

A.5 Semantic-level text prompts

WorldTraj's text annotations predominantly describe camera motion in cinematographic terms (e.g., "the camera moves left, pans right"). To assess whether WorldDirector can be extended to more semantic, intent-driven instructions, we use Qwen2.5 to rewrite a subset of WorldTraj prompts into semantic-level form, e.g.,

> *"The camera pans right and moves backward, then shifts left and down while tilting up, capturing a wide view of the store's wall decor and interior." → "go back, then move left and down while looking up at the store's wall decor."*

We then fine-tune WorldDirector on a mixture of the original kinematic prompts and these newly generated semantic-level prompts, and evaluate on both prompt types. We will release this fine-tuned checkpoint alongside our original model.

Tab. 9 show that the original model, trained only on kinematic prompts, degrades substantially when evaluated on semantic-level prompts, confirming that these are out-of-domain for kinematic-only training. After fine-tuning on the mixed kinematic-and-semantic-level prompts, semantic-prompt performance improves markedly while kinematic-prompt performance is maintained, demonstrating that WorldDirector readily extends to semantic-level instructions.

B Comparison to state-of-the-art methods

B.1 Details of evaluation metrics

Trajectory quality. We adopt the evaluation protocol from GenDoP Zhang et al. (2025b) and utilize the CLaTr embedding space for quantitative analysis. We Follow ET Courant et al. (2024) for the CLaTr embedding, and the VAE for the CLaTr embedding (different from our Cam-VAE) is trained on our WorldTraj.

- CLaTr-FID: We calculate the Fréchet Inception Distance (FID) Heusel et al. (2017) between the feature distributions of the GT and generated trajectories. Lower values indicate that the generated trajectories more closely resemble the statistical distribution of realistic camera motions.
- Density & Coverage: These metrics quantify the fidelity and diversity of the generated manifold. Density measures how many GT samples are near each generated sample, reflecting the "realism" and local concentration. Coverage evaluates the fraction of the GT manifold covered by the generated samples, indicating the "diversity" of the model's output.

Table 11: Comparison to state-of-the-art camera trajectory generation methods on ScanNet Dai et al. (2017). GenDoP* is trained on our dataset.

Methods	Trajectory quality			Text-camera consistency		World-camera consistency	
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
Director3D Li et al. (2024)	586.0005	0.5578	0.2	8.4	0.0405	0.5148	0.776
GenDoP Zhang et al. (2025b)	367.89	0.3978	0.4367	36.4056	0.0844	0.5362	0.419
GenDoP Zhang et al. (2025b)*	160.3883	0.5422	0.6	45.3692	0.0654	0.5587	**0.344**
WorldDirector (Ours)	**142.5628**	**0.6123**	**0.7324**	**48.1055**	**0.1279**	**0.6150**	0.351

Table 12: Comparison to state-of-the-art camera trajectory generation methods on ARKitScene Baruch et al. (2021). GenDoP* is trained on our dataset.

Methods	Trajectory quality			Text-camera consistency		World-camera consistency	
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑	Collision↓
Director3D Li et al. (2024)	539.3806	**0.7543**	0.1698	7.9596	0.0051	0.3781	3.1754
GenDoP Zhang et al. (2025b)	223.7258	0.3957	0.4437	39.2442	0.0551	0.4262	1.7612
GenDoP Zhang et al. (2025b)*	97.7809	0.5585	0.6144	44.8981	0.052	0.4364	1.7389
WorldDirector (Ours)	**90.5325**	0.5212	**0.6299**	**49.4189**	**0.0972**	**0.6369**	**0.9975**

Text-camera consistency. These metrics assess whether the generated camera trajectory aligns semantically with the provided text prompt. We Follow ET Courant et al. (2024) for the CLaTr embedding, and the VAE for the CLaTr embedding (different from our Cam-VAE) is trained on our WorldTraj.

- CLaTr-CLIP: Inspired by CLIP-based evaluation, we compute the cosine similarity between the embedding of the input text and the embedding of the generated trajectory within the CLaTr latent space. A higher score signifies a stronger semantic correlation.
- F1-score: We employ a rule-based motion tagging (introduced in Sec. 4 of the main manuscript) to obtain discrete action labels from the generated trajectories. The F1-score is calculated by comparing these predictions with the GT labels, assessing the precision and recall of the model's semantic execution.

World-camera consistency. We introduce a novel metric to evaluate how well the generated trajectory is grounded in the 3D world geometry.

- Visible Point IoU (VP-IoU): This metric measures the spatial alignment between the GT and generated camera paths within the 3D scene. We first identify the set of 3D point indices from the world mesh that are visible within the camera's frustum along the entire GT path (S_{GT}) and the generated path (S_{gen}). Visibility is determined via a tolerance-based z-buffer, to account for depth noise in the reconstructed point clouds. The VP-IoU is then defined as:
$$\text{VP-IoU} = \frac{|S_{\text{GT}} \cap S_{\text{gen}}|}{|S_{\text{GT}} \cup S_{\text{gen}}|}.$$
Unlike frame-wise metrics, VP-IoU evaluates the global consistency of the trajectory. A high VP-IoU confirms that the generated camera precisely captures the intended visual features and navigates the 3D space as planned, without violating geometric constraints.
- Collision rate: For each generated trajectory, we normalize the corresponding 3D scene to a common scale and identify a segment as colliding if at least k scene points fall within a fixed radius r of the camera position at that segment. The radius r is calibrated so that GT trajectories exhibit a collision rate less than 1%, providing a physically meaningful threshold that is robust to noisy floating points. We report the fraction of trajectory segments identified as colliding, averaged over the test set. Lower values indicate trajectories that more successfully avoid penetrating scene geometry.

To validate that VP-IoU reflects genuine human judgment of world-grounding quality rather than serving as a purely geometric proxy, we conduct a user study with 27 participants, as shown in Tab. 10, each comparing three trajectories of low/medium/high VP-IoU under matched world/text conditions, scored 1–5 (5=most preferred), yielding $N = 90$ ratings. We find a significant positive correlation between VP-IoU and human preference (Pearson $r = 0.4524$, $N = 90$, $p < 0.0001$), with preference rising across VP-IoU ranges (best-choice rate: $20.1\% \rightarrow 76.8\%$ from lowest to

Table 13: Long-horizon generation comparisons on the WorldTraj subset (734 samples). GenDoP* is trained on our WorldTraj dataset.

Metric	Method	Chunk 0	Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5	Mean
CLaTr-FID↓	GenDoP*	115.41	106.73	99.33	110.65	90.47	109.77	105.39
	Ours (unguided)	61.71	38.55	**35.6**	**54.03**	**27.3**	**38.16**	**42.56**
	Ours (guided)	**59.02**	**36.04**	37.61	59.31	28.59	40.96	43.59
CLaTr-CLIP↑	GenDoP*	50.54	49.84	50.20	46.54	51.32	50.53	49.83
	Ours (unguided)	**56.69**	**56.8**	**58.2**	**55.76**	**57.5**	**58.06**	**57.17**
	Ours (guided)	56.08	56.08	57.83	55.44	57.10	57.21	56.62
Collision rate↓	GenDoP*	2.2434	2.2136	2.8471	1.5632	2.0140	2.2974	2.1965
	Ours (unguided)	0.909	1.009	**0.586**	1.278	1.293	1.837	1.1519
	Ours (guided)	**0.0403**	**0.7408**	0.6879	**1.1629**	**0.9339**	**1.0446**	**0.7684**
BVR↓	GenDoP*	–	0.466	0.492	0.569	0.552	0.496	0.515
	Ours (unguided)	–	0.542	0.55	0.568	0.522	0.568	0.55
	Ours (guided)	–	**0.407**	**0.414**	**0.455**	**0.401**	**0.426**	**0.421**
BDC (°)↓	GenDoP*	–	40.6	**37.3**	52.9	56.0	39.2	45.2
	Ours (unguided)	–	45.8	48.2	51.9	46.2	48.9	48.2
	Ours (guided)	–	**39.2**	42.3	**45.2**	**42.8**	**38.8**	**41.7**

highest range; see supplementary material for the full breakdown). When asked for the reason behind their choice, participants cited text adherence most often (53.70%), followed by spatial grounding (29.60%) and smoothness (16.70%), confirming that VP-IoU – which specifically targets spatial grounding – captures one of several factors underlying human preference, consistent with its intended role as a targeted world-consistency metric.

B.2 More quantitative results

ScanNet and ARKitScene datasets. Tab. 11 and 12 provide comparisons against state-of-the-art methods on the ScanNet Dai et al. (2017) and ARKitScene Baruch et al. (2021) datasets, respectively. Consisting of real-world indoor reconstructions, ScanNet and ARKitScene are used exclusively for evaluation to validate the zero-shot generalization capabilities of our model on unseen real-world environments.

Autoregressive long-horizon extension. Although long-horizon generation is not a claimed contribution of WorldDirector (fixed 49-frame horizon), we explore extending it via a simple autoregressive (AR) chaining scheme, where the last camera of one chunk becomes the first camera of the next. Long-horizon rollout requires worlds large enough for extended exploration, so we use a subset of 734 WorldTraj samples meeting this criterion. Following AR video-generation evaluation conventions Zhao et al. (2025); Li et al. (2025b); Yang et al. (2025), we do not compare against a single long GT trajectory, since some drift is expected by construction; instead we track per-chunk quality and boundary continuity, using *boundary velocity ratio* (BVR. relative change in speed across the chunk boundary) and *boundary direction change* (BDC. angle between consecutive velocity vectors across the boundary).

To improve physical plausibility and boundary continuity within this AR rollout, we introduce a *training-free, test-time guidance*. At each denoising step, we estimate $\hat{x}_0$ via Tweedie's formula from the current noisy latent, decode it through the Cam-VAE to obtain a camera trajectory, and compute the following loss terms on this trajectory: a KNN-distance-based hinge collision loss for physical plausibility, and, at each new chunk's first denoising step, two additional boundary terms using the previous chunk's final velocity v_{prev} and the new chunk's initial velocity v_{new},

$$\mathcal{L}_{\text{velocity}} = \left(\log \frac{\|v_{\text{new}}\|}{\|v_{\text{prev}}\|}\right)^2, \qquad \mathcal{L}_{\text{direction}} = 1 - \cos(v_{\text{new}}, v_{\text{prev}}). \tag{1}$$

We use the gradient of these losses to correct $\hat{x}_0$ before re-encoding it back to ϵ for the next DDIM step. No retraining is required.

We compare against GenDoP* – itself an autoregressive model natively suited to chaining – under the identical protocol (6 chunks $\times$ 10 seconds/chunk = 60 seconds).

Across all six chunks, WorldDirector's quality remains disjoint from GenDoP*'s: our CLaTr-FID range (27.3–61.71) never overlaps GenDoP*'s (90.47–115.41), and likewise for CLaTr-CLIP

(55.76–58.20 vs. 46.54–51.32). Guidance leaves both metrics nearly unchanged from the unguided setting (mean CLaTr-FID 42.56 vs. 43.59; mean CLaTr-CLIP 57.17 vs. 56.62), while substantially improving collision rate, BVR, and BDC.

On collision rate, GenDoP* stays uniformly high (1.5632–2.8471). Even the unguided rollout achieves a 1.9× lower mean, and with guidance, every chunk of our rollout (max 1.1629) falls below GenDoP*'s best chunk, yielding a 2.9× lower mean overall (0.7684 vs. 2.1965). We note that collision rate does rise within our own rollout (0.909 at chunk 0 to 1.837 at chunk 5, reduced to 1.0446 with guidance), which is precisely the behavior dedicated AR training should address in future work.

Without guidance, our boundary continuity is slightly weaker than GenDoP*'s (BVR 0.55 vs. 0.515), unsurprising given that GenDoP is designed for autoregressive generation while ours is not. Our guidance reverses this (BVR 0.421, BDC 41.7° vs. 45.2°) while simultaneously achieving a substantially lower collision rate. Note that this guidance mechanism is specific to diffusion-based formulations such as ours: GenDoP*'s discrete-token autoregressive decoding has no iterative denoising loop or differentiable latent to apply it to. We view this training-free guidance as a promising starting point for AR long-horizon extension, to be combined with dedicated AR training in future work.

B.3 More qualitative results on WorldTraj

We provide additional qualitative comparisons with GenDoP Zhang et al. (2025b) and Director3D Li et al. (2024) in Fig. 11 on our dataset, WorldTraj. These examples further highlight the advantages of our method in following complex camera instructions while maintaining consistency with the 3D scene. Note that rendering artifacts may occur as the 3DGS is reconstructed from monocular videos, which inherently lack complete geometric information.

In the first example, the camera follows an instruction that requires moving requires coordinated camera translation and rotation around wanted subject such as table. Our method successfully generates a trajectory around the display table in counterclockwise while continuously keeping the table in the view and revealing surrounding products. In contrast, the baseline methods generate rotation in incorrect directions or generate unstable rolling motions that are inconsistent with the instruction.

In the second example, the instruction requires the focus to remain on a target object during camera movement. Our model correctly moves around the statue in a clockwise manner, while keeping the statue centered in the frame and simultaneously showcasing the surrounding area. However, the baseline methods fail to maintain focus on the statue and instead drift toward irrelevant parts of the scene.

Finally, the third example requires long-horizon navigation through a constrained environment. Our model navigates through a narrow corridor without colliding with the surrounding geometry and reaches the intended destination, where the target hallway is clearly captured. In contrast, the baseline methods do not follow the intended motion sequence and remain near the starting point without reaching the target scene.

B.4 More qualitative results on unseen datasets

We provide additional qualitative comparisons with GenDoP Zhang et al. (2025b) and Director3D Li et al. (2024) on TartanAir Wang et al. (2020) and ScanNet Dai et al. (2017) datasets in Fig. 12, to evaluate generalization beyond the training distribution. These examples demonstrate that our method generalizes well to unseen datasets while maintaining accurate instruction following and spatial consistency in diverse environments. Note that the point cloud renderings may appear sparse or incomplete due to imperfect scene reconstruction and limited viewpoint coverage in the original data.

In the first example from TartanAir Wang et al. (2020) dataset, the instruction requires coordinated camera movement to reveal different parts of the scene. Our method successfully moves backward while tilting upward, transitioning the view from the ground to the sky as intended. In contrast, the baseline methods fail to follow the instruction and continue focusing only on the ground.

In the second example from ScanNet Dai et al. (2017) dataset, the instruction requires sequentially capturing multiple target objects during camera movement. Our trajectory correctly follows the instruction by shifting focus from the sofa to the backpack, and then to the bicycle wheel and desk in the intended order. However, the baseline methods fail to follow the described motion and instead rotate toward irrelevant directions, capturing only the sofa or nearby walls.

Finally, the third example from ScanNet Dai et al. (2017) dataset requires vertical navigation while capturing different parts of the scene. Our method successfully moves upward and then downward, capturing both the upper and lower sections of the bathroom as intended. In contrast, the baseline methods fail to follow the instruction, either moving only upward and capturing only the upper section or colliding with surrounding walls.

B.5 User study

As summarized in Fig. 13, WorldDirector achieved a dominant preference of 83.0%, significantly outperforming GenDoP Zhang et al. (2025b) (9.8%) and Director3D Li et al. (2024) (7.2%). For a fair evaluation, we randomly sampled 8 test cases consisting of scenes and text prompts unseen during training. We invited 14 experts in computer vision and graphics to assess these cases, where each participant was presented with three videos rendered via 3DGS using trajectories from each method. All methods received the same text prompt, and experts were instructed to select the video that best aligned with the given instruction. The results confirm that our model generates far more realistic and instruction-compliant camera paths in novel environments.

B.6 Supplementary videos

Please refer to the attached supplementary video for additional qualitative results and demonstrations of our method. The videos show that our approach successfully executes the given text prompts while maintaining spatial consistency, ensuring that the generated motions are both contextually accurate and physically grounded within the world environment. For high-quality visualization, the generated camera trajectories of all methods are temporally interpolated to ensure smooth video rendering.

- `sota_compare_1.mp4`: The baseline models fail to follow instruction and captures scene inconsistently without knowing what to focus on. In contrast, our proposed model successfully follows the text instructions while maintaining spatial awareness and keeps focus on target vehicle.
- `sota_compare_2.mp4`: Our model effectively adapts the given instruction to the surrounding environment. The baseline models struggle to properly follow the instruction of how to move and what to capture. GenDoP Zhang et al. (2025b) model partially follows instruction but fails to keep focus on golf cart and surrounding building at the same time. On the other hand, WorldDirector recognizes the surrounding environment, identifies position where it can capture two target subjects simultaneously and properly follow the instruction.
- `sota_compare_3.mp4`: While the baseline methods fail to follow complex instruction combined with three motion patterns, WorldDirector precisely follows the text prompt while navigating the 3D scene. This demonstrates our model's superior ability to ground semantic instructions into accurate, scene-aware camera trajectories.
- `sota_compare_4.mp4`: The baseline models fail to follow the complex instruction and do not capture the target objects. In contrast, our model successfully follows the instruction and captures the blackboard and chair, which were not visible in the initial view. This demonstrates our model's strong ability to understand the surrounding world and accurately identify where to focus based on the given instruction.

B.7 Supplementary downstream task videos

Please refer to the attached supplementary video for the world-consistent video generation downstream application that could be aligned with our proposed method. The videos are generated with video-to-video models Pika Labs (2023); Runway AI (2023); Wan et al. (2025) conditioned on world-grounded video that is rendered with the generated camera trajectory of WorldDirector. For

Table 14: Ablation study on the world crop ratio for the world embedding on subset of WorldTraj test set.

World crop ratio	Trajectory quality			Text-camera consistency		World-camera consistency
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑
1.0	23.3937	0.9418	0.8175	43.0456	0.2118	0.6337
1.25 (Ours)	16.5335	0.9577	0.8392	**45.1494**	**0.2411**	**0.6464**
1.5	**16.0301**	**0.9629**	**0.8442**	45.1161	0.2361	0.6396

Table 15: Ablation on world-coordinate fusion strategy on a subset of the WorldTraj dataset.

Fusion strategy	Trajectory quality			Text-camera consistency		World-camera consistency
	CLaTr-FID↓	Density↑	Coverage↑	CLaTr-CLIP↑	F1-Score↑	VP-IoU↑
Fourier encoding + concat	18.9987	**0.9654**	0.8372	43.5053	0.2321	0.6393
Raw coordinates + gated fusion	19.6322	0.9559	**0.8475**	43.1007	0.2222	0.638
Raw coordinates + concat (Ours)	**16.5335**	0.9577	0.8392	**45.1494**	**0.2411**	**0.6464**

high-quality visualization, the generated camera trajectories are temporally interpolated to ensure smooth video rendering.

- `world-consistent_video_1.mp4`: This video demonstrates our method could be applied on world-consistent video generation task. WorldDirector generates world-grounded camera trajectory given text instruction and reconstructed 3D world, which is basketball court. With generated camera trajectory, world-grounded video can be rendered and injected into video-to-video generation model Pika Labs (2023) to generate world-consistent dynamic video. The final video includes a man playing a basketball strictly grounded in our reconstructed 3D world.
- `world-consistent_video_2.mp4`: Similarly, the final video in this video is generated under a different world and text condition, demonstrating the generalizability of our method and its potential to extend to diverse scenarios.

C Ablation Studies

Videos. We provide the supplementary video `ablation.mp4` to further illustrate the ablation studies discussed in Sec. 5.2 of the main manuscript. The video demonstrates qualitative results that are consistent with experimental results reported in the main text (Tab. 4 and Fig. 6), confirming the effectiveness of each proposed module in enhancing trajectory generation.

World crop ratio. Tab. 14 demonstrates that our proposed world crop ratio (second row) achieves the most favorable results compared to other scaling factors. Given that unconstrained 3D environments can be extensively large, processing the entire world point cloud is computationally prohibitive. To address this, we crop the world around the initial camera to include points within a radius of 1.25 times the average distance from the starting camera position to all points in the scene. As shown in the ablation results, a crop ratio that is too small provides insufficient geometric context, leading to suboptimal trajectory generation. Conversely, an excessively high crop ratio not only increases the computational overhead but also degrades the model's generalization capability by introducing irrelevant spatial information. Our findings suggest that the 1.25 ratio provides an optimal balance, ensuring that WorldDirector captures necessary environmental constraints while maintaining computational efficiency and robust performance across diverse 3D scenes.

Camera-aware token fusion mechanism. We compare our concatenation-based fusion against alternative strategies for injecting camera-relative coordinates into world tokens: Fourier encoding of the coordinates before concatenation, and a gated fusion of raw coordinates with the Concerto features. Tab. 15 shows that our simple concatenation of raw, camera-relative coordinates with world features achieves the best overall trade-off, notably leading on CLaTr-FID, text-camera alignment, and VP-IoU, while Fourier encoding and gated fusion each improve isolated metrics at the cost of

Table 16: APD across datasets (higher indicates more diverse generations). GenDoP* is trained on our WorldTraj dataset.

Methods	WorldTraj	TartanAir
Director3D Li et al. (2024)	0.3855	0.4357
GenDoP Zhang et al. (2025b)	0.1947	0.1771
GenDoP Zhang et al. (2025b)*	0.3365	0.3266
WorldDirector (Ours)	**0.4043**	**0.9083**

Table 17: VP-IoU (mean $\pm$ std across 3 random seeds) on WorldTraj. GenDoP* is trained on our WorldTraj dataset.

Method	VP-IoU
Director3D Li et al. (2024)	0.5239 ± 0.0003
GenDoP Zhang et al. (2025b)	0.5653 ± 0.0003
GenDoP Zhang et al. (2025b)*	0.5785 ± 0.0004
WorldDirector (Ours)	$\mathbf{0.6742 \pm 0.0005}$

others. This suggests raw coordinate concatenation preserves the precise metric information needed for spatial grounding, whereas additional encoding or gating introduces a bottleneck that trades off world-consistency for other properties.

D Additional visualizations

D.1 Latent space of Cam-VAE

In Fig. 14, we visualize the camera trajectories by decoding the linearly interpolated latent $\mathbf{z}_\alpha = (1-\alpha)\mathbf{z}_1 + \alpha\mathbf{z}_2$, where $\alpha \in [0, 1]$. $\mathbf{z}_1$ and $\mathbf{z}_2$ are encoded latents from camera trajectory samples in WorldTraj dataset. While α increases, the interpolated latent captures the continuous and smooth transitions from $\mathbf{z}_1$ to $\mathbf{z}_2$ in camera translation and rotation, which means the structured latent space of our Cam-VAE.

D.2 Diverse generation capability of Cam-DiT

The Cam-DiT exhibits the capability to generate diverse camera trajectories under identical text and world conditions. Tab. 16 shows our WorldDirector achieves the highest average pairwise distance (APD) between generated trajectories. Tab. 17 shows that, despite generating diverse trajectories across seeds (Tab. 16), WorldDirector attains the highest mean VP-IoU with low variance (std $=$ 0.0005), indicating that diverse samples remain consistently well-grounded in the 3D world rather than trading quality for diversity. In Fig. 15, we visualize the camera trajectories generated from Cam-DiT, varying only the random seed used to sample the initial noise of the denoising process. As illustrated in the figure, the Cam-DiT captures the camera descriptions of text condition (*moves left, pans right, and trucks left*) and the world geometry, while representing the distinct camera path and viewpoint at the last camera frame.

E Architecture details

Fig. 2 and Sec. 3 of the main manuscript describe the architecture of WorldDirector. We provide detailed descriptions of each component, including Cam-VAE and Cam-DiT.

E.1 Cam-VAE

The Cam-VAE consists of two components: encoder and decoder.

Encoder. The encoder takes the camera trajectory relative to the first frame: the identity matrix for the extrinsic matrix and the intrinsic focal length normalized to 1. The camera trajectory is represented as $\mathbf{x} \in \mathbb{R}^{T\times 11}$, where T is the number of camera time steps and 11 is the dimension of the camera parameter representation including a 6D rotation representation, 3D translation, and focal lengths (f_x, f_y). The encoder is composed of 9 stacked 1D convolution layers with ReLU activations. Two of the convolution layers perform downsampling using stride 2, resulting in a

Table 18: Computational cost comparison to state-of-the-art camera trajectory generation methods.

Methods	Conditions	Peak VRAM (GB)	Time (sec)	# of trainable params.
Director3D Li et al. (2024)	Text	11	**0.5**	**53M**
GenDoP Zhang et al. (2025b)	Text and RGB-D	**10**	3.45	1800M
WorldDirector (Ours)	Text and World	19	2.19	65M

latent representation $\mathbf{z} \in \mathbb{R}^{(1+T/4)\times D_{\mathbf{z}}}$. We set the number of camera time steps $T = 49$ and latent dimension $D_{\mathbf{z}} = 64$. The 49 cameras are temporally contiguous frames, not sparsely sampled across a longer sequence.

Decoder. The decoder takes the latent representation $\mathbf{z}$. The decoder is composed of 8 stacked 1D convolution layers with ReLU activations. Two nearest-neighbor upsampling layers are applied between the convolution layers to progressively increase the sequence length from $1 + T/4$ to T, producing the reconstructed camera trajectory $\mathbf{x} \in \mathbb{R}^{T\times 11}$.

E.2 CAM-DIT

The Cam-DiT denoises the noisy latent camera tokens $\mathbf{z}_i$ into $\mathbf{z}_{i-1}$, conditioned on text tokens and camera-aware world tokens at each denoising timestep i. The Cam-DiT consists of 8 DiT Peebles & Xie (2023) blocks with a token embedding dimension of 512 and 8 attention heads. Specifically, in a single DiT block, the noisy latent camera tokens act as the queries during one self-attention layer and two cross-attention layers. The two cross-attention layers alternately utilize text tokens and camera-aware world tokens as keys and values. Camera-aware world tokens are tokenized from the world, which is cropped to include points within 1.25 times the average distance from the first frame camera position to all points. This cropping strategy enables efficient world encoding and prevents the Cam-DiT from attending out-of-range regions in the world during camera trajectory generation. The denoising timestep embeddings are injected into the query tokens twice via AdaLN Peebles & Xie (2023): before the self-attention layer and before the cross-attention layer with the camera-aware world tokens.

F IMPLEMENTATION DETAILS

Cam-VAE. The Cam-VAE is trained using the AdamW optimizer with an initial learning rate 10^{-3}. The training is conducted on 4 NVIDIA RTX Pro 6000 GPUs with batch size 256 during 300 epochs, taking almost 6 hours. The weight of KLD loss is set as 10^{-3}.

Cam-DiT. We train Cam-DiT using AdamW optimizer with an initial learning rate 10^{-4}. The training is conducted on 8 NVIDIA RTX Pro 6000 GPUs with batch size 32 for a total of 96K iterations, taking approximately 36 hours. Before training, we first extract the camera-aware world tokens for memory and time efficiency.

Computational cost. Tab. 18 compares the inference efficiency of our method with state-of-the-art models. Despite taking a full 3D world as input, WorldDirector maintains a highly efficient profile with only 65M trainable parameters—nearly 27 times fewer than GenDoP's 1800M. Furthermore, our model achieves a faster inference time (2.19 sec) compared to GenDoP (3.45 sec), demonstrating a superior balance between computational cost and scene-aware generation capability.

G LIMITATIONS

Our collision-avoidance guidance is a soft, distance-based penalty rather than a hard physical constraint, so a small fraction of trajectories still exhibit residual penetration, particularly where the input point cloud is sparse. More broadly, WorldDirector operates purely in the kinematic space of camera poses, without explicit physical quantities such as velocity limits or contact dynamics, so plausibility is enforced only implicitly rather than guaranteed by a physics model.

Our AR extension also generates each chunk independently conditioned only on the previous chunk's final camera state, without explicit memory of earlier chunks. A streaming AR formulation that retains memory across chunks is a promising direction for future work.



Figure 8: Examples of the WorldTraj dataset. We present examples of our dataset construction pipeline on the DynamicVerse dataset, including video inpainting, 3D scene reconstruction, and camera trajectory extraction.

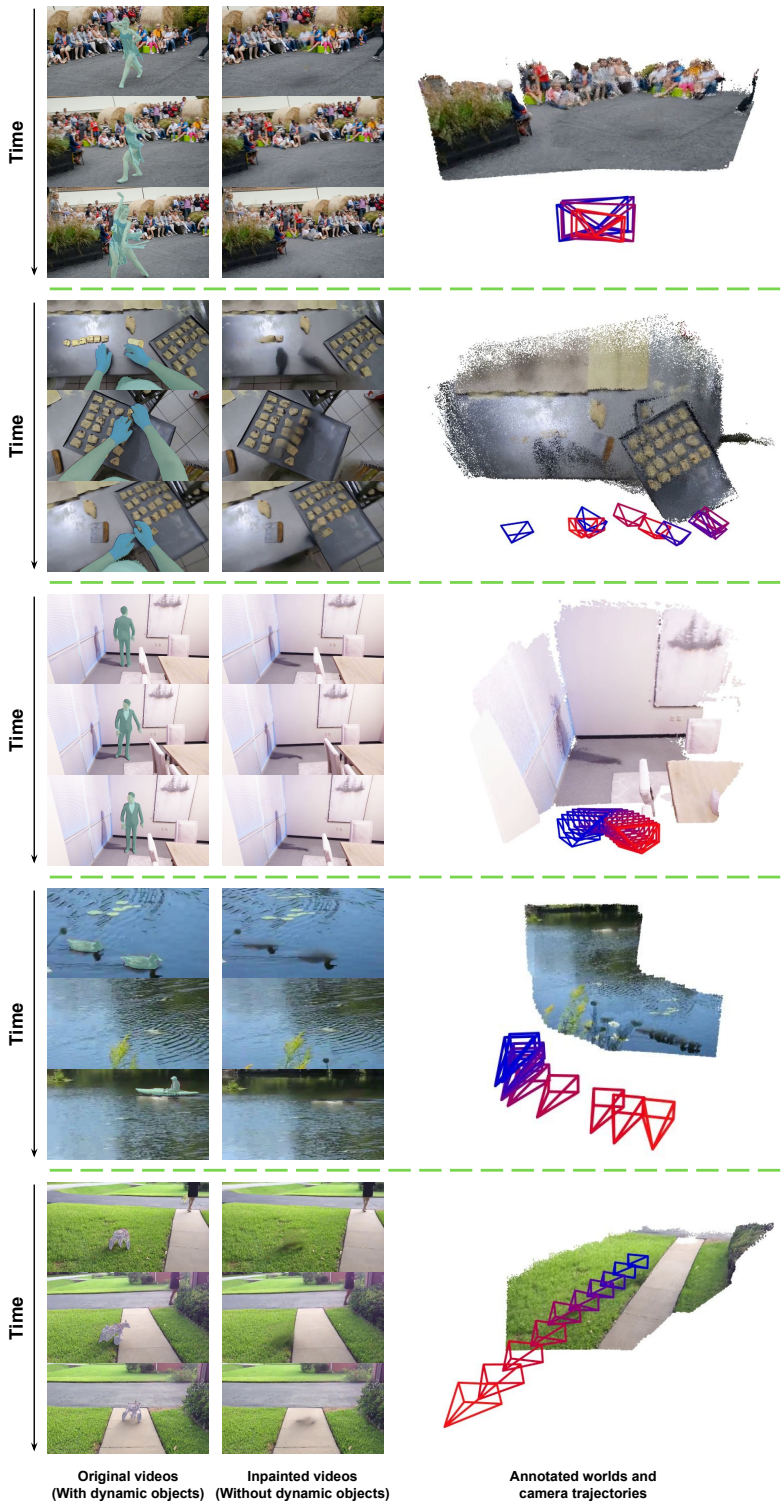

Figure 9: Additional examples of our WorldTraj dataset worlds and cameras constructed from DynamicVerse videos.

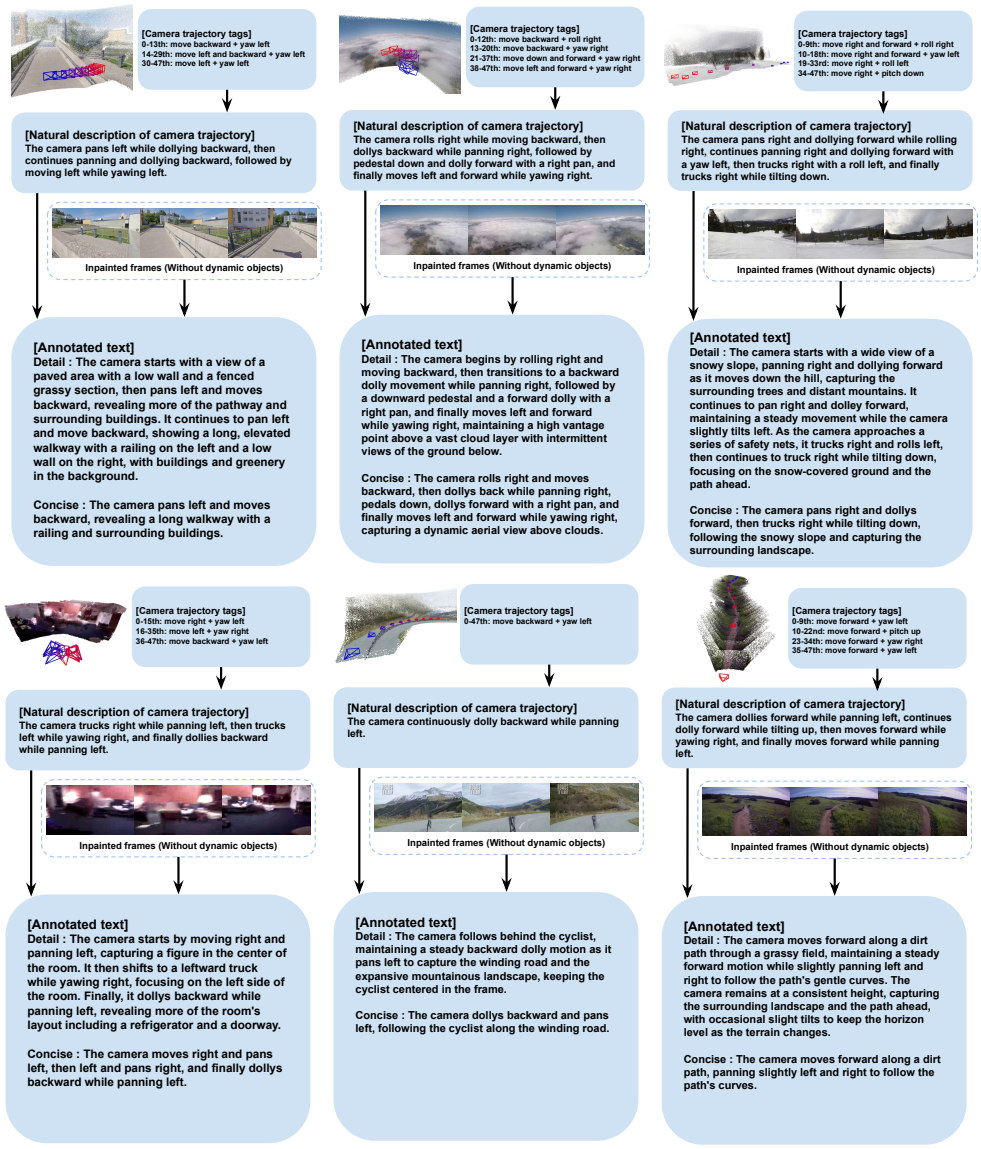

Figure 10: Examples of captions in our WorldTraj dataset. The captions describe the camera motion and scene context inferred from the annotated camera trajectories and the inpainted frames.

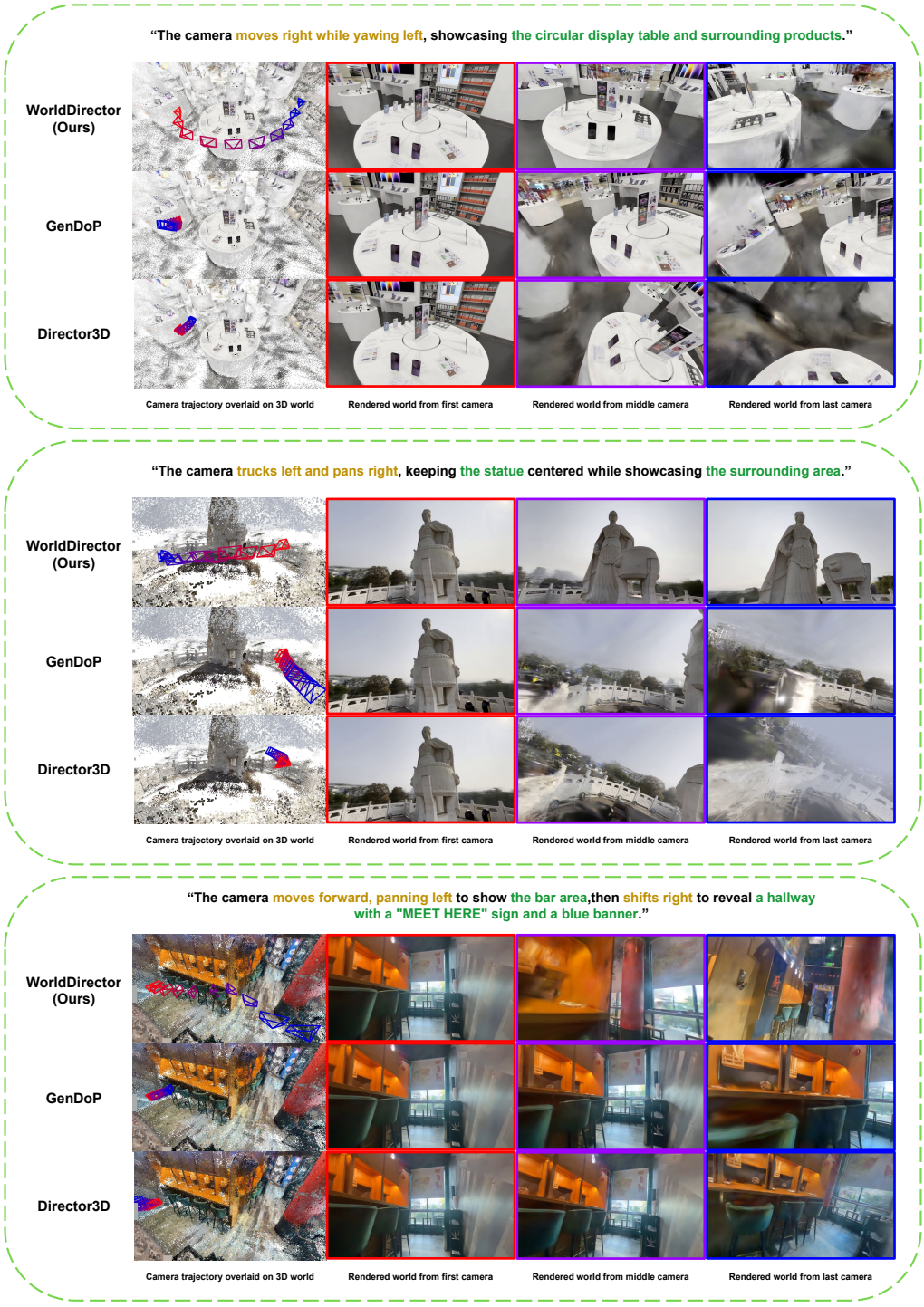

Figure 11: Qualitative comparison between our WorldDirector and state-of-the-art camera trajectory generation methods Li et al. (2024); Zhang et al. (2025b). The rightmost three columns show rendered world via 3DGS along the generated camera trajectories.

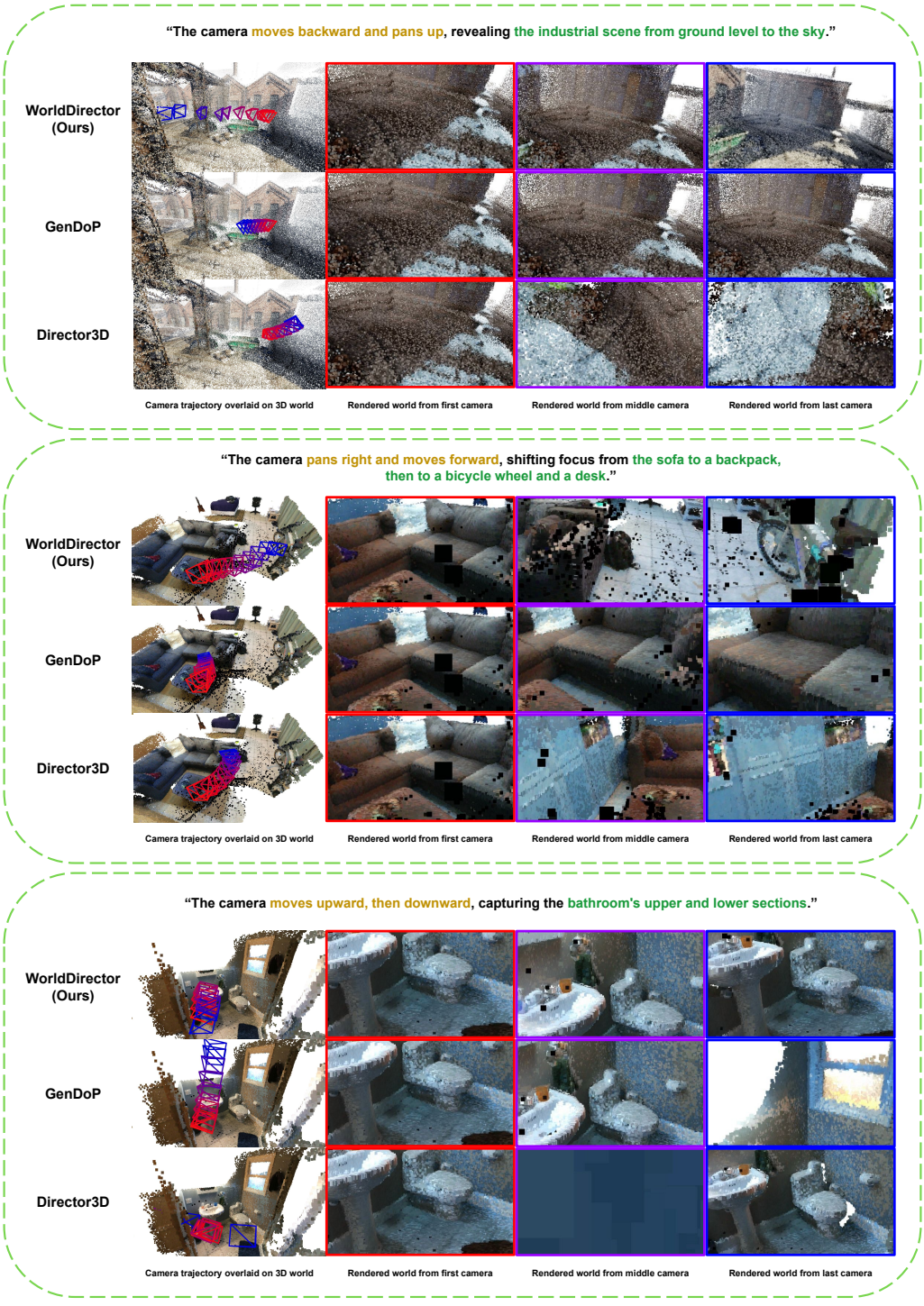

Figure 12: Qualitative comparison between our WorldDirector and state-of-the-art camera trajectory generation methods Li et al. (2024); Zhang et al. (2025b) on unseen datasets. The first example is from TartanAir Wang et al. (2020), while the second and third examples are from ScanNet Dai et al. (2017). The rightmost three columns visualize point cloud renderings from the generated camera trajectories.

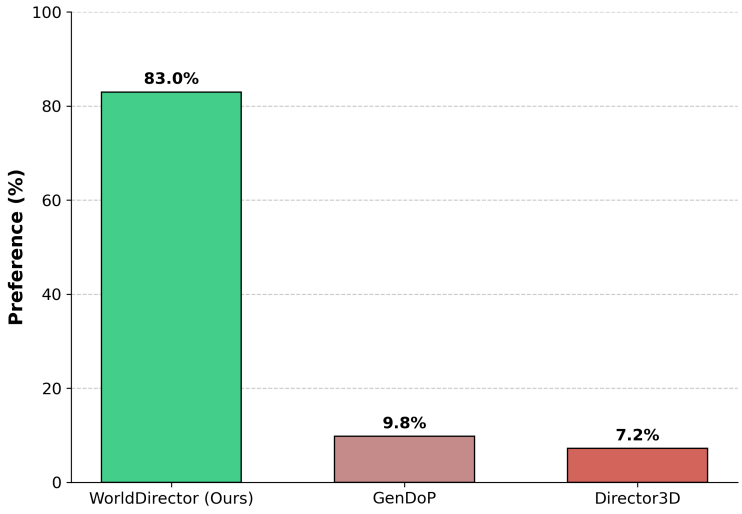

Figure 13: User study results. Preference rates for camera trajectory generation on randomly sampled unseen cases.

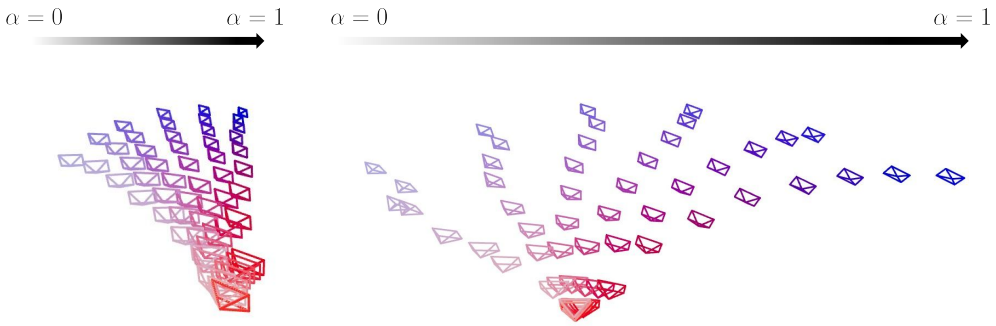

Figure 14: Structured latent space of our Cam-VAE. We visualize the camera trajectories decoded from the interpolated Cam-VAE latent representations $\mathbf{z}_\alpha = (1-\alpha)\mathbf{z}_1 + \alpha\mathbf{z}_2$, where $\alpha \in [0, 1]$. $\mathbf{z}_1$ and $\mathbf{z}_2$ are encoded latents from camera trajectory samples in WorldTraj dataset.



Figure 15: Diversity of camera trajectory generation. Cam-DiT produces diverse camera trajectories depending on the initial random noise under the same text and world conditions.