Who Does What and Where to Go: Orthogonal Alignment and Hierarchical Planning for Multi-Entity Trajectories

Zhang Wan[1,2], Yu Li[1,2⋆], Tianze Huang[1,2], Juan Cao[1,2], and Sheng Tang[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
liyu@ict.ac.cn

Abstract. Generating precise multi-entity 3D trajectories is fundamental for interactive video generation and embodied world models. However, existing text-to-trajectory mappings face two bottlenecks: severe feature leakage from coupled multi-modal inputs, and physical collisions caused by blind end-to-end 3D coordinate regression. We propose ProgTraj-Director, a framework driven by physical property disentanglement and hierarchical spatial planning. To resolve the "who does what" attribution ambiguity, our Structured Vision-Language Alignment (StructVLA) module disentangles static identity masks from dynamic spatial motion flows, binding them via orthogonal factorized tensor fusion in the latent space. This formulation encourages entity-specific semantic separation and mitigates cross-entity feature leakage. To bridge the "where to go" semantic-physical gap, we formulate trajectory generation as hierarchical spatial planning. The model first constructs a Geometry-Aware Bird's-Eye View (BEV) to resolve anti-collision topological boundaries. Guided by this structural prior, 3D trajectories are derived via progressive spatial denoising to improve geometric and temporal consistency. To support this research, we introduce Stepwise-ME, a large-scale dataset of over 31,000 interactive video clips (8M frames) with fine-grained stepwise annotations. Extensive experiments show that our approach enhances action attribution and spatial coherence, providing a reliable structural prior for downstream video generation. Code and dataset are available at https://github.com/wzbos-token/Who-Does-What-and-Where-to-Go.

Keywords: 3D Trajectory · Entity Binding · BEV Planning

1 Introduction

Recently, video generation is gradually transitioning from passive pixel synthesis toward physics-grounded Embodied World Models. A genuine world model requires more than extended generation duration; it demands the consistent simulation of complex multi-entity interactions (e.g., avoiding object interpenetration

⋆ Corresponding author.

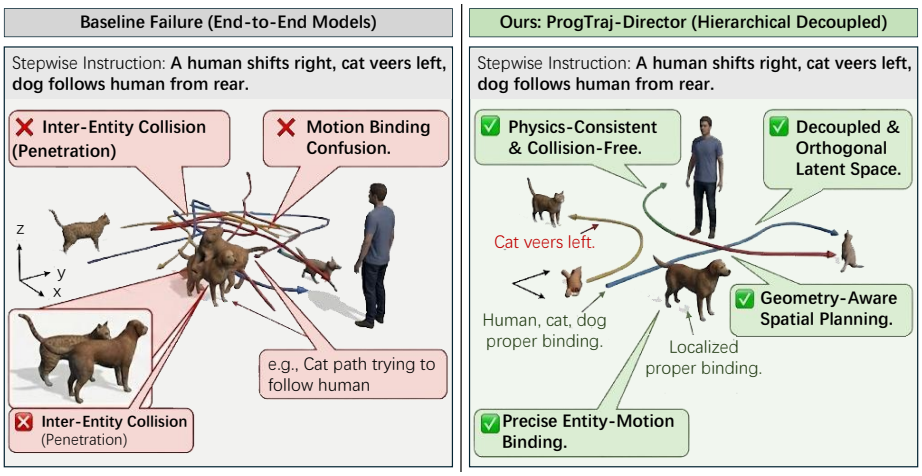

Fig. 1: Qualitative comparison between end-to-end baselines and our ProgTraj-Director. End-to-end baselines suffer from inter-entity collisions (penetration) and motion binding confusion. In contrast, our hierarchical decoupled framework achieves physics-consistent, collision-free trajectories with precise entity-motion binding via geometry-aware spatial planning.

in highly dynamic scenes). In such scenarios, purely pixel-level appearance fitting is fundamentally insufficient. The underlying backbone of any logical physical interaction is inevitably the continuous 3D motion trajectories of multiple entities. Therefore, precise multi-entity 3D trajectory generation has emerged as a crucial prerequisite for breaking visual generation limits, achieving highly controllable videos, and paving the way toward embodied world models.

Current text-based trajectory generation primarily relies on vision-language models (e.g., CLIP) to map abstract text to 3D trajectories. However, a critical limitation is often overlooked: *can natural language directly and precisely determine high-dimensional 3D coordinates?* Since text inherently lacks spatial topology and geometric concepts, attempting to directly bridge this "semantic-physical gap" frequently leads to object interpenetration and motion drifting.

Furthermore, multi-entity scenes introduce a more intractable challenge: *does the model truly know "who" is doing "what"?* Because conventional text encoders mix entire sentences without structured isolation, the visual features of one entity (e.g., a "cat") easily entangle with the action features of another (e.g., a "dog running left"). This makes the model highly susceptible to severe feature leakage and action misattribution (see Baseline Failure in Fig. 1).

Fundamentally, to achieve the physically consistent and collision-free generation demonstrated by our *ProgTraj-Director* (Fig. 1, right), solving multi-entity trajectory generation requires addressing two basic questions: 1) **Entity Binding (Who does what):** How to accurately assign action instructions to their corresponding visual entities without feature confusion? 2) **Spatial Planning**

(Where to go): How to translate spatial-agnostic semantics into collision-free, continuous trajectories?

To address entity binding, we argue that standard feature fusion struggles with semantic ambiguity. The solution lies in explicit physical property disentanglement: treating "entities" as static visual identities and "actions" as dynamic spatial displacements. Mixing them in a shared space causes entanglement. Thus, we propose the Structured Vision-Language Alignment (StructVLA) module. It features a static branch to extract entity visual masks (identity cues) and a dynamic branch utilizing a Spatial Motion Flow to encode directional action priors. Instead of simple concatenation, we apply orthogonal tensor fusion in the latent space. This mathematically encourages the perpendicularity of feature subspaces for unrelated entities, ensuring visual features are strictly bound to their corresponding actions, thereby substantially mitigating feature leakage.

For spatial planning, directly mapping semantics to 3D coordinates suffers from high ambiguity. We propose a Hierarchical Spatial Planning mechanism that uses a dimension-reduced geometric blueprint rather than directly regressing discrete coordinates. Before predicting 3D coordinates, our model constructs a Geometry-Aware Bird's-Eye View (BEV) on a 2D plane. This compact spatial representation pre-solves relative occupancy and anti-collision topological relations. We then formulate 3D trajectory generation as a Progressive Spatial Denoising Process guided by this BEV prior. Constrained by explicit spatial topology, the diffusion process transitions from unconstrained coordinate fitting to a structured restoration of collision-free, temporally coherent trajectories. This combination of 2D geometric planning and 3D denoising enables highly stable and spatially consistent multi-entity trajectory generation.

The main contributions of this paper are summarized as follows:

- We analyze feature leakage and the semantic-physical gap in multi-entity trajectory generation, showing that insufficient disentanglement and spatial reasoning lead to interaction collapse.
- We propose **ProgTraj-Director**, which combines StructVLA for entity-action disentanglement with a Geometry-Aware BEV prior for hierarchical 3D spatial planning.
- We construct **Stepwise-ME**, a large-scale dataset with over 31,000 interactive video clips and fine-grained stepwise annotations, and demonstrate improved spatial consistency and binding accuracy in extensive experiments.

2 Related Work

Trajectory generation. Early research in trajectory generation focused on geometric modeling [11] and rule-based systems [1,5]. With deep learning, methods like drone path prediction [18] emerged, though lacking fine control. In controllable filmmaking, Jiang et al. [7,8] improved camera path accuracy via Toric coordinates and keyframes. Recently, E.T. [3] joined camera and object path generation under text descriptions. Concurrently, latest works [6, 20] explicitly model multi-object semantics and dynamic reasoning. Even though these methods improved control,

they often conflate cinematic camera styles with complex object interactions, lacking explicit 3D spatial grounding for collision avoidance. We therefore propose a trajectory generation model for multi-object interaction. This model directly represents dynamic interactions, simultaneously generating multiple realistic 3D paths where each follows its own logic but is explicitly coordinated with the others.

Trajectory-based Applications. Recently, trajectory-guided video generation has gained popularity due to its interactivity [13, 16, 21]. While existing methods mostly focus on 2D patterns or camera poses [15, 19], 3DTrajMaster [17] enables effective control of multi-entity motions via 3D trajectories. To push this boundary, latest frameworks incorporate 3D-aware kinematic projections [24] for perspective-correct guidance. Furthermore, pioneering world models [26] demonstrate that synthesizing coherent trajectories serves as a fundamental proxy for simulating interactive world dynamics. Inspired by this paradigm, we aim to generate precise multi-entity 3D trajectories aligned with structured text. By establishing a reliable geometric blueprint, our approach effectively bridges the pervasive semantic-physical gap in current multimodal architectures. Ultimately, these physically plausible paths serve as indispensable structural priors for highly interactive video synthesis.

3 Preliminary and Motivation

Our approach is driven by a fundamental question: *why do current end-to-end multimodal models fail catastrophically when generating complex multi-entity 3D trajectories?* To delve into the root causes, we conducted empirical analyses on existing baselines using our Stepwise-ME dataset, revealing two critical bottlenecks hindering multi-entity interactions: **latent semantic entanglement** and **geometric uncertainty**.

Which entities are executing the actions? Our tests reveal a persistent **"static bias"** in mainstream CLIP-based models; as shown in Fig. 2(a), these models exhibit high sensitivity to static content (nouns) but remain unresponsive to motion verbs. This bias manifests as a logical **"role binding collapse"** (i.e., the inability to correctly assign specific actions to corresponding subjects). As illustrated in Fig. 2(b), swapping entity roles in an interaction (e.g., $Cat \leftrightarrow Dog$) results in negligible embedding changes, indicating a failure to anchor motions to specific IDs. Simultaneously, at the feature level, attention leakage occurs, where motion-related weights erroneously bleed into the spatial regions of inactive entities (Fig. 2(c)). This confirms that end-to-end models struggle in multi-entity scenarios due to a lack of thorough physical property disentanglement and orthogonal isolation in the latent space.

Can semantics directly regress 3D coordinates? Even after resolving "who does what," existing models typically attempt to map semantic features directly to high-dimensional 3D physical coordinates (x, y, z). To evaluate the reliability of this end-to-end mapping, we tracked the inter-entity collision rates and spatial geometric errors of baseline trajectories over time. As illustrated in Fig.

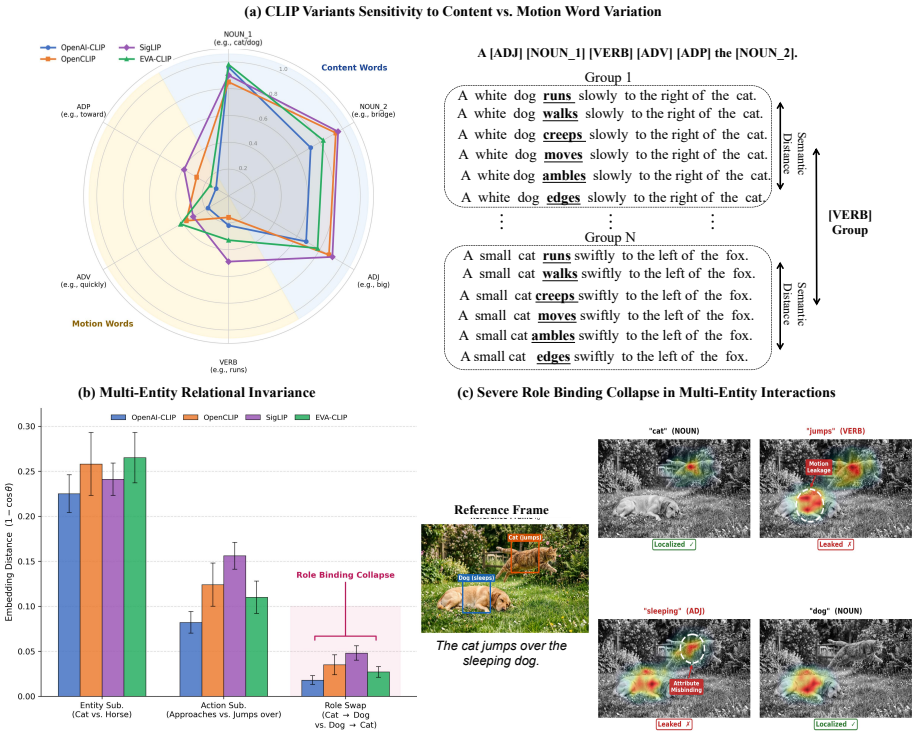

Fig. 2: Empirical analysis of semantic entanglement and static bias. (a) Radar charts show a "static bias" where CLIP variants favor static content words over dynamic motion verbs. (b) Relational tests reveal a role binding collapse when swapping entity roles (e.g., *Cat* ↔ *Dog*). (c) Attention heatmaps visualize the resulting semantic leakage and attribute misbinding.

3, although initial frames maintain reasonable relative positions, geometric errors accumulate exponentially as the sequence lengthens, particularly along the depth (Z) axis, leading to frequent trajectory overlaps and physical penetrations. Since abstract semantic embeddings inherently lack explicit topological boundaries, directly bridging the semantic-physical gap is highly unreliable. This inevitably requires a hierarchical spatial planning mechanism, utilizing a dimension-reduced geometric blueprint (e.g., 2D BEV) as an explicit anti-collision prior before 3D temporal elevation.

4 Methodology

4.1 Overall Architecture and Task Formulation

Given a multi-entity textual instruction $\mathcal{T}$ and a reference video frame $\mathcal{I}_{ref}$, our task is to generate temporally coherent and physically plausible 3D motion trajectories $\mathbf{X} = \{\mathbf{X}^1, \ldots, \mathbf{X}^N\}$, where $\mathbf{X}^n \in \mathbb{R}^{L\times 3}$ denotes the continuous 3D spatial coordinates of the n-th entity over L frames.

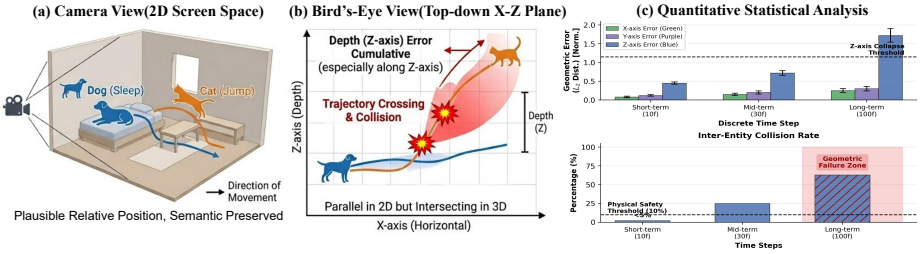

Fig. 3: The semantic-physical gap and geometric uncertainty. (a-b) Depth (Z-axis) ambiguity results in physical collisions despite plausible 2D projections. (c) Statistical analysis shows that geometric errors and inter-entity collision rates accumulate exponentially over longer sequences due to the lack of explicit spatial priors.

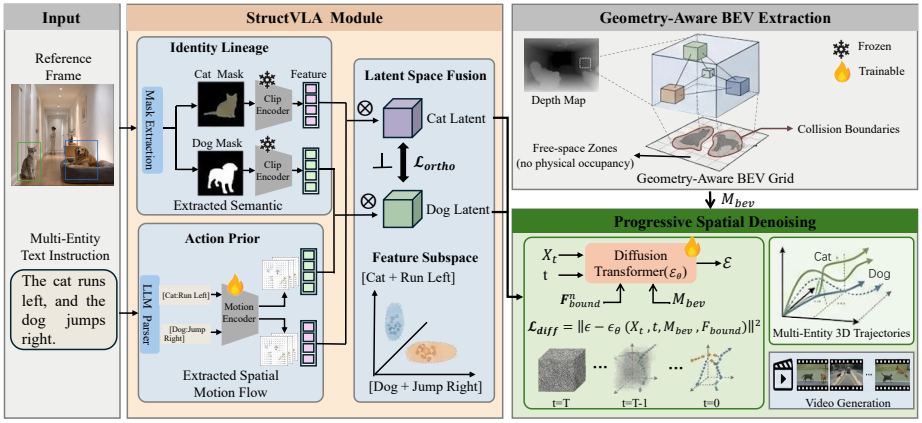

Fig. 4: Overview of the ProgTraj-Director framework. Our model integrates a StructVLA module for identity-action decoupling and a hierarchical spatial planning mechanism guided by Geometry-Aware BEV priors.

As illustrated in Fig. 4, the proposed ProgTraj-Director tackles the semantic-physical gap through a progressive reasoning pipeline. It consists of two pivotal stages: (1) **Structured Vision-Language Alignment (StructVLA)**: explicitly disentangles static identity and dynamic motion, binding them via orthogonal tensor fusion to reduce cross-entity feature leakage; and (2) **Hierarchical Spatial Planning**: extracts explicit depth geometry to construct a compact Bird's-Eye View (BEV) blueprint, subsequently guiding the 3D diffusion-based spatial denoising process toward more physically plausible spatial layouts.

4.2 Structured Vision-Language Alignment (StructVLA)

Conventional vision-language models inherently encode scene elements into highly entangled latent representations, failing to answer "who does what" in composi-

tional multi-entity scenes. To alleviate this, StructVLA promotes heterogeneous physical property disentanglement followed by orthogonal intent binding.

Static-Dynamic Heterogeneous Disentanglement. We explicitly bifurcate the representation extraction into a static identity branch and a dynamic motion branch. For the *Static Identity Branch*, to securely anchor the visual subject, we deploy an open-vocabulary segmentation paradigm (e.g., Grounding DINO combined with SAM) to extract a precise boolean mask $\mathcal{M}^n$ for each entity. The static identity feature is then derived via a visual encoder $\mathcal{E}_{\text{vis}}$:

$$\mathbf{f}_{\text{id}}^n = \mathcal{E}_{\text{vis}}\left(\mathcal{I}_{\text{ref}} \odot \mathcal{M}^n\right) \in \mathbb{R}^d. \tag{1}$$

For the Dynamic Action Branch, to overcome the static bias of pre-trained text encoders, we leverage a Large Language Model (LLM) to parse the raw instruction $\mathcal{T}$ into structured, verb-centric action tuples $\mathcal{T}_{\text{act}}^n$. Subsequently, we introduce a dedicated *Spatial Motion Prior Network* $\mathcal{E}_{\text{motion}}$, which maps the action tuple into a directional spatial motion feature:

$$\mathbf{f}_{\text{act}}^n = \mathcal{E}_{\text{motion}}\left(\mathcal{T}_{\text{act}}^n\right) \in \mathbb{R}^d. \tag{2}$$

This feature encodes the motion direction and kinematic prior for each entity.

Orthogonal Tensor Fusion. Having acquired the decoupled features, we discard naive concatenation and propose a factorized tensor fusion mechanism in the latent space. The bound multimodal intent $\mathbf{F}_{\text{bound}}^n$ is computed via the outer product of the heterogeneous features, projected by a learnable weight matrix $\mathbf{W}_{\text{f}}$:

$$\mathbf{F}_{\text{bound}}^n = \mathbf{W}_{\text{f}}\,\text{vec}\left(\mathbf{f}_{\text{id}}^n \otimes \mathbf{f}_{\text{act}}^n\right) \in \mathbb{R}^D. \tag{3}$$

To reduce motion leakage across multiple entities, we impose an orthogonal regularization. By penalizing the cosine similarity between the intent embeddings of distinct entities, we force their feature subspaces to be mutually perpendicular:

$$\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\left|\left\langle \mathbf{F}_{\text{bound}}^i, \mathbf{F}_{\text{bound}}^j \right\rangle\right|}{\left\|\mathbf{F}_{\text{bound}}^i\right\|_2 \left\|\mathbf{F}_{\text{bound}}^j\right\|_2}. \tag{4}$$

Minimizing $\mathcal{L}_{\text{ortho}}$ encourages the intent embeddings of different entities to occupy more separable latent subspaces, thereby promoting more reliable entity-action attribution and reducing cross-entity motion leakage.

Co-moving Entities and Complexity. It is worth noting that the orthogonal regularization is applied to the latent intent embeddings rather than directly to the generated trajectories. Therefore, it does not prevent different entities from executing correlated or co-moving motions. Instead, the regularization encourages entity-conditioned intent features to remain distinguishable before trajectory decoding, thereby reducing cross-entity motion leakage. In practice, the pairwise term in Eq. (4) is computed by a batched Gram matrix over all entity intent embeddings, resulting in $O(N^2)$ complexity with respect to the number of entities. Since typical interactive scenes contain only a small number of foreground entities, this cost is minor compared with the diffusion denoising process.

4.3 Hierarchical Spatial Planning

Bridging the semantic-physical gap via direct 3D coordinate regression is highly susceptible to physical collisions. We formulate the generation process as a hierarchical spatial planning mechanism, transitioning from explicit 2D topological mapping to fine-grained 3D temporal denoising.

Geometry-Aware BEV Blueprint. Given the reference frame $\mathcal{I}_{\text{ref}}$, we utilize a state-of-the-art monocular depth foundation model (e.g., Depth Anything) to predict a high-quality depth map $\mathcal{D}_{\text{ref}}$. Simultaneously, we extract dense visual features $\mathbf{F}_{\text{vis}} \in \mathbb{R}^{H \times W \times C}$ from $\mathcal{I}_{\text{ref}}$ using a pre-trained visual backbone. To translate the 2D image domain into a 3D physical space, we introduce a pseudo-camera intrinsic matrix $\mathbf{K}$. For the i-th pixel at coordinate (u, v), its corresponding 3D spatial coordinate $\mathbf{P}^i_{\text{3D}}$ is computed via unprojection:

$$\mathbf{P}^i_{\text{3D}} = \mathcal{D}_{\text{ref}}(u, v) \cdot \mathbf{K}^{-1} \left[u, v, 1\right]^\top . \tag{5}$$

Each unprojected 3D point inherits its corresponding 2D visual feature, denoted as $\mathcal{F}(i) = \mathbf{F}_{\text{vis}}(u, v)$. By traversing all pixels, we obtain a feature-rich dense 3D point cloud $\mathcal{P}$ of the entire scene. Subsequently, to reduce redundant information along the vertical direction, we perform voxel pooling and project the 3D point cloud features $\mathcal{F}$ onto the top-down X-Z plane, generating the 2D BEV grid $\mathbf{M}_{\text{bev}} \in \mathbb{R}^{H' \times W' \times C}$:

$$\mathbf{M}_{\text{bev}}(x, z) = \max_{i \in \Omega_{(x,z)}} \mathcal{F}(i), \tag{6}$$

where $\Omega_{(x,z)} = \{i \mid \pi_{\text{xz}}(\mathbf{P}^i_{\text{3D}}) = (x, z)\}$ denotes the set of point indices projected to the BEV grid (x, z), and $\pi_{\text{xz}}(\cdot)$ is the top-down projection function. Plausible 2D screen-space layouts may still suffer from depth ambiguity, trajectory crossing, and inter-entity collisions in 3D space.

Therefore, projecting scene geometry into a top-down BEV grid preserves depth-aware spatial topology and provides coarse occupancy cues for collision-aware trajectory planning. Thus, $\mathbf{M}_{\text{bev}}$ serves as a compact geometric prior for the subsequent 3D diffusion process.

Dynamic BEV Update and Implementation Details. Our BEV prior is constructed in a camera-aligned coordinate frame and is recomputed when a new reference frame or keyframe is available, allowing it to adapt to camera motion and visible scene geometry. For datasets without calibrated intrinsics, we use a pseudo-camera intrinsic matrix K based on the image resolution and a normalized focal length, with depth values clipped and normalized before unprojection. The resulting BEV mainly provides coarse topological cues, such as relative occupancy, free space, and inter-entity boundaries, rather than relying on perfect metric depth.

Progressive Spatial Denoising. With the multi-entity intent and the geometric blueprint established, we formulate the 3D trajectory synthesis as a diffusion-based generative process. The Diffusion Transformer (DiT) ϵ_θ predicts the added noise under dual conditioning. Crucially, we adopt a Decoupled Condition Injection

Table 1: Comparison of Stepwise-ME with other datasets. "Prompt Type" refers to text annotation forms: static descriptions and the proposed motion-centric prompts.

Dataset	Traj Type	Prompt Type	#Samp.	#Frame	Avg (s)
MVImgNet [22]	S. Ent./Scene-Cen.	Stat./Sing.	22K	6.5M	10
RealEstate10k [25]	S. Ent./Scene-Cen.	Stat./Sing.	27K	11M	5.5
DL3DV-10K [10]	S. Ent./Scene-Cen.	Stat./Sing.	10K	5.1M	85
CCD [8]	S. Ent./Track.	Stat./Sing.	25K	4.5M	7.2
E.T. [3]	S. Ent./Track.	Stat./Sing.	15K	1.3M	3.8
DataDop [23]	S. Ent./Free-Mov.	Stat./Mult.	29K	13M	14.4
Stepwise-ME (Ours)	**M. Ent./Free-Mov.**	**Mot./Step.**	**31K**	**8M**	**3.3**

Strategy: $\mathbf{F}_{\text{bound}}$ is injected to govern the *semantic direction*, while the BEV prior $\mathbf{M}_{\text{bev}}$ is integrated as an explicit spatial prior to guide the geometric layout.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{X}_0, \epsilon \sim \mathcal{N}(0,\mathbf{I}), t} \left[\left\| \epsilon - \epsilon_\theta \left(\mathbf{X}_t, t, \mathbf{F}_{\text{bound}}, \mathbf{M}_{\text{bev}} \right) \right\|_2^2 \right]. \tag{7}$$

This progressive spatial denoising translates the coarse-grained 2D BEV planning into continuous 3D physical trajectories with improved spatial coherence.

4.4 Training Objective

The overall training objective of ProgTraj-Director is formulated as the joint optimization of the spatial diffusion loss and the orthogonal intent regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{ortho}}. \tag{8}$$

Compared to prior methods burdened with heuristic proxy losses, our objective is mathematically succinct, entirely driven by representation orthogonality and explicit physical constraints.

5 Stepwise-ME Dataset

To fundamentally address the challenges of multi-entity attribution and spatial planning, we introduce **Stepwise-ME**, a large-scale 3D multi-entity motion dataset featuring stepwise temporal annotations.

5.1 Motivation and Fundamental Uniqueness

The primary motivation for constructing Stepwise-ME stems from the critical limitations of existing benchmarks. As compared in Table 1, current large-scale datasets (e.g., RealEstate10k, DL3DV-10K) are predominantly scene-centric, capturing camera ego-motion rather than dynamic object interactions. While datasets like CCD and E.T. provide object trajectory data, they are largely

restricted to single-entity tracking and rely on static, sentence-level global descriptions. These static prompts inherently fail to capture the sequential nature of complex interactions (e.g., "the cat turns left in step 1, then the dog approaches in step 2").

Stepwise-ME fundamentally bridges this gap. By providing *Multi-Entity*, *Free-Moving* trajectories paired with *Motion-centric, Stepwise* text prompts, it serves as the first dataset strictly designed to support fine-grained, spatio-temporally aligned multi-entity trajectory generation.

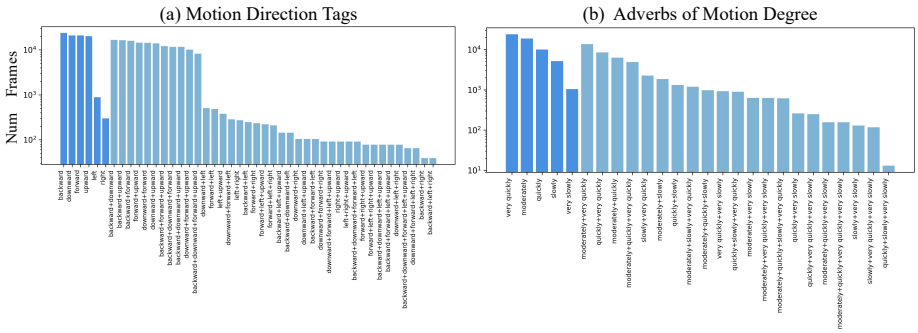

Fig. 5: Dataset Statistics. The figure illustrates the composition and distribution of 45 translation motions **(a)** and 26 speed motions **(b)**, emphasizing the complexity and diversity of trajectories in our dataset.

5.2 Scale and Statistical Complexity

Data Scale. Stepwise-ME comprises over 31,000 annotated video samples, totaling approximately 8 million frames across 12 synchronized views. To reduce temporal redundancy and computational overhead, the original videos are uniformly downsampled to 8 FPS during the trajectory extraction phase. With an average duration of 3.3 seconds per sample, it provides sufficient granularity for modeling continuous and complex temporal dynamics.

Motion Complexity. The core challenge of our dataset lies in its compositional diversity. As illustrated in Fig. 5, the stepwise annotations yield a highly complex motion vocabulary. We explicitly extract and decouple 45 distinct spatial translation combinations (Fig. 5a) and 26 motion speed modifiers (Fig. 5b). This combinatorial explosion of directions and speeds across multiple entities explicitly forces the generative models to learn precise identity-motion binding and rigorous spatial anti-collision planning, preventing them from merely relying on dataset bias or pattern memorization. For more comprehensive details regarding the dataset construction and statistics, please refer to the supplementary material.

6 Experiments

6.1 Implementation Details

Our ProgTraj-Director generates multi-entity 3D trajectories conditioned on step-wise text and multi-view observations. All experiments are conducted on a single NVIDIA A100 (40GB) GPU. We employ a 28-layer Diffusion Transformer (DiT) backbone with 16 attention heads. The model is trained on the full Stepwise-ME dataset (31K video samples, ∼8M frames). Both trajectories and textual instructions are tokenized and strictly aligned in the temporal domain prior to training.

Visualization Protocol. The illustrative trajectory scenes in the main paper are GPT-assisted visual renderings based on real trajectory data or model-generated 3D coordinates. They are used only for visualization purposes; all quantitative results are computed on raw coordinates, with raw coordinate-level visualizations provided in the supplementary material.

6.2 Quantitative Results

Evaluation Metrics. We evaluate three metric groups. Text-Traj. Alignment uses F1-score (F1↑) and CLaTr-CLIP (CLIP↑) [23]. Trajectory Quality uses Coverage (Cov.↑) for distributional diversity, CLaTr-FID (FID↓) [23] with MotionCLIP [12] features, and Collision Rate (CR↓) for inter-entity penetrations. Classification reports Accuracy (Acc.↑) for latent Entity-Action Binding.

Baselines and Protocols. We compare *ProgTraj-Director* against state-of-the-art baselines: CCD [8], E.T. [3], Director3D [9], LANGTRAJ [2], and Lang2Motion [4]. For a fair comparison on the multi-entity Stepwise-ME dataset, single-entity baselines follow an Independent Aggregation Protocol, where trajectories are generated individually before being projected into a joint 3D space. Notably, the Acc. metric is not applicable (N/A) to these baselines. Because they are fundamentally identity-agnostic and lack dedicated latent slots, they cannot perform the in-situ multi-entity binding inherently addressed by our StructVLA.

Results Analysis. As demonstrated in Table 2, *ProgTraj-Director* demonstrates competitive performance across the evaluated benchmarks. On the Stepwise-ME dataset, our method significantly outperforms the baseline, Lang2Motion, in both semantic alignment and distributional quality. Crucially, our method substantially reduces the Collision Rate (CR) compared to the baselines. This gap validates our motivation: while baselines generate trajectories without mutual awareness, leading to severe physical violations, our Geometry-Aware BEV Prior provides an explicit anti-collision blueprint that enforces spatial constraints. Furthermore, our superior classification accuracy (Acc.) proves that decoupled latent optimization effectively resolves complex motion attribution, ensuring that each entity strictly follows its respective instruction without feature leakage.

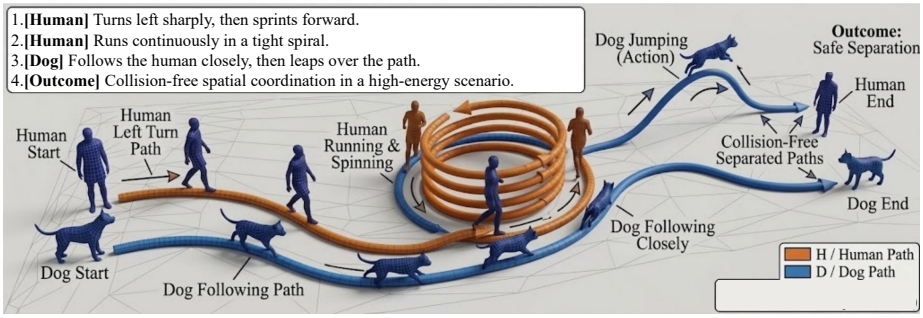

Fig. 6: Intention-guided multi-entity trajectory generation in high-dynamic scenarios. Our method produces coherent 3D trajectories for complex motions while maintaining strong spatial coordination.

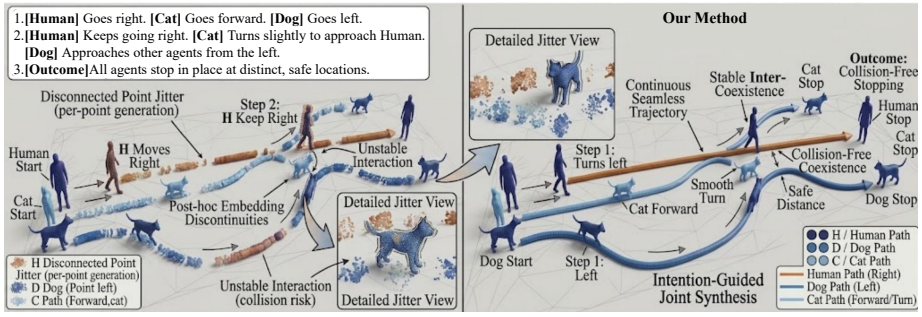

Fig. 7: Qualitative comparison with point-wise generation baselines. Our method produces smoother and more physically plausible trajectories aligned with step-wise instructions.

6.3 Qualitative Results.

Fig. 6 shows complex high-dynamic scenarios, such as tight spiral running and entity leaping. This demonstrates that *ProgTraj-Director* maintains strict spatial coordination and collision-free generation even during high-energy interactions. Fig. 7 shows a qualitative comparison. This demonstrates that, compared to the severe discontinuities and spatial jitter caused by baseline methods, our method stably produces smooth, continuous, and physically plausible 3D paths.

6.4 Ablation Studies

As shown in Table 3, each component is crucial for overall performance. (1) **BEV Prior:** Removing it spikes the Collision Rate (CR) from 7.3% to 19.8%, highlighting its necessity in providing explicit spatial constraints for collision avoidance. (2) **Action Prior:** Its exclusion drops classification accuracy (Acc.) from 0.96 to 0.68, confirming its essential role in precisely attributing motion features to specific entity IDs. (3) **Identity Lineage:** Without it, the F1-score

Table 2: Quantitative comparison of trajectory generation performance. We evaluate our method against state-of-the-art baselines on DataDop and Stepwise-ME.

Method	Dataset	Text-Traj. Align		Traj. Quality			Class.
		F1↑	CLIP↑	Cov.↑	FID↓	CR↓	Acc.↑
Evaluation on Single-Entity/General Dataset (DataDop)							
CCD [8]	DataDop	0.292	5.77	0.291	68.38	–	–
E.T. [3]	DataDop	0.344	2.08	0.033	69.37	–	–
Director3D [9]	DataDop	0.077	0.01	0.196	82.68	–	–
GenDoP [23]	DataDop	0.414	33.81	0.584	72.05	–	–
ProgTraj-Director (Ours)	DataDop	**0.416**	**35.12**	**0.602**	**69.10**	–	–
Evaluation on Multi-Entity Interaction Dataset (Stepwise-ME)							
GenDoP [23]	S-ME	0.421	34.69	0.590	69.34	38.5	–
NaviDIFF [14]	S-ME	0.539	37.28	0.635	67.21	24.2	–
LANGTRAJ [2]	S-ME	0.716	42.58	0.793	52.68	15.8	–
Lang2Motion [4]	S-ME	0.847	48.86	0.820	46.13	16.4	–
ProgTraj-Director (Ours)	**S-ME**	**0.875**	**53.11**	**0.913**	**43.65**	**7.3**	**0.96**

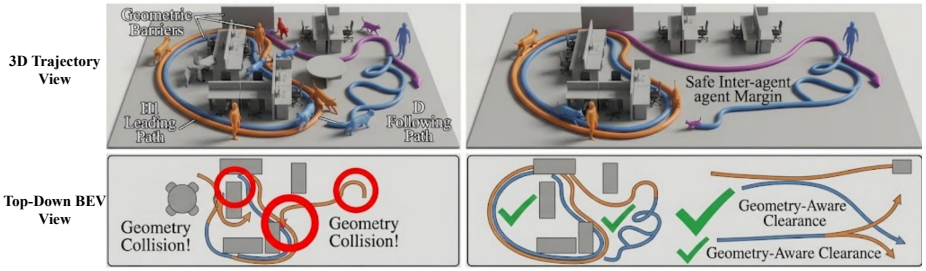

Fig. 8: Ablation of the Geometry-Aware BEV prior. Without BEV guidance, trajectories suffer from environmental and inter-agent collisions, while the full model better preserves spatial margins and reduces collisions.

decreases to 0.718 and FID worsens to 58.74, indicating its critical role in semantic coherence and trajectory realism across multi-step instructions. The "Vanilla Diffusion" baseline exhibits the worst performance across all metrics, highlighting the synergistic effect of our framework.

Furthermore, Fig. 8 qualitatively demonstrates the impact of the BEV Prior. Without it, the model produces trajectories suffering from severe environmental and inter-agent collisions (left). By integrating this prior, our framework better preserves safe inter-agent margins and produces smoother, lower-collision trajectories within complex spatial layouts (right).

6.5 Application: Structuring World Models for Long-Horizon Interactive Video

Recent video diffusion models can synthesize visually realistic content, but long-horizon multi-entity interaction remains difficult without explicit 3D spa-

Table 3: Ablation Study. We evaluate the contribution of each key component: Geometry-Aware BEV Prior , Action Prior, and Identity Lineage.

Variant	$\mathcal{M}_{bev}$	$\mathcal{A}_{prior}$	$\mathcal{L}_{id}$	Text-Traj		Traj Quality		Class.
				F1↑	CLIP↑	FID↓	CR↓	Acc.↑
Full Model	✓	✓	✓	**0.875**	**53.11**	**43.65**	**7.3**	**0.96**
w/o BEV Prior	×	✓	✓	0.862	51.45	54.21	19.8	0.94
w/o Action Prior	✓	×	✓	0.743	42.08	51.08	11.2	0.68
w/o Identity Lineage	✓	✓	×	0.718	39.52	58.74	9.5	0.82
Vanilla Diffusion	×	×	×	0.401	33.50	70.59	40.7	–

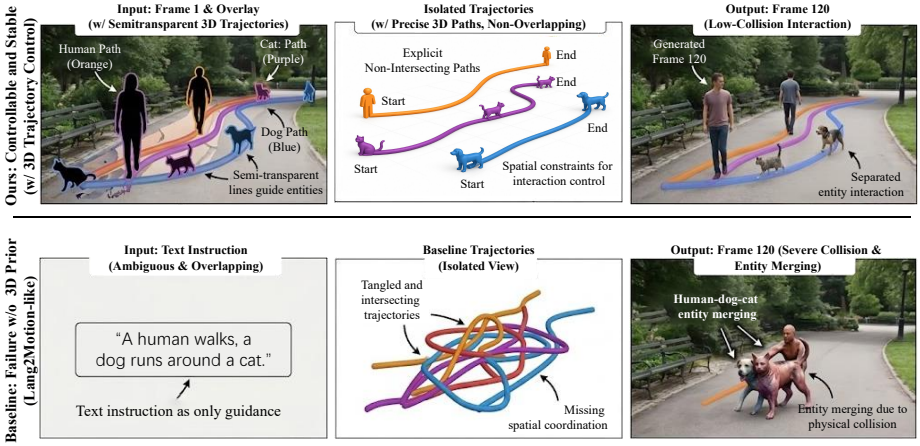

Fig. 9: Trajectory-guided versus text-only multi-entity video generation. Without 3D priors (Bottom), text-only generation suffers from tangled trajectories, missing spatial coordination, and entity merging. With 3D trajectory control (Top), ProgTraj-Director provides geometric guidance that improves entity separation and physical plausibility.

tial guidance. As illustrated in Fig. 9, text-only generation often suffers from semantic-physical collapse, including entity entanglement, identity drift, and severe collisions. Our *ProgTraj-Director* provides lower-collision 3D trajectories as geometric guidance for downstream video generators. By injecting these physically plausible paths, the generated videos better preserve entity separation and motion consistency over time. This suggests that structured trajectory priors can serve as an effective bridge toward more controllable embodied world models. More qualitative video results are provided in the supplementary material.

Capability Boundaries. As shown in Fig. 10, ProgTraj-Director still has clear degradation boundaries in challenging scenarios. Crowded interactions may compress trajectories and introduce short-term near-collisions, while severe occlusions or inaccurate masks can cause local binding errors. Long compositional instructions may also accumulate temporal lag and drift, suggesting that more robust perception and long-horizon planning remain important future directions.

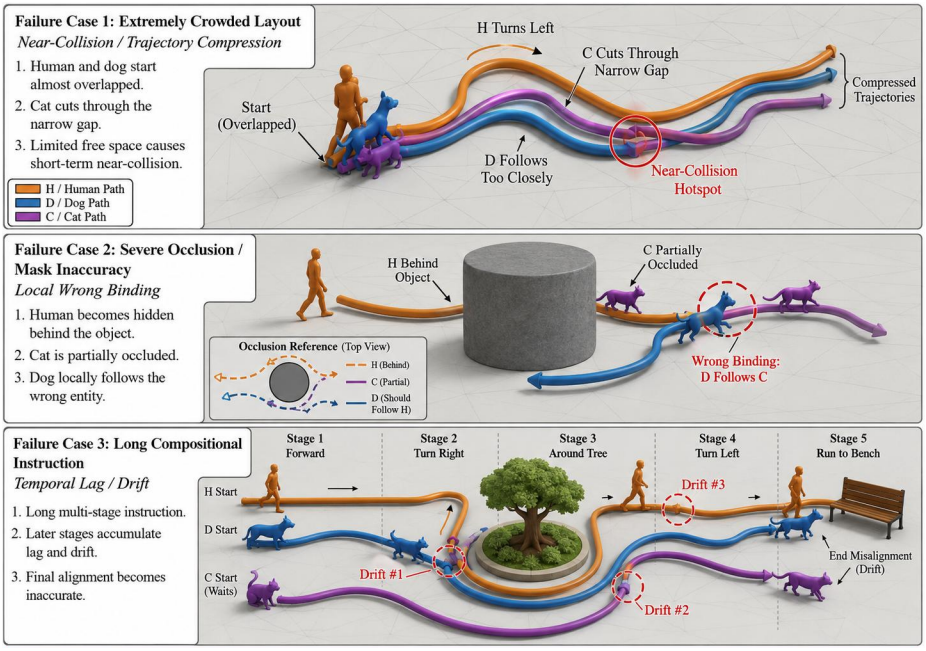

Fig. 10: Capability boundaries of ProgTraj-Director. Our method improves binding and spatial consistency, but may degrade under crowded scenes, severe occlusions or mask errors, and long ambiguous instructions.

7 Conclusion

In this paper, we propose ProgTraj-Director, a framework for multi-entity 3D trajectory generation toward Embodied World Models. To bridge the semantic-physical gap, we introduce StructVLA to disentangle static identity cues from dynamic motion intents, enabling more reliable identity-action binding in complex interactions. Furthermore, our Geometry-Aware BEV-guided hierarchical planning mechanism provides an explicit spatial prior for progressive 3D trajectory denoising, translating abstract text instructions into more physically plausible trajectories with fewer collisions. Extensive experiments on the Stepwise-ME dataset show that our method achieves favorable performance over existing baselines in semantic alignment, binding accuracy, trajectory quality, and collision reduction. By providing structured and physically grounded trajectory priors, ProgTraj-Director offers a practical step toward more controllable and interactive video synthesis in complex dynamic scenes.

Acknowledgement

This work was supported by the Institute Innovation Project of the Institute of Computing Technology, Chinese Academy of Sciences under Grant No. E561090.

References

1. Blinn, J.: Where am i? what am i looking at?(cinematography). In IEEE Computer Graphics and Applications **8**(4), 76–81 (2002)
2. Chang, W.J., Zhan, W., Tomizuka, M., Chandraker, M., Pittaluga, F.: Langtraj: Diffusion model and dataset for language-conditioned trajectory simulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26622–26631 (2025)
3. Courant, R., Dufour, N., Wang, X., Christie, M., Kalogeiton, V.: Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In: European Conference on Computer Vision. pp. 464–480. Springer (2024)
4. Galoaa, B., Bai, X., Ostadabbas, S.: Lang2motion: Bridging language and motion through joint embedding spaces. arXiv preprint arXiv:2512.10617 (2025)
5. Galvane, Q., Christie, M., Lino, C., Ronfard, R.: Camera-on-rails: automated computation of constrained camera paths. In: ACM SIGGRAPH Conference on Motion in Games. pp. 151–157 (2015)
6. Ji, L., Zhong, L., Wei, P., Li, C.: Posetraj: Pose-aware trajectory control in video diffusion. arXiv preprint arXiv:2503.16068 (2025)
7. Jiang, H., Wang, B., Wang, X., Christie, M., Chen, B.: Example-driven virtual cinematography by learning camera behaviors. ACM Trans. Graph. **39**(4), 45 (2020)
8. Jiang, H., Wang, X., Christie, M., Liu, L., Chen, B.: Cinematographic camera diffusion model. In: Computer Graphics Forum. vol. 43, p. e15055. Wiley Online Library (2024)
9. Li, X., Lai, Z., Xu, L., Qu, Y., Cao, L., Zhang, S., Dai, B., Ji, R.: Director3d: Real-world camera trajectory and 3d scene generation from text. Advances in Neural Information Processing Systems **37**, 75125–75151 (2024)
10. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
11. Liu, X., Zhang, T., Johnson-Roberson, M., Zhi, W.: Splatraj: Camera trajectory generation with semantic gaussian splatting. arXiv preprint arXiv:2410.06014 (2024)
12. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
13. Wan, Z., Tang, S., Wei, J., Zhang, R., Cao, J.: Dragentity: Trajectory guided video generation using entity and positional relationships. In: ACM International Conference on Multimedia. pp. 108–116 (2024)
14. Wang, W., Yu, C., Wang, Y., Min, B.C.: Human-robot cooperative distribution coupling for hamiltonian-constrained social navigation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 10808–10815. IEEE (2025)
15. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH. pp. 1–11 (2024)
16. Wu, W., Li, Z., Gu, Y., Zhao, R., He, Y., Zhang, D.J., Shou, M.Z., Li, Y., Gao, T., Zhang, D.: Draganything: Motion control for anything using entity representation. In: European Conference on Computer Vision. pp. 331–348. Springer (2024)
17. Xiao, F., Liu, X., Wang, X., Peng, S., Xia, M., Shi, X., Yuan, Z., Wan, P., Zhang, D., Lin, D.: 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In: In International Conference on Learning Representations (2024)

18. Yang, H., Xie, K., Huang, S., Huang, H.: Uncut aerial video via a single sketch. In: Computer Graphics Forum. vol. 37, pp. 191–199. Wiley Online Library (2018)
19. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH. pp. 1–12 (2024)
20. Yariv, G., Kirstain, Y., Zohar, A., Sheynin, S., Taigman, Y., Adi, Y., Benaim, S., Polyak, A.: Through-the-mask: Mask-based motion trajectories for image-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18198–18208 (2025)
21. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
22. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimgnet: A large-scale dataset of multi-view images. In: Computer Vision and Pattern Recognition. pp. 9150–9161 (2023)
23. Zhang, M., Wu, T., Tan, J., Liu, Z., Wetzstein, G., Lin, D.: Gendop: Auto-regressive camera trajectory generation as a director of photography. arXiv preprint arXiv:2504.07083 (2025)
24. Zhang, R., Zhou, J., Xu, Z., Liu, Z., Huang, J., Zhang, M., Sun, Y., Li, X.: Zo3t: Zero-shot 3d-aware trajectory-guided image-to-video generation via test-time training. arXiv preprint arXiv:2509.06723 (2025)
25. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
26. Zhu, Y., Feng, J., Zheng, W., Gao, Y., Tao, X., Wan, P., Zhou, J., Lu, J.: Astra: General interactive world model with autoregressive denoising. arXiv preprint arXiv:2512.08931 (2025)