DRIVING VIDEO GENERATION FROM VIEW-SPACE BLOCKS WORLD WITH OBJECT-LEVEL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generative models promise driving simulators that can render arbitrary traffic scenarios, but this requires both precise control over the scene and consistency across long rollouts. Existing driving generators condition only on traffic participants and road layout, leaving the static world, and with it the ego motion, underspecified, while frame-level memories grow quickly and cannot tell which cached tokens belong to which instance. We present **MemOIR**, a driving video diffusion model conditioned on a *view-space blocks-world representation* that renders every dynamic and static instance as a 3D block encoding its class, identity, and occlusion, curated automatically from raw driving videos. This representation densely specifies both agent motion and the ego trajectory, and enables object-level KV caching: each instance's tokens are cached from its most complete observations and re-embedded where its block projects when it reappears, attending to far fewer tokens than frame-level caching. **MemOIR** achieves the most accurate instance placement and ego-motion control among driving generators with competitive visual quality, faithfully places vehicles performing unseen maneuvers such as spinouts, and substantially improves the identity consistency of returning vehicles. Video results are available at https://memoir-wm-anon.github.io/.

1 INTRODUCTION

Video generative models are rapidly becoming a foundation for world simulation (Bruce et al., 2024; Agarwal et al., 2025), promising driving simulators that can render arbitrary traffic scenarios, including rare and safety-critical ones. Serving this role requires both precise control over what happens in the scene and consistency with what has already been generated as rollouts grow long.

For control, text prompts are too coarse, so prior work introduces explicit conditions such as dense per-frame maps (Zhang et al., 2023a; Chen et al., 2023), physics-based prompts (Gillman et al., 2025; Liu et al., 2024), and point trajectories (Geng et al., 2025). Driving generators (Gao et al., 2025b; Mei et al., 2024) combine separately encoded modalities, such as BEV maps, 3D boxes, and camera poses, that describe traffic participants and road layout but leave the surrounding static world unspecified in view space, limiting controllability, particularly over ego motion. For consistency, long videos are generated chunk by chunk, and existing memories cache past context at the frame level (Ji et al., 2025; Yu et al., 2025): their caches grow quickly and must be pruned, and they cannot tell which cached tokens belong to which instance.

We propose **MemOIR**, a driving video diffusion model conditioned on a unified *view-space blocks-world representation*: every instance, whether dynamic (e.g., vehicles) or static (e.g., buildings, vegetation), is abstracted as a 3D block and rendered into the camera view as a conditioning video that encodes its class, identity, and occlusion. This single representation enables two capabilities. First, it allows **precise controllability**: because the rendered blocks densely fill the view and move consistently with the camera, the condition specifies the ego trajectory as well as each agent's placement and motion, and removing static instances raises ego trajectory error by 40%. Second, it enables **object-level KV caching**, from which our model takes its name: token **Mem**orization of **O**bjects via **I**nstance-level **R**epresentation. Since every block carries an explicit identity and pose, we cache each instance's tokens from its most complete observations and, when it reappears, re-embed them where its block now projects, attending to roughly $24\times$ fewer tokens than frame-level caching.

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.

To achieve this, we introduce a curation pipeline that derives blocks worlds from raw driving videos with off-the-shelf segmentation and geometry models, requiring no manual 3D annotation (Sec. 3.1); train a control branch on a frozen video DiT to condition generation on the representation (Sec. 3.2); and equip it with object-level KV caching for consistent chunk-wise rollout (Sec. 3.3). On nuScenes, **MemOIR** achieves the most accurate instance placement and ego-motion control among driving generators with competitive visual quality, places vehicles performing unseen maneuvers such as spin-outs at their commanded locations in 98.1% of frames, and raises re-identification accuracy of returning vehicles from at most 11.0% to 45.6%. Our model opens door to wide range of applications: since the blocks world is all the model requires, new scenes can be composed directly from blocks, while an optional initial frame enables instance-level editing and extension of real footage.

2 Related Work

2.1 Driving-Scene Video Diffusion Models

Recent advances in video diffusion models have enabled high-quality video synthesis (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2024; Wan et al., 2025), motivating their use for controllable driving-scene generation. In this domain, structured conditions provide more explicit control over scene layout and motion than text prompts alone. The MagicDrive lineage (Gao et al., 2023; 2025a) uses BEV map, 3D bounding boxes and camera parameters for multi-view scene-street generation. Panacea (Wen et al., 2024) and DreamForge (Mei et al., 2024) also explore structured conditioning for controllable multi-view driving video generation. More Recently, Cosmos (Agarwal et al., 2025) and SWM Seo et al. (2026) provides video foundation models for world simulation, while OmniDreams (NVIDIA et al., 2026) targets action-conditioned driving simulation. However, these methods either focus on controlling when extending control to surrounding scene elements, use heterogeneous representations such as object bounding boxes and BEV maps rather than a single view-aligned scene representation.

2.2 Explicit Scene and Motion Control in Generative Models

To enable more precise control over generated scenes and motion, prior work introduces structured visual conditioning. One line of work uses dense per-frame signals, such as depth, edge, pose, or sketch maps, to guide generation (Zhang et al., 2023b; Chen et al., 2023; Wang et al., 2023; Kwak et al., 2026), extending the ControlNet paradigm (Zhang et al., 2023a). For 3D spatial control, CC3D (Bahmani et al., 2023) conditions scene generation on top-down semantic layouts derived from object bounding boxes. Another line of work focuses on motion control, using sparse point trajectories (Geng et al., 2025) or physics-based guidance such as force fields and rigid-body simulation (Gillman et al., 2025; Liu et al., 2024). To control human body pose and motion, prior work conditions video generation on human keypoint sequences (Hu et al., 2023; Zhang et al., 2024) or rendered SMPL motion sequences (Nam et al., 2025). Our method integrates spatial layout, object motion, and human pose conditions into a unified control model through a shared blocks-world representation, enabling fine-grained control over scene composition and instance-level motion.

2.3 Long-Term Consistency in Video Generation

World models require extended video generation to represent environments as they evolve over time. A common approach generates videos autoregressively, conditioning each new chunk on previously generated content (Song et al., 2025; Henschel et al., 2025; Chen et al., 2024; Yin et al., 2025; Huang et al., 2026b; Zhang et al., 2025b; Yang et al., 2025), as in the large-scale chunk-wise generator MAGI-1 (Sand AI, 2025) and the driving world model Epona (Zhang et al., 2025a). Since separate chunk-generation calls do not retain past attention states, consistency beyond the most recent frames requires explicit memory, typically realized by caching, compressing or routing past context (Guo et al., 2025; Yu et al., 2025). MemFlow (Ji et al., 2025) selectively retrieves historical KV features, while geometry-indexed memories (Xiao et al., 2026; Li et al., 2025b; Wu et al., 2026) retrieve whole past views by their camera overlap with a largely static scene. None of these associates cached features with explicit instance identities and per-instance geometry. Our view-space blocks representation provides both, enabling object-level KV caching that retrieves each instance's tokens and aligns them spatially, at over an order of magnitude less memory than full-frame caching.

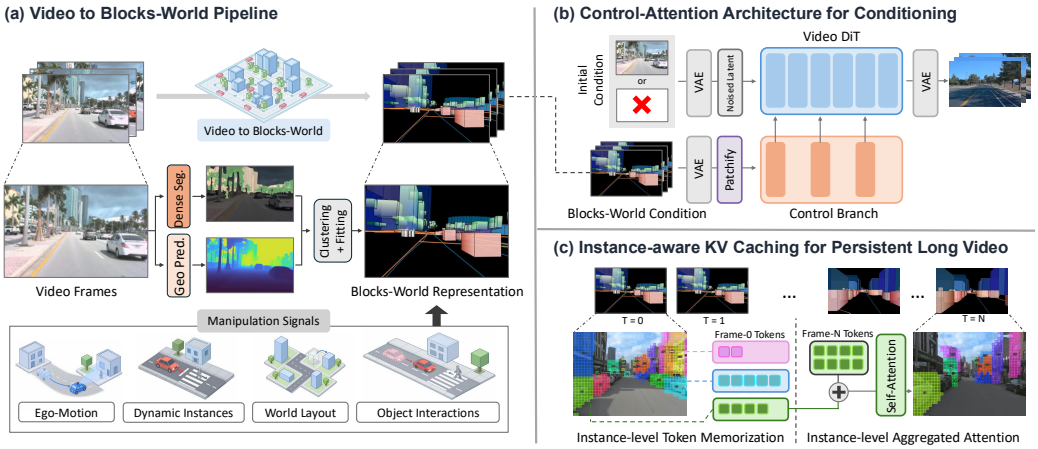

Figure 1: **Method Overview. (a)** Video to blocks-world pipeline: panoptic segmentation and depth information are aggregated into a per-frame blocks-world canvas video. **(b)** A frozen video DiT is conditioned on the canvas through a trainable control branch with zero-initialized residual injection. **(c)** Instance-aware KV caching and retrieval for persistent long-term generation.

3 METHOD

3.1 VIDEO-TO-3D BLOCKS-WORLD DATA CURATION AND CONSTRUCTION

To obtain our view-space blocks world representation, we present a curation pipeline that uses off-the-shelf models to process a video into blocks format with minimal sensor data. Our objective is to obtain a per-instance bounding box for every object constituting the given scene, including static objects such as buildings and vegetation, for which ground-truth 3D labels are scarce and noisy.

Geometry-based dense instance segmentation. Constructing the blocks-world representation requires dense, instance-level segmentation of every object across the video. We find that instance segmentation models such as SAM3 (Carion et al., 2026) are ill-suited for this task: they inevitably miss obscure or heavily occluded instances, and their inference time scales with the number of instances (Carion et al., 2026; Hou et al., 2020), making them impractical for instance-dense domains such as driving scenes (see Appendix Sec. B.4). Rather than relying on image-level instance proposals, we turn to Mask2Former (Cheng et al., 2022), a dense panoptic segmentation model that assigns a class to every pixel. We combine its class masks with the point map predicted by Depth Anything 3 (DA3) (Lin et al., 2025) and perform geometric clustering, which produces a surprisingly accurate instance-segmented point cloud of the scene.

Specifically, we use DA3 to obtain an aggregated per-class point cloud over the entire clip by lifting every class-masked pixel through the aligned pointmap into a canonical world frame via the camera extrinsics and ego poses, accumulating all frames into a single cloud. We apply watershed segmentation (Vincent & Soille, 1991) on a bird's-eye-view height map of the given point cloud, where the aggregated voxels are projected onto a top-down grid where each cell records its maximum height above the ground plane. Local height peaks are identified as instance seeds, and each voxel is assigned to the seed whose slope it lies on, following the watershed algorithm; small clusters are discarded. Since cluster identities live in the canonical world frame, projecting them back to each frame yields temporally consistent per-instance masks by construction without explicit tracking.

View-space block-based condition synthesis. To condition our video generation model, we abstract every instance into a 3D bounding box - an instance-level "block" - which we term a **blocks-world representation**. We render a per-frame conditioning video that encodes each block's projected location, class, instance identity, and inter-instance occlusion. Each instance cluster is fitted with a 3D block oriented along the nearest lane centerline, with extents set to the [0.08, 0.92] voxel quantiles to avoid inflation from noisy geometry. Each block's color is decided by two components: a class-level *hue band* and an instance-level *hue offset*. Each class maps to a base HSV hue h_c

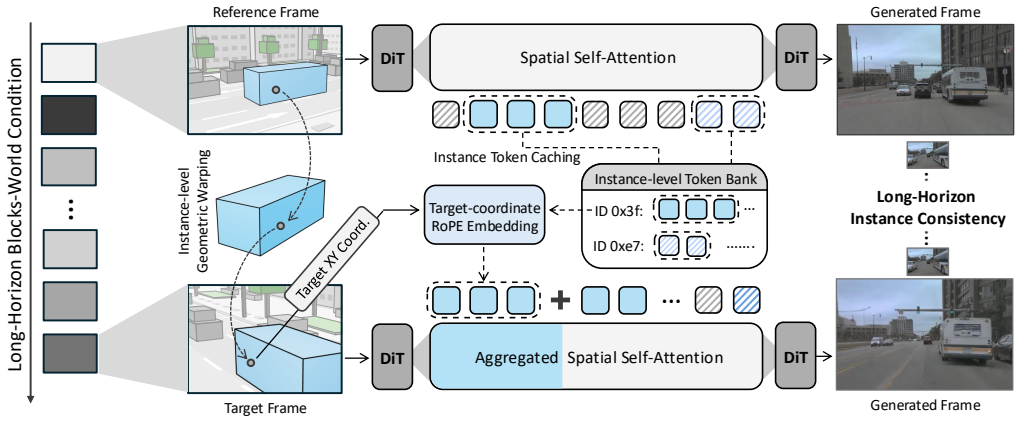

Figure 2: **Instance-level KV Caching.** Our instance-level KV caching leverages the blocks-world representation for efficient and effective long-horizon video generation.

(e.g. `vehicle` → red, `building` → blue), and each instance receives a deterministic $\pm 30°$ jitter around this base keyed to its instance ID (full conventions in the Appendix Sec. B.4). Class is thus recoverable from the broad hue band, while individual instances within a class are distinguishable by their precise hue. Finally, the blocks are projected through each frame's camera into the conditioning video via a global float z-buffer that accounts for inter-instance occlusion.

3.2 Model Architecture and Training

Our blocks-world representation effectively guides the model toward instance-level controllable, photorealistic video generation. As described in (b) of Fig. 1, we train a ControlNet-style (Zhang et al., 2023a) control branch on a frozen Cosmos-Predict2.5 (Ali et al., 2025) backbone, conditioned on the blocks-world video with the standard rectified-flow objective:

$$\mathcal{L} \;=\; \mathbb{E}_{x_0,\,\epsilon,\,\tau}\Big[\,\big\|\, v_\theta\big(x_\tau,\,\tau,\,C\big) \;-\; (\epsilon - x_0)\big\|_2^2\Big], \qquad x_\tau = (1-\tau)\,x_0 + \tau\,\epsilon, \tag{1}$$

where x_0 is the clean video latent, $\epsilon \sim \mathcal{N}(0, I)$, $\tau \in [0, 1]$ is the flow time, and C is the blocks-world video. Following our injection sweep analysis (Appendix Sec. B.2), we attach control blocks only to base layers 4–10, where our control signal is most effectively absorbed.

To obtain a model that handles both unconditional and conditional scenarios, we apply initial-frame dropout: with 50% probability, we provide the first latent frame of x_0 as a conditioning initial frame. This teaches the branch either to synthesize a clip from the blocks-world layout C alone (layout-only generation), or to extend a given observation forward, enabling image-to-video generation and video editing. This mechanism also enables long-horizon rollout: since the frozen backbone has a fixed temporal RoPE horizon with no extrapolation, we generate autoregressively in overlapping chunks, feeding each segment's final generated frame back as the clean initial frame of the next chunk, while our blocks-world representation preserves information such as ego-motion state and scene layout.

3.3 Long-term Object Consistency via Instance-level KV Caching

We propose instance-level KV caching, which maintains per-instance key-value entries across generation chunks by leveraging the known location and orientation of every block in our blocks-world representation. Our model loses instance identity across chunk boundaries: a vehicle that leaves the view and returns in a later chunk has a different appearance, a problem shared with the autoregressive video model discussed in Sec. 2.3. Frame-level KV caching (Yu et al., 2025) mitigates this at a high memory cost, yet remains prone to failure in scenes with numerous instances, where the model struggles to associate cached key-value pairs with the correct instance. As described in Fig. 2, our object-level KV caching instead leverages the explicit per-instance identity and geometry of our view-space blocks representation: the model caches each instance's tokens from its most visible observations in previous chunks and retrieves them whenever the instance reappears. Since each instance contributes only a handful of highly informative tokens rather than entire frames, the cache is compact enough to persist throughout generation, enabling instance-consistent long-horizon rollout.

Instance-level KV token memorization. Let $\mathcal{I}$ denote the set of instances (i.e., objects) present in a generated chunk, trivially acquired from its corresponding blocks-world representation. For each instance $i \in \mathcal{I}$, we select the tokens it owns at its most visible observations, determined by the box-surface completeness in each frame's rendered blocks-world. We store their post-norm, pre-RoPE keys and values captured at diffusion timestep $t = 0$, together with their box-local 3D surface coordinates, as the instance's bank entry $\mathcal{B}_i$:

$$\mathcal{B}_i \;=\; \big(\, K_i,\; V_i,\; P_i \,\big), \qquad K_i, V_i \in \mathbb{R}^{N_i \times d}, \qquad P_i \in [-1, 1]^{N_i \times 3}, \tag{2}$$

where N_i is the number of retained tokens for instance i, and each row of P_i is the normalized box-local 3D coordinate of the token. The bank $\mathcal{B} = \{\mathcal{B}_i\}_{i \in \mathcal{I}}$ persists across all subsequent chunks.

Instance-aware KV token reading. When generating chunk c, the set of recurring instances $\mathcal{I}_c \subseteq \mathcal{I}$ is read directly from the chunk's blocks-world canvas. For each such instance $i \in \mathcal{I}_c$, its stored keys are retrieved from $\mathcal{B}_i$ and re-embedded at the RoPE positions of their corresponding locations in the target frame, obtained by mapping P_i into the canonical world frame through the instance's box transform $\mathbf{T}_i^c$ and projecting the result into the token grid:

$$\hat{K}_i \;=\; \mathrm{RoPE}\Big(K_i,\; \big(\underbrace{t_{\mathrm{obs}} - \Delta}_{\text{temporal}},\; \underbrace{\lfloor \pi(\mathbf{T}_i^c P_i) \rceil}_{\text{spatial}}\big)\Big), \tag{3}$$

where $\pi(\cdot)$ denotes camera projection into token-grid coordinates, $\lfloor \cdot \rceil$ rounds to the nearest token cell coordinate, $t_{\mathrm{obs}} \in \{0, \ldots, T-1\}$ is the latent frame index of the observation within its source chunk, and Δ is the chunk offset, in latent frames, between the source and target chunks. As RoPE encodes only relative position, the temporal component allows the model to reason exactly *when* an instance was observed, while the spatial component tells *where* the instance is at the current frame.

The memory keys and values of all recurring instances are then concatenated with the chunk's own local self-attention tokens (Q_c, K_c, V_c) as follows:

$$\begin{aligned} Q &= Q_c, \qquad & K &= \big[\, K_c;\; \hat{K}_{i_1};\; \hat{K}_{i_2};\; \ldots;\; \hat{K}_{i_{|\mathcal{I}_c|}} \,\big], \\ & & V &= \big[\, V_c;\; V_{i_1};\; V_{i_2};\; \ldots;\; V_{i_{|\mathcal{I}_c|}} \,\big], \end{aligned} \tag{4}$$

and self-attention is computed with an instance-aware mask M that permits each query to attend only to the memory of instances that own it:

$$\begin{aligned} \mathrm{Attention}(Q, K, V) \;&=\; \mathrm{Softmax}\!\left(\frac{QK^\top}{\sqrt{d_k}} + M\right) V, \\ \text{where} \quad M_{qj} \;&=\; \begin{cases} 0 & \text{if } j \text{ is local } \textbf{or } q \text{ is owned by } \mathrm{inst}(j), \\ -\infty & \text{otherwise,} \end{cases} \end{aligned} \tag{5}$$

where d_k is the key dimensionality and $\mathrm{inst}(j)$ returns the instance of memory column j. In practice, the local self-attention tokens and the memory tokens are computed separately and merged via `log-sum-exp`, which is mathematically equivalent to Eq. 5 while retaining flash-attention efficiency on the local tokens.

We demonstrate the effectiveness of our instance-aware KV caching in Fig. 3, where we compare it against naïve chunk-by-chunk generation (no memory) and dense frame-level KV caching (no instance association). Naïve generation loses instance identity across chunk boundaries, and frame-level caching degrades when multiple same-class instances coexist, as it cannot disambiguate which tokens belong to which instance. Our approach preserves each instance's appearance even in crowded scenes, and because the per-instance bank persists

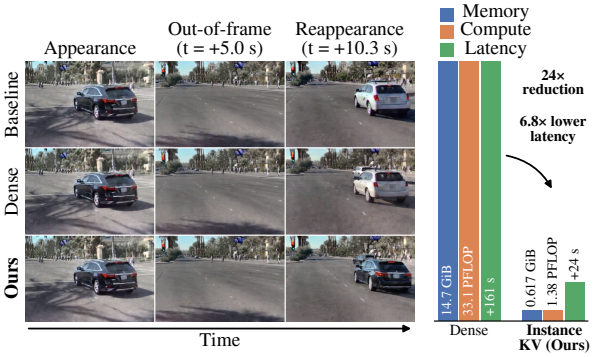

Figure 3: **Instance KV memory** preserves a vehicle's preserves through a 10.3 s absence, where naive generation and dense KV caching fail.

throughout generation, this consistency extends to instances that re-enter the scene many chunks later. As the figure also shows, our caching is far more efficient: compared to dense frame-level caching, it stores $\sim$93% fewer tokens and attends to $\sim$96% fewer, cutting memory and attention compute by 24$\times$ and making consistency across distant chunks practical.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation details. **MemOIR** is built on Cosmos-Predict2.5-2B (Ali et al., 2025), whose DiT, VAE and text encoder are kept frozen while the control branch of seven blocks attached to base blocks 4–10 is trained; the KV caching stage then fine-tunes the base blocks with the control branch frozen. We train on $\sim$12k scenes from Waymo Open (Sun et al., 2020), Argoverse 2 (Wilson et al., 2023), nuPlan (Caesar et al., 2021) training split's curated subset (3k), and the nuScenes (Caesar et al., 2020) training split, together with 3.3k Human3.6M (Ionescu et al., 2014) composites; nuScenes validation is never seen in training. Training runs for $\sim$27k iterations on 8 NVIDIA L40S GPUs with a global batch size of 24, on 25-frame clips at 1056×704; at inference, videos are generated in 93-frame chunks extended autoregressively. Full details are given in Appendix Sec. B.

Evaluation metrics. We evaluate generated videos along two key aspects: visual quality and controllability, the latter covering both the instances in the scene (other vehicles and pedestrians) and the ego motion. Since controllability is measured with off-the-shelf detectors and camera estimators whose own errors are non-zero, we apply the identical pipeline to the real videos to obtain a ground-truth ceiling (for instance metrics) or floor (for ego metrics), which isolates the generator's control from the errors of the measurement tools (Appendix C.1).

- **Video quality:** FVD (Unterthiner et al., 2018) and FID (Heusel et al., 2017) against the real clips of the same scenes.
- **Instance control:** Visible conditioning boxes are matched to YOLO11x (Khanam & Hussain, 2024) detections, giving mIoU, recall at IoU 0.3/0.5 (R@0.3/0.5), center error (CE) and the share of objects tracked over time (Track).
- **Ego control:** Using Depth-Anything-3 (Lin et al., 2025), we recover the ego-camera trajectory from the generated videos and measure its position (ATE), speed errors, and heading errors against the ground-truth camera poses.

Baselines. We compare against TrackDiffusion (Li et al., 2025a), Panacea (Wen et al., 2024), DreamForge (Mei et al., 2024) and MagicDrive-V2 (Gao et al., 2025b), run from public checkpoints with the same annotations and text prompting as ours (Appendix C.2).

4.2 NORMAL DRIVING VIDEO GENERATION

Table 1: **Evaluation on normal scenarios.** In-domain comparison on the nuScenes (Caesar et al., 2020) validation set (15 s clips), with the gray row giving the real-video detector ceiling.

Method	Controllability								Visual quality	
	Instance location			Instance motion		Ego-motion				
	mIoU↑	R@0.3↑	R@0.5↑	Track↑	CE↓	ATE(m)↓	Speed(m/s)↓	Heading(°)↓	FVD↓	FID↓
Real video	0.761	96.6	94.2	95.9	4.7	2.00	0.15	1.06	–	–
TrackDiffusion (Li et al., 2025a)	0.590	37.5	24.9	34.5	13.0	22.68	5.58	5.59	905.2	64.94
Panacea† (Wen et al., 2024)	<u>0.695</u>	56.3	50.8	59.7	<u>7.6</u>	12.10	2.69	**1.52**	323.0	**12.93**
DreamForge (Mei et al., 2024)	0.455	44.7	12.4	44.0	18.6	<u>6.03</u>	<u>0.92</u>	2.49	298.0	28.75
MagicDrive-V2 (Gao et al., 2025b)	0.671	<u>85.1</u>	<u>75.6</u>	<u>88.6</u>	8.4	9.29	1.45	<u>1.88</u>	**228.4**	25.99
Ours	**0.705**	**91.8**	**84.9**	**90.9**	**7.0**	**4.50**	**0.68**	1.91	<u>236.7</u>	<u>25.43</u>

We evaluate our generated results, 15-second front-camera videos for the 150 nuScenes validation scenes following standard convention (Gao et al., 2025b), measuring their visual quality and controllability. **MemOIR** is competitive in visual quality and outperforms all baselines on the controllability metric, as shown in Fig. 4 and Tab. 1, where our view-space blocks-world representation gives dense, richer information for the model to determine the precise speed and motion of the ego-motion as well as other instances that inhabit the scene.

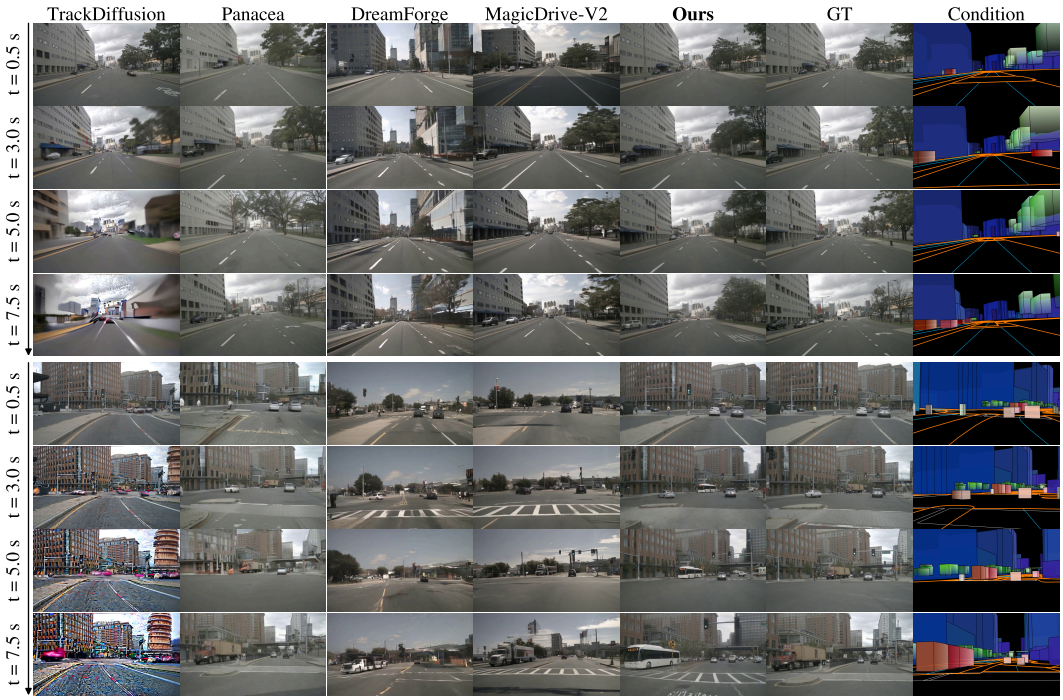

Figure 4: **Qualitative comparison in normal scenarios.** Generated frames of all methods on nuScenes validation scenes, shown with the blocks-world condition and the real video.

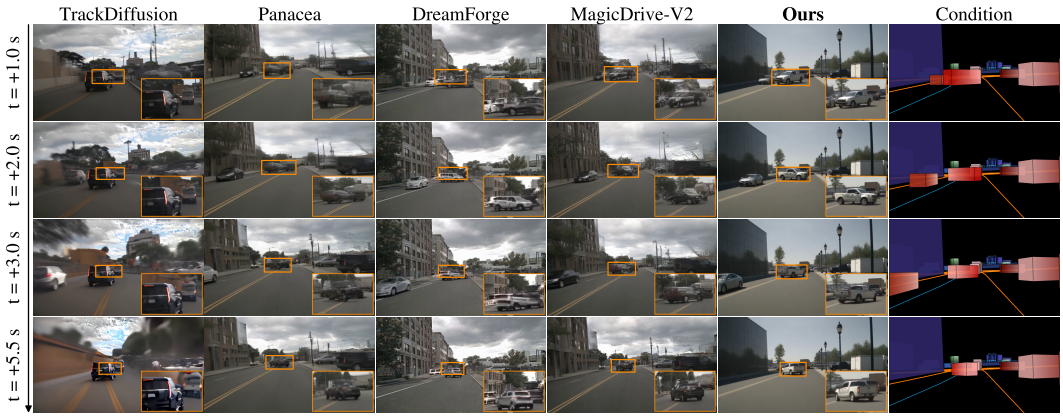

Figure 5: **Qualitative comparison in counterfactual scenarios.** Generated frames around the critical event of CARLA scenarios, shown with the blocks-world condition and the CARLA render.

4.3 Counterfactual Driving Video Generation

To probe controllability beyond the training distribution, we build 20 counterfactual scenes on nuScenes validation hosts by injecting out-of-distribution vehicle motions (spin-outs, loss-of-control trajectories, skids) recorded in CARLA (Dosovitskiy et al., 2017) into the real box annotations, either by having an existing vehicle take over the motion or by introducing a new vehicle that enters from outside the view. Map, camera, and ego motion are extracted from nuScenes validation data, and no method receives an initial frame. We measure the controllability metric of the generated videos and whether it follows the commanded orientation

Table 2: **Controllability in counterfactual scenarios.** Injected out-of-distribution vehicle motions (spin-outs, skids, control loss) in 20 counterfactual nuScenes val scenes.

Method	R@0.3(%)↑	R@0.5(%)↑	mIoU↑	Center Err.(%)↓	Heading Err.(°) [3D]↓
TrackDiffusion (Li et al., 2025a)	60.7	26.3	0.326	2.97	49.2
DreamForge (Mei et al., 2024)	92.1	29.2	0.443	2.75	12.9
Panacea† (Wen et al., 2024)	96.8	87.2	0.648	1.00	10.7
MagicDrive-V2 (Gao et al., 2025b)	97.8	93.6	0.673	0.86	7.4
Ours	**99.9**	**98.1**	**0.791**	**0.60**	**6.9**

(heading error from a 3D detector, using WildDet3D (Huang et al., 2026a)). TrackDiffusion and DreamForge frequently lose the maneuvering vehicle (R@0.5 of 26.3% and 29.2%), and even the strongest baseline, MagicDrive-V2, places it less accurately (mIoU 0.67). **MemOIR** renders the vehicle in 98.1% of frames with the highest IoU (0.79) and the lowest center and heading errors (Tab. 2, Fig. 5), showing superior controllability in out-of-domain motions that were unseen.

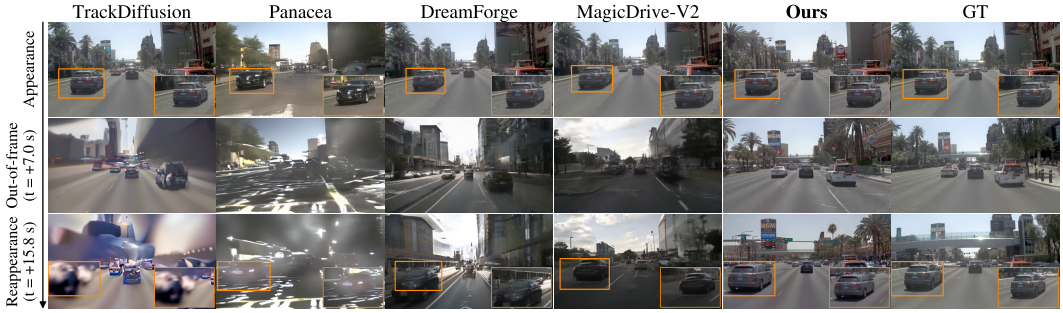

Figure 6: **Instance consistency after re-appearance.** A vehicle before leaving the view and after returning, generated by each method and compared with the real video.

4.4 Long-horizon Instance Consistency Evaluation

We evaluate instance consistency on 29 held-out nuPlan validation scenes in which a vehicle leaves the view for 2–15 s before returning (Appendix C.4). Given the real initial frames, we compare the returning vehicle with ground truth through vehicle re-identification (ReID Sim. and ReID Top-1) and object crop-level PSNR, SSIM and LPIPS (Sec. 4.1). **MemOIR** best preserves the instance's identity (ReID Sim. 0.310 vs. at most 0.178, Top-1 45.6% vs. at most 11.0%), as its instance-aware KV cache retrieves the stored tokens of the vehicle upon re-entry, whereas the baselines, which lack a per-instance memory, lose its identity (Tab. 3, Fig. 6).

Table 3: **Instance consistency.** Identity (vehicle ReID) and photometric consistency of disappearing-then-reappearing vehicles.

Method	ReID Sim.↑	ReID Top-1(%)↑	PSNR↑	SSIM↑	LPIPS↓
GT (VAE recon.)	0.873	99.7	31.50	0.904	0.086
TrackDiffusion (Li et al., 2025a)	0.136	11.0	10.45	0.313	0.720
Panacea† (Wen et al., 2024)	0.150	5.4	12.70	0.237	0.682
DreamForge (Mei et al., 2024)	0.178	5.7	12.56	0.237	0.627
MagicDrive-V2 (Gao et al., 2025b)	0.173	7.5	13.17	0.294	0.590
Ours	**0.310**	**45.6**	**14.74**	**0.379**	**0.512**

4.5 Ablation Studies

Stationary instances. We train two models from scratch for 1K iterations with identical recipes, differing only in whether the canvas contains stationary instances (i.e., buildings and vegetation), and evaluate ego-motion on nuScenes validation scenes with abrupt ego motion (Sec. 4.2). Removing them raises the ATE by 40% (Tab. 4), up to 3.9 m → 49.1 m in a U-turn facing a building. Unlike sparse, road-bound map lines, stationary instances densely fill the image with structures from which ego-motion is deducible, anchoring the camera trajectory. This show dense stationary instances are therefore essential for robust ego control.

Table 4: **Ablation on stationary objects.** Effect of stationary instances on ego-motion.

Condition	ATE (m)↓	Speed (m/s)↓	Heading (°)↓
w/o stationary	8.21	1.08	9.54
Full (ours)	**5.88**	**0.95**	**6.53**

Instance-level KV caching. We disable the memory read at inference on the re-appearance benchmark of Sec. 4.4, keeping the same model and initial frames. As shown in Tab. 5 and Fig. 7, the returning vehicle is then re-generated from scratch and all photometric metrics drop sharply, confirming that the instance-level memory preserves instance identity across chunks.

Table 5: **Ablation on instance-KV caching.** Memory components on the re-appearance benchmark.

Component	ReID Sim.↑	ReID Top-1 (%)↑	PSNR↑	SSIM↑	LPIPS↓
Baseline	0.245	26.1	14.21	0.367	0.522
+ Dense KV caching	0.257	27.3	14.52	**0.383**	0.522
+ Instance-level caching	0.297	38.1	14.40	0.364	0.520
+ Warped RoPE **(Ours)**	**0.310**	**45.6**	**14.74**	0.379	**0.512**

4.6 Persistent Long-horizon Driving Video Generation

We qualitatively compare **MemOIR** with two autoregressive long-horizon generators, Epona (Zhang et al., 2025a) and MAGI-1 (Sand AI, 2025) (Fig. 8). **MemOIR** remains coherent and layout-faithful over the full 55-second rollout, whereas both baselines degrade within 20 seconds and collapse once

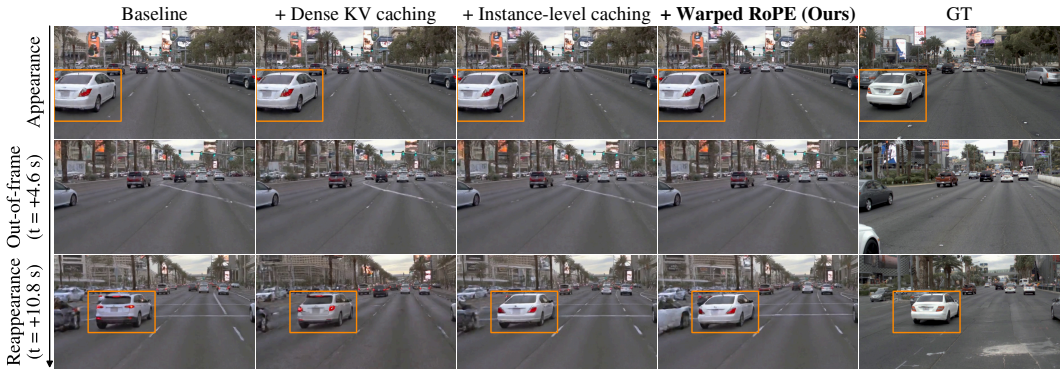

Figure 7: **Qualitative ablation on instance-level KV.** A white sedan leaves the view and returns about 8 s later. Instance consistency increases with each component.

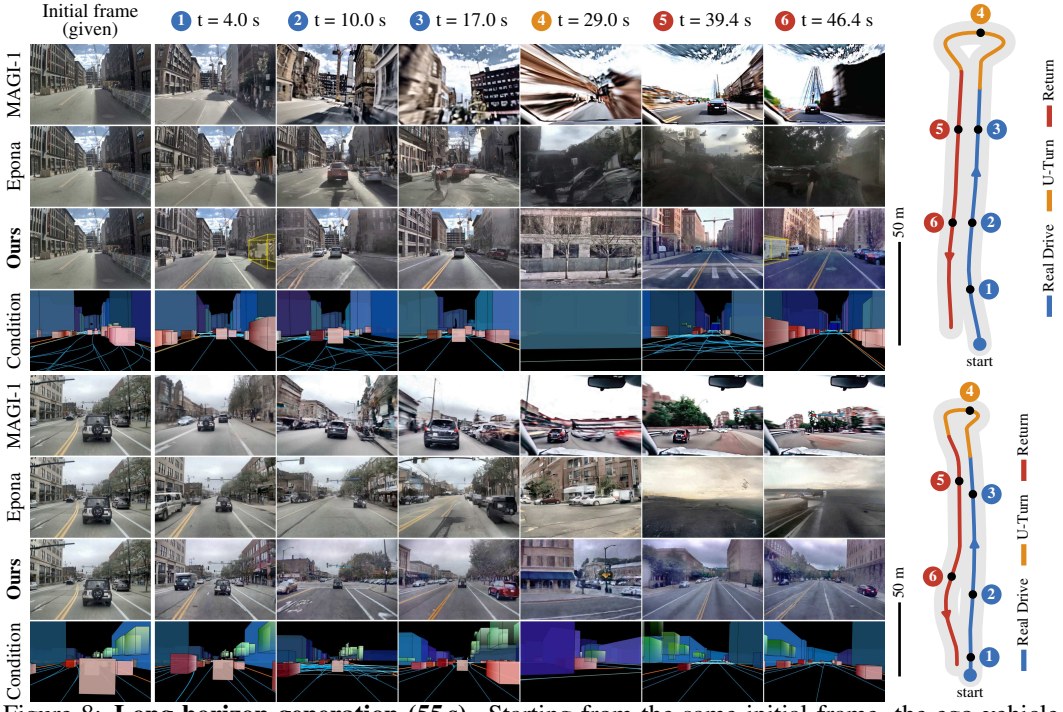

Figure 8: **Long-horizon generation (55 s).** Starting from the same initial frame, the ego vehicle drives along a real street for 20 s, makes a U-turn, and returns in the opposite lane (BEV, right; markers match the columns). MAGI-1 and Epona collapse after the U-turn, while ours stays consistent with the given layout (bottom rows) and remembers what it passed: the van highlighted in yellow (top) reappears at the same spot 42 s later.

the vehicle turns around. Instance-level memory is decisively effective: buildings keep their appearance across every chunk boundary, the white van passed at keyframe 1 reappears intact from the opposite side at keyframe 6, five chunks and over 40 seconds later, and the blue awning glimpsed at keyframe 3 is preserved when faced head-on during the U-turn at keyframe 4, two chunks later. This shows that instance-level memory provides a persistent world memory for long-horizon generation.

5 Conclusion

We presented **MemOIR**, a driving video diffusion model conditioned on a view-space blocks-world representation that abstracts every dynamic and static instance as a 3D block. This single, simple representation unlocks two capabilities at once: precise control over both scene agents and ego motion, and object-level KV caching that remembers each instance at a fraction of the cost of frame-level memory. Curated automatically from raw driving videos, blocks worlds scale without manual 3D annotation, and **MemOIR** delivers state-of-the-art controllability, faithful rare maneuvers, and persistent instance identity across long rollouts. We believe blocks-world conditioning offers a practical path toward controllable, persistent world models for driving simulation and beyond.

REPRODUCIBILITY STATEMENT

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

USE OF LARGE LANGUAGE MODELS

We used large language models (LLMs) as assistants for writing and evaluation experiment-level engineering. For writing, LLMs helped polish author-written text and draft figure captions and table layouts, which the authors revised. For experiments, an LLM-based coding assistant helped implement evaluation and plotting scripts, launch and monitor inference and metric jobs on our cluster, and assemble result tables and qualitative figures under the authors' direction. LLMs were not used to conceive the research questions, design the method, or choose the evaluation protocols; all such decisions, the interpretation of results, and the selection of reported examples were made by the authors, who verified every reported number and figure against the underlying outputs and take full responsibility for the paper's content.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. Cc3d: Layout-conditioned generation of compositional 3d scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7137–7147. IEEE, 2023.

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.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. URL https://arxiv.org/abs/2407.01392.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive-V2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28135–28144, 2025b.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17281–17291, 2025.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2568–2577. IEEE, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Weikai Huang, Jieyu Zhang, Sijun Li, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Matthew Wallingford, Rustin Soraki, Chris Dongjoo Kim, Shuo Liu, et al. Wilddet3d: Scaling promptable 3d detection in the wild. 2026a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026b.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Min-Seop Kwak, Junho Kim, Sangdoo Yun, Dongyoon Han, Taekyung Kim, Seungryong Kim, and Jin-Hwa Kim. Aligned novel view image and geometry synthesis via cross-modal attention instillation. In *International Conference on Learning Representations*, volume 2026, pp. 16029–16050, 2026.

Pengxiang Li, Kai Chen, Zhili Liu, Ruiyuan Gao, Lanqing Hong, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Tracklet-conditioned video generation via diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3539–3548. IEEE, 2025a.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025b.

Haotong Lin, Sili Chen, Junhao 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.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

Jianbiao Mei, Tao Hu, Xuemeng Yang, Licheng Wen, Yu Yang, Tiantian Wei, Yukai Ma, Min Dou, Botian Shi, and Yong Liu. Dreamforge: Motion-aware autoregressive video generation for multi-view driving scenes. *arXiv preprint arXiv:2409.04003*, 2024.

Hyelin Nam, Hyojun Go, Byeongjun Park, Byung-Hoon Kim, and Hyungjin Chung. Generating human motion videos using a cascaded text-to-video framework. *arXiv preprint arXiv:2510.03909*, 2025.

NVIDIA, Aarti Basant, Amlan Kar, Despoina Paschalidou, Fangyin Wei, Francesco Ferroni, Guillermo Garcia Cobo, Haithem Turki, Huan Ling, Jaewoo Seo, James Lucas, Jay Zhangjie Wu, Jialiang Wang, Jonathan Lorraine, Jun Gao, Kai He, Katarina Tothova, Kevin Xie, Michał Tyszkiewicz, Qi Wu, Riccardo de Lutio, Ruilong Li, Sanja Fidler, Seung Wook Kim, Tianchang Shen, Tianshi Cao, Tobias Pfaff, William Lew, Xindi Wu, Xuanchi Ren, Yifan Lu, Yuxuan Zhang, Zan Gojcic, and Zian Wang. Nvidia omnidreams: Real-time generative world model for closed-loop autonomous vehicle simulation. *arXiv preprint arXiv:2606.03159*, 2026.

Sand AI. Magi-1: Autoregressive video generation at scale, 2025. URL `https://arxiv.org/abs/2505.13211`.

Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, et al. Grounding world simulation models in a real-world metropolis. *arXiv preprint arXiv:2603.15583*, 2026.

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

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991. doi: 10.1109/34.87344.

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

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

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6902–6912, 2024.

Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation, 2025. URL https://arxiv.org/abs/2509.22622.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive Diffusion World Model for Autonomous Driving. In *International Conference on Computer Vision*, pp. 27220–27230, 2025a. doi: 10.1109/ICCV51701.2025.02527.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023b.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

A Appendix

B Additional Implementation Details

B.1 Training and Sampling

Backbone. Cosmos-Predict2.5-2B (Ali et al., 2025) is a 28-block rectified-flow DiT (width 2048, 16 heads, full spatio-temporal self-attention, 3D RoPE) operating in the latent space of a Wan2.1 VAE (Wan et al., 2025) (8× spatial and 4× temporal compression, 16 channels) with 2×2 patchification and Cosmos-Reason1.1-7B text embeddings. At 1056×704, the latent grid is 88×132, i.e., $44\times66 = 2{,}904$ tokens per latent frame; a 25-frame training clip spans 7 latent frames and a 93-frame inference chunk 24. The DiT, VAE and text path are frozen during control-branch training.

Optimization. We use FusedAdam (weight decay 0.1, $\beta = (0.9, 0.99)$) with warm-up and cosine learning-rate schedules, power EMA, gradient clipping at 1.0, activation checkpointing, and FSDP over 8 NVIDIA L40S (46 GB) GPUs with a global batch size of 24. Training proceeds in three stages totaling ∼20.5k iterations: 7.5k iterations on Waymo Open and Argoverse 2 (peak learning rate 8×10^{-5}), 11.75k iterations with nuPlan added (peak 6×10^{-5}), and a final warm restart on the full mix (peak 4×10^{-5}, 200 warm-up iterations), sampled as nuScenes 35%, Waymo Open and Argoverse 2 40%, Human3.6M augmentation 15%, and nuPlan 10%; the checkpoint after 1,250 iterations of the final stage, selected on held-out generations, is used in all experiments. The training data comprises Waymo Open (996 scenes), Argoverse 2 (729), nuPlan (6,199; Las Vegas, Boston, Pittsburgh), nuScenes train (700), and 3,311 Human3.6M composites (App. B.5).

Objective and conditioning. We use the plain rectified-flow loss (Eq. 1) with a logit-normal time distribution (shift 5). The real first frame is provided as an image condition, at timestep 0.1, in 50% of training clips, so a single model supports both layout-only and first-frame-conditioned generation. Text and canvas conditions are each dropped with probability 0.2. Source frames are used verbatim without temporal resampling, and the fps label is randomized in $[16, 20]$ during training.

Captions. Per-scene captions are generated by Cosmos-Reason1 (Agarwal et al., 2025) and encoded with Cosmos-Reason1.1. At evaluation, the prompt carries the city, a road description, the time of day, and the weather, i.e., the same information given to the baselines.

Sampling. We sample with EMA weights in bf16 using UniPC with 14 steps and classifier-free guidance 7.0 with the trained default negative prompt. Each pass generates 93 frames, the temporal RoPE extent of the base model's training, as longer single passes deform. Longer videos are generated in consecutive chunks overlapping by one frame, the last generated frame of each chunk serving as the conditioning frame of the next.

B.2 Control Branch

The blocks-world video is encoded by the frozen VAE and patchified into control tokens. The control branch consists of seven blocks initialized as copies of base blocks 4–10, with zero-initialized input and output projections; the output of each control block is added to the hidden state after its corresponding base block. Only the control branch is trained.

Control signal injection layer sweep. Our ControlNet branch consists of trainable copies of base DiT blocks whose zero-initialised output projections add a control residual to selected blocks of the frozen 28-block backbone, so the choice of injection sites decides where in the network the blocks-world layout can steer generation. We ablate this choice with five layouts, each trained from scratch under an identical protocol (same base weights, data, schedule and seed; control blocks copied from the base blocks they feed): the default VACE-style *Spaced-14* layout (blocks $0, 2, \ldots, 26$), and four contiguous, compute-matched seven-block layouts placed at increasing depth, *First-7* (0–6), *Early-mid-7* (4–10), *Mid-7* (10–16) and *Late-7* (21–27). Depth matters strongly. Late-7, which only reaches the last quarter of the network, never learns to follow the layout: it drifts to unrelated streets, invents vehicles and replaces pedestrians with implausible objects, and on the validation clip it trails

every other layout by about 5 dB PSNR across all matched checkpoints (17.8 vs. 22.2–23.1 dB; LPIPS 0.43 vs. 0.22–0.25).

Among the early and middle placements, Early-mid-7 learns control fastest: at a matched early checkpoint (3k iterations) it is the most faithful to the real video on pedestrian-heavy scenes (PSNR 19.4 dB, LPIPS 0.310, vs. 16.8 dB / 0.426 for Spaced-14 and 17.5 dB / 0.384 for First-7), and over training it attains the best LPIPS of all layouts on the validation clip (0.219). With enough training, the early and middle layouts and Spaced-14 converge to similar photometric quality (21.3–21.5 dB, LPIPS 0.255–0.257 at the final checkpoints, vs. 20.0 dB for First-7). Spaced-14 reaches the lowest validation loss (0.064 vs. 0.082 for Early-mid-7), but it uses twice as many control blocks. We therefore adopt Early-mid-7 (Tab. A1): it matches the 14-block layout in output quality at half the control parameters, trains about 16 % and samples about 17 % faster, and places control where layer-relevance analyses of diffusion transformers find structure to be decided—in the early-to-middle blocks, before the late blocks that mostly refine texture.

Table A1: **Ablation on control injection layers.** ControlNet branches injected into different blocks of the 28-block base DiT, compared on Waymo pedestrian-heavy scenes and the training validation clip. Photometric agreement with the real video (frame 0, the given frame, excluded). Matched: every arm evaluated at the same training iteration.

Injection layout	Control		Iter 3k (matched)			Final ckpt (81k–96k)			Val clip, 2.5k–77.5k (matched)		Val loss
			3 scenes			3 scenes			mean over 31 ckpts		
	Blocks	#	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	LPIPS↓	33–44k↓
Spaced-14	0,2,…,26	14	16.76	0.597	0.426	<u>21.38</u>	<u>0.714</u>	**0.255**	**23.08**	<u>0.221</u>	**0.0640**
First-7	0–6	7	17.47	0.624	0.384	20.00	0.682	0.290	22.17	0.246	<u>0.0817</u>
Early-mid-7 (Ours)	4–10	7	**19.41**	<u>0.650</u>	**0.310**	21.27	0.710	<u>0.257</u>	<u>22.74</u>	**0.219**	0.0822
Mid-7	10–16	7	<u>19.20</u>	**0.662**	<u>0.327</u>	**21.45**	**0.720**	<u>0.257</u>	22.69	0.223	0.0849
Late-7	21–27	7	16.16	0.564	0.476	20.49	0.692	0.276	17.78	0.432	0.0911

B.3 Instance-Aware KV Caching

Writing. Keys and values are captured in all 28 base self-attention blocks after the per-head QK/V RMSNorm and before RoPE, in a no-gradient forward pass of the clean latent at timestep 0, and stored in bf16; the control branch is not involved. A DiT token covers 2×2 latent cells over the raw-frame support of its latent slice (frame 0 alone, then 4 raw frames); its ownership by instance i is the fraction of these support cells whose z-buffered dominant instance is i. A token is written only if this fraction is at least $\tau_{\text{write}} = 0.5$, purest tokens first.

Observation selection. Each latent slice of an instance is scored by completeness, the mean over visible box faces of the fraction of 4×4 surface bins seen (bucketed to 0.1), and then by visible token area: score $= 10^4 \cdot \mathrm{round}(\text{completeness}, 1) + \min(\text{area}, 9999)$. Up to four slices are selected per instance and chunk; each additional slice must show a new dominant box face or lie at least two latent frames from every accepted slice, and cover at least $0.3\times$ the area of the best slice. A slice is eligible only if the instance lies within 75 m of the camera, and vegetation is never written.

Budgets and bank management. Each observation stores at most 512 tokens, and at most 8,192 tokens are written per chunk across all instances, allocated by priority score × class weight × novelty (class weight 1.0 for dynamic and 0.3 for static classes; novelty 1.5 for instances not yet in the bank). An observation exceeding the remaining budget is uniformly thinned, to at least 4 tokens, rather than dropped. The bank keeps at most four observations per instance: a new observation is appended if fewer are stored, and otherwise replaces the weakest one if its score, multiplied by 1.25 when it shows a box face absent from all retained observations, exceeds the weakest score. At most 128 instances are held, evicting the least recently seen, then lowest-scored, first.

Reading and warping. The read set is every instance that owns at least one cell of the target chunk's correspondence rasters and has a bank entry; no appearance matching is involved. Instances are visited in descending target area, and whole observations are read while they fit a budget of 4,096 tokens. Each stored token carries the mean box-local coordinate $(u, v, w) \in [-1, 1]^3$ of its owned support cells. Its spatial RoPE position is assigned per target latent frame: for every frame in which the instance is visible, the coordinate is transformed by the instance's box pose in that frame, projected through the camera including radial and tangential distortion, and snapped to the nearest

token cell, so that each query frame attends the memory under the instance's current position; tokens with depth below 0.05 m or outside the grid are discarded in that frame. Every stored observation keeps its own temporal position (below) and no deduplication is applied, so several observations of an instance may land on the same target cell. The box warp places 98.6% of tokens within one cell at a 15-frame gap and 91.4% at a 40-frame gap, compared with 87% and 82% for an affine warp.

Temporal positions. The temporal RoPE coordinate of a memory token is $t_{\text{obs}} - \Delta$ (Eq. 3), on the same fps-scaled temporal axis as the local tokens. At inference, consecutive 93-frame chunks overlapping by one frame give $\Delta = 23$, placing the previous chunk's tokens at slots -23 to 0, where slot 0 is the shared overlap frame. During training, reference windows ending 15–60 frames before the target place memory tokens at slots -21 to -4.

Attention. Query q may attend memory column j only if instance $\text{inst}(j)$ owns at least $\tau_{\text{read}} = 0.3$ of q's support cells, so a token on an instance boundary may see the memory of both instances; background tokens attend local keys only. The mask is implemented as an additive bias of 0 or -10^4 (bf16-safe), and allowed memory logits additionally receive a fixed bias of $+3$, at training and inference alike. The local block runs FlashAttention (Dao et al., 2022) over the chunk's own tokens and the memory block runs a memory-efficient kernel with the bias; the two are merged exactly through their `log-sum-exp` statistics, with an exact backward pass.

Training. The KV model fine-tunes the self-attention (QKV and output projections), MLP and adaLN parameters of all 28 base blocks (1.53B parameters) with the control branch frozen, at learning rate 2×10^{-5} with a global batch size of 8 on 4 L40S GPUs, on Waymo Open and Argoverse 2 clips sampled such that an instance re-appears 15–60 frames after it was last seen. To force the model to rely on the memory, the current keys of eligible cells of returning instances are removed and the loss on returning-instance cells is up-weighted, in two phases: 2,000 iterations with 80% of the eligible cells dropped in every sample and a loss weight of 5, followed by a warm restart in which dropout is applied to 60% of the samples at a per-cell rate annealed linearly from 0.8 to 0.3 over 3,000 iterations with a loss weight of 3; the checkpoint after 1,250 iterations of the second phase is used. Neither the key removal nor the loss weighting is applied at inference.

Memory footprint. Measured on the 30 re-appearance clips of App. C.4 on one L40S, a 93-frame chunk has 69,696 local tokens per layer. The instance bank holds on average 4,520 tokens (0.97 GiB of keys and values over all 28 layers in bf16) of which 2,887 are read per chunk (0.62 GiB on the GPU), adding 4.1% to the attention FLOPs and 9% to the chunk sampling time (276 s $\rightarrow$ 300 s, plus a 22 s write pass). Caching every token of the previous chunk instead reads 68,970 tokens (14.7 GiB), doubles the attention FLOPs, and raises the chunk time by 58% and the peak GPU memory from 28 GiB to 40 GiB.

B.4 Blocks-World Canvas Construction

Rendering. Canvases are rendered at 1920×1280 on a black background with a global floating-point z-buffer, Sutherland–Hodgman near-plane clipping and back-face culling, so that inter-instance occlusion is exact. Box edges are drawn as 2 px black wires only where an adjacent face is front-facing, depth-tested with tolerance $\max(0.3, 0.03z)$. Dynamic instances are vehicles (red), pedestrians (white) and cyclists (yellow); stationary instances are buildings (blue, footprint prisms), vegetation (green, boxes) and traffic signs (orange). Each class occupies a hue band and each instance a unique hue within it, so no two instances in a scene share a color. Heading is encoded by a front-to-back shading gradient (front darkened by 0.22, back lightened by 0.60, opacity 0.90), fitted per pixel from the z-buffer. Vegetation boxes use a vertical opacity ramp from 0.15 at the base to 0.95 at the top, and crowns covering more than 40% of the frame are culled per frame. HD-map road edges (orange), lane lines (blue) and crosswalks (white) are drawn as depth-occluded lines; lane centerlines are excluded.

Geometry lifting. Per-frame Mask2Former (Cheng et al., 2022) panoptic segmentation (Cityscapes classes, with tracking) provides class masks, and Depth-Anything-3 (Lin et al., 2025) provides per-frame pointmaps, aligned to metric scale by a Huber-robust scale-and-shift fit against the ground plane or, where available, sparse LiDAR returns. Class-masked pixels, after a small

per-segment boundary erosion to suppress edge bleed, are lifted into the world frame through the camera extrinsics and ego poses and accumulated over the clip. As depth error grows with range, far observations are superseded when the ego later passes closer: world space is binned into 2.5 m cells, the closest observation distance $d_{\min}$ is recorded per cell, and a point is kept only if its observation distance is at most $\max(d_{\min}+8\,\text{m},\ 1.4\,d_{\min})$, capped at 35 m for cells ever observed within that range. The pruned cloud is voxelized at 0.3 m, keeping one mean point per voxel.

Instance clustering. Voxels are rasterized onto a 0.4 m bird's-eye-view grid of maximum height, normalized by the per-scene ground level (5th percentile of per-cell minimum height) and Gaussian-smoothed ($\sigma = 1.2$). Local maxima at least 2.5 m apart and 0.8 m above ground seed a marker-controlled watershed on the inverted height map, and clusters under 40 voxels are discarded. Each cluster then passes a distance-adaptive outlier filter: range-normalized 12-NN statistical outlier removal, a minimum cluster size decaying as $1/d^2$ with distance, and a rootedness constraint of at most 8 m above the local ground along the ego path.

Box fitting and road priors. The box yaw is snapped to the tangent of the nearest HD-map lane centerline, and the extents are the $[0.08, 0.92]$ quantiles of the cluster's voxels in the yaw-rotated frame; a strict min/max would inflate the volume by $\sim 1.9\times$ through the sparse depth-noise fringe. Ground-truth road geometry further constrains the static layer: the ground height is taken from the local road surface, and building points inside the HD-map road edges within 80 m of the camera are purged, since the drivable surface cannot contain a building.

Buildings. For training set construction, where building instance accuracy is crucial, we use existing map information to aid our block representation construction. Where a footprint map exists, buildings are extruded OpenStreetMap or Overture footprint prisms within 600 m of the ego path; minimum-area rectangles are avoided, as they swallow the road and the ego. Heights come from USGS 3DEP airborne lidar in US cities and Google Open Buildings in Singapore. Where no surveyed height is available, the height is triangulated from the video by intersecting each footprint's facade columns with the Mask2Former roofline and the Depth-Anything-3 depth of the pixels the footprint owns, using lower bounds when the roof is unobserved and nearest-neighbor filling otherwise. For Waymo Open, which has no georeference, buildings are reconstructed from the video with the same lifting, refined with LiDAR facade fits and multi-view silhouette heights, and consolidated from facade slices into building instances.

Per-dataset preprocessing. Argoverse 2 frames are undistorted and center-cropped to 1920×1280 (3:2). Waymo lens distortion is reproduced by rendering an overscanned pinhole canvas and applying the inverse distortion map. nuScenes 3D boxes, annotated at 2 Hz, are interpolated to the 12 Hz camera sweeps (linear centers, SLERP rotations, per-sweep ego poses). As nuScenes ships no geographic anchor, OpenStreetMap footprints are registered per scene; footprints with median Mask2Former support below 0.35 across frames are removed as phantoms, and heights are clamped to the observed facade top.

B.5 HUMAN3.6M AUGMENTATION

To improve pedestrian generation and enable pose-level control, we synthesize 3,311 training clips in which Human3.6M (Ionescu et al., 2014) subjects move through nuScenes training scenes. Single-person clips of subjects S1, S5, S7 and S8 are matted with MatAnyone and composited onto a frozen first frame of a real scene, so that only the humans (2–4 per clip) move. Each subject's ground trajectory is recovered by back-projecting its foot contact onto the scene's ground plane through the real camera, yielding a torso-centered 3D box consistent with the scene geometry. The box is rendered into the canvas like any other pedestrian, and the ground-truth COCO-17 skeleton is drawn inside it for pose-level control. The set is sampled at 15% during the final training stage.

C Evaluation Details

C.1 Metrics

Video quality. FVD is the content-debiased FVD with I3D features on 256×256 center crops; FID uses clean-fid with InceptionV3 features on individual frames. Both are computed at set level, comparing the generated clips with the real clips of the same scenes, and exclude frame 0 for every method. Every method is scored at MagicDrive-V2's pixel density (half of the raw 1600×900 camera): our 1920×1280 output is resized to 675×450, Panacea to 800×400 and TrackDiffusion to 800×450, with each real reference resized identically.

Instance controllability. The conditioning 3D boxes are projected into the camera, and a box is kept on a frame only if at least 60% of it is visible under a z-buffer of all scene objects, buildings and vegetation, at least 90% lies inside the image, and it is at least 25 px tall; the target is its visible part. Generated frames are parsed per frame by YOLO11x (Khanam & Hussain, 2024) at 1280 px. Vehicles match a detection at IoU ≥ 0.3; pedestrians match if the detection center lies inside the box and its height is within $\pm$50%. Track is the share of conditioned objects matched in at least 50% of their visible frames. Each method is additionally compared with a ceiling computed on the real video processed exactly like its own output (resolution and crop), so that lower-resolution baselines are not penalized for resolution (App. D).

Ego-motion controllability. The camera trajectory is recovered by Depth-Anything-3 from all frames in a single pass, anchored at the first frame, and brought to metric scale with the known camera height. ATE is the position RMSE (in m and in % of path length) and speed error is computed over 1 s windows; we additionally report the final heading error and the turn-rate error. We selected this estimator by benchmarking Depth-Anything-3, VGGT (Wang et al., 2025), VGGT-Omega, and a feature-matching baseline (SIFT with essential-matrix estimation) on real clips, where Depth-Anything-3 with camera-height scale recovered the trajectories most accurately (3.4% of path).

C.2 Baseline Settings

Table A2: **Baseline settings.** Conditions and generation settings used for each baseline.

Method	Condition	Setting
TrackDiffusion	2D box tracks	Public SVD-based release (YouTube-VIS weights), with our projected boxes as its 2D tracks; 12 Hz.
Panacea†	BEV layout sequence, text, real image	Released Panacea+ checkpoint; 2 Hz at 512×256.
DreamForge	BEV map, 3D boxes, text	DiT version (DriveArena); full-length 12 Hz multi-view generation in 25-frame autoregressive chunks with 10-frame overlap.
MagicDrive-V2	BEV map, 3D boxes, text	Official checkpoint; full-length 12 Hz multi-view generation, front camera scored.

C.3 Counterfactual Benchmark

The benchmark contains 20 counterfactual scenes built on real nuScenes validation clips: only the box list changes, while the map, the cameras and the ego motion stay real. In 10 scenes (mode A) a real vehicle takes over an out-of-distribution motion from a trigger frame on, its real boxes being deleted from the trigger; in the other 10 (mode B) a new vehicle enters from outside the view and performs the motion. The motions (Tab. A3) are spin-outs recorded in CARLA (Dosovitskiy et al., 2017) (total heading change capped at $400°$), NPC control-loss takes from public CARLA scenario libraries replayed with ScenarioRunner, and synthetic skids. An injected motion is scaled to the host vehicle's speed (3.9–11 m/s at onset), follows the terrain height, keeps at least 8 m from the ego, and must lie on the drivable area with its whole box inside the image in at least 80% of the frames after the onset; slides decelerate at 2–6 m/s^2 with a turn rate of at most $150°$/s. Pedestrians

are removed from these scenes. Each scene is one 93-frame window at 12 Hz with the trigger at frame 12. Every method generates without a real first frame, ours from the layout, map and text alone and the baselines in their stock image-free settings, since with the real first frame given, our model locks onto the real vehicle's appearance and pose instead of the injected motion. Baseline inputs are obtained by editing the injected boxes into the official nuScenes 12 Hz annotation rows; round-tripping the real boxes through the same conversion reproduces the official rows to within 8 mm and $0.4°$.

Motion following is scored with the WildDet3D (Huang et al., 2026a) 3D vehicle detector on each method's front view, given that method's camera intrinsics. The detection with the highest 2D IoU (≥ 0.3) with the projected injected box is taken as the generated vehicle, and frames count from the onset while more than half of the projected box lies inside the image. We report the detection rate, the 2D IoU, the ground-plane center error, and the heading error made front/back-symmetric, $\min(e, 180° - e)$, because the detector's orientation axis is reliable while its front/back sign flips on side and rear views. MagicDrive-V2 emits 89 of the 93 frames, so every method is scored on the common span.

Table A3: **Counterfactual scenario composition.** The 20 scenes cover three motion families, each injected into 10 mode-A (real vehicle takes over) and 10 mode-B (new vehicle enters) hosts in total.

Motion	#	Heading change	Source
Spin-out	8	360–400°	CARLA Town10HD recordings (5 distinct takes)
NPC control loss	5	79–137°	CARLA scenario libraries via ScenarioRunner (3 takes)
Skid	7	70–100°	synthetic constant-deceleration slide

C.4 Re-appearance Benchmark

The benchmark contains 30 target vehicles in 29 nuPlan clips from Las Vegas that are never used in training. Each vehicle is classified per frame as visible, partial or out of view from its projected box, a z-buffer of box hulls, and Mask2Former vehicle pixels. Clips are selected automatically such that the target is clearly visible, leaves the view for 2.0–15.1 s (mean 9.0 s; 23 of 30 gaps $\geq$4 s), and returns. Vehicles leave by exiting the frame (15), by being driven past or behind the ego (11), or by occlusion from another object (4). Metrics are computed on the crop of the target's projected box in the frames after its return.

C.5 Appearance Control and Free-View Generation

Appearance control uses an identical canvas and seed (0), guidance 7.0, and 93 frames at 1056×704; only the text embedding changes, with the trained default negative prompt. Free-view canvases are rendered with the same renderer and color code as the driving canvases, from virtual cameras following *orbit* (45 m altitude, 70 m radius), *low orbit* (25 m, 45 m), *span* (a straight pass at 55 m over 260 m), *low span* (30 m over 180 m), *survey* (110 m, pitched $-72°$), *crane* (rising from 5 to 42 m while pitching from $-8°$ to $-48°$), and low *FPV* flights at 4–18 m following the road network. No aerial or free-flying footage is used in training.

D Additional Experiments

Visual quality over time. Tab. A4 reports FID and FVD in 4 s windows with an equal number of samples per window (FID on 900 frames, FVD on 150 16-frame clips). **MemOIR** starts far ahead, as the given real first frame anchors the first seconds (FID 28.7, FVD 221.9 in 0–4 s), and its quality drifts as the generation moves away from that frame (FID 60.8, FVD 375.2 in the last window), whereas MagicDrive-V2, which receives no image, is flat over the clip (FID 44–47, FVD 340–351). **MemOIR** keeps the lowest FVD over the first 8 s and the second-lowest thereafter. TrackDiffusion loses the scene after the first window (FID 46.2 $\rightarrow$ 132.8, FVD 501.8 $\rightarrow$ 1176.7) and DreamForge degrades toward the end of the clip (FVD 388.7 $\rightarrow$ 519.6). Per-window values exceed the set-level values of Tab. 1, as each window has fewer samples.

Table A4: **Visual quality over time.** FID and FVD in 4 s windows on nuScenes validation, where Panacea is evaluated with FID only due to its 2 Hz output. Bold: best, underline: second best.

Method	FID↓				FVD↓			
	0–4 s	4–8 s	8–12 s	12–15 s	0–4 s	4–8 s	8–12 s	12–15 s
DreamForge (Mei et al., 2024)	51.2	49.6	50.3	56.5	388.7	377.2	434.0	519.6
MagicDrive-V2 (Gao et al., 2025b)	46.1	45.4	44.4	47.3	343.3	340.1	**350.8**	**345.1**
Ours	**28.7**	47.5	53.5	60.8	**221.9**	**299.4**	363.7	375.2

Detailed controllability. Tab. A5 breaks the controllability of Tab. 1 down by class and reports the full set of ego-motion errors. **MemOIR** follows the most vehicles and pedestrians and places them most precisely at the strict IoU threshold; Panacea, which receives a real image, keeps the most accurate pedestrian shapes and the lowest heading error, and DreamForge has the lowest turn-rate error, while **MemOIR** halves the camera position error of the closest baseline.

Table A5: **Detailed controllability in normal scenarios.** Per-class object control and full ego-motion errors on the same nuScenes clear-day scenes, where Heading and Turn denote the final heading and turn-rate errors of the recovered camera trajectory. Bold: best, underline: second best.

Method	Object control										Ego control				
	Vehicles					Pedestrians									
	Track↑	mIoU↑	R@0.3↑	R@0.5↑	CE↓	Track↑	mIoU↑	R@0.3↑	R@0.5↑	CE↓	ATE (m)↓	ATE (%)↓	Speed (m/s)↓	Heading (°)↓	Turn (°/s)↓
Real video	95.9	0.761	96.6	94.2	4.7	73.5	0.594	69.5	51.1	6.2	2.00	2.8	0.15	1.06	0.24
TrackDiffusion (Li et al., 2025a)	34.5	0.590	37.5	24.9	13.0	10.1	0.475	7.6	2.8	14.9	22.68	55.8	5.58	5.59	1.44
Panacea† (Wen et al., 2024)	59.7	0.695	56.3	50.8	7.6	13.6	**0.553**	15.3	9.6	**7.9**	12.10	15.2	2.69	**1.52**	1.03
DreamForge (Mei et al., 2024)	44.0	0.455	44.7	12.4	18.6	12.7	0.406	7.4	1.1	29.2	6.03	8.6	0.92	2.49	**0.60**
MagicDrive-V2 (Gao et al., 2025b)	88.6	0.671	85.1	75.6	8.4	32.2	0.505	30.3	14.5	10.5	9.29	13.6	1.45	1.88	0.74
Ours	**90.9**	**0.705**	**91.8**	**84.9**	**7.0**	**53.3**	0.510	**50.3**	**24.8**	11.2	**4.50**	**6.1**	**0.68**	1.91	0.62

Ceiling-normalized tracking. Tab. A6 normalizes Track by each method's own real-video ceiling, computed at its output resolution and crop. **MemOIR** reaches 94.7% of its ceiling on vehicles and 72.5% on pedestrians; even under this correction, the strongest baseline follows 92.4% of vehicles and only 43.5% of pedestrians relative to its ceiling.

Table A6: **Ceiling-normalized tracking.** Track as a percentage of each method's real-video ceiling on nuScenes clear-day scenes. Bold: best, underline: second best.

Method	Vehicles↑	Pedestrians↑
TrackDiffusion (Li et al., 2025a)	35.8	13.5
Panacea† (Wen et al., 2024)	62.1	19.3
DreamForge (Mei et al., 2024)	45.9	17.1
MagicDrive-V2 (Gao et al., 2025b)	92.4	43.5
Ours	**94.7**	**72.5**

Rain and night scenes. Tab. A7 reports controllability on the 24 rain and 15 night scenes of the nuScenes validation set, which are excluded from the clear-day tables. **MemOIR** keeps the lowest camera errors in both conditions, while MagicDrive-V2 tracks vehicles best at night, where the real-video ceiling itself drops to 91.5%. Ego control degrades most for the baselines at night: their camera errors grow to 22–87% of the path, against 14% for **MemOIR** and 5% for the real video.

Table A7: **Controllability in rain and night scenes.** Vehicle tracking and ego-motion errors on the nuScenes validation scenes excluded from Tab. 1. Bold: best, underline: second best.

Method	Rain (24 scenes)			Night (15 scenes)		
	Track↑	ATE (%)↓	Speed (m/s)↓	Track↑	ATE (%)↓	Speed (m/s)↓
Real video	87.3	2.4	0.22	91.5	5.1	0.43
TrackDiffusion (Li et al., 2025a)	17.1	57.4	7.26	29.7	86.9	11.25
Panacea† (Wen et al., 2024)	26.2	14.9	3.18	59.5	68.3	13.75
DreamForge (Mei et al., 2024)	21.7	8.8	1.14	25.0	<u>22.2</u>	**2.77**
MagicDrive-V2 (Gao et al., 2025b)	**84.3**	<u>7.0</u>	<u>0.97</u>	**94.6**	34.3	4.60
Ours	<u>81.4</u>	**6.1**	**0.81**	<u>72.9</u>	**14.3**	<u>2.86</u>