LOC2DRIVE: BEYOND VISUAL HISTORY WITH EXPLICIT 3D WORLD STATE FOR LONG-HORIZON DRIVING VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon driving-video generation is vulnerable to autoregressive error accumulation because generated frames are repeatedly reused as future context. Yet some errors arise before propagation: future views may reveal static scene content that the current visual history has never observed. Turns and disocclusions make this gap especially clear, exposing road topology or appearance outside the previous field of view or behind occluders. We present **Loc2Drive**, which supplies missing static evidence from route-queryable satellite imagery. At each ego pose, Loc2Drive queries a heading-aligned satellite crop and reconstructs a shared 3D world state. The state combines semantic occupancy for geometry and visibility, a height-aware 3D feature volume that preserves appearance cues, and BEV road semantics learned with HD-map supervision. Projecting this state through six calibrated cameras yields pixel-aligned RGB, depth, semantic, and pseudo-HD-map conditions for video generation. Satellite evidence is re-queried along the route rather than inferred recurrently from generated frames; specified dynamic agents remain on a separate 3D-box path, while the video prior handles unconstrained appearance and dynamics. Without an external HD map at inference time, Loc2Drive reduces FVD by 14.7% at 16 frames and FVD_{128} by 29.7% at 128 frames relative to UniMLVG with HD maps. With no visual references, it further reduces FVD_{32} by 23.1%. Explicit world-aligned reconstruction also outperforms direct 2D satellite fusion.

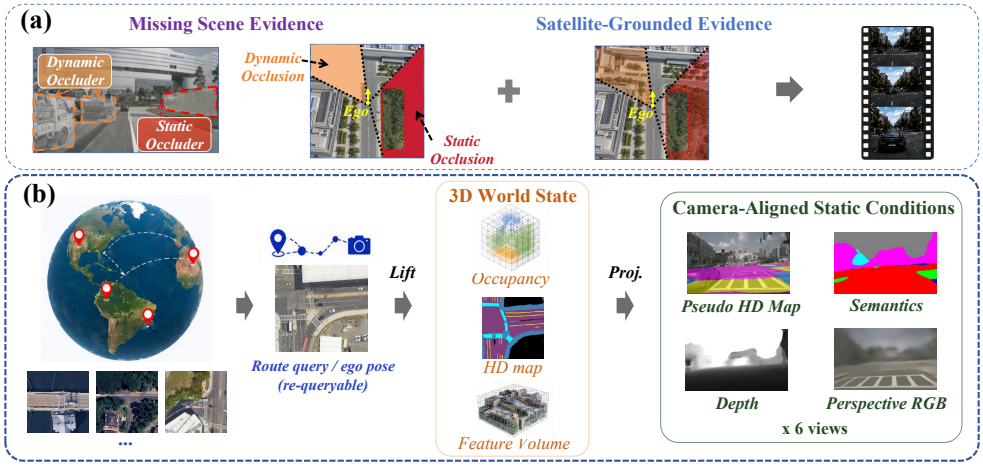


Figure 1: **Motivation and satellite-to-camera evidence path.** (a) Finite visual history and dynamic or static occlusion can leave newly revealed scene content unsupported; satellite imagery can expose the overhead-observable static structure before it must be generated. (b) A route/ego-pose query retrieves a satellite crop, which Loc2Drive lifts into the shared 3D world state $\mathcal{Z} = \{\mathbf{O}, \mathbf{V}, \mathbf{H}\}$. Projecting the same state through six calibrated cameras produces pixel-aligned RGB, depth, semantic, and pseudo-HD-map conditions for the video generator. Specified dynamic agents remain on the separate 3D-box pathway.

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.

1 INTRODUCTION

Driving-video generators and video world models can now synthesize controllable, multi-view scenes over increasingly long horizons (Kim et al., 2021; Gao et al., 2024a; Wang et al., 2024b; Gao et al., 2024b; Chen et al., 2025c). Long-horizon failures are commonly framed as autoregressive error accumulation: generated mistakes re-enter the conditioning context and propagate across rollout windows (Mei et al., 2024; Gao et al., 2024b; Chen et al., 2025c). This explains persistence, but not how strongly the initial hypothesis was constrained. Errors can arise from limited predictive modeling or missing conditioning evidence; either can propagate once generated outputs become future context. Figure 1(a) illustrates this missing-evidence case, while Figure 1(b) shows the Loc2Drive intervention: re-query overhead evidence, reconstruct a shared 3D state, and project it into pixel-aligned conditions for all six cameras.

We distinguish re-observing scene state supported by visual history from revealing previously unobserved state. In the first case, ego motion mainly changes how known static content is viewed, so prediction remains strongly constrained. In the second, the model must complete scene content with no direct visual evidence, leaving the future underdetermined by its conditioning. A vehicle turning at an intersection, for example, can expose post-turn road topology and appearance that were outside the egocentric field of view or hidden around the street corner. More generally, finite reference span, transient vehicles, and static occluders such as buildings, trees, and walls can create the same visibility gap, later resolved through disocclusion. We call the gap between the evidence needed to constrain a future window and the evidence available in its current conditioning a *future-scene observability deficit*. Appendix C.6 provides a descriptive Occ3D analysis of instantaneous onboard sensing incompleteness.

Such a deficit leaves the static scene underconstrained: road topology, intersection layout, and surrounding appearance can remain ambiguous even when the ego trajectory is fixed. The generator must instantiate one hypothesis, which autoregression can then propagate once generated frames become future references; the ambiguity therefore precedes accumulation. The hidden content, rather than the occluder itself, is what matters: transient traffic or static structures can hide persistent geographic state that remains observable from overhead.

Viewed as conditioning, visual history acts as short-term perceptual memory: it preserves recently observed evidence but cannot contain static content that never entered the cameras. We call the complementary interface *geographically addressable generation*: a coordinate and route specify where the sequence unfolds, while satellite imagery provides persistent, location-addressable evidence beyond the current views. At each route position, Loc2Drive lifts a heading-aligned crop into $\mathcal{Z} = \{\mathbf{O}, \mathbf{V}, \mathbf{H}\}$, where $\mathbf{O}$ captures geometry/visibility, $\mathbf{V}$ preserves height-aware appearance cues, and $\mathbf{H}$ encodes HD-map-supervised road semantics. Six-camera projection yields pixel-aligned RGB, depth, semantic, and pseudo-HD-map conditions. Control–State Fusion then selects satellite residuals using the frozen original-control residual and current DiT state before gated injection. Requerying supplies static evidence before missing content must be generated, while 3D boxes specify dynamic agents and the video prior handles unconstrained appearance and dynamics.

Across matched evaluations, Loc2Drive improves the standard three-reference setting even without an inference-time external HD map, retains its advantage from 16 to 128 generated frames, and remains effective with zero visual references. Direct 2D satellite cross-attention reaches 5.74/35.13 FID/FVD, whereas explicit world-aligned reconstruction reaches 5.31/29.77. Point-addressed qualitative queries further show geographic binding without target-location street-view retrieval.

Our contributions are:

- We distinguish the constrained prediction of previously observed scene state from the underdetermined completion of scene content missing from visual evidence, identifying limited future-scene observability as a complementary source of long-horizon uncertainty before autoregressive propagation.
- We introduce Loc2Drive, which lifts route-queryable satellite evidence into an explicit 3D state retaining occupancy geometry, height-aware appearance features, and HD-map-supervised road semantics, projects it into six-camera pixel-aligned conditions, and uses Control–State Fusion for prior-preserving injection.

- We validate this interface under standard conditioning, long autoregressive rollouts, reduced/zero visual-reference budgets, direct-2D versus explicit-state fusion, and point-addressed geographic queries.

2 Related Work

2.1 Long-Horizon Video Generative World Models

Driving-video generation spans controllable simulation (Kim et al., 2021; Yang et al., 2020; Chen et al., 2021; Swerdlow et al., 2024; Gao et al., 2024a), forecasting/planning world models (Wang et al., 2024b; Gao et al., 2024b; Hu et al., 2022; 2023; Min et al., 2024; Yang et al., 2025), and controllable multi-view video generators (Wang et al., 2024a; Li et al., 2024b; Wen et al., 2024; Zhao et al., 2025; Jiang et al., 2024; Mei et al., 2024; Chen et al., 2025c). Recent systems extend resolution, synchronization, geometric anchoring, or horizon (Gao et al., 2025; Deng et al., 2024; Bai et al., 2025; Chen et al., 2025b); M^4World further demonstrates minute-scale multi-view streaming (Cheng et al., 2026). Loc2Drive targets a different bottleneck: future views can reveal static scene state absent from visual history, while the missing evidence can be queried from geography before generation. Dynamic agents remain separately controllable through 3D boxes.

2.2 Implicit and Structured World-State Representations

Video-centric world models encode appearance and dynamics in high-dimensional visual representations (Hu et al., 2023; Gao et al., 2024b; Wang et al., 2024b), whereas structured approaches expose geometry and layout explicitly. DriveWorld separates temporal dynamics from static scene context (Min et al., 2024); OccWorld forecasts semantic 3D occupancy and ego motion (Zheng et al., 2024); WoVoGen uses an explicit temporal world volume (Lu et al., 2024); DiST-4D uses metric depth as a view-consistent geometric representation (Guo et al., 2025); DrivingSphere constructs an occupancy-based 4D world for closed-loop visual synthesis (Yan et al., 2025); and GenieDrive predicts 4D occupancy as a physics-informed state for subsequent multi-view video generation (Yang et al., 2026). Loc2Drive instead builds a coordinate-indexed state that combines occupancy geometry $\mathbf{O}$, a height-aware 3D feature volume $\mathbf{V}$ carrying appearance cues, and road semantics $\mathbf{H}$; the state is refreshed from overhead observations and projected into camera-aligned evidence before video generation.

2.3 Geographic Priors as Externally Queryable World State

Satellite-to-ground work covers cross-view synthesis (Regmi & Borji, 2018; Shi et al., 2022; Li et al., 2024a), trajectory-conditioned generation (Li et al., 2021; Qian et al., 2026a; Ze et al., 2026), renderable 3D scene generation (Qian et al., 2023; Li et al., 2024c; Hua et al., 2025; Qian et al., 2026b), HD-map construction (Gao et al., 2024c; Mazumder & Flohr, 2026; Huang et al., 2025; Lyu et al., 2025), and satellite-assisted occupancy prediction (Chen et al., 2025a). Closest neighbors differ in the source and interface of geographic evidence. UniMLVG uses map, box, and visual-reference conditions (Chen et al., 2025c); SR-AD retrieves satellite and target-location street-view imagery and injects geographic features through cross-attention (Jia et al., 2026); retrieval-grounded city simulation uses GPS-indexed street-view lookahead (Seo et al., 2026); and SatDreamer360 uses satellite-derived tri-plane ray querying (Ze et al., 2026). Loc2Drive treats satellite-only imagery as persistent, location-addressable external spatial memory, resolving each query into a co-registered state that combines occupancy geometry, 3D appearance features, and road semantics before calibrated six-view projection. Appendix A gives a broader comparison.

3 Method

3.1 Overview

Loc2Drive supplies externally observed static evidence before it must be completed by the video prior. Figure 1(b) summarizes route query $\rightarrow$ satellite crop $\rightarrow$ shared 3D state $\rightarrow$ six-view conditions, and Figure 2 gives the detailed architecture. *Stage I* reconstructs $\mathbf{O}$ for geometry/visibility,

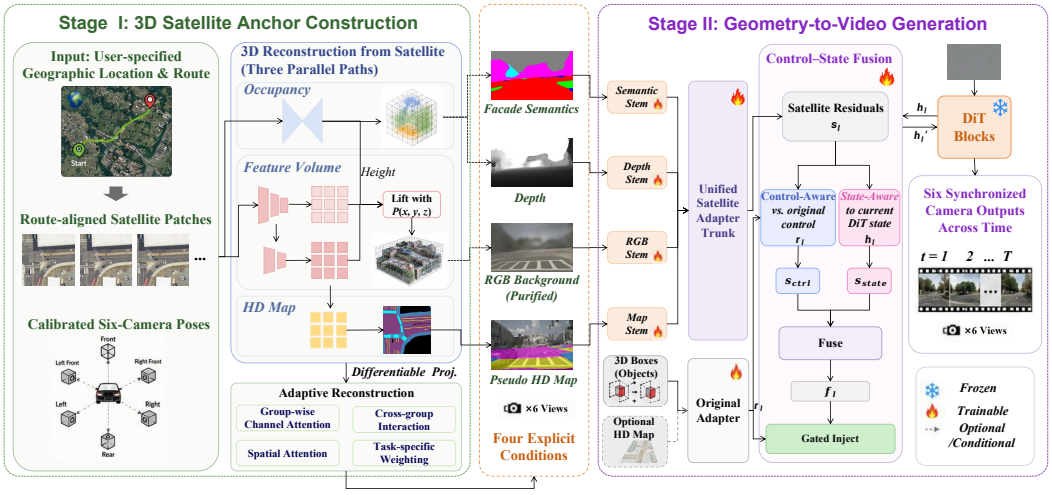


Figure 2: **Architecture of Loc2Drive.** Stage I lifts each route-indexed satellite crop into the shared world state $\{\mathbf{O}, \mathbf{V}, \mathbf{H}\}$ and projects four pixel-aligned conditions to all six views. Stage II parses modalities with separate stems, interacts in a shared satellite trunk, and uses Control–State Fusion for control-aware and state-aware selection followed by gated DiT injection. 3D boxes retain dynamic-agent control on the separate pretrained path. "HD Map" denotes Stage-I supervision; "Optional HD Map" is the inference-time external map.

$\mathbf{V}$ for lifted appearance cues, and $\mathbf{H}$ for fine road structure, then projects RGB, depth, semantic, and pseudo-HD-map conditions to each camera. Because the state comes from overhead rather than street-level reference pixels, it can represent static content hidden in the current street-level views without inheriting transient foreground occluders. *Stage II* follows a parse-first, interact-later design: separate stems feed a shared satellite trunk, then Control–State Fusion selects residuals using the original-control residual and current DiT state before gated injection. The pretrained UniMLVG/OpenDWM multi-view DiT and original control adapter remain frozen (Peebles & Xie, 2023; Yang et al., 2024; Chen et al., 2025c); 3D boxes provide dynamic-agent state when available, while the video prior synthesizes final appearance and unconstrained dynamics. Target-location visual references, an inference-time external HD map, and 3D boxes remain optional.

3.2 Externally Queryable Geographic State and Generation Regimes

Let the planned route be $\tau = \{(\mathbf{p}_t, \psi_t)\}_{t=1}^{T}$, where $\mathbf{p}_t$ is the ego location and ψ_t the route heading. The route specifies ego progression only; it does *not* provide lane boundaries, road width, neighboring branches, intersection geometry, or surrounding road topology. Visual references $\mathcal{R}$ provide observed appearance, the optional inference-time external HD map $\mathcal{M}$ provides structured road evidence, 3D boxes $\mathcal{B}_t$ provide dynamic-layout control, and satellite observations provide geographically indexed overhead evidence. These conditions play distinct, complementary roles. No text condition is used in any reported evaluation. Throughout the paper, "inference-time external HD map" refers only to $\mathcal{M}$; HD-map annotations still supervise the internal satellite-to-BEV road-semantic predictor during Stage I. At timestamp t, a geographic query retrieves a heading-aligned satellite patch

$$I_{\mathrm{sat},t} = Q(\mathbf{p}_t, \psi_t, s), \tag{1}$$

where Q is the geographic query operator and s is the physical crop scale. Given the queried sequence in Eq. (1), Loc2Drive generates synchronized videos

$$\widehat{\mathcal{V}}_{1:T}^{1:N_{\mathrm{cam}}} = G\big(\{I_{\mathrm{sat},t}\}_{t=1}^{T}, \{\Pi_t^c\}_{t,c}; \{\mathcal{B}_t\}_{t=1}^{T}, \mathcal{M}, \mathcal{R}\big)\,, \tag{2}$$

for $N_{\mathrm{cam}} = 6$ synchronized cameras, where G denotes the video generator and $\Pi_t^c = (K^c, T_{w\to c,t})$ contains camera intrinsics and the world-to-camera pose. The 3D boxes $\mathcal{B}_t$, inference-time external HD map $\mathcal{M}$, and visual references $\mathcal{R}$ are optional; the strict setting removes all three, and no text

condition is supplied. Equation (2) specifies a conditional generation interface rather than a deterministic reconstruction target: the satellite-derived state constrains overhead-observable geometry, road structure, and appearance cues, 3D boxes constrain dynamic agents when provided, and residual street-level appearance is synthesized by the video prior. The exact regimes and route records are given in Appendix C.1 (*Query Regimes and Route Record*).

3.3 Explicit 3D World-State Reconstruction from Overhead Observations

Three-path differentiable lifting. For clarity, we omit the route index t in the remainder of Stage I. Given $I_{\mathrm{sat}} \in \mathbb{R}^{H\times W\times 3}$, the model constructs a shared $X \times Y \times Z$ scene representation. *Path A* predicts semantic occupancy $\mathbf{O}$ with a U-shaped encoder and 3D decoder, together with auxiliary height logits D_{sat} and semantic logits S_{sat}. *Path B* preserves overhead appearance through shallow texture features F_{tex} and deeper semantic-geometric features F_{bev}. For each overhead cell, the height distribution is peak-normalized and modulated by satellite semantics:

$$\begin{aligned} p_h(z \mid x, y) &= \mathrm{softmax}_z\, D_{\mathrm{sat}}(x, y, z), \\ \bar{p}_h(z \mid x, y) &= \frac{p_h(z \mid x, y)}{\max_{z'} p_h(z' \mid x, y) + \epsilon}, \\ \mathbf{V}_0(x, y, z) &= \bar{p}_h(z \mid x, y)[F_{\mathrm{bev}}(x, y); F_{\mathrm{tex}}(x, y)], \\ m_h(x, y, z) &= \tanh(\phi_h([\mathrm{softmax}(S_{\mathrm{sat}}); \bar{p}_h])), \\ \mathbf{V}(x, y, z) &= \mathbf{V}_0(x, y, z) + m_h(x, y, z) \odot \mathbf{V}_0(x, y, z). \end{aligned} \tag{3}$$

Equation (3) spreads each overhead feature across plausible heights while preserving the strongest height hypothesis; semantic modulation adapts the vertical profile to the local category evidence. At each voxel, occupancy $\mathbf{O}$ and appearance feature $\mathbf{V}$ are co-registered and later sampled at identical world coordinates. *Path C* predicts an HD-map-supervised BEV road-semantic layer $\mathbf{H}$, giving the shared state $\mathcal{Z} = \{\mathbf{O}, \mathbf{V}, \mathbf{H}\}$. The three components serve different roles: $\mathbf{O}$ supplies semantic geometry and visibility, $\mathbf{V}$ carries height-aware appearance/texture evidence beyond occupancy labels, and $\mathbf{H}$ encodes fine ground-plane driving structure. $\mathbf{H}$ is an *internal satellite prediction* trained with HD-map annotations during Stage I and is distinct from the optional inference-time external HD map $\mathcal{M}$. Accordingly, inference without an external HD map removes $\mathcal{M}$ but not this training supervision. Full lifting details and voxel layout are in Appendix B.1.

Shared-3D projection to pixel-aligned six-view conditions. For compactness, let $\mathcal{P}$ denote the complete Stage-I projection-and-reconstruction operator. For calibrated camera c,

$$\mathbf{A}^c = \left(\mathbf{A}^c_{\mathrm{rgb}}, \mathbf{A}^c_{\mathrm{depth}}, \mathbf{A}^c_{\mathrm{sem}}, \mathbf{A}^c_{\mathrm{road}}\right) = \mathcal{P}(\mathcal{Z}, \Pi^c), \qquad c \in \{1, \ldots, N_{\mathrm{cam}}\}. \tag{4}$$

Here $\mathbf{A}^c$ contains four mutually pixel-aligned condition images—perspective RGB, depth, scene semantics, and the satellite-derived pseudo-HD map—rather than final generated video frames. Equation (4) makes all cameras query the same $\mathcal{Z}(\mathbf{x})$ and differ only in projection, so corresponding rays share one world-coordinate representation rather than six independent overhead-to-view translations. A key design is to reuse the semantic-occupancy *air* class as an occupancy-derived soft-visibility signal rather than learning an unrelated physical density field. For camera pixel (u, v), let $a_k = 1 - o_k^{(\mathrm{air})}$ be the occupied probability at ordered ray sample k, m_k indicate whether the sample lies inside the valid voxel lattice, and $\mathbf{t}_k$ be the co-registered feature sampled from $[\mathbf{O}; \mathbf{V}]$. The implemented soft opacity is

$$\widetilde{a}_k = a_k m_k\, \sigma(6(a_k - 0.1)), \qquad T_k = \prod_{j<k}(1 - \widetilde{a}_j), \qquad \widehat{\mathbf{F}}^c_{\mathrm{rend}}(u, v) = \sum_{k=1}^{N_{\mathrm{r}}} T_k \widetilde{a}_k \mathbf{t}_k. \tag{5}$$

Free-space voxels remain transparent while occupied support contributes to the rendered condition and attenuates farther samples. These are intermediate geometric/appearance visibility weights, not a radiometrically calibrated extinction model. The same transmittance weights underlie feature, expected-depth, and scene-semantic rendering; remaining ray probability is completed explicitly as described in Appendix B.1. The BEV road-semantic layer $\mathbf{H}$ follows the UniMLVG/OpenDWM ground-plane map projection convention and is projected with the full calibrated camera intrinsics and extrinsics. Shared-3D implementation checks are provided in Appendix F.1 (*Shared-3D Consistency Checks*).

Adaptive reconstruction. The renderer exposes heterogeneous occupancy, semantic, depth, deep-appearance, and shallow-texture groups. Rather than forcing one mixture to serve every output, Loc2Drive first normalizes each group, lets the groups exchange spatial evidence, and then learns task-specific mixtures for RGB, depth, and scene semantics; $\mathbf{H}$ is projected directly as the pseudo-HD-map condition. This parse–share–specialize design keeps the shared 3D evidence common while allowing each rendered modality to emphasize the cues it needs. Architecture and weighting equations are given in Appendix B.2.

3.4 Route-Aligned Re-Queryable World-State Integration

Coordinate-indexed evidence along route progression. We retain the baseline's route progression and attach a geographic observation to each 10-fps sampled position $\mathbf{p}_t$. Loc2Drive queries the corresponding satellite crop, reconstructs $\mathcal{Z}_t$, and projects it with the calibrated pose Π_t^c. Each route position therefore triggers a fresh external read rather than a recurrent update from generated references, so $\mathcal{Z}_t$ does not directly inherit their visual errors or transient foreground occlusions. The state can expose stable road, building, vegetation, and layout evidence hidden from the current street-level support before that static content must be instantiated as a generative hypothesis and reused as visual history.

Parse-first integration and Control–State Fusion. Four modality-specific stems encode RGB, depth, scene semantics, and pseudo-HD maps before a shared trunk produces satellite residuals $\{\mathbf{s}_\ell\}$. At DiT layer ℓ, a *control-aware* branch conditions on the frozen original-control residual $\mathbf{r}_\ell$, while a *state-aware* branch conditions on the current hidden state $\mathbf{h}_\ell$; their selected residuals fuse into $\mathbf{f}_\ell$. Separate gates then update

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + (1 + \tanh \mathbf{g}^r_\ell) \odot \mathbf{r}_\ell + \sigma(\mathbf{g}^f_\ell) \odot \mathbf{f}_\ell. \tag{6}$$

Zero-initialized fusion and a suppressed satellite gate start training close to the pretrained hidden-plus-control update. The two selection equations, gate prediction, stem/trunk dimensions, and initialization are given in Appendix B.3.

3.5 Geographic Background and Separate Dynamic Control

Occluder type and hidden-content type are distinct: a moving vehicle may hide a static road, and a tree or building may hide another static region. Reference-conditioned generation therefore receives images in which transient agents and persistent background can be visually entangled. Loc2Drive removes this dependency from the static branch: satellite observations reconstruct the persistent geographic background across these visibility gaps, while 3D boxes specify dynamic agents. Because the 2019 ground observations and newer satellite imagery are asynchronous, we mask dynamic annotations and cast shadows, filter night frames, and prevent transient foreground content from entering the static world state. The resulting background condition is therefore invariant to whether a transient vehicle happened to occlude the scene in a reference frame. This factorization keeps background conditioning independent of transient foreground and makes the box pathway the explicit source of specified agent layout rather than requiring agent information to pass through the background condition. The satellite maps follow the SA-Occ / Occ3D-nuScenes-SatExt source (Chen et al., 2025a) and are queried densely along 10-fps interpolated routes. Box edits affect only the dynamic-control path; details are in Appendices D.4 and B.5.

3.6 Training Objectives

Stage I jointly optimizes occupancy, perspective RGB, depth, semantics, satellite-derived pseudo-HD-map reconstruction, and pose-aligned temporal state consistency:

$$\begin{aligned}\mathcal{L}_{\mathrm{I}} = \lambda_{\mathrm{occ}}\mathcal{L}_{\mathrm{occ}} + \lambda_{\mathrm{rgb}}\mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}} \\ + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{map}}\mathcal{L}_{\mathrm{map}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}.\end{aligned} \tag{7}$$

Equation (7) groups the implementation losses by supervision target: semantic occupancy, masked RGB reconstruction, depth, perspective semantics, the satellite-derived road-semantic condition, and temporal alignment. The code uses constituent losses and target-specific multipliers within these groups; the exact terms and implementation weights are given in Appendices B.4–B.5. Stage

Table 1: Driving-video generation on nuScenes. B: 3D boxes; H: inference-time external HD map; Sat.: satellite-derived state from route-indexed imagery; Str.: retrieved target-location street imagery. The upper block gives literature-reported context under heterogeneous protocols; values follow the cited works or subsequent comparison tables and are not controlled head-to-head results. The lower block is our matched comparison, where FID aggregates 32 generated frames and FVD uses the standard 16-frame window.

Method	Cond.	FID↓	FVD↓	mAP$_{obj}$ ↑	mIoU$_{road}$ ↑	mIoU$_{veh}$ ↑
Oracle	–	–	–	35.56	73.67	31.86
DriveGAN	B+H	73.4	502.3	–	–	–
DriveDreamer	B+H	52.6	452.0	–	–	–
MagicDrive	B+H	19.1	218.1	12.30	61.05	27.01
MagicDrive-V2	B+H	–	94.84	–	–	–
Drive-WM	B+H	15.8	122.7	20.66	65.07	27.19
DriveDreamer-2	B+H	11.2	55.7	–	–	–
DreamForge	B+H	14.61	209.9	13.01	65.27	28.36
DiVE	B+H	–	94.6	24.55	–	–
SR-AD	B+H+Sat.+Str.	5.60	29.97	–	–	–
UniMLVG	B+H	5.82	36.11	22.50	70.81	29.12
Loc2Drive (no ext. H)	B+Sat.	5.44	30.80	22.77	71.27	**29.39**
Loc2Drive (full)	B+H+Sat.	**5.31**	**29.77**	**22.82**	**71.41**	29.34

II freezes the pretrained DiT and original control adapter and trains only the satellite adapter and fusion modules with the baseline flow-matching objective.

4 Experiments

4.1 Protocol

We evaluate on nuScenes with the UniMLVG preprocessing, backbone, and scene-level train/validation split. Controlled comparisons match validation clips, sampling seeds, denoising/guidance, resolution, rollout overlap, metric checkpoints, and six-view packing. One reference is one synchronized six-camera timestamp; all 3/1/0-reference results use the same fixed checkpoint, and Loc2Drive metrics average three stochastic evaluations.

Each flow-matching window predicts 16 frames. FVD denotes the standard 16-frame window; FVD$_{32/64/128}$ chains 2/4/8 windows, reusing the final three generated frames as the next references. FID aggregates 32 generated frames. FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) measure frame- and video-level distribution quality, while fixed perception models probe object and road structure. All quantitative rows retain the same 3D-box control. We evaluate standard conditioning, longer rollouts, reduced reference budgets, explicit-versus-2D fusion, and point-addressed queries; each horizon/reference budget is its own matched setting. Appendices C.2–C.3 give the full protocol and frame indexing.

4.2 Standard Context: Complementing Road-Map Control

Table 1 summarizes the standard-context comparison. Under the standard three-reference setting, Loc2Drive without an inference-time external HD map reaches 5.44/30.80 FID/FVD, outperforming UniMLVG's 5.82/36.11 even though UniMLVG receives its external HD map, with comparable or better fixed object/road/vehicle probes. Supplying the external map to Loc2Drive as an additional condition lowers FVD further to 29.77. Together, the no-H and +H variants show that the satellite-derived state contributes location-specific scene evidence beyond road topology alone while remaining complementary to external map control when both are available.

4.3 Long-Horizon Generation

With three visual references, Loc2Drive outperforms UniMLVG at every evaluated horizon (Table 2(a)). Without the inference-time external HD map, it reduces FVD by 14.7% on the matched 16-frame subset and FVD$_{128}$ by 29.7% on the separately eligible 128-frame subset; with the map,

Table 2: Long-horizon and reference-budget stress tests with a fixed checkpoint. All quantitative rows retain the same 3D-box control; +H denotes an inference-time external HD map and Sat. the satellite-derived state. Each autoregressive window predicts 16 new frames from the previous window's last three generated frames, so $FVD_{32/64/128}$ evaluates continuous 2/4/8-window rollouts. Comparisons are matched within each reference budget and horizon; changing the reference count also changes the inherited target start index.

(a) Long-horizon sweep with three references

Method	FID	FVD	FVD_{32}	FVD_{64}	FVD_{128}
UniMLVG (+H)	5.82	36.11	67.25	87.56	88.65
Loc2Drive (Sat., no H)	5.44	30.80	57.91	67.28	62.33
Loc2Drive (Sat.+H)	**5.31**	**29.77**	**56.26**	**66.70**	**61.44**

(b) Reference-budget sweep

	1 reference			0 references		
Method	FID	FVD	FVD_{32}	FID	FVD	FVD_{32}
UniMLVG (+H)	6.53	41.41	60.71	19.34	106.01	130.67
Loc2Drive (Sat., no H)	6.25	35.33	51.04	16.31	75.19	100.44
Loc2Drive (Sat.+H)	**6.19**	**34.91**	**49.89**	**15.86**	**73.96**	**98.11**

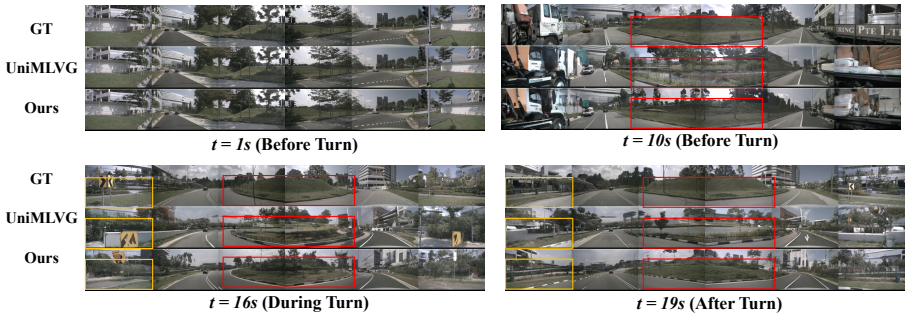


Figure 3: **From missing scene evidence to autoregressive propagation.** The methods are similar before the turn. As post-turn static content with weaker visual support is revealed, UniMLVG develops a spurious branch and topology distortion that persist after generated frames re-enter the reference history; Loc2Drive continues to receive route-indexed satellite-grounded static evidence. UniMLVG receives its standard inference-time external HD map.

the reductions are 17.6% and 30.7%. Because the eligible clip set changes with rollout length, these are separate within-horizon comparisons (Appendix C.2). As generated hypotheses increasingly replace observed history, the persistent gain is consistent with independently refreshed satellite-grounded static evidence during rollout.

4.4 When Visual References Disappear

Reducing the visual-reference budget decreases how much future scene state is directly constrained, increasing the missing scene evidence that must otherwise be completed from the model prior. With one reference, Loc2Drive remains better than UniMLVG. With no visual references, Loc2Drive without an inference-time external HD map lowers FVD by 29.1% and FVD_{32} by 23.1% relative to UniMLVG even though UniMLVG still receives its external map. Thus, even when direct visual evidence disappears, the satellite-grounded state continues to provide static geographic evidence while the same 3D-box pathway keeps dynamic-layout conditioning fixed. The matched comparison and inherited target-index offset are detailed in Appendices C.2–C.3.

Table 3: Effect of the explicit-state design and geographic integration on generation quality. The upper panel removes components of the reconstructed state; the lower panel compares direct 2D fusion with reconstruction-and-projection variants. Full ablations appear in Appendix E.1.

(a) Explicit state and reconstruction

Variant	FID↓	FVD↓
Full explicit state	**5.31**	**29.77**
No height-aware feature volume	5.75	32.41
No road-semantic condition	5.58	31.88
No adaptive reconstruction	5.61	31.92
No background purification	5.67	33.64

(b) Geographic integration

Variant	FID↓	FVD↓
No satellite injection	5.82	36.11
2D satellite cross-attention	5.74	35.13
Expanded input layer	5.44	30.47
Parallel satellite adapter	5.33	31.20
Split-to-unified adapter	**5.31**	**29.77**



Figure 4: **Point-addressed grounding without conventional target-location conditions.** With visual references, external HD map, 3D boxes, and text removed, Loc2Drive follows the queried geography and empty dynamic layout. Appendix D.3 reports all six cities.

4.5 What Evidence Matters, and How Should It Enter?

Table 3 separates *what* the satellite path should retain from *how* it should enter the generator. Direct 2D cross-attention reaches 5.74/35.13 FID/FVD, versus 5.31/29.77 for the explicit occupancy–appearance–road pipeline; removing the feature volume, road semantics, or background purification degrades performance to 5.75/32.41, 5.58/31.88, and 5.67/33.64. For integration, the split-to-unified adapter reaches 5.31/29.77, outperforming expanded-input and parallel variants. Control–State Fusion further outperforms the best simple spatial selector (5.48/32.21) and channel-wise post-selection scaling (5.34/30.71; Appendix E.1).

4.6 Point-Addressed Geographic Grounding

We additionally probe point-addressed queries outside the Boston/Singapore geographic-pairing set with no target-location visual references, inference-time external HD map, 3D boxes, or text (Figure 4). Both methods retain the same pretrained prior and route/camera specification; only Loc2Drive receives coordinate-indexed satellite state. In these examples, UniMLVG can remain plausible yet generic or route-inconsistent, whereas Loc2Drive better follows the requested route and visible overhead structure; empty-layout cases also more cleanly suppress traffic residuals. Without synchronized six-view ground truth for arbitrary sites, we evaluate geographic binding and condition compliance qualitatively.

5 Discussion and Conclusion

Loc2Drive complements short-term visual memory with persistent, location-addressable satellite evidence. Rather than waiting for an underconstrained static hypothesis to enter the autoregressive history, it reconstructs geometry, 3D appearance, and road semantics from overhead observations and projects them into six calibrated views. Control–State Fusion selectively injects these residuals while preserving the original control path, and specified agents remain on the separate 3D-box pathway. Across matched evaluations, explicit reconstruction and selective integration improve standard, long-horizon, and zero-reference generation. Re-queryable evidence therefore grounds persistent geography before propagation, while the video prior remains responsible for unconstrained appearance and dynamics (Appendix F.2).

Reproducibility Statement

The main paper specifies the task interface, datasets, training stages, model components, and matched evaluation protocol. Appendix B provides reconstruction, fusion, temporal-consistency, and training details; Appendices C.1–C.5 define query/reference/horizon protocols; and Appendix F records geometric checks, data provenance, scope, and limitations. We will release code, geographic query/route metadata, alignment and preprocessing scripts, and redistributable derived annotations. The underlying Google Earth imagery follows the SA-Occ/SatExt source and remains subject to the imagery provider's access and redistribution terms.

AI Use Statement

Generative AI tools were used for manuscript drafting and readability editing, literature search and reference checking, iterative discussion of research framing and experimental design/organization, interpretation of already-computed results, consistency checking of mathematical notation, and editing of scientific-figure labels. They were not used to generate the reported experimental measurements. All AI-assisted text, references, mathematical notation, claims, and research decisions were reviewed and verified by the authors, who take responsibility for the submission.

References

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Chen Chen, Zhirui Wang, Taowei Sheng, Yi Jiang, Yundu Li, Peirui Cheng, Luning Zhang, Kaiqiang Chen, Yanfeng Hu, Xue Yang, and Xian Sun. SA-Occ: Satellite-assisted 3d occupancy prediction in real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27021–27030, 2025a.

Jiacheng Chen, Ziyu Jiang, Mingfu Liang, Bingbing Zhuang, Jong-Chyi Su, Sparsh Garg, Ying Wu, and Manmohan Chandraker. AutoScape: Geometry-consistent long-horizon scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25700–25711, 2025b.

Rui Chen, Zehuan Wu, Yichen Liu, Yuxin Guo, Jingcheng Ni, Haifeng Xia, and Siyu Xia. UniMLVG: Unified framework for multi-view long video generation with comprehensive control capabilities for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25453–25463, 2025c.

Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. GeoSim: Realistic video simulation via geometry-aware composition for self-driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7230–7240, 2021.

Ke Cheng, Hanqiao Ye, Lei Shi, Yahui Liu, Yunhan Shen, Jingtao Dong, Zhenke Wang, Wenxuan Ao, Weixiang Xu, Kaining Huang, and Shuhan Shen. M^4World: A multi-view multimodal driving world model for interactive object manipulation and minute-long streaming. *arXiv preprint arXiv:2607.14005*, 2026.

Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657513.

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. In *International Conference on Learning Representations*, 2024a.

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*, pp. 28135–28144, 2025.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2906.

Wenjie Gao, Jiawei Fu, Yanqing Shen, Haodong Jing, Shitao Chen, and Nanning Zheng. Complementing onboard sensors with satellite maps: a new perspective for HD map construction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11103–11109. IEEE, 2024c.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, and Hao Zhao. DiST-4D: Disentangled spatiotemporal diffusion with metric depth for 4d driving scene generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27231–27241, 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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Anthony Hu, Gianluca Corrado, Nicolas Griffiths, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, and Jamie Shotton. Model-based imitation learning for urban driving. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1505.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Tongyan Hua, Lutao Jiang, Ying-Cong Chen, and Wufan Zhao. Sat2City: 3d city generation from a single satellite image with cascaded latent diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27978–27988, 2025.

Bingyuan Huang, Guanyi Zhao, Qian Xu, Yang Lou, Yung-Hui Li, and Jianping Wang. SATMapTR: Satellite image enhanced online hd map construction. *arXiv preprint arXiv:2512.11319*, 2025.

Xiaosong Jia, Chenhe Zhang, Yule Jiang, Songbur Wong, Zhiyuan Zhang, Chen Chen, Shaofeng Zhang, Xuanhe Zhou, Xue Yang, Junchi Yan, and Yu-Gang Jiang. Spatial retrieval augmented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17787–17797, 2026.

Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, Kun Zhan, Peng Jia, and Miao Zhang. DiVE: DiT-based video generation with enhanced control. *arXiv preprint arXiv:2409.01595*, 2024.

Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. DriveGAN: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5820–5829, 2021.

Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhimeng Zheng, Zilong Huang, Dahua Lin, and Conghui He. CrossViewDiff: A cross-view diffusion model for satellite-to-street view synthesis. *arXiv preprint arXiv:2408.14765*, 2024a.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. DrivingDiffusion: Layout-guided multi-view driving scenarios video generation with latent diffusion model. In *European Conference on Computer Vision*, pp. 469–485. Springer, 2024b. doi: 10.1007/978-3-031-73229-4_27.

Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Rongjun Qin, Marc Pollefeys, and Martin R. Oswald. Sat2Vid: Street-view panoramic video synthesis from a single satellite image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12436–12445, 2021.

Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R. Oswald. Sat2Scene: 3d urban scene generation from satellite images with diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7141–7150, 2024c.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pp. 329–345. Springer, 2024. doi: 10.1007/978-3-031-72989-8_19.

Hongyu Lyu, Julie Stephany Berrio Perez, Yaoqi Huang, Kunming Li, Mao Shan, and Stewart Worrall. Online high-definition map construction for autonomous vehicles: A comprehensive survey. *Journal of Sensor and Actuator Networks*, 14(1):15, 2025. doi: 10.3390/jsan14010015.

Kanak Mazumder and Fabian B. Flohr. Satmap: Revisiting satellite maps as prior for online hd map construction. In *Pattern Recognition: 28th International Conference, ICPR 2026, Proceedings, Part XV*, pp. 185–197. Springer, 2026.

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.

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, and Bin Dai. DriveWorld: 4d pre-trained scene understanding via world models for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15522–15533, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2Density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3683–3692, 2023.

Ming Qian, Bin Tan, Qiuyu Wang, Xianwei Zheng, Hanjiang Xiong, Gui-Song Xia, Yujun Shen, and Nan Xue. Seeing through satellite images at street views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5692–5709, 2026a. doi: 10.1109/TPAMI.2026.3652860.

Ming Qian, Zimin Xia, Changkun Liu, Shuailei Ma, Wen Wang, Zeran Ke, Bin Tan, Hang Zhang, and Gui-Song Xia. Sat3DGen: Comprehensive street-level 3d scene generation from single satellite image. In *International Conference on Learning Representations*, 2026b.

Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3501–3510, 2018.

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

Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10009–10022, 2022. doi: 10.1109/TPAMI.2022.3140750.

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*, pp. 2446–2454, 2020.

Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4):3578–3585, 2024. doi: 10.1109/LRA.2024.3368234.

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.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024a. doi: 10.1007/978-3-031-73195-2_4.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024b.

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, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Tianyi Yan, Dongming Wu, Wencheng Han, Junpeng Jiang, Xia Zhou, Kun Zhan, Cheng-zhong Xu, and Jianbing Shen. DrivingSphere: Building a high-fidelity 4d world for closed-loop simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27531–27541, 2025.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14662–14672, 2024.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. ReSim: Reliable world simulation for autonomous driving. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5591.

Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, and Henrik Kretzschmar. SurfelGAN: Synthesizing realistic sensor data for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11118–11127, 2020.

Zhenya Yang, Zhe Liu, Yuxiang Lu, Liping Hou, Chenxuan Miao, Siyi Peng, Bailan Feng, Xiang Bai, and Hengshuang Zhao. GenieDrive: Towards physics-aware driving world model with 4d occupancy guided video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35680–35690, 2026.

Xianghui Ze, Beiyi Zhu, Zhenbo Song, Jianfeng Lu, and Yujiao Shi. SatDreamer360: Multiview-consistent generation of ground-level scenes from satellite imagery. In *International Conference on Learning Representations*, 2026.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10412–10420, 2025. doi: 10.1609/aaai.v39i10.33130.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024. doi: 10.1007/978-3-031-72624-8_4.

A SCOPE OF EXPLICIT WORLD-STATE GROUNDING

Loc2Drive uses a coordinate–route query and route-aligned satellite observations to construct an explicit external state for synchronized six-view video generation. Conceptually, visual history provides short-term perceptual memory for observed scene state, while route-indexed satellite imagery provides persistent, location-addressable spatial evidence for static content outside that memory. Finite reference span, egocentric field of view, and dynamic or static occlusion can create this missing scene evidence; turning is an intuitive case because a rapid viewing-direction change can expose post-turn road topology and appearance that were not supported by the current view. The key distinction under occlusion is between the occluder and the hidden content: a moving vehicle may hide a static road or building, while a static occluder may hide another part of the static scene. Much of this persistent geographic content can still be observed from overhead. Loc2Drive resolves that satellite-grounded static component into semantic occupancy for geometry/visibility, a height-aware 3D feature volume that preserves satellite texture and appearance cues, and HD-map-supervised BEV road semantics for fine driving structure, then projects the shared state through six calibrated cameras to obtain pixel-aligned RGB, depth, semantic, and pseudo-HD-map conditions before the corresponding content must be completed from the video prior. Dynamic agent state is supplied through the separate 3D-box pathway when available; remaining unconstrained appearance is synthesized by the video prior. The interface requires neither per-query optimization nor target-location street-view retrieval.

The state is externally re-queryable and independent of generated-reference recursion. We evaluate it under long autoregressive rollouts, reduced-reference budgets, direct-2D versus explicit-state controls, and point-addressed geographic queries. Appendix C.6 provides complementary Occ3D context on instantaneous sensing incompleteness; this statistic does not by itself identify whether hidden content is static/dynamic or overhead-visible.

The closest neighboring systems solve related parts of the problem with different state interfaces. Table 4 summarizes the capability boundary relevant to Loc2Drive. UniMLVG provides controllable long-horizon six-view generation with reference, map, and dynamic conditions (Chen et al., 2025c). MagicDrive-V2 scales controllable multi-view driving-video generation to higher resolution and longer sequences (Gao et al., 2025); AutoScape uses iteratively generated RGB-D keyframes as long-horizon geometry anchors (Chen et al., 2025b); M^4World targets stable minute-long multi-view streaming (Cheng et al., 2026); and Vista studies generalizable long-horizon driving world modeling (Gao et al., 2024b). WoVoGen explicitly predicts a temporal world volume of HD maps and occupancy for multi-camera generation (Lu et al., 2024), DiST-4D uses metric depth as a geometric representation for temporal and spatial synthesis (Guo et al., 2025), and GenieDrive uses predicted 4D occupancy as a physics-informed state for subsequent video generation (Yang et al., 2026). SR-AD (Jia et al., 2026) retrieves geographic imagery along future trajectories and injects geographic features through cross-attention, including target-location street-view information. Loc2Drive instead turns satellite-only geographic evidence into an externally re-queryable heterogeneous state that combines semantic occupancy, lifted 3D appearance/texture features, and HD-map-supervised road semantics, projects that state into synchronized six-camera conditions before diffusion, and tests the interface under long autoregressive rollouts, reduced visual-reference budgets, and point-addressed queries.

For context, SR-AD reports 5.60 FID / 29.97 FVD under its source-paper UniMLVG-lineage protocol; our controlled method comparisons are against the matched UniMLVG baseline. Loc2Drive reaches 5.31/29.77 without target-location street-view retrieval. Our representation control gives 5.82/36.11 without satellite input, 5.74/35.13 with direct 2D satellite cross-attention, and 5.31/29.77 with the explicit co-registered reconstruction-and-projection pipeline. The gain therefore supports the explicit co-registered reconstruction-and-projection pipeline as a whole over direct 2D fusion under the same Stage-II training setup.

Table 4: Capability boundary relative to the closest neighboring directions.

Neighbor	Existing capability	Loc2Drive distinction
UniMLVG	Controllable long-horizon six-view generation with map, boxes, and visual references	Route-indexed satellite state independent of generated references; reduced-reference and long-rollout stress tests
SR-AD	Future-route geographic retrieval and DiT cross-attention using satellite/street-view geographic imagery	No target-location street-view retrieval; satellite evidence is reconstructed into occupancy, height-aware appearance, and road semantics before camera projection
SatDreamer360	Satellite-to-ground generation with triplane geometry and ray-based feature querying	Integration into a pretrained controllable six-camera driving-video prior with explicit O/V/H role separation
GenieDrive	Predict 4D occupancy as a physics-informed state for multi-view video generation	Externally query and refresh a heterogeneous static state that retains 3D appearance features and road semantics in addition to occupancy
Satellite-to-3D / occupancy	Recover or predict structured 3D scene information from geographic priors	Use the structured state as a conditioner for long-horizon video generation rather than as the final output

B Additional Method and Training Details

This section follows the implemented pipeline from world-state construction to reconstruction, adapter fusion, temporal consistency, and training. Section B.6 summarizes the notation.

B.1 World-State Construction

Figure 2 summarizes the three-path world-state construction. The semantic occupancy and feature volumes use the same Occ3D-aligned lattice over $[-40, 40]^2$ m with $200 \times 200 \times 16$ cells and 0.4-m cubic voxels; the vertical support is $z \in [-1.0, 5.4]$ m. For an overhead cell (x, y), the height head predicts

$$p_h(z \mid x, y) = \mathrm{softmax}_z \, D_{\mathrm{sat}}(x, y, z), \tag{8}$$

and peak normalization gives

$$\bar{p}_h(z \mid x, y) = \frac{p_h(z \mid x, y)}{\max_{z'} p_h(z' \mid x, y) + \epsilon}. \tag{9}$$

Let F_{bev} and F_{tex} denote deep semantic-geometric and shallow texture features. The peak-normalized base volume and semantic modulation are

$$\mathbf{V}_0(x, y, z) = \bar{p}_h(z \mid x, y)[F_{\mathrm{bev}}(x, y); F_{\mathrm{tex}}(x, y)], \tag{10}$$

$$\begin{aligned} m_h(x, y, z) &= \tanh(\phi_h([\mathrm{softmax}(S_{\mathrm{sat}}); \bar{p}_h])) \,, \\ \mathbf{V}(x, y, z) &= \mathbf{V}_0(x, y, z) + m_h(x, y, z) \odot \mathbf{V}_0(x, y, z). \end{aligned} \tag{11}$$

The occupancy tensor $\mathbf{O}$ and feature tensor $\mathbf{V}$ share the lattice and are sampled jointly during rendering. Path C predicts the BEV semantic layer $\mathbf{H}$ from satellite features; it is supervised by HD-map annotations in Stage I, predicted from satellite imagery at inference, and is distinct from an inference-time external HD map.

Occupancy-derived soft visibility. For a calibrated camera ray, the implementation samples $N_{\mathrm{r}} = 89$ depths uniformly between 1 and 64 m. Let $o_k^{(\mathrm{air})}$ be the semantic-occupancy air probability, $a_k = 1 - o_k^{(\mathrm{air})}$, and $m_k \in \{0, 1\}$ indicate whether the k-th sample lies inside the voxel lattice. The renderer uses

$$\widetilde{a}_k = a_k m_k \, \sigma(6(a_k - 0.1)) \,, \qquad T_k = \prod_{j<k} (1 - \widetilde{a}_j), \qquad w_k = T_k \widetilde{a}_k, \tag{12}$$

$$\widehat{\mathbf{F}}^{c}_{\mathrm{rend}}(u,v) = \sum_{k=1}^{N_{\mathrm{r}}} w_k \mathbf{t}_k. \tag{13}$$

The approximately 0.716-m sampling interval is a ray-query resolution, not a physical extinction coefficient. Free-space support remains transparent and occupied support attenuates farther samples; the weights are intermediate soft-visibility weights for generating camera-aligned conditions rather than radiometrically calibrated volume rendering. The same weights form the in-volume component of feature, depth, and semantic rendering. The road-semantic tensor $\mathbf{H}$ follows the UniMLVG/OpenDWM ground-plane map convention and is projected using the full calibrated camera intrinsics and extrinsics.

Remaining-probability completion. The renderer retains ray probability that does not terminate inside the voxel grid,

$$r(u,v) = \prod_{k=1}^{N_{\mathrm{r}}} (1 - \widetilde{a}_k(u,v)).$$

For semantic rendering, the remaining mass is assigned to the explicit air class. If $\mathbf{o}_k \in \mathbb{R}^{C_{\mathrm{sem}}}$ is the sampled semantic probability over C_{sem} classes and $\mathbf{e}_{\mathrm{air}}$ is the air one-hot vector,

$$\widetilde{\mathbf{p}}_{\mathrm{sem}} = \sum_{k=1}^{N_{\mathrm{r}}} w_k \mathbf{o}_k + r\mathbf{e}_{\mathrm{air}}, \qquad \widehat{\mathbf{S}} = \log \frac{\widetilde{\mathbf{p}}_{\mathrm{sem}}}{\sum_{j=1}^{C_{\mathrm{sem}}} \widetilde{p}_{\mathrm{sem},j} + \epsilon}.$$

The resulting log-probabilities can be used as semantic logits because their softmax recovers the normalized semantic distribution (up to the numerical ϵ term). For depth, remaining mass is placed immediately after the last valid in-volume sample, or at the farthest sampled depth if the ray never enters the lattice. For appearance reconstruction, with $a = \sum_k w_k$ and $\mathbf{f}_{\mathrm{hit}} = \sum_k w_k \mathbf{f}_k$,

$$\overline{\mathbf{f}}_{\mathrm{hit}} = \begin{cases} \mathbf{f}_{\mathrm{hit}} / \max(a, 0.15), & a > 0.4, \\ \mathbf{f}_{\mathrm{hit}}, & \text{otherwise}, \end{cases} \qquad \widehat{\mathbf{f}}_{\mathrm{rend}} = \overline{\mathbf{f}}_{\mathrm{hit}} + r\mathbf{f}_{\mathrm{sky}},$$

where $\mathbf{f}_{\mathrm{sky}}$ is a learnable background feature. This normalization operates on decoder input features rather than physical radiance.

B.2 Adaptive Reconstruction: Parse–Share–Specialize

The projector forms five groups: semantic/air probabilities, expected depth, high-level satellite semantics, low-level texture, and satellite-derived pseudo-HD-map features. Optional satellite RGB adds linear RGB and a validity mask. Semantic, high-level, low-level, and satellite-derived pseudo-HD-map groups use separate channel attention; depth and optional RGB pass directly to cross-group fusion.

Parse. For each attended group g, channel weights are computed by global pooling and a 1×1 projection,

$$\mathbf{X}^{\mathrm{p}}_g = \mathbf{X}_g \odot \sigma\big(\mathrm{Conv}^{g}_{1\times 1}(\mathrm{GAP}(\mathbf{X}_g))\big). \tag{14}$$

This preserves modality-specific statistics before interaction.

Share. The parsed and pass-through groups are concatenated, reweighted jointly by cross-group channel attention, and processed by a multi-scale spatial module with dilations $\{1, 6, 12, 18\}$:

$$\mathbf{X}^{\mathrm{s}} = \mathcal{A}^{\{1,6,12,18\}}_{\mathrm{sp}}\big(\mathcal{A}_{\mathrm{cross}}([\mathbf{X}^{\mathrm{p}}_g]_{g\in\mathcal{G}})\big). \tag{15}$$

The multi-scale receptive field connects sparse projected evidence across the perspective view.

Specialize. After splitting $\mathbf{X}^{\mathrm{s}}$ back into group slices, each task $q \in \{\mathrm{RGB}, \mathrm{depth}, \mathrm{semantic}\}$ uses learned task–group logits $v^{(q)}_g$:

$$\bar{v}^{(q)}_g = \frac{\exp(v^{(q)}_g)}{\sum_{j\in\mathcal{G}} \exp(v^{(q)}_j)}, \qquad \widetilde{\mathbf{X}}^{(q)} = [\bar{v}^{(q)}_g \mathbf{X}^{\mathrm{s}}_g]_{g\in\mathcal{G}}, \tag{16}$$

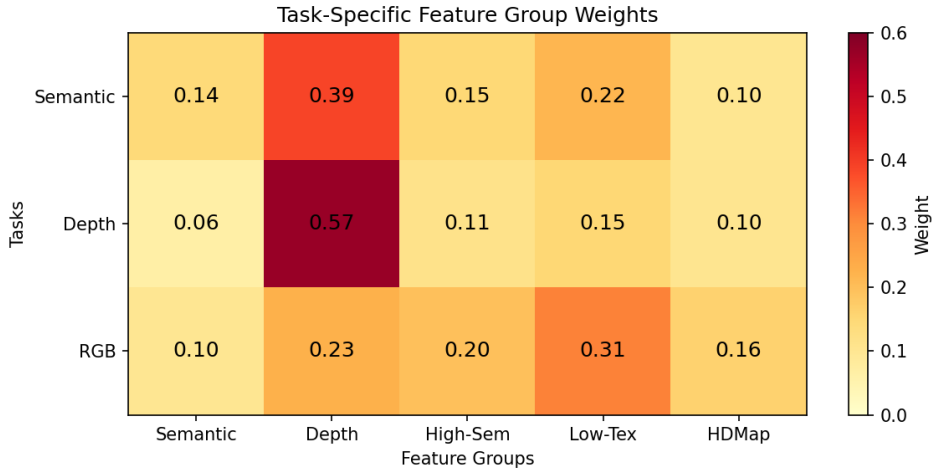


Figure 5: *
Task-specific feature-group weights after Stage-I training. Rows are output tasks and columns are feature groups; each row is softmax-normalized.

$$\widehat{\mathbf{Y}}^{(q)} = D_q(\widetilde{\mathbf{X}}^{(q)}). \tag{17}$$

Each task scales and concatenates the group slices before its own decoder, allowing RGB, depth, and semantics to learn different mixtures. The satellite-derived pseudo-HD-map condition remains the direct projection of $\mathbf{H}$.

Learned task specialization. The following heatmap reports the normalized task–group weights after Stage-I training.

The learned mixtures are task-specific. Depth relies most on rendered depth (0.57); RGB favors low-level texture (0.31) while retaining depth (0.23) and high-level semantics (0.20); semantics relies most on depth (0.39), consistent with surface localization complementing categorical evidence. Nonzero weights across multiple groups indicate that each task draws on more than one cue, in line with the adaptive-reconstruction ablation in the main paper.

Geometry-preserving output heads. Depth is refined around the rendered estimate rather than predicted from scratch. With bilinear upsampling $\uparrow$ and refinement head R_d, the outputs are

$$\begin{aligned}
\widehat{\mathbf{S}}^c &= D_{\text{sem}}(\widetilde{\mathbf{X}}^{(\text{semantic})}),\\
\widehat{\mathbf{d}}^c &= \text{clip}\Big(\uparrow \mathbf{d}^c_{\text{rend}} + a_d \tanh R_d(\widetilde{\mathbf{X}}^{(\text{depth})}), 1, 64\Big)\,,\\
\widehat{\mathbf{I}}^c &= \sigma\Big(D_{\text{RGB}}(\widetilde{\mathbf{X}}^{(\text{RGB})})\Big)\,,
\end{aligned}$$

where a_d is initialized to 20 m and $\tanh$ bounds the residual to $[-|a_d|, |a_d|]$. RGB is constrained to the valid image range, and depth remains tied to occupancy-rendered geometry.

B.3 Split-to-Unified Adapter and Control–State Fusion

Multimodal adapter implementation. Stage II orders the inputs as semantics (3 channels), the satellite-derived pseudo-HD-map condition (3), depth (1), and RGB (3). After $8\times$ pixel unshuffle, one modality-specific AdapterBlock maps RGB to 384 channels and each auxiliary modality to 192. A 1×1 projection then maps their concatenation to a 1536-channel trunk with six shared AdapterBlocks (`split_layers` $= 0$), each aligned with one DiT injection layer.

For modality $m \in \{\mathrm{RGB}, \mathrm{sem}, \mathrm{road}, \mathrm{depth}\}$, the adapter is

$$\begin{aligned}
\mathbf{e}_m &= A_m(\mathrm{PU}_8(\mathbf{X}_m)), \\
\mathbf{z}_0 &= \mathrm{SiLU}(\mathrm{GN}(W_{1\times1}[\mathbf{e}_{\mathrm{RGB}}; \mathbf{e}_{\mathrm{sem}}; \mathbf{e}_{\mathrm{road}}; \mathbf{e}_{\mathrm{depth}}])), \\
\mathbf{z}_\ell &= A_\ell^{\mathrm{shared}}(\mathbf{z}_{\ell-1}), \qquad \mathbf{s}_\ell = \mathrm{reshape}_{B,T,N_{\mathrm{cam}}}(\mathbf{z}_\ell), \quad \ell = 1, \ldots, 6.
\end{aligned}$$

Here A_m are the modality stems and A_ℓ^{shared} are the six shared blocks. The stem-merging projection $W_{1\times1}$ is initialized from $\mathcal{N}(0, 0.01^2)$. The six satellite residuals follow the pretrained UniMLVG/OpenDWM condition-adapter schedule and are consumed by the first six DiT transformer blocks (0-based blocks 0–5). These residual-injection locations are distinct from the cross-view/temporal module locations used inside the DiT.

At DiT layer ℓ, let $\mathbf{h}_\ell$ be the current hidden state, $\mathbf{r}_\ell$ the residual from the frozen original control adapter, and $\mathbf{s}_\ell$ the satellite residual. Control–State Fusion uses two selectors:

$$\begin{aligned}
\mathbf{s}_\ell^{\mathrm{ctrl}} &= [1 - \sigma(\phi_{\mathrm{ctrl}}([\mathbf{r}_\ell; \mathbf{s}_\ell]))] \odot \mathbf{s}_\ell, \\
\mathbf{s}_\ell^{\mathrm{state}} &= \sigma(\phi_{\mathrm{state}}([\mathbf{h}_\ell; \mathbf{s}_\ell])) \odot \mathbf{s}_\ell, \\
\mathbf{f}_\ell &= \psi([\mathbf{s}_\ell^{\mathrm{ctrl}}; \mathbf{s}_\ell^{\mathrm{state}}]).
\end{aligned} \tag{18}$$

The control-aware branch suppresses satellite components already explained by the original control path, whereas the state-aware branch retains satellite evidence useful to the current DiT hidden state. Their fused residual is then integrated with separate gates for the original and satellite paths:

$$\begin{aligned}
[\mathbf{g}_\ell^r; \mathbf{g}_\ell^f] &= \phi_g([\mathbf{h}_\ell; \mathbf{r}_\ell; \mathbf{f}_\ell]), \\
\mathbf{h}'_\ell &= \mathbf{h}_\ell + (1 + \tanh \mathbf{g}_\ell^r) \odot \mathbf{r}_\ell + \sigma(\mathbf{g}_\ell^f) \odot \mathbf{f}_\ell.
\end{aligned} \tag{19}$$

Selector weights are zero-initialized with bias -4, the cross-fusion projection ψ is zero-initialized, and the satellite gate uses bias -6. The initial update is therefore close to $\mathbf{h}_\ell + \mathbf{r}_\ell$, preserving the pretrained control path while the satellite branch learns.

B.4 Pose-Aligned Temporal State Consistency

Temporal consistency is applied to Stage-I satellite features. Adjacent 3D features are collapsed along height,

$$\overline{\mathbf{V}}_t(x, y) = \max_z \mathbf{V}_t(x, y, z). \tag{20}$$

The previous feature is warped into the current ego frame using the relative pose. The valid $[-40, 40]^2$ m overlap is eroded with a 5×5 kernel to remove boundary interpolation. Let M_t be the mask, $N_t = \sum M_t$, and let D_{align} denote the channel dimension of $\overline{\mathbf{V}}_t$. Then

$$E_t = \mathrm{stopgrad}\left(\frac{\sum M_t \|\overline{\mathbf{V}}_t\|_2^2}{D_{\mathrm{align}} N_t + \epsilon}\right). \tag{21}$$

We use energy-normalized MSE with an empirical energy-adaptive margin,

$$\mathcal{L}_{\mathrm{feat}} = \left[\frac{\sum M_t \|\overline{\mathbf{V}}_t - \mathcal{W}_{t-1\to t}(\overline{\mathbf{V}}_{t-1})\|_2^2}{D_{\mathrm{align}} N_t (E_t + \epsilon)} - 0.01 E_t\right]_+. \tag{22}$$

Here $\overline{\mathbf{V}}_t$ denotes the height-collapsed dimensionless latent feature and E_t is a stop-gradient feature-energy statistic; $0.01E_t$ is an empirical energy-adaptive margin rather than a physical error threshold. Variance and spatial-diversity terms prevent feature collapse:

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{feat}} + 0.5[0.2 - \mathrm{Var}_{M_t}(\overline{\mathbf{V}}_t)]_+ + 0.5[0.1 - \mathrm{Std}_{xy, M_t}(\overline{\mathbf{V}}_t)]_+. \tag{23}$$

The previous branch is stop-gradient, providing a stable pose-aligned target.

B.5 Background Purification and Training

Dynamic objects, cast shadows, and night scenes are excluded or downweighted from static-background supervision so that transient foreground content does not contaminate the persistent geographic state. This does not discard the static scene hidden behind a dynamic occluder: that

background is supplied from the overhead observation rather than copied from the occluded street-level pixels. The valid support combines the source-validity, static-region, daylight, brightness, and shadow masks with explicit logical conjunction,

$$M = M_{\text{valid}} \odot M_{\text{static}} \odot M_{\text{day}} \odot M_{\text{bright}} \odot (1 - M_{\text{shadow}}).$$

Day/night filtering first uses image-level brightness statistics and then checks the timestamp/location-derived solar elevation, reducing sensitivity to a single appearance threshold. Cast-shadow purification begins from a global brightness criterion and applies region-adaptive road thresholds together with dynamic-object masking; these masks are used only for supervision construction and are not inference conditions.

RGB supervision applies semantic-class-aware intensity normalization. For class probability $p_c(\mathbf{u})$ and masked means $\mu_c^{\text{pred}}, \mu_c^{\text{gt}}$,

$$a_c = \text{clip}\left(\frac{\mu_c^{\text{gt}}}{\text{stopgrad}(\mu_c^{\text{pred}}) + \epsilon}, 0.6, 1.6\right),$$
$$a(\mathbf{u}) = \sum_c p_c(\mathbf{u}) a_c, \qquad \widetilde{\mathbf{I}}(\mathbf{u}) = a(\mathbf{u})\widehat{\mathbf{I}}(\mathbf{u}).$$

The implemented reconstruction loss uses

$$\mathcal{L}_{\text{RGB}} = \mathcal{L}_{\text{chroma}} + 2\mathcal{L}_{\text{SSIM}} + 0.5\mathcal{L}_{\text{grad}} + 0.1\mathcal{L}_{\text{int}} + 0.5\mathcal{L}_{\text{class}},$$

where chroma is the masked color reconstruction term, SSIM measures local structural similarity on sufficiently valid neighborhoods, $\mathcal{L}_{\text{grad}}$ matches image gradients, $\mathcal{L}_{\text{int}}$ penalizes masked intensity mismatch, and $\mathcal{L}_{\text{class}}$ is the class-color prior induced by the semantic normalization above.

Equation (7) in the main paper groups the implementation losses by supervision target rather than implying one scalar per displayed group. For the reported Stage-I checkpoint, the camera occupancy loss is multiplied by 5.0; perspective depth and semantic projection losses are each multiplied by 6.0; the satellite-derived road-semantic loss is multiplied by 0.5; and pose-aligned temporal feature consistency uses weight 10.0. The RGB sub-loss weights are those shown above. Stage I is optimized with AdamW for 500 epochs from an initial learning rate of 10^{-3}. Stage II follows the UniMLVG initialization (Chen et al., 2025c; Yang et al., 2024; Caesar et al., 2020; Sun et al., 2020; Wilson et al., 2021); the DiT and original control adapter are frozen, and only the satellite adapter and fusion modules are trained for four epochs at 8×10^{-5} on 16 RTX 4090 GPUs.

B.6 Notation Summary

Table 5 collects the symbols used across the main and supplementary derivations.

C Evaluation Protocol and Query Definitions

C.1 Query Regimes and Route Record

A geographic query is an ordered route sampled at model timestamps,

$$\tau = \{(\mathbf{p}_t, \psi_t)\}_{t=1}^{T},$$

where $\mathbf{p}_t$ is longitude–latitude and ψ_t is the navigation heading. This route record specifies ego progression only: it does not contain lane boundaries, road width, neighboring branches, intersection geometry, or surrounding road topology. Camera calibration defines the six views; crop scale and size define the satellite observation. One visual reference denotes one synchronized six-camera timestamp. No text condition is used in any reported evaluation. In the 3/1/0-reference regimes, the satellite-derived state remains available while the visual-reference evidence budget is reduced. Table 6 summarizes these inference regimes. In the strict point-addressed regime, target-location visual references, the inference-time external HD map, 3D boxes, and text conditions are absent, leaving the coordinate–route query, calibrated cameras, and satellite-derived world state for Loc2Drive. Queried routes are stored as an ordered GeoJSON LineString and a frame-aligned CSV described in Section D.3.

Table 5: Symbols shared by the main paper and supplement.

Symbol	Meaning	Symbol	Meaning
$\tau = \{(\mathbf{p}_t, \psi_t)\}_{t=1}^{T}$	route coordinates and headings	$Q, G, \mathcal{P}$	query, video-generation, and Stage-I projection/reconstruction operators
$I_{\mathrm{sat},t}$	heading-aligned satellite crop	$N_{\mathrm{cam}}, \Pi_t^c$	number of cameras and calibrated camera projection
$\mathcal{B}_t, \mathcal{M}, \mathcal{R}$	optional boxes, inference-time external HD map, and references	$\mathbf{O}, \mathbf{V}, \mathbf{H}$	occupancy, feature volume, and BEV semantics
$\mathbf{A}^c$	four pixel-aligned camera-condition maps	$C_{\mathrm{sem}}, C_{\mathrm{img}}, D_{\mathrm{align}}$	semantic classes, image channels, and temporal-alignment feature channels
N_{r}	samples per calibrated camera ray	$\mathbf{V}_0, m_h$	peak-normalized base volume and semantic modulator
$o_k^{(\mathrm{air})}, a_k, \widetilde{a}_k, T_k, w_k$	air probability, occupied probability, soft opacity, transmittance, ray weight	$\mathcal{G}, v_g^{(q)}, D_q$	rendered groups, task logits, and task decoder
$\mathbf{h}_\ell, \mathbf{r}_\ell, \mathbf{s}_\ell, \mathbf{f}_\ell$	hidden state, original-control residual, satellite residual, and fused satellite residual	$\overline{\mathbf{V}}_t, M_t, \mathcal{W}_{t-1\to t}$	height-collapsed feature volume, overlap mask, and pose warp
λ_*	conceptual Stage-I supervision-group weights		

Table 6: Inference regimes used in the main paper. One reference denotes one synchronized six-camera timestamp.

Regime	Satellite-derived state	Target-location visual references	Inference-time external HD map	3D boxes
Matched continuation	Yes	3	Optional	Yes
Reduced-reference continuation	Yes	1	Optional	Yes
Reference-free continuation	Yes	0	Optional	Yes
Strict point-addressed setting	Yes	0	No	No

The 3/1/0-reference evaluations use the same Loc2Drive checkpoint. UniMLVG's inherited training regime natively includes a reference-free video-generation task and partially visible reference conditions, so zero-reference evaluation does not introduce an unseen conditioning mode. At inference the zero-reference window contains no visual-reference latent. "No inference-time external HD map" refers only to the inference input: HD-map annotations still supervise the internal satellite-to-BEV semantic predictor during Stage I.

C.2 Matched Evaluation, Inherited Configuration, and Reference Budgets

Controlled UniMLVG/Loc2Drive comparisons match sampling seeds, denoising and guidance settings, resolution, validation clips, rollout overlap, metric checkpoints, and six-view packing. Stage I and Stage II use the original nuScenes scene-level split, and every derived satellite sample inherits the split of its source scene. The 3/1/0-reference settings use one fixed Loc2Drive checkpoint and differ in visual-reference availability; all quantitative rows in Table 2 retain the same 3D-box control. Reported Loc2Drive metrics are the mean of three stochastic evaluations from that same fixed checkpoint. For each repeat, paired methods use the corresponding matched sampling seed. For Loc2Drive, the strict point-addressed regime retains only the coordinate–route query, satellite-derived external world state, and calibrated cameras; no visual, map, box, or text condition is supplied. Table 7 records the inherited and Loc2Drive-specific configuration used in these controlled comparisons.

Table 7: Inherited and added configuration in the controlled UniMLVG–Loc2Drive comparison.

Item	Inherited/matched	Loc2Drive-specific addition
Backbone and initialization	UniMLVG pretrained video/control model	Satellite Stage-I state and Stage-II satellite adapter/fusion
Data split	Original nuScenes scene-level train/validation split	Satellite crops inherit source-scene split
Sampling/evaluation	Resolution, sampler/denoising, guidance, clip IDs, seeds, overlap, metric checkpoints, six-view packing	Satellite conditions aligned to the same sampled route poses
Reference budget	Same 3/1/0 six-view reference timestamps	Same fixed Loc2Drive checkpoint across budgets
Training	Baseline pretrained initialization and flow-matching objective	Stage I geometric/semantic supervision; four-epoch Stage-II fine-tuning

These reference-budget and horizon sweeps are complementary stress tests of the proposed mechanism. Reducing references decreases the amount of scene state directly constrained by visual evidence, while longer rollouts increase repeated reuse of generated hypotheses as future context. They are interpreted within each matched setting rather than as a causal decomposition of modeling error, missing evidence, occlusion, and autoregressive propagation.

C.3 Autoregressive Windows and FVD/FID Definitions

The flow-matching generator predicts 16 new frames per window. A standard window consumes three reference frames and generates the next 16 frames. Unless otherwise specified, FVD refers to this standard 16-frame window. During autoregressive rollout, the final three generated frames of the previous window become references for the next window; FVD_{32}, FVD_{64}, and FVD_{128} are computed on the continuous 32/64/128 generated frames obtained from 2/4/8 autoregressive windows rather than by averaging multiple standard-window FVD scores. The default FID evaluation aggregates 32 generated frames. This protocol makes the proposed chain explicit: limited visual support can create missing scene evidence; the resulting future scene is underdetermined; the generator instantiates one hypothesis; and the final generated frames then become conditioning context for later windows, allowing that hypothesis to propagate. Loc2Drive supplies satellite-grounded static evidence before this completion step when the hidden static scene is visible from overhead.

FVD follows the UniMLVG/OpenDWM evaluator with the same I3D checkpoint (`i3d_pretrained_400.pt`) for all horizons. Frames are resized to 224×224 before I3D feature extraction. Six-view packing treats each synchronized camera as a video sample rather than mosaicking views: with $N_{\text{cam}} = 6$ and RGB channel count $C_{\text{img}} = 3$, a tensor of shape $B \times T \times N_{\text{cam}} \times C_{\text{img}} \times H \times W$ is reshaped to $(BN_{\text{cam}}) \times T \times C_{\text{img}} \times H \times W$ before metric extraction. The camera order used throughout visualizations and data records is front-left, front, front-right, rear-right, rear, and rear-left.

The inherited nuScenes loader constructs target timestamps at 10 fps and matches each to the nearest sample in the 12-Hz interpolated stream. With synchronization checking enabled, the allowed mismatch is half a target-frame period, i.e., 50 ms at 10 fps. Changing the reference count also changes the target start index under this inherited evaluator. Consequently, the absolute FVD values for 3-reference and 1-reference settings are not interpreted as a monotonic context-ablation curve; comparisons are matched and interpreted within each reference budget. This explains why the measured FVD_{32} can be numerically lower in the 1-reference column even though the standard FVD degrades when references are removed.

C.4 Why the Quantitative Horizon Stops at 128 Frames

nuScenes scenes last approximately 20 seconds (Caesar et al., 2020), or about 200 synchronized timestamps at 10 fps. Not every validation start can support a 128-frame generated rollout after accounting for the initial references and scene boundary. For each reported horizon we therefore use the largest nuScenes-validation subset that can complete that continuous rollout; FVD_{128} is evaluated on the maximum subset eligible for eight 16-frame autoregressive windows. A 256-frame rollout would exceed nearly all scenes and would require either discontinuous scene concatenation or a much more selective subset. We therefore stop quantitative evaluation at 128 frames and use the approximately 19-second qualitative rollouts in Section D.1 for later behavior. This is primarily a dataset-length constraint rather than an architectural limit.

C.5 Post-Turn Qualitative Case and Horizon Definition

Table 2 reports horizon-wise validation evaluation on the horizon-specific eligible subsets defined above. Figures 3, 6, and 7 separately visualize matched post-turn cases selected to illustrate the structural failure discussed in the main text. Within each qualitative comparison, references, route, dynamic controls, and sampling seed are matched; the main-paper Figure 3 supplies UniMLVG with its standard inference-time external HD map.

Cross-horizon comparability. The eligible clip set shrinks with rollout length because not every nuScenes validation start has enough remaining frames. All methods share exactly the same clips within a given horizon, so each reported column is a controlled method comparison. Across horizons, both temporal length and subset composition change; we therefore interpret Table 2(a) as separate within-horizon comparisons rather than a causal isolation of horizon alone.

C.6 Keyframe-Level Camera Visibility on Occ3D-nuScenes

Table 8 reports the resulting descriptive visibility statistic. We use Occ3D annotations corresponding to the nuScenes validation split at their native annotated keyframes. Occ3D provides `mask_camera`, `mask_lidar`, and semantic occupancy; class 17 denotes free space. We use these annotations only as a descriptive measure of per-timestamp sensing incompleteness. The statistic aggregates field-of-view limitations and occlusion in the street-level cameras; it does not indicate whether the hidden content itself is static or dynamic, nor whether that static content is visible from overhead. It is therefore not a future-disocclusion or generation-error metric.

Let $\mathcal{C}_i = \{v : \texttt{mask_lidar}_i(v) = 1,\ y_i(v) \neq 17,\ \texttt{mask_camera}_i(v) = 1\}$ be camera-visible LiDAR-valid occupied voxels at keyframe i. Dataset-level occupied-scene visibility is

$$R_{\text{occ}} = \frac{\sum_i |\mathcal{C}_i|}{\sum_i |\{v : \texttt{mask_lidar}_i(v) = 1,\ y_i(v) \neq 17\}|} = 65.22\%. \tag{24}$$

Thus 34.78% of LiDAR-valid occupied support is not directly visible to the current six-camera rig at the same annotated timestamp. For scale only, visible LiDAR-valid occupied voxels occupy

$$R_{\text{grid}} = \frac{\sum_i |\mathcal{C}_i|}{N \cdot 200 \cdot 200 \cdot 16} = 3.069\% \tag{25}$$

of the complete fixed annotation grid. R_{grid} is not a camera-frustum coverage statistic: its denominator includes free and unoccupied grid cells, and we use it only as an appendix scale reference.

Table 8: Descriptive per-timestamp camera observability on Occ3D-nuScenes validation keyframes.

Statistic	Validation-set ratio
Six-camera visible / LiDAR-valid occupied support	65.22%
Visible LiDAR-valid occupied support / full annotation grid	3.069%

D Additional Qualitative Evidence

D.1 Additional Long-Horizon Rollouts

Figure 6 separates the rollout into three phases.

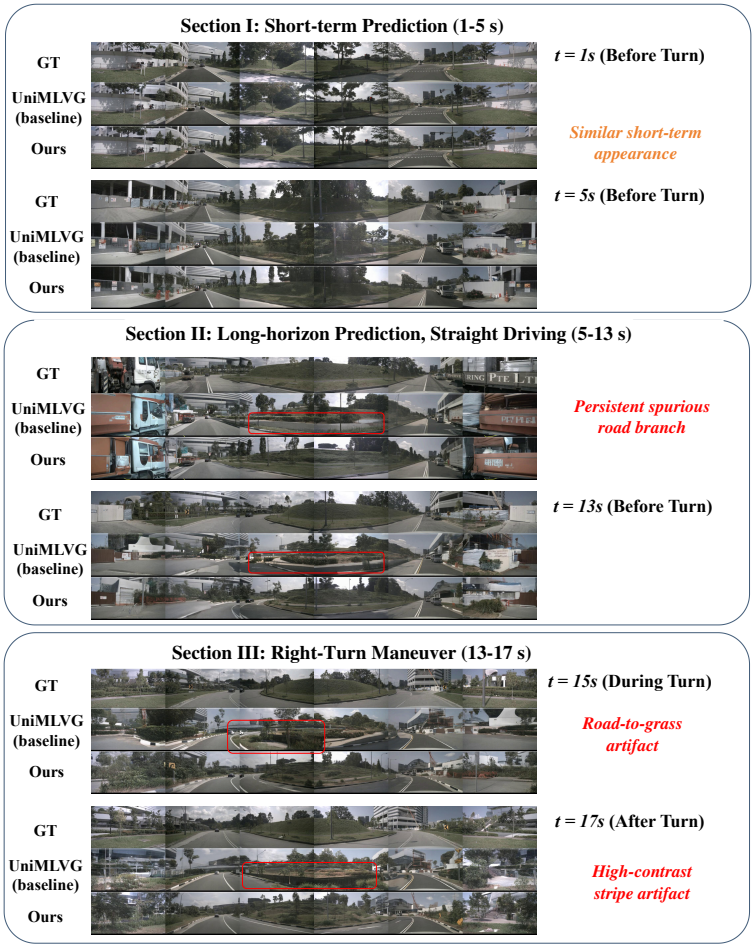


Figure 6: Additional long-horizon comparison under a shifted starting reference. The three groups show short-term prediction, straight driving, and the turn. UniMLVG evolves from similar short-term appearance to a false road branch and later road-to-grass/high-contrast artifacts; Loc2Drive preserves the route-aligned layout.

Short-term prediction (1–5 s). Both methods remain visually close to ground truth during the early rollout (1–5 s), despite Loc2Drive's metric advantage.

Long straight driving (5–13 s). Around 10–13 s, a false road branch appears in the baseline and persists across subsequent frames. Loc2Drive preserves the intended continuation while updating route-aligned geographic evidence.

Turning maneuver (13–17 s). After the turn, the baseline drift becomes more severe: road becomes grass near 15 s and high-contrast artifacts appear by 17 s. Loc2Drive keeps a coherent road layout. This rollout uses a different starting reference from the same main-paper scenario; the specific baseline failure changes, whereas Loc2Drive preserves the route structure.

D.2 Enlarged Post-Turn Comparison

Figure 7 enlarges the matched main-paper example. Under identical references, controls, and seed, the baseline carries a false branch through the turn, whereas Loc2Drive preserves the queried continuation.

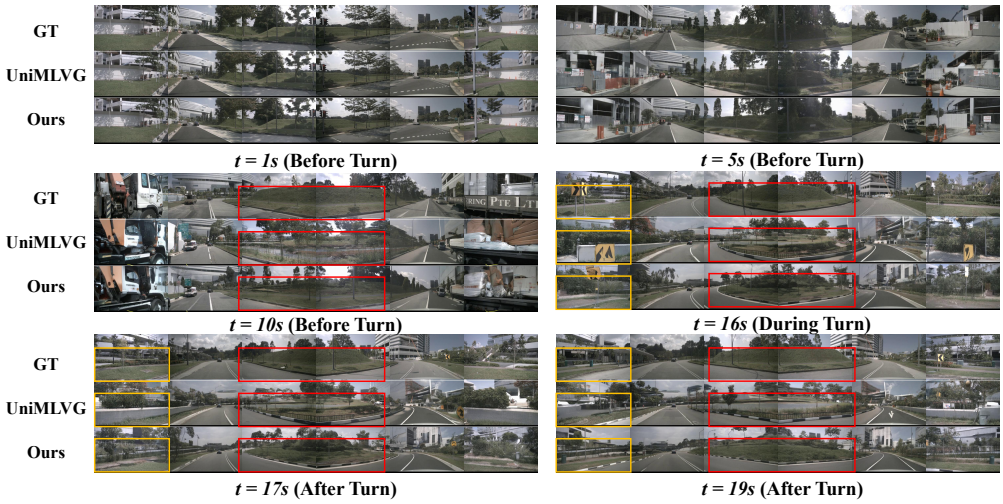


Figure 7: Enlarged matched post-turn rollout used for the qualitative case study. Identical references, route, dynamic controls, and seed show the effect of route-aligned satellite-derived geographic conditioning.

D.3 Exploratory Cross-City Geographic Queryability and Route Metadata

Appendix Figure 9 provides the complete six-city qualitative overview covering London, New York, Shanghai, Dubai, Paris, and Tokyo under the strict coordinate–route setting, together with route metadata and the geographic split. Loc2Drive's geographic pairing/fine-tuning uses Boston and Singapore, while both methods retain the same pretrained video initialization. The experiment probes point-addressed generation: a coordinate–route query retrieves external geographic evidence that binds the shared visual prior to the requested physical site. The strict setting uses no target-location visual references, inference-time external HD map, 3D boxes, or text condition. OpenDV-YouTube pretraining is geographically broad, so the experiment evaluates binding to the requested site rather than city-level absence from generic pretraining. Each route uses an ordered drivable-route GeoJSON LineString and a frame-aligned CSV,

$$\{\mathrm{frame_idx}, \mathrm{X}, \mathrm{Y}, \mathrm{angle}, \mathrm{arc_length_m}\},$$

where X/Y are longitude/latitude, angle is the north-referenced heading (x: East, y: North), and $\mathrm{arc_length_m}$ is distance from the route origin.

Point-addressed query set and geographic coverage. Beyond Boston and Singapore, which provide Loc2Drive-specific geographic pairing, we qualitatively probe point-addressed queries in London, New York, Shanghai, Dubai, Paris, and Tokyo. They span Europe, East Asia, North America,

and the Middle East and include irregular historical roads, dense Asian networks, a North-American grid, and wide planned roads; London and Tokyo also add left-hand-traffic environments. Figure 8 shows the split used for Loc2Drive-specific geographic training, and Table 9 summarizes the urban diversity. Because synchronized target-scene ground truth is unavailable for these cities, the experiment evaluates whether generated road layout and surrounding structure qualitatively follow the queried route and satellite-visible evidence rather than reporting pixel-level geographic fidelity.

Table 9: Additional cities used for point-addressed qualitative queries; Loc2Drive-specific geographic pairing uses Boston and Singapore. Coordinates are representative city-center locations in longitude–latitude order; the queried routes lie within the corresponding urban regions.

City	Region	Representative coordinate	Urban and road characteristics
London	Europe	$(-0.1276,\ 51.5074)$	Historical core, irregular road network, left-hand traffic
Paris	Europe	$(2.3522,\ 48.8566)$	Radial streets, complex intersections, historical urban fabric
Shanghai	East Asia	$(121.4737,\ 31.2304)$	Dense CBD, large urban roads, elevated infrastructure
Tokyo	East Asia	$(139.6917,\ 35.6895)$	Dense local network, narrow roads, left-hand traffic
New York	North America	$(-74.0060,\ 40.7128)$	Regular grid network and high-rise street canyons
Dubai	Middle East	$(55.2708,\ 25.2048)$	Wide multilane roads and modern planned districts

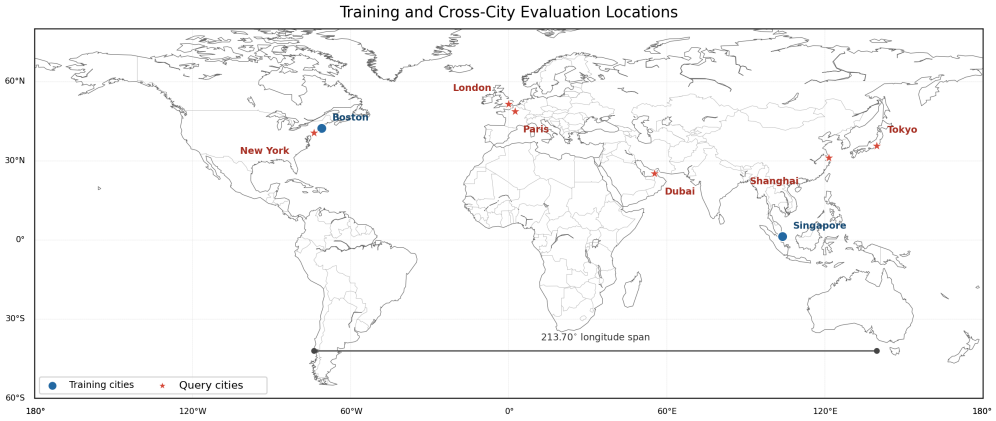


Figure 8: Geographic distribution for point-addressed evaluation. Singapore and Boston provide Loc2Drive-specific geographic pairing; the six additional query cities span Europe, East Asia, North America, and the Middle East.

Unified satellite and route construction. All cities use the same four-stage pipeline: QGIS annotation, metric route sampling, frame-aligned interpolation, and per-frame satellite cropping.

QGIS extent and route annotation. For each city, QGIS exports a route-covering georeferenced GeoTIFF and an ordered drivable-route LineString in GeoJSON. All later steps use one Python pipeline.

Metric sparse trajectory. The LineString is transformed to local metric coordinates, sampled every ~ 5 m, and converted back to WGS84. The heading between consecutive metric points is

$$\theta_i = \mathrm{mod}\left[\frac{180}{\pi}\,\mathrm{atan2}\big(x^m_{i+1} - x^m_i,\ y^m_{i+1} - y^m_i\big)\,, 360\right].$$

The navigation convention is north 0° and clockwise-positive.



Figure 9: Complete six-city point-addressed qualitative overview for London, New York, Shanghai, Dubai, Paris, and Tokyo. The strict setting uses no target-location visual references, inference-time external HD maps, 3D boxes, or text condition. Loc2Drive produces plausible road scenes compatible with the queried route, visible overhead structure, and empty dynamic-layout condition, providing qualitative evidence that the location-indexed state interface binds generation to the queried geography.

Frame-aligned interpolation. For speed v in km/h and interpolation rate f in Hz, the target arc-length interval is

$$\Delta s = \frac{v/3.6}{f}.$$

We use $v = 30$ km/h and $f = 12$ Hz, giving $\Delta s \approx 0.694$ m. Positions use cubic-spline interpolation with a linear fallback for fewer than four points. Headings are interpolated through sine/cosine components, recovered by $\mathrm{atan2}$, and smoothed over five frames; the loader then selects the matched 10-fps timestamps.

Per-frame satellite conditions. At each selected timestamp, longitude, latitude, and heading define a 400×400 crop at 0.2 m/pixel (80×80 m). The same satellite branch is used without city-specific tuning; strict evaluation supplies no target-location visual references, inference-time external HD map, 3D boxes, or text condition.

The resulting data flow is

georeferenced satellite extent and ordered route
→ GeoTIFF and GeoJSON LineString
→ 5-m metric route samples
→ 12-Hz interpolation and matched 10-fps timestamps
→ per-frame satellite crops
→ Loc2Drive geographic conditions.

Cross-city qualitative comparison. Appendix Figure 9 provides the complete six-city overview. Across these strict queries without target-location visual references, inference-time external HD maps, 3D boxes, or text conditions, UniMLVG often produces plausible but generic or route-inconsistent roads and can retain transient traffic residuals despite the empty dynamic-layout condition. Loc2Drive uses the satellite query to produce scenes more compatible with the requested route and visible overhead structure while more cleanly respecting the absence of specified agents in the shown examples. This is consistent with the factorized conditioning design: persistent background is supplied by a purified overhead state, whereas specified dynamic content enters only through the box pathway. Without synchronized ground truth or a dedicated object-compliance metric, we interpret these observations as qualitative evidence of geographic binding and dynamic-condition compliance rather than a standalone quantitative controllability result.

D.4 Geographic Background and Separate Dynamic Control

Stage I separates persistent geographic background from optional dynamic box control. This separation is especially important under occlusion: a vehicle may hide a static road or building in the street-level reference, yet the satellite branch can still represent that background from overhead without inheriting the foreground vehicle. It receives no boxes; dynamic annotations and cast-shadow masks prevent transient foreground from entering background supervision. Consequently, static background consistency does not depend on which transient agents happened to occlude the reference image. Conversely, dynamic agents—including agents not visible in the current references—are specified through the box pathway when provided. This makes the box pathway the explicit source of specified dynamic layout and reduces a route by which transient agent content can leak into the static condition. Figure 10 visualizes this factorization, so box edits change dynamic control without removing the route-indexed static background.



Figure 10: Geographic-background construction. Dynamic objects and cast shadows are masked from satellite-to-background supervision; boxes enter through a separate optional path.

E Ablation Controls and Interpretation

The controls below expand the main-paper ablations across geographic evidence, representation, encoding, and fusion. Table 10 adds temporal-consistency and satellite-vintage controls.

E.1 Completed Mechanistic Ablations

The controls mirror three questions in the main argument: what information the explicit state must retain, how the external state should enter the pretrained video prior, and why a world-aligned recon-

struction is preferable to direct 2D geographic fusion. Controls A–F vary one factor at a time under matched training, inference, and evaluation. Component ablations use the standard single-window protocol and report frame-level FID and FVD; Table 2 separately evaluates the long-horizon behavior of the complete system.

A. What information must the explicit state retain? The full model obtains 5.31/29.77 FID/FVD. Removing the feature volume causes the largest FID loss (5.75/32.41), while removing background purification causes the largest FVD loss (5.67/33.64), consistent with keeping transient foreground out of the persistent state. Removing the satellite-derived pseudo-HD-map condition (5.58/31.88) or adaptive reconstruction (5.61/31.92) also degrades FID/FVD.

B. How should multiple rendered modalities enter the video model? Without satellite injection, performance is 5.82/36.11. Expanding the pretrained adapter input to stack all four rendered conditions gives 5.44/30.47; a parallel satellite adapter improves to 5.33/31.20. Four fully separated modality adapters reach 5.37/31.43, whereas modality-specific stems followed by a shared trunk perform best at 5.31/29.77.

C. What is gained by explicit 3D reconstruction? No satellite input gives 5.82/36.11. The direct 2D satellite cross-attention control follows the SR-AD-style formulation (Jia et al., 2026): the noised video latent supplies the query Q, satellite/BEV geographic features with positional encoding supply the key K, and the corresponding geographic features supply the value V. It injects these features directly into the video-generation branch without constructing the explicit 3D state, reaching 5.74/35.13. The explicit geometry–feature-volume–road-semantic reconstruction-and-projection pipeline reaches 5.31/29.77 and does so without retrieving target-location street-view imagery. The result favors the complete explicit co-registered pipeline over direct 2D fusion under the matched Stage-II setting. The direct 2D control uses the same Stage-II learning rate, four-epoch schedule, pretrained video generator, and evaluation protocol as the explicit-state variant.

D. Do more rendered conditions automatically help? No. With only first-layer channel expansion, RGB is the strongest single condition (5.48/30.17), but stacking depth, semantics, and the satellite-derived pseudo-HD-map condition gives non-monotonic results (5.51/30.22, 5.48/30.58, and 5.44/30.47). The full adapter and fusion design achieves the best joint FID/FVD among these variants at 5.31/29.77.

E. Why use control-aware and state-aware selection? Direct addition, scalar, channel-wise, and spatial gating obtain 5.68/32.20, 5.67/32.80, 5.63/32.76, and 5.48/32.21. Control–State Fusion reaches 5.31/29.77; the control-aware selector suppresses satellite evidence already covered by the original control path, while the state-aware selector retains evidence useful to the current DiT state.

F. How should the selected satellite residual be injected? After satellite selection, independent addition gives 5.44/31.13, while global and channel-wise scaling give 5.41/30.84 and 5.34/30.71. Concatenating the hidden state, original residual, and satellite residual to predict separate path gates performs best at 5.31/29.77, while initialization preserves the pretrained hidden-plus-control path.

E.2 Additional Temporal and Cross-Vintage Ablations

Table 10 varies only temporal consistency or satellite vintage under matched settings.

Table 10: Additional ablations on temporal consistency and satellite-vintage supervision. The mixed-vintage setting randomly samples satellite–video pairs from both 2025 and 2026 imagery.

Panel	Variant	FID↓	FVD↓
G. Temporal consistency	Without temporal consistency	5.37	33.01
	With pose-aware temporal consistency	**5.31**	**29.77**
H. Satellite-vintage supervision	2025 satellite only	**5.27**	30.31
	2026 satellite only	5.33	30.19
	Randomly sampled from 2025 and 2026	5.31	**29.77**

G. Temporal consistency. Removing temporal consistency degrades FID from 5.31 to 5.37 and FVD from 29.77 to 33.01. The larger change in FVD is consistent with a stronger effect on temporal video quality; the objective itself operates by aligning adjacent geographic states, as detailed in Section B.4.

H. One-vintage versus mixed-vintage supervision. The 2025-only setting gives the best FID (5.27), while 2026-only yields 5.33/30.19. Random sampling from both years gives the best FVD (29.77) with 5.31 FID. Because both vintages are asynchronous with the 2019 nuScenes videos, mixed-year training is intended to reduce reliance on year-specific lighting, foliage, vehicles, and shadows; the FVD result is consistent with this motivation.

F Reproducibility and Scope

F.1 Shared-3D Consistency Checks

All cameras sample the same world-space state $\mathcal{Z}_t$; only Π_t^c changes across views. We verify that overlapping cameras return the same pre-projection voxel features, that feature/depth/semantic rendering shares transmittance weights, and that the satellite-derived pseudo-HD-map projection uses the same world-to-camera pose.

F.2 Data Provenance, Scope, and Limitations

Loc2Drive performs externally grounded generative synthesis. Visual history constrains observed scene state, while finite reference span, field of view, and occlusion can leave newly exposed static content underdetermined. The route specifies ego progression and the queried overhead observation supplies static geographic evidence for that missing content. Satellite imagery does not recover transient dynamic state; when available, 3D boxes specify dynamic agents and layout. Fine-grained street-level appearance, weather, illumination, and unboxed dynamics remain generative.

Zero-reference results measure behavior when observation-derived visual-reference state is removed, while the point-addressed study evaluates qualitative geographic binding rather than recovery of a unique target appearance. OpenDV-YouTube pretraining is geographically broad and its public city metadata are incomplete; the cross-city study therefore tests whether point-addressed satellite evidence binds generation to the queried site.

The underlying geographic imagery is the georeferenced Google Earth imagery used by the SA-Occ / Occ3D-nuScenes-SatExt construction (Chen et al., 2025a). SatExt provides imagery aligned to occupancy/keyframe samples; Loc2Drive queries the same underlying satellite maps more densely along interpolated routes for the 10-fps video pipeline. HD-map annotations supervise the internal satellite-to-BEV predictor during Stage I, but an inference-time external HD map is not required. The satellite-derived road-semantic layer follows the same local ground-plane map representation used by the UniMLVG/OpenDWM control interface; strongly graded or multi-level roads are therefore not fully represented by this 2D layer, although camera projection uses the full calibrated intrinsics and extrinsics.

Satellite availability, resolution, freshness, and registration quality vary geographically. The 2019 nuScenes videos are paired with newer satellite imagery, so road changes or misregistration can create conflicts. Overhead imagery supplies otherwise hidden static geographic structure but does not fully constrain fine-grained vertical or street-level appearance. Loc2Drive therefore externalizes overhead-observable static state, uses 3D boxes for dynamic agents when available, and leaves remaining appearance and dynamics to the generative prior. The empty-box examples additionally suggest cleaner dynamic-condition compliance as a qualitative secondary observation.