DYNAMIC AIRSPACE: A BENCHMARK FOR 4D SEMANTIC OCCUPANCY GENERATION FOR UAV

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are becoming important for UAV embodied intelligence, as they enable agents to anticipate future environmental states before taking actions. However, many existing UAV world models still formulate world prediction primarily as 2D video generation, which overlooks the inherently 3D nature of UAV motion. Unlike ground agents, UAVs move freely in 3D space, and their future prediction should therefore capture 3D dynamic airspace, particularly changes in altitude along the vertical axis. Thus, in this paper, we present UAVX, a platform collecting large-scale data of dynamic 3D airspace for UAV world modeling. Specifically, we formulate future airspace prediction as 4D semantic occupancy generation, where the 3D occupancy and semantic state of the environment are predicted over time. We simulate UAV motion with 6-DoF control in large-scale 3D environments. As UAVX moves along 6-DoF trajectories, our system continuously records UAV observations together with the corresponding 4D semantic occupancy evolution, forming a large-scale benchmark for learning and evaluating future airspace prediction. Experimentally, we benchmark six representative methods on UAVX. The results show that existing methods generally struggle with 4D semantic occupancy generation under complex 6-DoF UAV motion, with temporal consistency remaining a major challenge. Even the best-performing method, GenieDrive, achieves only 65.31% on the temporal consistency metric. Meanwhile, our benchmark shows that methods that explicitly model the temporal evolution of UAV motion tend to achieve better performance, suggesting that motion-aware temporal modeling is a promising direction for future UAV world models.

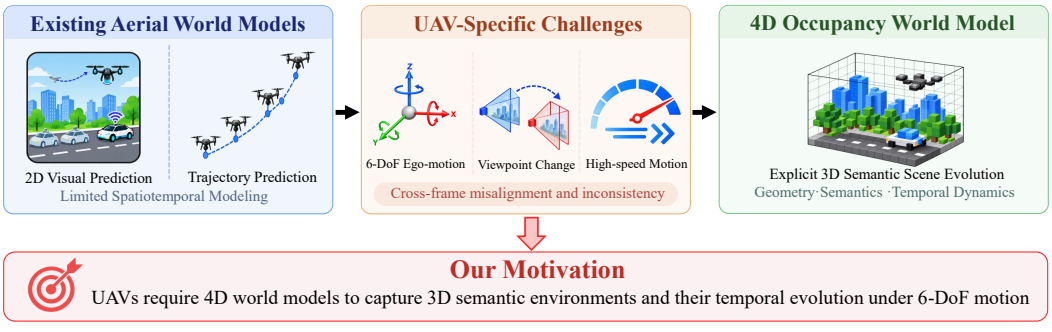


Figure 1: **Motivation for 4D UAV world modeling.** Existing prediction paradigms cannot fully capture 3D dynamic airspace under complex 6-DoF UAV motion. Rapid viewpoint changes make spatial alignment and temporal modeling difficult. These challenges motivate 4D semantic occupancy for explicit 3D structure and temporal evolution.

1 INTRODUCTION

World models capture the evolution of environmental states under agent actions (Ha & Schmidhuber, 2018b). They enable embodied agents to anticipate future states and support planning and decision-making. Recent advances in generative and temporal modeling have extended world models beyond low-dimensional state prediction. They can now predict visual observations (Bruce et al., 2024;

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.

Brooks et al., 2024) and model structured 3D scenes (Zheng et al., 2024; Bian et al., 2025). Such capabilities are particularly important for UAV embodied intelligence (Zhao et al., 2025; Zhang et al., 2025). Unlike ground agents, UAVs operate freely in 3D space and experience continuous changes in position, altitude, orientation, and viewpoint. Effective UAV world models therefore need to reason about free space, static structures, and dynamic actors in the surrounding airspace. Predicting how this 3D airspace evolves is therefore essential for flight planning, simulation, training, and safety analysis.

Scene representation is central to UAV world modeling. A suitable representation should capture not only scene appearance, but also 3D geometry, spatial occupancy, semantic structure, and temporal evolution. Visual prediction describes how future observations may appear, but it does not model the structure of the surrounding 3D space. Trajectory prediction focuses on the motion of the UAV itself and provides limited information about environmental evolution. In contrast, 4D semantic occupancy represents 3D geometry and semantic categories as temporally evolving voxel sequences (Ma et al., 2024; Bian et al., 2025). It describes what occupies the space, where each element is located, and how the scene changes over time (Zheng et al., 2024; Wang et al., 2025). This makes 4D semantic occupancy a natural representation for modeling future 3D dynamic airspace.

Learning coherent 4D semantic occupancy for UAVs remains challenging. UAVs can translate along all three spatial axes and rotate in roll, pitch, and yaw (Zhao et al., 2025). Such 6-DoF motion leads to rapid viewpoint changes and large spatial variation across observations (Zhang et al., 2025). Vertical motion further changes the observed 3D structure, while high-speed flight increases inter-frame variation. Meanwhile, onboard sensors often provide only local and sparse observations (Gross et al., 2026). Effective modeling therefore requires temporal alignment of observations while accounting for UAV ego-motion. Models also need to distinguish motion-induced changes from actual changes in the environment (Wang et al., 2025). As illustrated in Figure 1, these factors make spatial alignment and temporal consistency difficult. They motivate a dedicated benchmark for learning and evaluating 4D UAV world models.

UAVX provides a platform for 4D UAV world modeling in dynamic 3D airspace. It supports full 6-DoF UAV control in large-scale 3D urban environments. UAVX collects data along continuous flight trajectories with substantial changes in position, altitude, orientation, and viewpoint.Along these trajectories, it synchronously records sensor observations, UAV poses, dynamic actor states, and semantic occupancy. These records are used to construct temporally aligned 4D semantic occupancy sequences. The resulting sequences preserve spatial and temporal continuity along each flight trajectory. They provide a foundation for learning and evaluating future airspace prediction under complex aerial motion.

We systematically evaluate six representative methods on UAVX. All methods are evaluated using unified data splits, voxel representations, and evaluation protocols. The results show that accurate and temporally coherent 4D prediction remains difficult under complex 6-DoF motion. The main failure modes include spatial misalignment, incomplete geometry and semantics, and temporal error accumulation. Methods that explicitly model UAV motion and temporal evolution tend to achieve better performance. These findings highlight the importance of motion-aware temporal modeling for future UAV world models.

2 Related Work

World models learn predictive representations of environments and model their evolution under different actions (Ha & Schmidhuber, 2018a; Hafner et al., 2019a; 2023). Early methods focused on latent dynamics for model-based planning (Hafner et al., 2019a;b), while recent generative models synthesize future visual observations (OpenAI, 2024; Bruce et al., 2024). However, pixel-level prediction does not explicitly preserve 3D geometry, semantic structure, or physical consistency (Xu et al., 2024; Kong et al., 2025). Therefore, structured representations such as occupancy and scene flow have been adopted in embodied world models (Liu et al., 2019; Wei et al., 2023; Huang et al., 2023). Compared with image-space predictions, these representations provide explicit spatial states that can be aligned across time and queried for planning (Ma et al., 2024; Zheng et al., 2024). They also offer a common interface for jointly modeling scene geometry, semantic information, and object motion under action-conditioned prediction (Zheng et al., 2024; Kong et al., 2025). Recent studies have further extended world modeling from visual prediction to structured 3D and 4D scene

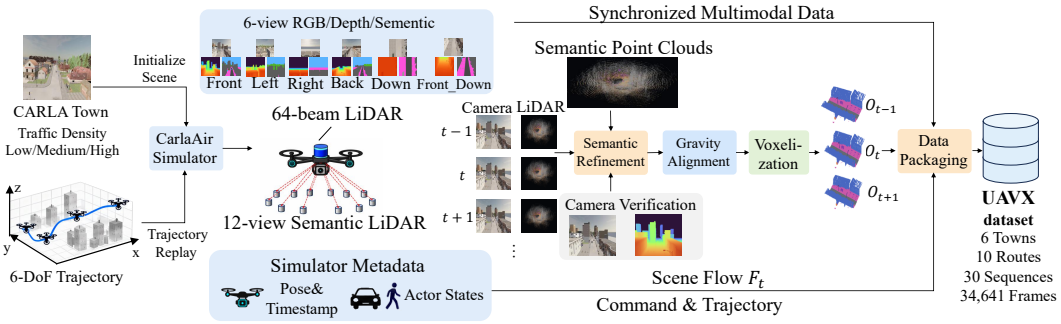


Figure 2: **UAVX platform.** Across different towns and traffic densities, the UAV collects multi-modal data along 6-DoF trajectories, which are then processed through semantic refinement, gravity alignment, and voxelization to construct 4D semantic occupancy sequences.

generation, establishing a connection between predictive modeling and temporally evolving spatial representations (Kong et al., 2025; Wang et al., 2024; Bian et al., 2025).

4D occupancy generation aims to synthesize 3D geometry, semantic labels, and temporal evolution as temporally aligned semantic occupancy sequences (Kong et al., 2025). Early occupancy methods mainly focused on single-frame 3D semantic reconstruction from multi-view images or depth observations, using voxel-based and tri-perspective representations to preserve geometric and semantic information (Song et al., 2017; Cao & de Charette, 2022; Huang et al., 2023; Wei et al., 2023). Recent approaches extend this setting to future occupancy prediction and long sequence generation. Cam4DOcc formulates future occupancy prediction as a camera-based forecasting problem, while OccWorld predicts future occupancy states directly in occupancy space (Ma et al., 2024; Zheng et al., 2024). OccSora further investigates long-sequence 4D semantic occupancy generation with generative modeling, and DynamicCity learns compact latent representations for synthesizing temporally evolving occupancy sequences (Wang et al., 2024; Bian et al., 2025). UniOcc provides a unified framework and evaluation protocol for occupancy prediction and forecasting across different settings (Wang et al., 2025). However, under complex 6-DoF UAV motion, existing methods generally struggle to generate accurate and temporally coherent 4D semantic occupancy sequences, highlighting the importance of explicitly modeling UAV motion and its temporal evolution.

3 The UAVX Platform

UAVX is a data collection platform for future airspace prediction under continuous 6-DoF UAV motion.We implement UAVX in CARLA-Air (Zeng et al., 2026) and replay UAV trajectories in large-scale 3D environments while synchronously collecting sensor observations, UAV poses, and scene states. As illustrated in Figure 2, we further construct temporally aligned 4D semantic occupancy sequences together with the corresponding UAV trajectories and flight commands, providing the data required for learning and evaluating future airspace prediction.

Problem definition. UAV world modeling requires a representation of 3D space and its temporal evolution. UAVs can move freely with 6-DoF motion, including translation along three spatial axes and rotation in roll, pitch, and yaw, exposing them to different spatial structures and viewpoints. Meanwhile, vehicles, pedestrians, and other dynamic actors continuously change the surrounding environment. Future airspace prediction must therefore capture both 3D spatial occupancy and its evolution over time. We formulate future airspace prediction as 4D semantic occupancy generation. A sequence of length T is represented as $\mathcal{O}_{1:T} = \{\mathcal{O}_t\}_{t=1}^{T}$, where $\mathcal{O}_t$ denotes semantic occupancy state at time t. This representation describes 3D geometry, semantic categories, and temporal evolution. In UAVX, each occupancy frame is associated with the corresponding UAV pose, allowing scene evolution to be modeled under continuous aerial motion.

3.1 6-DoF Simulation

CARLA-Air (Zeng et al., 2026) provides a unified simulation environment that combines large-scale CARLA urban scenes and dynamic traffic with multirotor UAV simulation. Building on this environment, UAVX replays UAV trajectories and synchronously records the data required for future

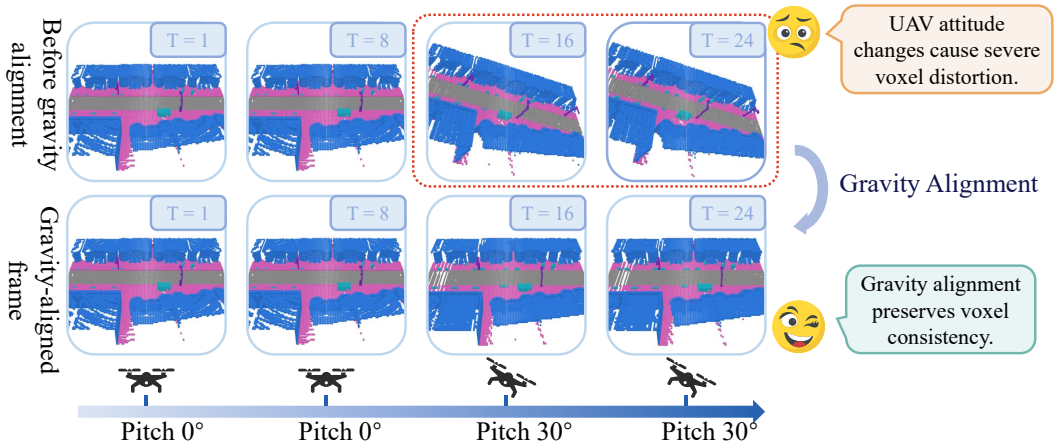


Figure 3: **Effect of gravity alignment.** UAV attitude changes, especially variations in pitch and roll, can cause severe distortion in UAV-centered voxel representations. Gravity alignment removes these attitude variations while preserving UAV position and heading, resulting in more consistent voxel structures across time.

airspace prediction. At time step t, the UAV state is represented as $\boldsymbol{\xi}_t = (x_t, y_t, z_t, \phi_t, \theta_t, \psi_t)$, where (x_t, y_t, z_t) denotes position, and ϕ_t, θ_t, and ψ_t denote roll, pitch, and yaw, respectively. The corresponding rigid-body pose is

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3), \tag{1}$$

where $\mathbf{t}_t = (x_t, y_t, z_t)^\top$ is the UAV translation and $\mathbf{R}_t \in SO(3)$ describes its orientation. Under the ZYX Euler-angle convention, $\mathbf{R}_t = \mathbf{R}_z(\psi_t)\mathbf{R}_y(\theta_t)\mathbf{R}_x(\phi_t)$, where $\mathbf{R}_x$, $\mathbf{R}_y$, and $\mathbf{R}_z$ are rotations around the corresponding axes. A trajectory of length T is represented by $\tau_{1:T} = \{\boldsymbol{\xi}_t\}_{t=1}^{T}$. Using the first frame as the reference, the relative UAV pose is

$$\Delta\mathbf{P}_{1\to t} = \mathbf{P}_1^{-1}\mathbf{P}_t. \tag{2}$$

Here, $\Delta\mathbf{P}_{1\to t}$ represents the UAV ego-motion from the first frame to time t, including both translation and rotation. The resulting relative pose sequence provides a consistent geometric description of UAV motion across the trajectory, which is used to relate observations across time.

UAVX includes multiple UAV trajectories across different towns. To further diversify the dynamic environment, each trajectory is replayed under low, medium, and high traffic densities, while the UAV trajectory itself is kept unchanged. During replay, UAVX synchronously records six-view RGB, depth, and semantic images, an onboard LiDAR scan, and the UAV pose at each time step. We also record the poses, velocities, and 3D bounding boxes of surrounding vehicles and pedestrians to describe their temporal evolution. In addition, twelve semantic LiDARs are used to provide dense observations for constructing the corresponding semantic occupancy. All data at each time step are aligned to the same CARLA frame.

3.2 4D Occupancy Construction

As shown in Figure 2, UAVX constructs 4D semantic occupancy sequences from synchronized observations and scene states. At each time step, point clouds from different viewpoints are transformed into a common reference frame and fused. Synchronized semantic images and depth maps are then used to refine unreliable semantic labels. A label is updated only when its projection and depth are consistent with the observations and no semantic conflict occurs across views. This process preserves the spatial locations of existing points.

Before voxelization, we apply gravity alignment to handle UAV attitude changes. Let $\mathbf{q}$ be a 3D point in the world coordinate system. Its coordinates in the gravity-aligned frame at time t are

$$\mathbf{q}_t^g = \mathbf{R}_z(\psi_t)^\top (\mathbf{q} - \mathbf{t}_t), \tag{3}$$

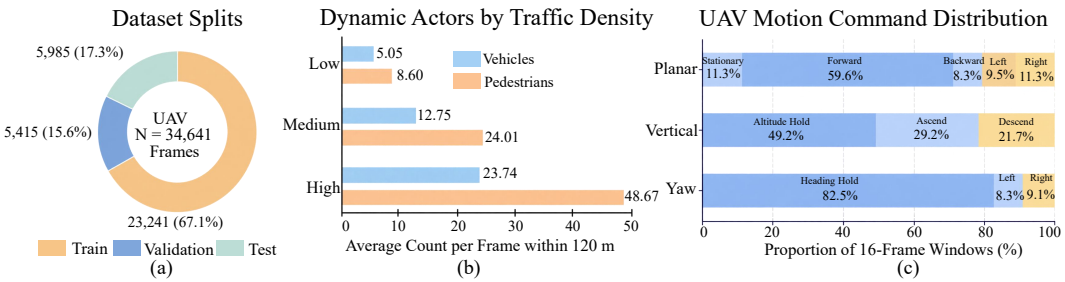


Figure 4: **Dataset Statistics of UAVX.** (a) Frame distribution across the town-disjoint training, validation, and test sets. (b) Frame-averaged numbers of vehicles and pedestrians within 120 m of the UAV across all 30 route–density sequences, grouped by traffic density. (c) Distributions of planar, vertical, and yaw commands over 11,397 unique 16-frame trajectory windows.

where $\mathbf{q}_t^g$ denotes the point in the gravity-aligned frame. This frame follows the UAV position and heading but does not inherit its roll and pitch. As illustrated in Figure 3, gravity alignment reduces voxel distortion caused by UAV attitude changes and preserves spatial consistency across time. The aligned points are then discretized into a UAV-centered semantic voxel grid, where LiDAR ray tracing is used to estimate free space and regions without observational support remain invalid.

Multi-frame static information further supplements sparsely observed structures, while synchronized actor states provide correspondences for moving objects and scene flow. The occupancy grids, validity masks, relative poses, and scene flow are then organized into temporally aligned 4D semantic occupancy sequences. Each sequence is associated with the corresponding UAV trajectory and flight commands. Based on this construction pipeline, UAVX contains 34,641 synchronized frames across different towns, flight routes, and traffic conditions. Figure 4 summarizes the dataset statistics, including dynamic actor distributions and UAV motion commands such as ascent and descent.

4 Evaluation Protocol

To comprehensively evaluate 4D semantic occupancy models for UAV world modeling, we organize the evaluation into four complementary dimensions: Reconstruction Metrics measure the preservation of spatial geometry and semantic categories; Temporal Consistency Metrics assess scene stability under full 6-DoF UAV ego-motion and legitimate object motion; Generation Distribution Quality Metrics evaluate the realism, similarity, coverage, and diversity of generated occupancy distributions; and Conditional Generation and Control Metrics measure trajectory following, command following, and physical validity. Together, these metrics characterize spatial reconstruction, temporal consistency, generation distribution quality, and controllability under a unified protocol with the same semantic taxonomy, voxel grid, clip definition, and data split.

4.1 Reconstruction Metrics

This subsection focuses on two complementary reconstruction metrics that expose failure modes not captured by voxel overlap alone. Surface $F@1v$ measures boundary displacement and thin-structure loss at the physical voxel resolution, whereas Change IoU measures whether the model preserves actual occupancy changes after compensating for UAV ego-motion.

Surface F@1v. For each frame, let P_t and G_t denote the predicted and ground-truth occupied/free interface point sets. With $d(p, G_t) = \min_{g \in G_t} \|p - g\|_2$ and $\tau = \max(v_x, v_y, v_z)$, the one-voxel surface score is the mean frame-level harmonic mean of the directed surface precision P_t^{1v} and recall R_t^{1v}, whose detailed definitions are given in Appendix B:

$$F@1v = \frac{1}{N_{\text{frames}}} \sum_t \frac{2P_t^{1v} R_t^{1v}}{P_t^{1v} + R_t^{1v}}. \tag{4}$$

Change-region IoU. Let W_t warp frame $t+1$ into frame t using the frozen relative SE(3) pose, and define the ground-truth and predicted change masks by $\Delta_t(Y) = \mathbf{1}[Y_t \neq W_t(Y_{t+1})]$. The

change-region score is

$$\mathrm{ChgIoU}_t = \mathrm{IoU}\big(\Delta_t(\hat{Y}), \Delta_t(Y)\big), \tag{5}$$

computed over the 15 adjacent pairs of a 16-frame clip and restricted to their common valid region. The detailed derivations, masking rules, and aggregation procedures are provided in Appendix B.

4.2 Temporal Consistency Metrics

This subsection focuses on three complementary temporal metrics. BG-Cons. measures the stability of static background classes after full 6-DoF ego-motion compensation. FGSC measures whether moving objects preserve their shape after translation and rotation are removed. SemFlicker measures semantic changes in regions that should remain stable. Following UniOcc (Wang et al., 2025), all three metrics use the UAVX taxonomy and full relative $\mathrm{SE}(3)$ alignment.

BG-Cons. Let Ω_t be the common valid region after warping, S the frozen set of static classes, and B_t^s the voxel mask for static class s. The class-aware background consistency is

$$\text{BG-Cons.} = \frac{1}{T-1}\sum_{t=0}^{T-2}\frac{1}{|S|}\sum_{s\in S}\mathrm{IoU}_{\Omega_t}\left(B_t^s, \mathrm{Warp}(B_{t+1}^s, T_{t\leftarrow t+1})\right). \tag{6}$$

FGSC. For dynamic objects, direct overlap would incorrectly penalize legitimate motion. Let I_t^k be the voxel mask of dynamic instance k and K_t the set of instances associated across frames t and $t+1$. After removing translation and rotation, the foreground shape consistency is

$$\mathrm{FGSC} = \frac{1}{T-1}\sum_{t=0}^{T-2}\frac{1}{|K_t|}\sum_{k\in K_t}\mathrm{IoU}\big(I_t^k, \mathrm{Align}(I_{t+1}^k)\big), \tag{7}$$

which measures shape preservation rather than positional overlap.

SemFlicker. Define the stable region as $\mathrm{Stable}_t = \Omega_t \cap \{Y_t = W_t(Y_{t+1})\} \cap \mathrm{StaticRegion}_t$. The semantic flicker score is

$$\mathrm{SemFlicker} = \frac{1}{T-1}\sum_{t=0}^{T-2}\frac{\sum_{v\in\mathrm{Stable}_t}\mathbf{1}[\hat{Y}_t(v) \neq \mathrm{Warp}(\hat{Y}_{t+1})(v)]}{|\mathrm{Stable}_t|}. \tag{8}$$

The detailed derivations of the motion transform, class macro-averaging, instance association, stable-region construction, and edge-case handling are provided in Appendix B.

4.3 Generation Distribution Quality Metrics

Generation distribution quality. We evaluate realism, distributional similarity, coverage, and diversity using fixed 2D and 3D representations of generated occupancy clips. For 2D, each clip is rendered with a fixed semantic colormap and camera: InceptionV3 provides 2048-dimensional features for FID and KID and town-class probabilities for Inception Score (IS), while VGG-16 provides 512D features for Precision and Recall. For 3D, a semantic MinkowskiUNet provides 256D stage4 features for FID, KID, Precision, and Recall, as well as voxel-class probabilities for IS. FID and KID measure similarity to the reference distribution (lower is better); Precision measures sample fidelity, Recall measures coverage, and IS reflects classifiability and class diversity (higher is better). Because covariance estimates for Precision and Recall are unstable in high dimensions with limited samples, we apply PCA only to those metrics, retaining 16 dimensions for 2D and 10 for 3D; FID and KID use the original features. The renderer, feature extractors, weights, preprocessing, sample sets, and clip aggregation are fixed, with metric definitions and computation details in Appendix B.8.

4.4 Conditional Generation and Control Metrics

Trajectory conditioning. We evaluate trajectory-conditioned generation using standard trajectory errors. Position ADE and FDE quantify average and final 3D displacement drift, respectively; velocity MAE measures errors in the three translational velocity components; and wrapped Yaw MAE

Table 1: **Summary of representative methods on the UAVX benchmark.** Results are grouped by reconstruction, generation, and conditional control, covering semantic fidelity, distributional quality, temporal consistency, and adherence to commands and trajectories.

Method	Venue	Recon.	Generation					Conditional control			
		mIoU (%) ↑	mIoU (%) ↑	FID_{3D} ↓	KID_{3D} ↓	BG-Cons. (%) ↑	FGSC (%) ↑	Mean axis (%) ↑	Joint (%) ↑	ADE (m) ↓	FDE (m) ↓
OccSora (Wang et al., 2024)	arXiv '24	8.59	0.58	31.625	0.679	28.30	46.08	46.39	21.58	2.618	5.216
OccWorld (Zheng et al., 2024)	ECCV '24	35.61	18.35	2.476	**0.019**	52.07	48.13	65.61	57.64	**1.527**	**3.095**
SemCity (Lee et al., 2024)	CVPR '24	66.55	0.56	12.142	0.165	0.50	15.32	30.53	1.72	3.681	6.396
DynamicCity (Bian et al., 2025)	ICLR '25	56.82	17.42	**1.587**	**0.019**	40.66	52.49	60.22	39.27	1.960	3.886
UniScene (Li et al., 2025)	CVPR '25	53.66	8.73	11.720	0.275	43.78	44.24	58.92	40.27	1.665	3.232
GenieDrive (Yang et al., 2026)	CVPR '26	**67.02**	**38.20**	2.499	0.034	**61.59**	**69.03**	**73.14**	**64.76**	1.834	3.807

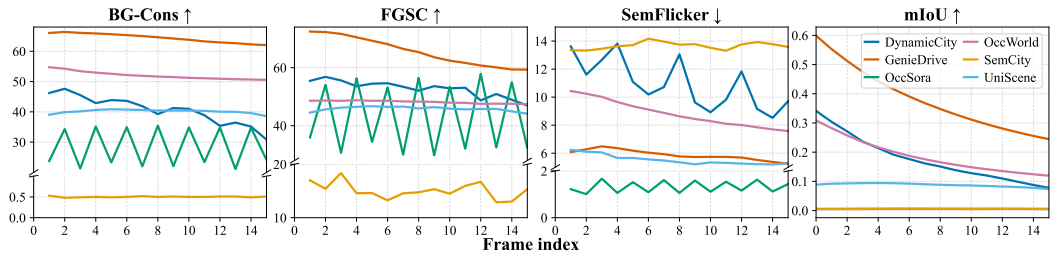


Figure 5: **Frame-wise temporal consistency and semantic fidelity on the UAVX test set.** From left to right, the four panels show BG-Cons., FGSC, SemFlicker, and semantic mIoU across the 16-frame horizon. BG-Cons. and FGSC measure static background and dynamic-object shape stability after motion compensation, whereas SemFlicker measures semantic label changes in stable regions.

measures the shortest angular discrepancy. Lower values indicate more accurate trajectory following. We retain the full 3D position and orientation because altitude and attitude changes are central to UAV motion, and report these established metrics without introducing a new definition.

Command conditioning. We evaluate command-conditioned generation by testing whether generated occupancy follows the requested planar, vertical, and yaw commands. A frozen three-head recognizer computes per-command accuracy, macro accuracy for each axis, and Joint exact match, where all three axis labels must be correct for the same clip. Per-command results are macro-averaged to prevent frequent commands from dominating; the raw axis and Joint accuracies are then chance-normalized using rates $1/5$, $1/3$, $1/3$, and $1/45$, respectively, and the mean normalized axis score is weighted equally with the normalized Joint score. The recognizer is trained only on ground-truth training clips, while collision, out-of-bounds, kinematic-violation, and instance-interpenetration rates are reported separately as physical-validity diagnostics. The exact normalization, aggregation, and edge-case handling are detailed in Appendix B.7.

5 Experiments

We organize the evaluation into three complementary tasks: occupancy reconstruction measures current-scene semantic and geometric fidelity together with cross-frame stability; future occupancy generation assesses paired semantic fidelity, temporal consistency, and distributional realism and coverage; and conditional control tests command recognition and trajectory adherence. These questions define the structure of the chapter. Table 1 first summarizes the compared methods together with the metrics used in each of the three tasks. Building on this overview, Section 5 answers the questions in the same order: how accurately the current scene is represented, how coherently its future is synthesized, and how reliably the model follows motion commands.

Reconstruction. Reconstruction asks whether each method can represent the current 3D semantic scene accurately and preserve it across frames. Accordingly, Table 2 evaluates semantic overlap (mIoU and bIoU), surface geometry ($F@1v$), change localization (ChgIoU), and temporal stability through background consistency (BG-Cons.), foreground shape consistency (FGSC), and semantic flicker. GenieDrive is strongest on the main semantic and geometric measures (67.02 mIoU, 73.10 bIoU, and 85.65 $F@1v$), while DynamicCity is competitive on background consistency. SemCity has the lowest flicker, but its static single-frame design cannot establish temporal modeling. To-

Table 2: **Reconstruction performance on the UAVX test set.** We compare semantic overlap, geometric fidelity, change localization, and temporal consistency in reconstructed occupancy.

Method	mIoU (%) ↑	bIoU (%) ↑	$F@1v$ ↑	ChgIoU ↑	BG-Cons. ↑	FGSC ↑	SemFlicker ↓
OccSora (Wang et al., 2024)	8.59	22.94	19.70	2.83	34.30	45.98	8.38
OccWorld (Zheng et al., 2024)	35.61	53.29	59.62	15.03	54.24	52.77	8.43
SemCity (Lee et al., 2024)	66.55	51.05	66.16	26.80	61.11	64.50	**4.20**
DynamicCity (Bian et al., 2025)	56.82	64.93	78.30	18.11	**61.65**	66.18	7.14
UniScene (Li et al., 2025)	53.66	62.63	75.64	21.94	54.47	61.94	7.96
GenieDrive (Yang et al., 2026)	**67.02**	**73.10**	**85.65**	**29.75**	56.28	**67.18**	6.63

Table 3: **Generation performance on the UAVX test set.** We compare semantic fidelity and temporal consistency under each method's conditioning setting. SemCity lacks temporal modeling and consequently performs poorly on this task.

Method	Condition	Semantic fidelity		Temporal consistency		
		mIoU (%) ↑	bIoU (%) ↑	BG-Cons. ↑	FGSC ↑	SemFlicker ↓
OccSora (Wang et al., 2024)	Trajectory	0.58	2.88	28.30	46.08	**1.32**
OccWorld (Zheng et al., 2024)	History	18.35	36.04	52.07	48.13	8.82
SemCity (Lee et al., 2024)	Single-frame	0.56	4.40	0.50	15.32	13.67
DynamicCity (Bian et al., 2025)	History + trajectory	17.42	29.47	40.66	52.49	10.96
UniScene (Li et al., 2025)	BEV layout	8.73	12.77	43.78	44.24	5.06
GenieDrive (Yang et al., 2026)	Current occupancy + control	**38.20**	**52.98**	**61.59**	**69.03**	5.88

gether, these results motivate our view of UAV reconstruction as both a 3D representation problem and a cross-frame stability problem.

Generation. Generation evaluates whether a model can synthesize a plausible future rather than only reconstruct the current scene. We therefore examine three complementary views: Table 3 pairs semantic fidelity (mIoU and bIoU) with temporal consistency (BG-Cons., FGSC, and SemFlicker) under each method's native conditioning interface; Figure 5 tracks these temporal measures and mIoU across the 16-frame horizon; and Table 4 measures distributional quality in 2D and 3D feature spaces using IS, FID, KID, precision, and recall. This organization connects per-clip fidelity with temporal evolution and distribution-level realism and coverage.

GenieDrive is the strongest paired generator, reaching 38.20 mIoU, 52.98 bIoU, 61.59 BG-Cons., and 69.03 FGSC, while OccWorld and DynamicCity reach 18.35 and 17.42 mIoU, respectively. The frame-wise curves in Figure 5 show horizon degradation, especially for DynamicCity (34.13 to 7.77 mIoU) and GenieDrive (59.88 to 24.45), whereas UniScene has lower fidelity but relatively stable sequences. Table 4 reveals a complementary distributional trade-off: DynamicCity has the best completed 3D FID and precision, GenieDrive has the strongest 3D recall, and UniScene remains weaker in feature-space distance despite usable paired fidelity. SemCity remains a static reference, so its index-wise scores provide chance-level diagnostics rather than evidence of 4D forecasting.

Conditional control. Generation quality is not sufficient for a controllable world model. We therefore evaluate command recognition across the planar, vertical, yaw, and joint axes, together with trajectory recovery from generated occupancy. For methods without explicit motion conditions, these measurements are diagnostic and do not directly measure adherence to an input command. Conditional-control evaluation reveals that controllability is not captured by a single aspect of generation: Figure 6 shows that GenieDrive attains the strongest command recognition, with 73.14 mean-axis and 64.76 joint accuracy, whereas Table 5 shows that OccWorld achieves the lowest ADE/FDE and velocity MAE and UniScene obtains the lowest yaw MAE. Together, these complementary results provide a more complete characterization of controllability by capturing both command understanding and motion preservation in generated occupancy.

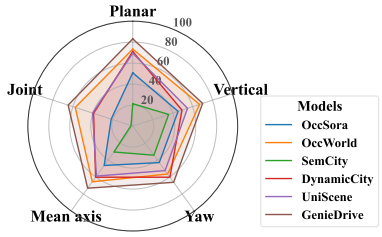


Figure 6: **Recognition accuracy (%).** Higher values indicate more reliable recognition across control dimensions.

Efficiency. Table 6 shows that efficiency is stage-dependent rather than determined by a single model-size ranking. Reconstruction is comparatively lightweight: UniScene achieves the lowest latency and highest throughput, whereas GenieDrive has the smallest inference-memory footprint. 4D occupancy generation is substantially more demanding, with temporal pipelines introducing additional computation and memory overhead.GenieDrive provides the most compact and lowest-latency generation path, while OccSora achieves the highest generation throughput at the cost of a

Table 4: **Distributional quality of 4D occupancy generation .** We report 2D and 3D IS, FID, KID, and distribution precision and recall to assess realism, distributional similarity, and coverage.

Method	IS_{2D} ↑	FID_{2D} ↓	KID_{2D} ↓	P_{2D} ↑	R_{2D} ↑	IS_{3D} ↑	FID_{3D} ↓	KID_{3D} ↓	P_{3D} ↑	R_{3D} ↑
OccSora (Wang et al., 2024)	1.008	616.812	4.609	0.222	0.000	**2.635**	31.625	0.679	0.107	0.225
OccWorld (Zheng et al., 2024)	1.010	48.829	0.162	0.943	0.060	1.256	2.476	**0.019**	0.887	0.440
SemCity (Lee et al., 2024)	**1.014**	25.253	**0.059**	0.719	0.169	1.322	12.142	0.165	0.391	0.546
DynamicCity (Bian et al., 2025)	1.007	**16.885**	0.062	**0.991**	0.068	1.129	**1.587**	**0.019**	**0.907**	0.586
UniScene (Li et al., 2025)	1.007	304.984	1.827	0.752	0.000	1.727	11.720	0.275	0.361	0.113
GenieDrive (Yang et al., 2026)	1.010	31.520	0.107	0.964	**0.422**	1.201	2.499	0.034	0.868	**0.601**

Table 5: **Trajectory recovery performance on the UAVX test set.** We compare position, velocity, and orientation errors between trajectories recovered from generated occupancy and ground truth.

Method	ADE (m) ↓	FDE (m) ↓	Velocity MAE (m/s) ↓	Yaw MAE (°) ↓	roll MAE (°) ↓	pitch MAE (°) ↓
OccSora (Wang et al., 2024)	2.618	5.216	1.415	2.52	**2.52**	1.96
OccWorld (Zheng et al., 2024)	**1.527**	**3.095**	**1.044**	2.040	2.545	1.999
SemCity (Lee et al., 2024)	3.681	6.396	4.560	3.74	3.31	3.20
DynamicCity (Bian et al., 2025)	1.960	3.886	1.279	2.408	2.566	**1.950**
UniScene (Li et al., 2025)	1.665	3.232	1.172	**1.569**	2.596	2.088
GenieDrive (Yang et al., 2026)	1.834	3.807	1.109	1.931	2.536	2.058

much larger model. DynamicCity incurs the largest memory footprint in both stages. These results indicate that practical deployment involves a trade-off between throughput, memory usage, and generative capacity.

Table 6: **Reconstruction and generation efficiency.** FLOPs and latency are measured on 16-frame clips; reconstruction includes encoding and decoding, while generation samples and decodes voxels.

Stage	Method	Params (M) ↓	FLOPs (T) ↓	Train mem. (GB) ↓	Infer. mem. (GB) ↓	Latency (s/clip) ↓	Throughput (clips/s) ↑
Reconstruction	OccSora (Wang et al., 2024)	162.54	3.06	12.49	3.35	0.15	6.566
	OccWorld (Zheng et al., 2024)	14.61	1.30	14.96	2.01	0.39	2.556
	SemCity (Lee et al., 2024)	**0.50**	11.73	9.18	5.83	0.50	1.985
	DynamicCity (Bian et al., 2025)	1.76	8.09	41.99	12.04	0.53	2.052
	UniScene (Li et al., 2025)	12.40	**0.99**	14.47	2.57	**0.10**	**10.532**
	GenieDrive (Yang et al., 2026)	2.26	1.35	**5.04**	**0.47**	0.61	1.649
Generation	OccSora (Wang et al., 2024)	113.99	5.06	5.87	2.72	1.51	**3.401**
	OccWorld (Zheng et al., 2024)	71.00	4.91	5.68	1.15	1.59	0.629
	SemCity (Lee et al., 2024)	59.43	352.99	**1.83**	6.01	9.12	0.110
	DynamicCity (Bian et al., 2025)	45.72	72.19	16.18	10.78	1.58	0.863
	UniScene (Li et al., 2025)	63.19	33.31	10.81	2.79	1.41	0.726
	GenieDrive (Yang et al., 2026)	**2.26**	**1.35**	5.04	**0.47**	**0.61**	1.649

These results indicate that semantic fidelity, temporal consistency, and conditional control capture different aspects of 4D occupancy modeling. DynamicCity remains competitive on distributional precision, while the static SemCity reference is useful for calibrating single-frame behavior but cannot establish temporal forecasting or control.

Limitations. Overall, UAVX demonstrates that 4D occupancy generation under aerial motion remains challenging and that semantic fidelity, temporal consistency, and control adherence represent complementary capabilities. As a simulation-based benchmark, its conclusions are necessarily bounded by the gap between the benchmark setting and the diversity of real-world UAV operation. Future work should broaden the benchmark and develop models that are more robust, scalable, and uncertainty-aware.

6 Conclusion

We introduced UAVX, a benchmark for 4D semantic occupancy generation under full 6-DoF UAV motion. It provides unified evaluation of generation quality, temporal consistency, trajectory recovery, command adherence, and efficiency under a common protocol. Results reveal complementary strengths across methods and persistent long-horizon challenges. UAVX establishes a reproducible foundation for advancing robust and controllable aerial world models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, to assist in writing and editing software code, and to edit the manuscript for clarity, grammar, and readability. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or evaluate research methodology or experiments, clean or reformat datasets, perform qualitative or thematic analysis, or interpret results. All other required disclosure tasks are not applicable to this work. We reviewed and verified all AI-assisted translations, code, and manuscript edits for accuracy, correctness, and consistency with the research. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data were generated in a simulation platform; no real-world UAV flights, human participants, personal data, or identifiable locations were involved. The simulator and its assets were used in accordance with their applicable licenses. The released materials are intended for research and benchmarking only and do not provide operational flight-control or real-world surveillance capabilities.

REPRODUCIBILITY STATEMENT

We will release immutable scene-level split manifests, preprocessing and label mapping code, metric implementations, baseline configurations, checkpoint and feature-extractor hashes, and per-sample evaluation outputs. Appendix A lists the metadata required to reproduce every table.

REFERENCES

Hengwei Bian, Lingdong Kong, Haozhe Xie, Liang Pan, Yu Qiao, and Ziwei Liu. Dynamiccity: Large-scale 4d occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. OpenAI technical report.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Markus Gross, Sai B. Matha, Aya Fahmy, Rui Song, Daniel Cremers, and Henri Meeß. Occufly: A 3d vision benchmark for semantic scene completion from the aerial perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21474–21485, 2026.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018a.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2451–2463. Curran Associates, Inc., 2018b. URL https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*, 2019a.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Junjie Huang et al. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Lingdong Kong, Yu Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025. URL https://arxiv.org/abs/2509.07996.

Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bohan Li, Jiazhe Guo, Hongsi Liu, Yingshuang Zou, Yikang Ding, Xiwu Chen, Hu Zhu, Feiyang Tan, Chi Zhang, Tiancai Wang, Shuchang Zhou, Li Zhang, Xiaojuan Qi, Hao Zhao, Mu Yang, Wenjun Zeng, and Xin Jin. Uniscene: Unified occupancy-centric driving scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xingyu Liu, Charles R. Qi, and Leonidas J. Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

OpenAI. Video generation models as world simulators, 2024. Technical report.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Lening Wang, Wenzhao Zheng, Yilong Ren, Han Jiang, Zhiyong Cui, Haiyang Yu, and Jiwen Lu. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*, 2024.

Yuping Wang, Xiangyu Huang, Xiaokang Sun, Mingxuan Yan, Shuo Xing, Zhengzhong Tu, and Jiachen Li. Uniocc: A unified benchmark for occupancy forecasting and prediction in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25560–25570, 2025.

Yiming Wei et al. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *arXiv preprint arXiv:2405.05173*, 2024.

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*, 2026.

Tianle Zeng, Hanxuan Chen, Yanci Wen, and Hong Zhang. Carla-air: Fly drones inside a carla world – a unified infrastructure for air-ground embodied intelligence. *arXiv preprint arXiv:2603.28032*, 2026. URL https://arxiv.org/abs/2603.28032.

Weichen Zhang, Peizhi Tang, Xin Zeng, Fanhang Man, Shiquan Yu, Zichao Dai, Baining Zhao, Hongjin Chen, Yu Shang, Wei Wu, Chen Gao, Xinlei Chen, Xin Wang, Yong Li, and Wenwu Zhu. Aerial world model for long-horizon visual generation and navigation in 3d space. *arXiv preprint arXiv:2512.21887*, 2025. URL https://arxiv.org/abs/2512.21887.

Baining Zhao, Rongze Tang, Mingyuan Jia, Ziyou Wang, Fanhang Man, Xin Zhang, Yu Shang, Weichen Zhang, Wei Wu, Chen Gao, Xinlei Chen, and Yong Li. Airscape: An aerial generative world model with motion controllability. In *Proceedings of the 33rd ACM International Conference on Multimedia*. ACM, 2025. doi: 10.1145/3746027.3758180. URL https://doi.org/10.1145/3746027.3758180.

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*, 2024.

A Benchmark Protocol and Reproducibility Specification

The benchmark release and each reported result should specify the following fields. Paths are relative to the repository root. The values below define the current UAVX validation protocol; the feature-weight hashes are recorded as pending until the trained extractor checkpoints are archived with the release.

```
coordinate_frame: gravity-aligned yaw; origin=drone_position;
  x=forward, y=left, z=world_up; roll/pitch excluded from voxel grid
voxel_range_m: [-25.6, -25.6, -30.0, 25.6, 25.6, 10.0]
voxel_size_m: [0.4, 0.4, 0.625]
grid_shape: [128, 128, 64]
sampling_hz: 10.0
sequence_frames: 16
sequence_seconds: 1.5
empty_id: 0
ignore_ids: [11]
condition_type: track-specific
  options: reconstruction|unconditional|history|layout|trajectory|command
feature_extractor_2d: InceptionV3 pool3 + VGG16 last-conv pool;
  ImageNet1K_V1 initialization with UAVX train fine-tuning;
feature_extractor_3d: SPVNAS MinkowskiUNet stage4 global-average-pool;
num_eval_samples: 5940
random_seeds: [0, 1, 2]
```

B Evaluation Metric Details

This appendix gives the detailed construction of the reconstruction and temporal-consistency metrics used in Sections 4. All operations are performed inside the common valid region after applying the frozen voxel mask and coordinate convention.

B.1 Surface $F@1v$

For frame t, we extract the occupied/free interface from the predicted and ground-truth voxel grids, obtaining point sets P_t and G_t. Surface voxels adjacent to invalid in-grid voxels are removed; the grid exterior is treated as free. The voxel resolution defines the tolerance $\tau = \max(v_x, v_y, v_z)$, so a point is considered matched when its nearest counterpart lies within one physical voxel step. For nonempty surface pairs, the directed surface precision and recall are

$$\begin{cases} P_t^{1v} = \dfrac{1}{|P_t|} \displaystyle\sum_{p \in P_t} \mathbf{1}\left[\min_{g \in G_t} \|p - g\|_2 \le \tau\right], \\ R_t^{1v} = \dfrac{1}{|G_t|} \displaystyle\sum_{g \in G_t} \mathbf{1}\left[\min_{p \in P_t} \|g - p\|_2 \le \tau\right]. \end{cases} \tag{9}$$

Here $\mathbf{1}[\cdot]$ is the indicator function. To compute P_t^{1v}, we query the nearest ground-truth surface point in G_t for each predicted point $p \in P_t$. That point contributes one to the numerator when the nearest-neighbor distance is at most τ and contributes zero otherwise; dividing by $|P_t|$ therefore gives the fraction of predicted surface points that are supported by the reference surface. The computation of R_t^{1v} is the reverse directed test: for each reference point $g \in G_t$, we query its nearest predicted point in P_t and divide the number of matches by $|G_t|$. These are directed nearest-neighbor tests rather than one-to-one assignments, so multiple points on one surface may use the same nearest counterpart. Precision penalizes predicted surface points without a nearby reference point, whereas recall penalizes missing reference surface points. Their harmonic mean for frame t is

$$F_t^{1v} = \frac{2P_t^{1v}R_t^{1v}}{P_t^{1v} + R_t^{1v}}, \tag{10}$$

with $F_t^{1v} = 0$ when exactly one surface is empty or the denominator is zero; it is N/A only when both surfaces are empty. The reported clip score is the macro-average over valid frames,

$$F@1v = \frac{1}{|\mathcal{T}_{\text{valid}}|} \sum_{t \in \mathcal{T}_{\text{valid}}} F_t^{1v}, \tag{11}$$

followed by macro-averaging over clips. This construction makes the metric sensitive to boundary shifts, over-thick surfaces, and small structures even when volumetric IoU remains relatively high.

B.2 Geometric Reconstruction Metrics

Binary occupancy precision and recall are computed over all valid test voxels. If TP_{occ}, FP_{occ}, and FN_{occ} denote pooled occupied voxel counts, OccP measures the fraction of predicted occupied voxels that are correct, while OccR measures the fraction of reference occupied voxels that are recovered. These ratios are reported as percentages and are distinct from the feature-distribution P_{2D}/R_{2D} and P_{3D}/R_{3D}, which measure generated-sample fidelity and coverage in feature space. The thresholded surface scores $F@1v$ and $F@2v$ use the corresponding voxel-based tolerances defined by the evaluation protocol.

$$\text{OccP} = \frac{TP_{\text{occ}}}{TP_{\text{occ}} + FP_{\text{occ}}}, \qquad \text{OccR} = \frac{TP_{\text{occ}}}{TP_{\text{occ}} + FN_{\text{occ}}}. \tag{12}$$

For each frame, we take the centers of valid occupied voxels with at least one free 6-neighbor as the predicted and ground-truth surfaces P_t and G_t. Surface voxels adjacent to invalid in-grid voxels are discarded, and the grid exterior is treated as free. Using physical voxel spacings (v_x, v_y, v_z), we calculate Euclidean nearest-neighbor distances in metres. Surface completeness (SComp) averages the distance from each reference surface point in G_t to its nearest predicted point in P_t, whereas surface accuracy (SAcc) averages the reverse distance from each predicted point to G_t. Frame-level Chamfer–L_1 is the mean of SAcc and SComp. Frames with either surface empty have undefined distance metrics and are excluded; defined frame values are averaged within each clip, then averaged over clips with defined values. SComp exposes missing reference surfaces, whereas SAcc exposes unsupported predicted surfaces. Unlike thresholded $F@1v$, these distances retain the magnitude of each localization error; lower is better.

$$\begin{aligned} \text{SComp}_t &= \frac{1}{|G_t|} \sum_{g \in G_t} \min_{p \in P_t} \|g - p\|_2, \\ \text{SAcc}_t &= \frac{1}{|P_t|} \sum_{p \in P_t} \min_{g \in G_t} \|p - g\|_2. \end{aligned} \tag{13}$$

B.3 Change-region IoU

Let Y_t and $\hat{Y}_t$ be the ground-truth and predicted semantic occupancy grids at frame t. The relative transform between adjacent ego frames is computed from the world-to-ego poses as

$$T_{t \leftarrow t+1} = T_{\text{world} \leftarrow \text{ego},t}^{-1} T_{\text{world} \leftarrow \text{ego},t+1}. \tag{14}$$

The operator $W_t(\cdot)$ applies this full $\mathrm{SE}(3)$ transform using nearest-neighbor label warping, removes out-of-bounds voxels, and returns the common valid region Ω_t. We then define change masks for any occupancy sequence Z as

$$\Delta_t(Z) = \mathbf{1}\left[Z_t \neq W_t(Z_{t+1})\right] \cap \Omega_t. \tag{15}$$

The per-pair score is the intersection-over-union of predicted and reference change regions,

$$\mathrm{ChgIoU}_t = \frac{|\Delta_t(\hat{Y}) \cap \Delta_t(Y)|}{|\Delta_t(\hat{Y}) \cup \Delta_t(Y)|}, \tag{16}$$

and the clip-level result is the macro-average over the 15 adjacent frame pairs with a non-empty valid union. Full-pose alignment is important here: otherwise, apparent changes caused only by translation, altitude, roll, pitch, or yaw would be incorrectly counted as scene dynamics.

B.4 BG-Cons.

For adjacent ego-centric frames, the world-to-ego poses define the relative transform

$$T_{t\leftarrow t+1} = T^{-1}_{\mathrm{world}\leftarrow\mathrm{ego},t} T_{\mathrm{world}\leftarrow\mathrm{ego},t+1}. \tag{17}$$

The warp operator applies this full $\mathrm{SE}(3)$ transform with nearest-neighbor labels, removes out-of-bounds voxels, and returns Ω_t. For each static class s in the frozen set S, we compute

$$b_{t,s} = \mathrm{IoU}_{\Omega_t}\left(B^s_t, \mathrm{Warp}(B^s_{t+1}, T_{t\leftarrow t+1})\right). \tag{18}$$

Here:

- B^s_t is the binary mask of static class s in frame t;
- B^s_{t+1} is the binary mask of static class s in frame $t+1$;
- $\mathrm{Warp}(\cdot, T_{t\leftarrow t+1})$ applies the relative-pose transform to the second mask and maps it into the coordinate system of frame t;
- IoU_{Ω_t} is the intersection-over-union computed only inside the common valid region Ω_t.

The reported BG-Cons. is the macro-average of $b_{t,s}$ over static classes and the $T-1$ adjacent pairs. Class-wise averaging prevents large road or building regions from dominating thin static categories such as poles and barriers.

B.5 FGSC

For each dynamic instance k, the evaluator first associates instances between adjacent frames using the frozen class, overlap, and distance thresholds. Only successfully matched vehicle and pedestrian instances enter K_t. The predicted and reference masks are then rasterized in a canonical object frame: centroids remove translation and principal axes remove rotation. The aligned mask overlap is

$$f_{t,k} = \mathrm{IoU}\left(I^k_t, \mathrm{Align}(I^k_{t+1})\right). \tag{19}$$

The foreground shape consistency over the valid adjacent pairs is

$$\mathrm{FGSC} = \frac{1}{|\mathcal{T}_{\mathrm{valid}}|} \sum_{t\in\mathcal{T}_{\mathrm{valid}}} \frac{1}{|K_t|} \sum_{k\in K_t} f_{t,k}. \tag{20}$$

Here:

- I^k_t is the voxel mask of dynamic instance k in frame t;
- $\mathrm{Align}(I^k_{t+1})$ is the next-frame instance mask transformed into the same canonical object frame as I^k_t;
- $\mathrm{IoU}(\cdot,\cdot)$ is the intersection-over-union of the two aligned instance masks;
- K_t is the set of instances with a valid association and a non-empty aligned mask union for the pair $(t, t+1)$;
- $\mathcal{T}_{\mathrm{valid}}$ is the set of adjacent frame-pair indices for which K_t is non-empty.

For each valid adjacent pair, the evaluator first averages $f_{t,k}$ over the matched instances in K_t, and then averages these pair-level scores over $\mathcal{T}_{\mathrm{valid}}$. Thus, a pair with no valid matched instance is excluded from the macro-average rather than assigned zero. This construction penalizes shape collapse, stretching, and local flicker while not penalizing a physically valid object translation.

B.6 SEMFLICKER

Before measuring changes in the predicted semantic labels, we remove regions where the ground-truth scene genuinely changes and retain only regions that should remain stable. Specifically, the stable region is

$$\mathrm{Stable}_t = \Omega_t \cap \{Y_t = W_t(Y_{t+1})\} \cap \mathrm{StaticRegion}_t. \tag{21}$$

Here:

- Ω_t is the common valid region after aligning adjacent frames by the relative ego-motion;
- Y_t is the ground-truth semantic occupancy grid at frame t;
- $W_t(Y_{t+1})$ is the ground-truth semantic grid at frame $t+1$ warped into the coordinate system of frame t;
- $\mathrm{StaticRegion}_t$ is the static-scene region in frame t.

Thus, a voxel enters Stable_t only when the current ground-truth label agrees with the ego-motion-compensated label from the next frame. Within this region, the per-pair semantic flicker rate is

$$\mathrm{SemFlicker}_t = \frac{\sum_{v \in \mathrm{Stable}_t} \mathbf{1}\left[\hat{Y}_t(v) \neq W_t(\hat{Y}_{t+1})(v)\right]}{|\mathrm{Stable}_t|}, \tag{22}$$

where $\hat{Y}_t$ and $\hat{Y}_{t+1}$ are the predicted semantic occupancy grids, and the same ego-motion warp is applied to the prediction at frame $t+1$. The numerator counts stable voxels whose predicted semantic label changes after this warp, while the denominator is the total number of stable voxels. The clip-level SemFlicker is the macro-average over the adjacent pairs with a non-empty stable region:

$$\mathrm{SemFlicker} = \frac{1}{|\mathcal{T}_{\mathrm{valid}}|} \sum_{t \in \mathcal{T}_{\mathrm{valid}}} \mathrm{SemFlicker}_t, \tag{23}$$

where $\mathcal{T}_{\mathrm{valid}} = \{t : |\mathrm{Stable}_t| > 0\}$. Lower values indicate fewer temporal label changes. Newly revealed space, invalid voxels, and legitimate object motion do not contribute to this metric.

B.7 COMMAND RECOGNITION ACCURACY

Command recognition evaluates whether a generated occupancy clip follows the requested planar, vertical, and yaw commands. A frozen three-head recognizer is applied to each generated clip. For clip i, let $\mathbf{y}_i = (y_{i,\mathrm{p}}, y_{i,\mathrm{v}}, y_{i,\mathrm{y}})$ be the requested command tuple and let $\hat{\mathbf{y}}_i = (\hat{y}_{i,\mathrm{p}}, \hat{y}_{i,\mathrm{v}}, \hat{y}_{i,\mathrm{y}})$ be the recognizer output. The three command axes and their class sets are

$$\mathcal{A} = \{\mathrm{p}, \mathrm{v}, \mathrm{y}\}, \qquad |\mathcal{C}_{\mathrm{p}}| = 5, \quad |\mathcal{C}_{\mathrm{v}}| = 3, \quad |\mathcal{C}_{\mathrm{y}}| = 3, \tag{24}$$

where the corresponding class names are listed in the per-command result table. The joint command space is their Cartesian product and therefore has $5 \times 3 \times 3 = 45$ possible tuples.

For an axis $a \in \mathcal{A}$ and a requested class $c \in \mathcal{C}_a$, define the subset of clips requesting that class and its size by

$$\mathcal{I}_{a,c} = \{i \in \{1, \ldots, N\} : y_{i,a} = c\}, \qquad N_{a,c} = |\mathcal{I}_{a,c}|. \tag{25}$$

The class-conditional recognition accuracy is the fraction of those clips for which the recognizer returns the requested class:

$$a_{a,c} = \frac{1}{N_{a,c}} \sum_{i \in \mathcal{I}_{a,c}} \mathbf{1}[\hat{y}_{i,a} = c]. \tag{26}$$

The table reports $100a_{a,c}$ in percent. This conditioning ensures that a frequent command does not hide poor recognition of a less frequent command. If $N_{a,c} = 0$, that class is unavailable and is excluded from the corresponding macro average rather than assigned zero. Let $\mathcal{C}_a^+ = \{c \in \mathcal{C}_a : N_{a,c} > 0\}$ denote the available classes. The raw macro accuracy for axis a is

$$A_a = \frac{1}{|\mathcal{C}_a^+|} \sum_{c \in \mathcal{C}_a^+} a_{a,c}. \tag{27}$$

Under the frozen test split all command classes are present, so $\mathcal{C}_a^+ = \mathcal{C}_a$.

Joint accuracy tests the complete command tuple on the same clip. Define

$$j_i = \prod_{a \in \mathcal{A}} \mathbf{1}[\hat{y}_{i,a} = y_{i,a}], \qquad A_{\text{joint}} = \frac{1}{N} \sum_{i=1}^{N} j_i. \tag{28}$$

Thus, a clip contributes one to A_{joint} only when its planar, vertical, and yaw labels are all correct; correctness on separate axes is not combined under an independence assumption.

Because the axes have different numbers of classes, chance-level performance is different for each raw accuracy. The chance rates are

$$c_{\text{p}} = \frac{1}{5}, \qquad c_{\text{v}} = \frac{1}{3}, \qquad c_{\text{y}} = \frac{1}{3}, \qquad c_{\text{joint}} = \frac{1}{45}. \tag{29}$$

For any raw accuracy x and its chance rate c, chance-rate normalization maps chance to zero and perfect recognition to 100:

$$Q(x; c) = 100 \,\text{clip}\left(\frac{x - c}{1 - c}, 0, 1\right). \tag{30}$$

We first average the three normalized axis scores and then combine that mean with the normalized Joint score:

$$S_{\text{axis}} = \frac{1}{3} \sum_{a \in \{\text{p,v,y}\}} Q(A_a; c_a), \qquad S_{\text{cmd}} = \frac{1}{2} S_{\text{axis}} + \frac{1}{2} Q(A_{\text{joint}}; c_{\text{joint}}). \tag{31}$$

The equal weights give the axis-wise behavior and complete tuple adherence the same influence. Higher A_a, A_{joint}, and S_{cmd} indicate more reliable command following. The recognizer is trained only on ground-truth training clips and is frozen during evaluation; its validation accuracy is reported separately as a calibration ceiling.

B.8 Feature spaces and distributional evaluation

This appendix specifies the feature spaces and evaluation protocol used for the distribution metrics in Section 4. Real and generated occupancy use the same frozen renderer, extractor weights, preprocessing, and sample protocol. For each 16-frame clip, frame-level features are averaged into one clip feature before computing distribution statistics; the current test protocol contains 5,940 stride-1 clips. For 2D evaluation, valid non-Free occupancy voxels are rendered with a fixed semantic colormap. InceptionV3 provides a 2048-dimensional global-average-pool feature for FID/KID and town-class probabilities for Inception Score, while VGG-16 provides a 512-dimensional global-average-pool feature for Precision/Recall. For 3D evaluation, a semantic MinkowskiUNet autoencoder with SPVNAS topology provides a 256-dimensional stage4 feature for FID/KID and Precision/Recall, together with mean voxel-class probabilities for Inception Score. Inception Score, FID, and KID use their standard definitions cited in the main text; this appendix specifies their representations and protocol.

Precision and Recall use a real-only PCA projection to stabilize covariance estimation. The PCA mean and principal directions are fitted on real features and then applied unchanged to both real and generated features. We retain 16 dimensions for the 2D VGG space and 10 dimensions for the 3D space. These dimensions are fixed by a split-half self-check on real data at target coverage 0.95. FID and KID use the original, unprojected features. We use the standard Gaussian-ellipsoid Precision and Recall protocol without changing its metric definition. In the retained feature space, we estimate unbiased sample covariances and use a 0.95 chi-square coverage threshold. Precision is the fraction of generated samples inside the real-data ellipsoid, whereas Recall is the fraction of real samples inside the generated-data ellipsoid. The two scores are reported separately for the 2D and 3D spaces; higher values indicate better fidelity and coverage, respectively. If a required covariance is rank deficient, the corresponding metric is reported as unavailable rather than changing the protocol with an implicit regularizer.

C SUPPLEMENTARY EXPERIMENTS

C.1 RECONSTRUCTION RESULTS

Per-class Semantic Reconstruction. Figure 7 visualizes the per-class semantic reconstruction IoU under the frozen 12-class taxonomy. Free and unknown voxels are excluded from semantic mIoU; the rightmost bars report the free-region IoU separately.

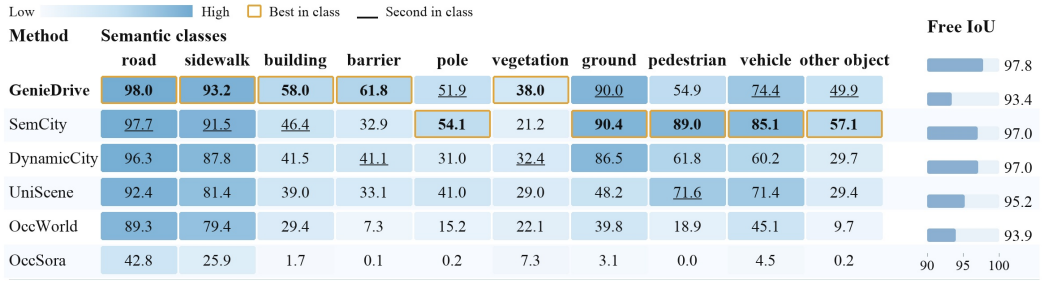


Figure 7: **Per-class semantic reconstruction IoU (%) on the UAVX test set.** The heatmap reports class-wise IoU under the frozen semantic taxonomy, while the right panel reports Free IoU separately. Higher values indicate better reconstruction; color intensity encodes IoU, orange boxes mark the best value in each class, and underlining marks the second-best value.

Geometric Reconstruction Results. Table 7 reports the geometric reconstruction diagnostics defined in Appendix B.2. The table decomposes binary occupancy and surface localization using the fixed voxel validity mask, physical distances, and threshold conventions specified there.

Table 7: **Geometric reconstruction diagnostics on the UAVX test set.** We compare occupied-voxel precision and recall with surface localization and voxel-tolerance scores.

Method	OccP (%) ↑	OccR (%) ↑	SAcc (m) ↓	SComp (m) ↓	Chamfer–L_1 (m) ↓	$F@1v$ (%) ↑	$F@2v$ (%) ↑
OccSora (Wang et al., 2024)	68.16	25.70	1.785	3.762	2.774	19.70	34.91
OccWorld (Zheng et al., 2024)	64.21	75.80	0.884	0.978	0.931	59.62	75.63
SemCity (Lee et al., 2024)	52.82	**93.84**	1.069	0.484	0.777	66.16	77.35
DynamicCity (Bian et al., 2025)	81.55	76.12	0.352	0.501	0.427	78.30	91.88
UniScene (Li et al., 2025)	**88.06**	68.44	**0.250**	0.734	0.492	75.64	88.75
GenieDrive (Yang et al., 2026)	86.35	82.65	0.280	**0.355**	**0.318**	**85.65**	**94.38**

C.2 GENERATION RESULTS

Per-class Semantic Generation. Figure 8 visualizes per-class semantic generation IoU on the UAVX test set under the frozen 12-class taxonomy. Free and unknown voxels are excluded from semantic mIoU, while the rightmost bars report Free IoU separately.

Low High Best in class Second in class

Method	Per-frame semantic generation mIoU (higher is better)															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OccSora	0.54	0.55	0.57	0.60	0.59	0.60	0.60	0.60	0.59	0.59	0.60	0.60	0.59	0.58	0.55	0.54
OccWorld	30.83	28.24	25.74	23.48	21.68	20.13	18.75	17.58	16.55	15.65	14.84	14.16	13.52	12.95	12.42	11.94
SemCity	0.57	0.55	0.53	0.54	0.57	0.58	0.60	0.57	0.57	0.58	0.53	0.56	0.57	0.58	0.57	0.56
DynamicCity	34.13	30.38	27.14	23.65	21.40	19.12	17.64	16.17	15.09	13.81	12.81	12.00	10.98	9.85	8.77	7.77
UniScene	8.88	9.20	9.30	9.40	9.45	9.37	9.23	9.00	8.82	8.68	8.59	8.43	8.23	8.04	7.78	7.38
GenieDrive	**59.88**	**55.28**	**51.26**	**47.50**	**44.28**	**41.53**	**39.07**	**36.81**	**34.79**	**32.90**	**31.15**	**29.56**	**28.10**	**26.78**	**25.57**	**24.45**

Figure 9: **Frame-wise semantic generation mIoU (%) on UAVX clips.** The heatmap reports semantic fidelity at each frame position in the 16-frame horizon; higher values indicate better generation quality.

Low High Best in class Second in class

Method	Semantic classes										Free IoU
	road	sidewalk	building	barrier	pole	vegetation	ground	pedestrian	vehicle	other object	
GenieDrive	**70.96**	**67.81**	**37.51**	**31.66**	**21.34**	**25.17**	**62.78**	**13.77**	**26.22**	**24.77**	95.93
OccWorld	55.18	49.36	15.34	1.40	3.33	16.18	24.43	2.04	13.00	3.26	93.04
DynamicCity	40.64	38.17	15.47	3.34	4.64	14.89	37.54	4.06	11.52	3.90	93.60
UniScene	14.99	14.92	12.77	1.17	6.06	1.97	8.60	8.98	13.69	4.20	92.80
OccSora	3.27	1.86	0.14	0.01	0.02	0.21	0.25	0.00	0.02	0.02	92.64
SemCity	1.85	1.43	0.60	0.03	0.05	1.23	0.39	0.00	0.01	0.06	89.09

85 90 95 100

Figure 8: **Per-class semantic generation IoU (%) on UAVX test set.** The heatmap reports class-wise IoU under the frozen semantic taxonomy, while the right panel reports Free IoU separately. Higher values indicate better generation fidelity; color intensity encodes IoU, orange boxes mark the best value in each class, and underlining marks the second-best value.

Per-frame Semantic Generation mIoU. Figure 9 compares semantic generation mIoU across the six methods at each of the 16 frame positions. The heatmap retains the per-frame values, showing how semantic fidelity changes over the generated sequence; orange boxes mark the best method at each position, and underlining marks the second-best method.

Table 8: **Per-command recognition accuracy (%) on the UAVX test set.** Columns report class-conditional recognition accuracy for the planar, vertical, and yaw command axes; higher values indicate better recognition.

Method	Planar					Vertical			Yaw		
	Still	Forward	Backward	Strafe-left	Strafe-right	Level	Ascend	Descend	Straight	Left	Right
OccSora (Wang et al., 2024)	45.29	51.64	7.59	74.67	74.61	57.46	50.72	28.19	90.83	25.55	12.50
OccWorld (Zheng et al., 2024)	36.67	92.70	65.37	82.86	89.40	**89.46**	45.05	**66.40**	92.21	42.10	35.12
SemCity (Lee et al., 2024)	**60.87**	14.98	0.00	21.33	9.49	56.74	32.12	19.06	41.64	55.15	6.10
DynamicCity (Bian et al., 2025)	30.37	**99.34**	41.67	84.57	**99.56**	56.56	**80.23**	11.17	96.69	**56.62**	27.38
UniScene (Li et al., 2025)	49.28	74.92	29.44	**96.76**	96.69	62.47	57.59	44.68	94.18	40.07	23.07
GenieDrive (Yang et al., 2026)	55.31	93.78	**83.70**	94.29	88.74	88.80	60.22	61.08	**97.10**	56.07	**45.54**

C.3 Temporal Consistency Results

Per-frame Temporal Consistency. Figure 10 reports the adjacent-frame temporal diagnostics as heatmaps. Column t compares frames $t-1$ and t; column 0 is undefined because there is no preceding frame.

C.4 Per-command recognition accuracy

This appendix reports accuracy for each individual command class in Table 8. Accuracy is computed over samples whose requested label is that command; the chance baseline is the reciprocal of the number of classes on the corresponding axis. The command columns are ordered as Planar (Still, Forward, Backward, Strafe-left, Strafe-right), Vertical (Level, Ascend, Descend), and Yaw (Straight,

Low High Best in class Second in class

Method	Per-frame temporal consistency for BG-Cons (higher is better)															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OccSora	--	23.75	34.31	21.28	35.15	23.29	34.96	22.02	35.43	22.06	34.84	23.40	34.82	21.13	34.79	24.17
OccWorld	--	54.73	54.24	53.43	52.95	52.57	52.14	51.90	51.64	51.48	51.23	51.08	50.94	50.77	50.67	50.58
SemCity	--	0.53	0.48	0.49	0.50	0.49	0.50	0.52	0.50	0.51	0.50	0.50	0.51	0.51	0.49	0.51
DynamicCity	--	46.21	47.62	45.53	42.89	43.88	43.63	41.87	39.26	41.27	40.99	38.89	35.38	36.45	35.15	30.84
UniScene	--	39.04	39.89	40.16	40.63	40.85	40.75	40.61	40.43	40.43	40.42	40.32	40.06	40.00	39.59	38.59
GenieDrive	--	**66.04**	**66.39**	**66.07**	**65.88**	**65.63**	**65.34**	**65.02**	**64.63**	**64.21**	**63.80**	**63.26**	**62.93**	**62.67**	**62.28**	**62.03**

Low High Best in class Second in class

Method	Per-frame temporal consistency for FGSC (higher is better)															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OccSora	--	35.98	54.06	30.64	56.33	34.43	53.19	30.00	56.50	29.80	53.44	32.08	57.88	32.53	54.96	32.45
OccWorld	--	48.60	48.66	48.59	48.77	48.60	48.59	48.37	48.32	48.22	47.95	47.88	47.55	47.60	47.56	47.42
SemCity	--	16.99	15.41	18.35	14.59	14.61	13.25	14.57	14.75	15.34	14.50	15.96	16.72	12.89	13.02	15.34
DynamicCity	--	55.45	56.84	55.61	53.59	54.49	54.64	53.42	52.07	53.64	52.89	53.10	48.74	50.93	48.94	46.98
UniScene	--	44.51	45.63	46.24	46.51	46.73	46.46	46.61	45.93	46.43	45.90	45.63	45.78	45.80	44.96	44.14
GenieDrive	--	**72.45**	**72.26**	**71.65**	**70.47**	**69.28**	**68.10**	**66.46**	**65.36**	**63.61**	**62.42**	**61.64**	**60.75**	**60.11**	**59.37**	**59.34**

Low High Best in class Second in class

Method	Per-frame temporal consistency for SemFlicker (lower is better)															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
OccSora	--	**1.22**	**1.01**	**1.67**	**1.06**	**1.52**	**1.09**	**1.61**	**1.06**	**1.59**	**1.11**	**1.54**	**1.16**	**1.63**	**1.12**	**1.44**
OccWorld	--	10.44	10.24	10.02	9.65	9.35	9.11	8.88	8.63	8.44	8.29	8.10	8.01	7.85	7.70	7.58
SemCity	--	13.35	13.32	13.45	13.63	13.71	14.17	13.96	13.74	13.78	13.52	13.31	13.76	13.93	13.78	13.58
DynamicCity	--	13.62	11.61	12.66	13.82	11.09	10.19	10.72	13.05	9.62	8.92	9.82	11.83	9.16	8.52	9.71
UniScene	--	6.24	6.11	6.06	5.66	5.66	5.56	5.48	5.35	5.21	5.34	5.31	5.26	5.22	5.20	5.28
GenieDrive	--	6.08	6.28	6.49	6.36	6.18	6.04	5.93	5.77	5.73	5.74	5.74	5.69	5.52	5.38	5.25

Figure 10: **Frame-wise temporal consistency on UAVX clips.** From top to bottom, the panels report BG-Cons., FGSC, and SemFlicker at each frame position. Higher values are better for BG-Cons. and FGSC, whereas lower values are better for SemFlicker. Column 0 is undefined because no preceding frame is available.

Left, Right). Each cell is the class-conditional match rate. The axis scores reported in the main text are macro averages of the corresponding command columns; Joint remains the exact-match rate requiring all three axis labels to be correct.