HGWM: HIERARCHICAL 3D GAUSSIAN WORLD MODEL WITH 4D OCCUPANCY FORECASTING FOR END-TO-END AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end autonomous driving requires a world state that preserves explicit three-dimensional geometry, evolves coherently over time, and remains directly accessible to the planner. Existing driving world models commonly rely on dense grids or latent representations, making future geometry and its connection to planning less explicit. We propose the *Hierarchical Gaussian World Model* (HGWM), which formulates semantic 3D Gaussian primitives as an explicit recurrent world state shared by 4D occupancy forecasting and end-to-end trajectory planning. Rather than evolving all primitives uniformly, HGWM decomposes the scene into comparatively stable background and actor-centric foreground fields and models their distinct intra- and cross-field interactions through geometry-aware hierarchical communication and sparse local refinement. A residual transition operator jointly evolves Gaussian geometric, feature, opacity, and semantic attributes, while sparse expert routing accommodates heterogeneous foreground dynamics. Crucially, the same autoregressively evolved Gaussian state is rendered into future semantic occupancy and directly queried to generate multi-modal ego trajectories, coupling scene forecasting and planning through a shared geometric state. Under the reported camera-only nuScenes protocol, HGWM achieves 25.49% semantic mIoU and 49.86% binary IoU at 3 s, together with an average planning L2 error of 0.28 m and a 0.11% collision rate. It further reaches a PDMS of 92.1 on NAVSIM and a Driving Score of 82.6 with a 61.8% Success Rate on Bench2Drive. These results support hierarchical Gaussian evolution as an explicit shared world-state formulation for future scene prediction and trajectory planning.

1 INTRODUCTION

A driving policy must choose an action before the future scene is observed. End-to-end systems therefore benefit from a state that is useful not only for perception, but also for predicting how geometry and actors will change. Planning-oriented methods such as UniAD, VAD, and SparseDrive learn strong scene-to-action representations (Hu et al., 2023b; Jiang et al., 2023; Sun et al., 2025), while driving world models predict future observations or latent states (Hu et al., 2023a; Wang et al., 2024b). Yet the interface between these two objectives remains consequential: if forecasting and planning operate on different or weakly coupled states, accurate scene prediction need not translate into a planning-relevant representation.

The central observation of this work is that future occupancy and ego motion should be read from the same predicted world state. Such a state must satisfy three requirements. It should expose 3D geometry, evolve coherently over a multi-step horizon, and remain compact enough for direct planning queries. Dense voxel or BEV grids make occupancy explicit but allocate computation to empty and occupied space alike. Compact latent tokens are efficient, but the geometry being rolled out is less directly inspectable. Semantic 3D Gaussians offer a middle ground: each primitive carries a center, anisotropic covariance, opacity, feature, and semantic distribution, so the state is sparse and geometrically grounded while remaining differentiable.

Representing the scene is only the first step. Static layout and actor-centric dynamics obey different priors, and treating every primitive uniformly can displace background structure or blur moving-

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

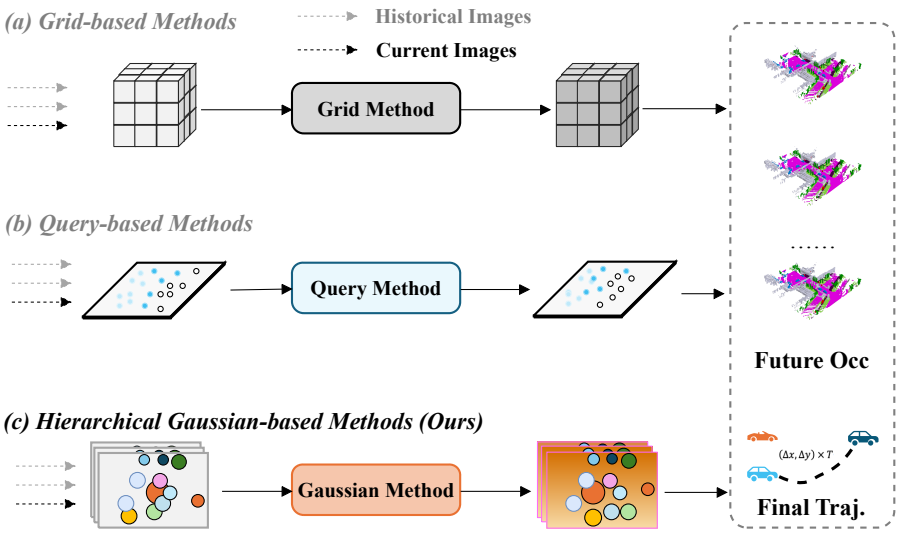


Figure 1: Internal world-state representations. Grid and token pipelines can also support planning; the schematic contrasts how explicitly the rollout geometry is exposed. HGWM evolves sparse semantic Gaussians and uses the same predicted state for occupancy and planning.

object boundaries. HGWM therefore partitions its Gaussian state into background and foreground fields. A hierarchical block models four directed relations within and across these fields. Dynamic-Static Decoupled Spatial-Aware Graph Attention (DS-SAGA) then refines foreground primitives from sparse local context, and a Sparse Mixture-of-Experts (SMoE) represents heterogeneous actor dynamics. Residual updates roll the state forward for six 0.5 s steps. At every step, an occupancy renderer and an independent ego query read the same state.

This design makes the empirical argument testable in stages. We first assess whether the shared state supports planning, then whether it preserves future scene geometry, and finally whether the proposed interaction and regularization mechanisms explain the observed behavior. Across nuScenes, NAVSIM, and Bench2Drive, HGWM supports both scene-level forecasting and action-level evaluation under the reported protocols.

Our contributions are:

- an explicit recurrent semantic-Gaussian state whose autoregressive rollout is shared by 4D occupancy forecasting and trajectory planning;
- a field-aware transition with typed background–foreground interaction, sparse DS-SAGA refinement, and sparse experts for heterogeneous foreground dynamics;
- evaluation of future geometry, open-loop planning, non-reactive planning, and closed-loop driving, together with controlled mechanism and rollout diagnostics.

2 Related Work

End-to-end autonomous driving. Spatial-temporal learning (Hu et al., 2022), planning-oriented perception (Hu et al., 2023b), vectorized scene representations (Jiang et al., 2023), sparse driving representations (Sun et al., 2025), diffusion planners (Liao et al., 2025), and unified transformers (Jia et al., 2025) increasingly couple perception and action. WorldRFT (Yang et al., 2026) and SafeDrive (Kim et al., 2026) further study reinforcement fine-tuning and safety reasoning. HGWM is complementary: it makes the predicted state shared by forecasting and planning explicit.

Driving world models and occupancy. GAIA-1 (Hu et al., 2023a), DriveDreamer (Wang et al., 2024a), and Drive-WM (Wang et al., 2024b) synthesize future video or latent features, whereas

SLEDGE (Chitta et al., 2024) and GUMP (Hu et al., 2024) emphasize structured scenario generation. OccWorld (Zheng et al., 2024a), RenderWorld (Yan et al., 2024), and GaussianWorld (Zuo et al., 2025) directly model occupancy evolution. HGWM instead uses one explicit rollout for both occupancy and planning.

Gaussian scene representations. 3D Gaussian Splatting (Kerbl et al., 2023) represents geometry with continuous anisotropic primitives. GaussianFormer (Huang et al., 2024) and GaussianFormer-2 (Huang et al., 2025) adapt them to semantic occupancy, while GaussianAD (Zheng et al., 2024b) and GaussianFusion (Liu et al., 2025) extend Gaussian representations toward planning and multi-sensor driving. HGWM focuses on field-aware recurrent evolution and shared occupancy/planning readouts.

3 Hierarchical Gaussian World Model

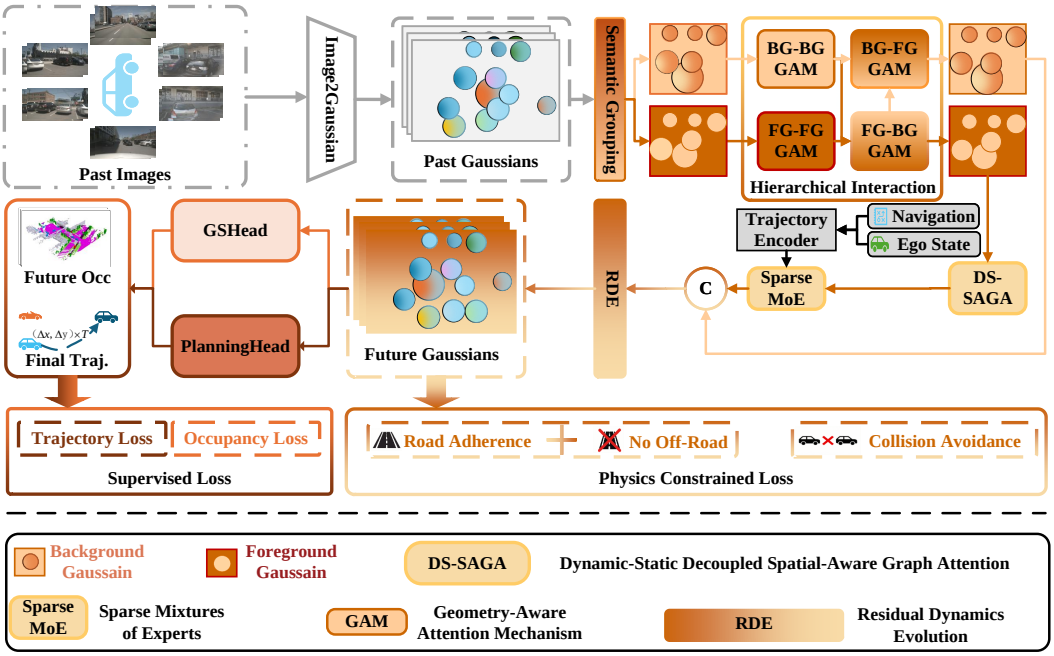


Figure 2: HGWM overview. Three ego-aligned surround-view frames initialize an explicit Gaussian state. Hierarchical interaction and residual dynamics roll it forward; each predicted state is rendered as semantic occupancy and read by an ego query for trajectory planning. NAVSIM and Bench2Drive add a LiDAR feature pyramid only at initialization.

3.1 Shared World State and Task

At time t, the camera-only nuScenes model consumes a three-frame, six-view history $\mathcal{I}_{t-2:t} = \{I_\ell^v : \ell \in \{t-2, t-1, t\}, v = 1, \ldots, 6\}$ aligned to the current ego frame using historical ego poses. Given ego-motion and navigation context u_t, it predicts semantic occupancy $\{\hat{Y}_{t+\tau}\}_{\tau=1}^{H}$ and $M = 5$ candidate trajectories for the active command, where $H = 6$ and $\Delta t = 0.5\,\text{s}$. NAVSIM and Bench2Drive fuse image and LiDAR feature pyramids before Gaussian initialization; the subsequent interaction, evolution, and decoding modules are unchanged. nuScenes uses cameras only at inference; occupancy labels and instance-token/3D-box associations provide training supervision.

The recurrent state is $S_t = \{g_{i,t}\}_{i=1}^{N}$ with $N = 6400$ and

$$g_i = \{\mu_i, s_i, q_i, o_i, h_i, z_i\}, \tag{1}$$

where $\mu_i, s_i \in \mathbb{R}^3$ are the center and positive anisotropic scale, $q_i \in \mathbb{R}^4$ is a unit quaternion, o_i is an opacity logit, $h_i \in \mathbb{R}^{256}$ is the state feature, and $z_i \in \mathbb{R}^C$ contains class logits. Thus

$\alpha_i = \sigma(o_i)$ and $\Sigma_i = R(q_i)\,\mathrm{diag}(s_i^2)R(q_i)^\top$. Following the GaussianFormer family (Huang et al., 2024; 2025), multi-scale features predict ray-conditioned occupied anchors, supplement them with learnable anchors, and refine both by multi-view deformable attention. Ray-conditioned samples place primitives in regions supported by image evidence, while learnable anchors cover regions without high-confidence ray samples. Centers are mapped to the metric perception range, scales are constrained positive, rotations begin at the identity quaternion, and opacity is initialized in logit space. Image2Gaussian is optimized jointly with the transition and readout modules rather than frozen after initialization.

A foreground classifier gives $p_{i,r}^{\mathrm{fg}} = \sigma(\phi_{\mathrm{fg}}(h_{i,r}))$. Eight actor classes form the foreground; all remaining semantic classes form the background. A threshold of 0.5 yields fields F_r and B_r. Reapplying the semantic and grouping heads after each transition allows field membership to change while the primitive budget remains fixed. The eight classes are car, truck, bus, trailer, construction vehicle, pedestrian, bicycle, and motorcycle. Foreground targets are obtained from valid semantic-occupancy voxels at the predicted primitive centers; ignored voxels are excluded from the grouping loss. This is a semantic field partition, not a separate motion-state label.

3.2 Field-Aware Hierarchical Interaction

HGWM communicates geometry and context before advancing the state. We first standardize primitive centers using training-set center statistics, $\bar{\mu}_i = (\mu_i - m_\mu) \oslash (s_\mu + \epsilon)$. For query i and key j, geometry-aware attention uses

$$e_{ij} = \frac{(h_i W^Q)(h_j W^K)^\top}{\sqrt{D}} - \gamma \|\bar{\mu}_i - \bar{\mu}_j\|_2^2, \qquad \gamma = \mathrm{softplus}(\hat{\gamma}) + \epsilon. \tag{2}$$

The learned positive coefficient prevents distance from becoming a reward and calibrates metric separation against feature similarity. Each stage normalizes its logits over its own key set and adds the value-weighted message to the query through a residual LayerNorm update. A layer applies four directed relations in sequence,

$$B \leftarrow B, \quad F \leftarrow F, \quad F \leftarrow B, \quad B \leftarrow F, \tag{3}$$

separating structural background context, actor interaction, environmental influence on actors, and actor-conditioned background semantics. Here the left-hand field provides queries and receives the update; the right-hand field provides keys and values. Each stage consumes the most recently updated source features, so $F \leftarrow B$ uses the refined background and the final $B \leftarrow F$ uses the refined foreground.

The directed stages exchange field-level context, after which DS-SAGA refines local actor neighborhoods. For each foreground query, it retrieves the $K = 16$ nearest primitives from $B \cup F$ and forms a geometric embedding $\phi_k = \mathrm{MLP}([\Delta x_k, \Delta y_k, \|\Delta\mu_k\|_2, \sin\theta_k, \cos\theta_k])$, where $\theta_k = \mathrm{atan2}(\Delta y_k, \Delta x_k)$. Neighbor weights α_k are obtained by applying Eq. equation 2 within this sparse neighborhood. Separate value projections then aggregate background and foreground contributions:

$$\begin{aligned} h_{\mathrm{bg}} &= \sum_{k=1}^{K} \alpha_k V_{\mathrm{bg}}(h_k, \phi_k)(1 - p_k^{\mathrm{fg}}), \\ h_{\mathrm{fg}} &= \sum_{k=1}^{K} \alpha_k V_{\mathrm{fg}}(h_k, \phi_k) p_k^{\mathrm{fg}}, \\ h_q^{\mathrm{DS}} &= \mathrm{LN}(h_q + \sigma(W_g[h_q, h_{\mathrm{bg}}, h_{\mathrm{fg}}]) \odot W_p[h_{\mathrm{bg}}, h_{\mathrm{fg}}])\,. \end{aligned} \tag{4}$$

The soft field probabilities avoid treating a near-threshold primitive as exclusively foreground or background during aggregation. The gated residual combines both contexts without replacing the query feature. For heterogeneous actor dynamics, a router then activates two of four foreground experts:

$$h_i' = h_i + \sum_{e \in \mathrm{Top2}(w_i)} \frac{w_{i,e}}{\sum_{j \in \mathrm{Top2}(w_i)} w_{i,j}} \mathrm{MLP}_e([h_i; f_{\mathrm{ego}}]), \quad w_i = \mathrm{softmax}(h_i W_r). \tag{5}$$

A load-balancing loss discourages routing collapse. The expert input includes f_{ego}, which encodes ego speed, acceleration, and navigation command. Refined foreground and background features are reunited before the residual transition.

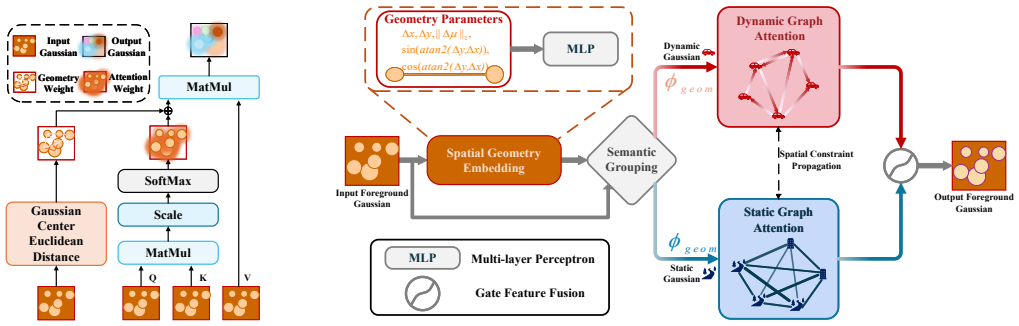


Figure 3: Left: geometry-aware attention calibrates feature similarity with a positive 3D-distance penalty. Right: DS-SAGA separates sparse local background and foreground context before gated foreground refinement.

3.3 Residual Evolution and Shared Readouts

At every predicted step, the current field partition and interaction stack are re-evaluated. A three-layer predictor maps the resulting primitive features to residuals $[\Delta\mu_i, \Delta\ell_i, \Delta q_i, \Delta o_i, \Delta h_i]$. The scale increment $\Delta\ell_i \in \mathbb{R}^3$ is anisotropic, while the rotation increment is four-dimensional. One transition is

$$\begin{aligned} \mu_i^+ &= \mu_i + \Delta t\, \Delta\mu_i, & s_i^+ &= s_i \odot e^{\Delta\ell_i}, \\ q_i^+ &= \mathrm{qnorm}(q_i \otimes \mathrm{qnorm}(\mathbf{e}_0 + \Delta q_i)), & o_i^+ &= o_i + \Delta o_i, \\ h_i^+ &= h_i + \Delta h_i, & z_i^+ &= \phi_{\mathrm{sem}}(h_i^+). \end{aligned} \tag{6}$$

Here $\mathbf{e}_0 = (1,0,0,0)^\top$ is the identity quaternion and qnorm normalizes a nonzero quaternion, returning $\mathbf{e}_0$ for a zero vector. Thus a zero rotation residual preserves the current orientation. The sigmoid of o_i^+ keeps opacity in $[0,1]$, and the semantic head regenerates logits from the evolved feature. Applying this operator autoregressively produces $S_{t+1:t+H}$ without substituting ground-truth future states. The fixed set of 6400 primitives can change appearance and field membership, but cannot create a newly observed actor or delete an instance during the rollout.

At voxel center x, Gaussian density $\rho_i(x) = \alpha_i \exp[-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i)]$ gives

$$\hat{O}(x) = 1 - \exp\left(-\sum_i \rho_i(x)\right), \qquad \hat{P}(c \mid x, \mathrm{occ}) = \frac{\sum_i \rho_i(x)\, \mathrm{softmax}(z_i)_c}{\sum_i \rho_i(x)}. \tag{7}$$

In parallel, an independent ego query reads the full state at every horizon, $\tilde{h}_{\mathrm{ego},t+\tau} = \Psi_{\mathrm{ego}}(\phi_{\mathrm{ego}}(u_t), H_{t+\tau})$, and five heads decode candidate ground-plane waypoints in the current ego frame. This query is separate from the N scene Gaussians, and no unavailable future ego pose is supplied as a decoder input. A probability head selects the highest-probability candidate at inference. Occupancy and every trajectory candidate therefore depend on the same predicted state rather than independent forecast and planning latents.

3.4 Training Objective

Semantic cross-entropy supervises valid occupied voxels, whereas binary BCE supervises occupancy over the full valid volume. The grouping BCE uses the class-derived foreground targets described above. For motion supervision, initial non-ego actor boxes and instance tokens associate foreground primitives with tracks; the loss matches each primitive's predicted displacement to its actor's tracked displacement, masking missing tracks. For planning, the lowest-error of five candidates receives waypoint regression and probability-classification supervision. The drivable-area term samples a ground-plane distance transform at that matched trajectory. The overlap term penalizes anisotropic Gaussian overlap only for pairs of distinct tracked non-ego actors; it is not an ego-collision loss. A stationarity penalty limits background-center drift while leaving feature and opacity updates available. Together with MoE load balancing, these terms give

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{occ}} + \lambda_{\mathrm{plan}}\mathcal{L}_{\mathrm{plan}} + \lambda_{\mathrm{mot}}\mathcal{L}_{\mathrm{mot}} + \lambda_{\mathrm{grp}}\mathcal{L}_{\mathrm{grp}} \\ + \lambda_{\mathrm{drv}}\mathcal{L}_{\mathrm{drv}} + \lambda_{\mathrm{coll}}\mathcal{L}_{\mathrm{coll}} + \lambda_{\mathrm{bg}}\mathcal{L}_{\mathrm{bg}} + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}. \end{aligned} \tag{8}$$

The separate Appendix gives the full loss equations, masks, and weights.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS AND SENSOR PROTOCOLS

We evaluate three benchmark-specific protocols. nuScenes (Caesar et al., 2020) uses the standard 700/150 train/validation split and camera-only inference; semantic occupancy and instance-token/3D-box associations provide training supervision. NAVSIM (Dauner et al., 2024) evaluates non-reactive planning over logged nuPlan scenarios, whereas Bench2Drive (Jia et al., 2024) evaluates 220 closed-loop CARLA routes covering 44 interaction types. The latter two configurations use cameras and LiDAR. Because the protocols differ in sensors, supervision, and interaction, we interpret results within each benchmark rather than as direct cross-benchmark rankings.

4.1.2 METRICS

For 4D occupancy forecasting, we report semantic mean intersection-over-union (mIoU) and binary occupied-space IoU at 1, 2, and 3 s. For nuScenes planning, we follow VAD (Jiang et al., 2023) and report waypoint L2 error and collision rate (CR) at the same horizons. NAVSIM reports No At-Fault Collisions (NC), Drivable Area Compliance (DAC), Time-to-Collision (TTC), Comfort, Ego Progress (EP), and the aggregate PDM Score (PDMS). Bench2Drive reports Driving Score (DS), route Success Rate (SR), and scenario-level capability success rates.

4.2 IMPLEMENTATION DETAILS

On nuScenes, a ResNet-50 processes three aligned six-view frames at 360×640. We train for 24 epochs on eight RTX 4090 GPUs with total batch size 8, AdamW, initial learning rate 5×10^{-5}, 500-step warm-up, and cosine decay to 10^{-6}. Main results and ablations are averaged over seeds $\{0, 1, 2\}$, selecting one checkpoint per seed by the lowest validation $\mathcal{L}$. The complete model has 74.8M parameters and 236.4G FLOPs; single-GPU FP16 inference takes 58.7 ms/frame. Full configurations and evaluation details appear in the separate Appendix.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

4.3.1 NUSCENES PLANNING

Table 1 compares camera-based methods under their reported auxiliary supervision. HGWM obtains 0.28 m average L2 and 0.11% average collision rate. The table makes heterogeneous supervision explicit, and the open-loop metrics should not be read as closed-loop safety guarantees.

Table 1: Open-loop planning on nuScenes validation. Averages are recomputed from displayed horizons; WorldRFT reports only its published averages.

Method	Input	Aux. sup.	L2 (m)↓				CR (%)↓			
			1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD (Hu et al., 2023b)	Cam.	Map/box/mot./occ.	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
CoIRL-AD (Zheng et al., 2025a)	Cam.	None	0.29	0.59	1.00	0.63	0.06	0.10	0.37	0.18
LAW (Li et al., 2025a)	Cam.	None	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
World4Drive (Zheng et al., 2025b)	Cam.	None	0.25	0.52	0.92	0.56	0.08	0.17	0.42	0.22
GaussianAD (Zheng et al., 2024b)	Cam.	Occ./map/box	0.40	0.38	0.61	0.46	0.49	0.38	0.61	0.49
SSR (Li & Cui, 2025)	Cam.	None	0.18	0.36	0.63	0.39	0.07	0.15	0.23	0.15
WorldRFT (Yang et al., 2026)	Cam.	None	0.22	0.44	0.77	0.48	0.00	0.00	0.16	0.05
HGWM	Cam.	Occ./box/mot.	**0.16**	**0.27**	**0.40**	**0.28**	0.06	**0.09**	0.18	**0.11**

4.3.2 4D SEMANTIC OCCUPANCY FORECASTING

On nuScenes, semantic mIoU is 30.11%, 27.78%, and 25.49% at 1, 2, and 3 s; binary IoU is 52.89%, 51.48%, and 49.86% at the same horizons. The decline with horizon reflects accumulated forecast uncertainty. The complete horizon-wise results are tabulated in the separate Appendix. Figure 4 shows that static layout remains recognizable while errors concentrate around moving objects and occluded regions.

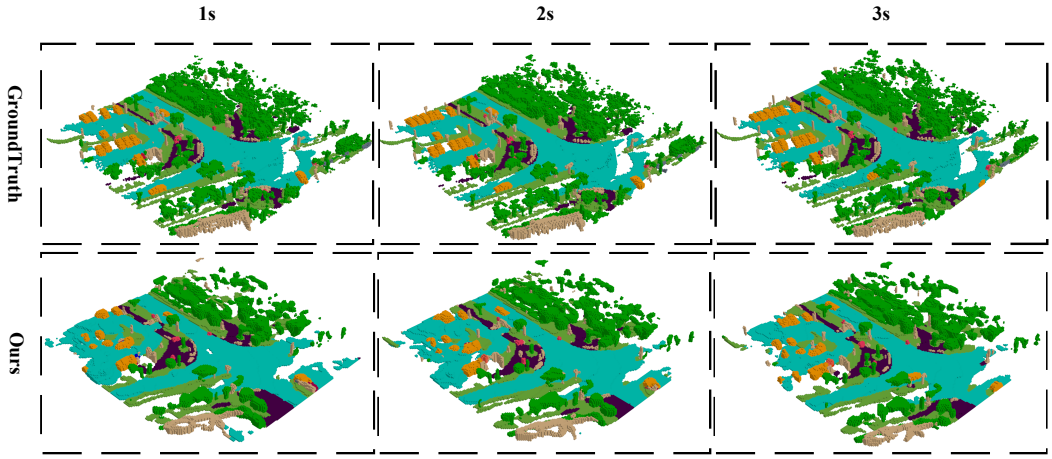


Figure 4: Illustrative 4D occupancy forecasts on nuScenes. Static layout remains recognizable across the horizon, while errors accumulate around moving-object boundaries and occluded regions. The examples complement aggregate evaluation.

4.3.3 NAVSIM AND BENCH2DRIVE PLANNING

Table 2 compares the reported NAVSIM planning results. HGWM attains 92.1 PDMS, 3.3 points above the listed camera–LiDAR GaussianFusion configuration and 0.5 points above SafeDrive. It reaches 99.3 DAC and 97.9 TTC, although SafeDrive has a higher NC score (99.5 versus 99.0). These differences are descriptive: the table mixes camera-only and camera–LiDAR inputs and does not control all training or supervision choices.

Table 2: NAVSIM planning results under the reported configurations. Higher is better for every metric; sensor inputs are listed for context.

Method	Input	NC↑	DAC↑	TTC↑	Comfort↑	EP↑	PDMS↑
LAW (Li et al., 2025a)	Camera	97.2	93.3	91.9	100.0	78.8	83.8
World4Drive (Zheng et al., 2025b)	Camera	97.4	94.3	92.8	100.0	79.9	85.1
WorldRFT (Yang et al., 2026)	Camera	97.8	96.8	94.0	100.0	81.7	87.8
CoIRL-AD (Zheng et al., 2025a)	Camera+LiDAR	98.6	96.8	95.5	100.0	81.0	88.2
WoTE (Li et al., 2025b)	Camera+LiDAR	98.5	96.8	94.9	100.0	81.9	88.3
GaussianFusion (Liu et al., 2025)	Camera+LiDAR	98.3	97.2	94.6	100.0	83.0	88.8
SafeDrive (Kim et al., 2026)	Camera+LiDAR	99.5	99.0	97.2	100.0	84.3	91.6
HGWM	Camera+LiDAR	99.0	99.3	97.9	100.0	86.0	92.1

On Bench2Drive, HGWM obtains an 82.6 Driving Score and a 61.8% route Success Rate. These closed-loop results use camera–LiDAR input, whereas the nuScenes configuration is camera-only; the separate Appendix reports the Bench2Drive baseline and capability tables.

4.4 Ablation Study

The cumulative component study (separate Appendix) reduces nuScenes mean L2 from 0.45 to 0.28 m and CR from 0.24% to 0.11%. Adding grouping and hierarchical interaction, DS-SAGA, sparse experts, and physical objectives successively improves the displayed configuration, but these conditional increments are not isolated causal effects. Controlled variants show that combining grouping with hierarchical interaction outperforms either alone; the four directed interaction stages provide complementary gains. The selected foreground threshold 0.5 and neighborhood size $K = 16$ give the best displayed results among their tested values. The separate Appendix reports the complete tables, interaction-depth and routing sensitivity, objective ablations, and rollout diagnostics.

5 Conclusion and Limitations

HGWM evolves a shared semantic-Gaussian state for 4D occupancy forecasting and ego planning. On nuScenes, it achieves 25.49% semantic mIoU and 49.86% binary IoU at 3 s, with 0.28 m average planning L2 error and 0.11% collision rate. It also achieves 92.1 PDMS on NAVSIM and an 82.6 Driving Score with 61.8% Success Rate on Bench2Drive. Ablations support the contributions of field-aware interaction, sparse refinement, expert routing, and physical regularization.

The fixed set of 6400 Gaussians cannot explicitly create or remove actors. Sensor and supervision protocols differ across benchmarks, so their results are not directly comparable. The current ablations do not isolate the shared state's causal effect on planning, and the reported latency was measured on an RTX 4090 rather than automotive hardware.

AI use statement

The authors did not use generative AI tools in the preparation of this manuscript or in any part of the research.

References

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.

Kashyap Chitta, Daniel Dauner, and Andreas Geiger. SLEDGE: Synthesizing driving environments with generative models and rule-based traffic. In *European Conference on Computer Vision*, pp. 57–74. Springer, 2024.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

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*, 2023a.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pp. 533–549. Springer, 2022.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023b.

Yihan Hu, Siqi Chai, Zhening Yang, Jingyu Qian, Kun Li, Wenxin Shao, Haichao Zhang, Wei Xu, and Qiang Liu. Solving motion planning tasks with a scalable generative model. In *European Conference on Computer Vision*, pp. 386–404. Springer, 2024.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as gaussians for vision-based 3D semantic occupancy prediction. In *European Conference on Computer Vision*, pp. 376–393. Springer, 2024.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic gaussian superposition for efficient 3D occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27477–27486, 2025.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0025.

Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified transformer for scalable end-to-end autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Jungho Kim, Jiyong Oh, Seunghoon Yu, Hongjae Shin, Donghyuk Kwak, and Jun Won Choi. SafeDrive: Fine-grained safety reasoning for end-to-end driving in a sparse world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24854–24864, 2026.

Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via BEV world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27137–27146, 2025b.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12037–12047, 2025.

Shuai Liu, Quanmin Liang, Zefeng Li, Boyang Li, and Kai Huang. GaussianFusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1603.

Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8795–8801. IEEE, 2025.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

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.

Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. RenderWorld: World model with self-supervised 3D label. *arXiv preprint arXiv:2409.11356*, 2024.

Pengxuan Yang, Ben Lu, Zhongpu Xia, Chao Han, Yinfeng Gao, Teng Zhang, Kun Zhan, Xianpeng Lang, Yupeng Zheng, and Qichao Zhang. WorldRFT: Latent world model planning with reinforcement fine-tuning for autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11649–11657, 2026. doi: 10.1609/aaai.v40i14.38149.

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, 2024a.

Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. GaussianAD: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024b.

Xiaoji Zheng, Ziyuan Yang, Yanhao Chen, Yuhang Peng, Yuanrong Tang, Gengyuan Liu, Bokui Chen, and Jiangtao Gong. CoIRL-AD: Collaborative-competitive imitation-reinforcement learning in latent world models for autonomous driving. *arXiv preprint arXiv:2510.12560*, 2025a.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28632–28642, 2025b.

Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. GaussianWorld: Gaussian world model for streaming 3D occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6772–6781, 2025.