OCCGSGEN: PERSISTENT GAUSSIAN WORLD GENERATION THROUGH OCCUPANCY-CONSTRAINED STATE UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable driving scene generation can provide diverse visual observations for training and evaluating autonomous driving systems. A central challenge is to generate newly revealed regions while maintaining consistent geometry and appearance across cameras and over time. We present OccGSGen, a framework for generating persistent Gaussian worlds from semantic occupancy (OCC), camera poses, and tracked object motion. The framework builds a shared 3D scene by integrating appearance from a video diffusion model through incremental updates constrained by occupancy, visibility, and surface correspondence. A persistent appearance atlas separates surface appearance from rendering primitives, allowing additional Gaussians to extend coverage while reusing the appearance of known surfaces. We supervise newly stored appearance through later observations of the same surfaces. Jointly trained on three forward-facing cameras, the model is evaluated on nuScenes through video quality, revisit consistency, and the growth of the Gaussian world. With an optional video completion module, our generated videos achieve FID 8.33 and FVD 70.42 on the official nuScenes validation split.

1 INTRODUCTION

Controllable scene generation offers a way to create visual data for autonomous driving under specified road layouts, actor motions, and camera trajectories. Semantic occupancy (OCC) describes where scene elements are located, while tracked motion specifies how dynamic objects move; their visual appearance remains to be generated. The same structural conditions can correspond to multiple plausible appearances, but a particular generated scene should retain its appearance across cameras and time. As the vehicle advances, the scene must incorporate newly revealed regions and remain available when earlier regions are observed again. For simulation and evaluation, this calls for a representation that supports both flexible camera queries and continued construction of a shared world.

Recent diffusion models have made substantial progress in generating driving observations from structural controls (Gao et al., 2024; 2025; Li et al., 2025). Spatial approaches, including GEN3C and WorldSplat, connect generation to a 3D cache or a Gaussian scene, making generated content available to subsequent camera queries (Ren et al., 2025; Zhu et al., 2026). Models with memory further improve retrieval of past observations and continuation along camera trajectories (Wu et al., 2026; Yu et al., 2026). For a world that continues to grow, new predictions must also be integrated with what is already represented. Later views contain both newly exposed regions and surfaces generated earlier; their updates must distinguish what needs to be created from what should be retained.

The difficulty is that a change in rendering support does not necessarily correspond to new scene content. A roadside facade, for example, may already have an assigned appearance yet require additional Gaussians to cover it from a new viewpoint. Other regions are previously unseen and require both geometry and appearance. Keeping the existing representation fixed leaves missing support unresolved, while independently assigning appearance to every added primitive can give the same surface inconsistent appearances. Extending a scene therefore requires separate decisions

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.

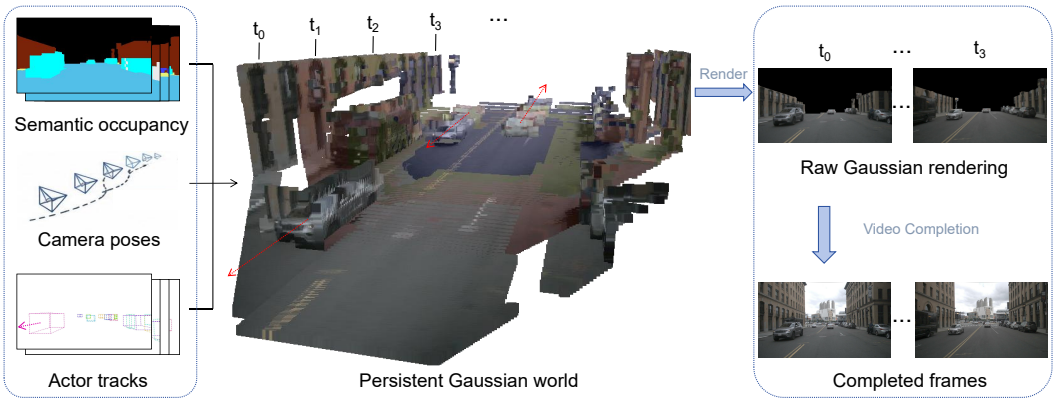


Figure 1: **Occupancy-conditioned Gaussian world generation.** Left: semantic occupancy, camera poses, and actor tracks provide the structural inputs. Center: the world is constructed along the illustrated sequence, with anchors $t_0, \ldots, t_3$. Right: raw Gaussian renders at the indicated times are paired with completed frames below. The video completion module operates on these renders without writing back to the world.

about where to add rendering support and where to initialize appearance. Persistent appearance can therefore be stored at the surface level rather than per Gaussian.

We present OccGSGen, a framework for generating persistent Gaussian worlds from semantic occupancy, camera poses, and tracked motion, without reference RGB at inference. Our approach separates structural decisions about Gaussian updates from the generation and storage of surface appearance. A shared video diffusion prior proposes appearance, while occupancy, visibility, and correspondence determine how candidate Gaussians are integrated into the existing world. A deterministic composer applies the permitted additions and actor motion to the saved state. Rendered history then conditions subsequent generation, so initialization and continuation use the same world representation and prediction modules.

To retain appearance as this representation grows, we maintain an appearance atlas that serves as surface memory, using world coordinates for static surfaces and object-local coordinates for tracked actors. We interpolate stored appearance within local surface patches for newly accepted Gaussians, allowing multiple Gaussians to share a surface record. Accepted primitives initialize appearance only where no compatible record exists. For a moving actor, this keeps stored appearance attached to the object as its position and rendering support change. Each accepted Gaussian stores its assigned appearance and can be rendered directly from the saved state. This separates the lifetime of a surface's appearance from that of the primitives covering it. Figure 1 illustrates the resulting progression from structural controls to an evolving scene and its observations.

Persistence also changes what an appearance prediction should be trained to explain. An assignment that fits its initial view may be reused under different visibility and coverage after subsequent world updates. We introduce *future read supervision* to train initial predictions through this later use. Later observations of the same surface supervise the initial predictions through the same memory lookup used during generation. Later observations therefore supervise the initial appearance values retrieved from memory, while inherited memory and structural decisions remain fixed.

We evaluate OccGSGen on nuScenes at the original and novel camera views, and examine the consistency of revisited scene details. Our ablation study shows that appearance reuse improves revisit consistency, future read supervision improves appearance quality, and the Gaussian acceptance rule controls the growth of the Gaussian representation. An optional video completion module fills missing regions in raw renders, including those outside the finite occupancy volume, without modifying the saved world or the history used for continuation. On the official nuScenes validation split, our completed outputs achieve a multi-camera FID of 8.33 and FVD of 70.42. We assess their perceptual quality separately from the coverage and persistence of the underlying Gaussian world.

2 Related Work

Driving world models. Driving world models synthesize observations under specified scene layouts, actor motions, and camera trajectories (Gao et al., 2024; Wang et al., 2024; Wen et al., 2024; Gao et al., 2025). Occupancy-conditioned methods provide denser spatial guidance through semantic volumes (Li et al., 2024; Lu et al., 2024; Li et al., 2025). Related video world models support continuation through retrieval of past observations, spatial caches, and memory conditioned on camera poses (Xiao et al., 2025; Wu et al., 2026; Yu et al., 2026). Closing the Loop further uses geometric correspondence to improve revisit consistency (Ma et al., 2026). Self Forcing trains on sampled histories, while Self Gradient Forcing uses future losses to supervise historical representations (Huang et al., 2025; Zhuang et al., 2026). Our work connects these two sources of evidence in an explicit Gaussian world: rendered history conditions generation, and subsequent observations supervise the appearance stored for reuse.

3D scene reconstruction and generation. Neural radiance fields and Gaussian splatting are widely used for novel-view synthesis (Mildenhall et al., 2020; Kerbl et al., 2023), with methods for driving scenes and feed-forward reconstruction extending these representations to dynamic settings (Yan et al., 2024; Chen et al., 2026a). Generative approaches combine scene representations with visual priors: MagicDrive3D and InfiniCube construct scenes from generated observations, whereas WorldSplat decodes diffusion latents into 4D Gaussians (Gao et al., 2026; Lu et al., 2025; Zhu et al., 2026). GaussianDWM++ further connects Gaussian features to language-grounded understanding, editing, and generation (Deng et al., 2026). GEN3C, LSM-World, and PERSIST maintain spatial state for continued generation (Ren et al., 2025; Wang et al., 2026; Garcin et al., 2026). Separating geometry from appearance also has precedents in Neural Textures and Texture-GS (Thies et al., 2019; Xu et al., 2024). We focus on incremental world construction, where structural evidence governs geometric updates, surface records support appearance reuse, and later reads supervise initial assignments.

3 Method

We formulate scene generation as updates to a persistent Gaussian world. Our central design separates the growth of rendering support from the assignment of surface appearance. Structural evidence determines which Gaussians can enter the world; an appearance atlas determines whether their appearance is reused or initialized. We then supervise initial appearance predictions through later observations of the same surfaces. Figure 2 shows how these components form one generation and continuation process.

3.1 World Representation and Generation

Our inputs are semantic occupancy, calibrated cameras, and tracked actor motion. Reference RGB from the three forward-facing cameras supplies joint training supervision and is not required at inference. At an occupancy timestamp t, the saved world is $W_t = (\mathcal{G}_t, \mathcal{A}_t)$. The Gaussians $\mathcal{G}_t$ store geometry, opacity, color, semantics, and actor identity. The appearance atlas $\mathcal{A}_t$ serves as persistent surface memory, storing appearance independently of the primitives that render it. Generation and rendering take the form

$$W_t = \mathrm{Composer}(W_{t-1}, U_t), \qquad \widehat{I}_t^v = \mathcal{R}(\mathcal{G}_t, P_t^v). \tag{1}$$

Here U_t comprises Gaussian proposals and update permissions; P_t^v is the query camera. The deterministic composer processes the front, front-left, and front-right views in order. Each reads the current state $W_{\mathrm{in}} = (\mathcal{G}_{\mathrm{in}}, \mathcal{A}_{\mathrm{in}})$, including preceding views' updates; W_t denotes the shared state after all three updates.

We group the supplied controls into blocks $\mathcal{C}_b$. History H_b contains RGB, alpha, and depth rendered from the previous boundary world and is empty at initialization. For timestamp t in block b, a video prior and an appearance lifting network produce

$$\widetilde{I}_b \sim p_\theta(\cdot \mid \mathcal{C}_b, H_b), \qquad \widetilde{\mathbf{c}}_{tj} = f_\phi(\widetilde{I}_b, \mathcal{S}_t, W_{\mathrm{in}}; j). \tag{2}$$

Here $\mathcal{S}_t$ is the scaffold defined by occupancy rays and j indexes candidates. The appearance lifting network f_ϕ predicts candidate Gaussian colors from sampled video appearance, occupancy features,

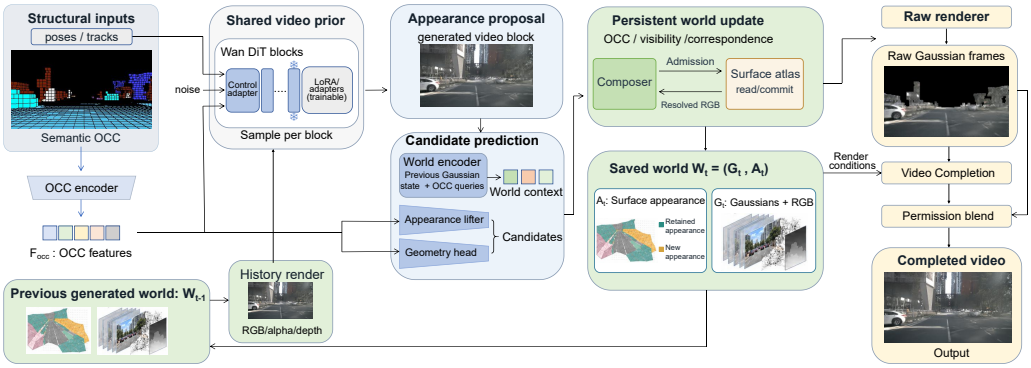


Figure 2: **Overview of our framework.** OCC features and rendered history condition the shared video prior. Generated appearance, structural features, and context from the previous world guide candidate prediction. The appearance atlas supplies stored colors, and the composer accepts candidates and stores new appearance records before saving $W_t = (\mathcal{G}_t, \mathcal{A}_t)$. On the right, video completion combines video diffusion with mask-based blending. The bottom path carries RGB, alpha, and depth rendered from the saved world to condition continuation. Completed video is used only as an output.

and the current world state. A separate geometry head predicts bounded offsets from occupancy rays, Gaussian shape, opacity, and confidence. We use a frozen Wan2.2-TI2V-5B backbone (Team Wan et al., 2025) with LoRA (Hu et al., 2022) and structural adapters. The prior samples a block jointly from its supplied controls; the composer then applies updates in timestamp order. Initialization and continuation use the same representation and prediction modules. Appendix A specifies the block schedule and model details.

3.2 Occupancy-Constrained World Updates

Occupancy constrains where candidates may be placed, while correspondence and visibility determine whether they add useful support. We decide whether to accept each Gaussian independently of its proposed color. Correspondence requires agreement in position and semantics, together with a matching valid track identity for actors. Rendered depth and alpha describe current view coverage; projected Gaussian footprints additionally measure the support already supplied by existing primitives. A static candidate is accepted only when both coverage and projected support are insufficient. Neither low alpha nor a depth discrepancy alone is sufficient to add new Gaussians.

This distinction matters for known surfaces: a candidate may correspond to an existing surface yet fill a gap in its rendered coverage. Rejecting every such candidate prevents coverage from improving, whereas accepting candidates solely from low alpha can accumulate overlapping support. Our rule permits additions according to structural evidence and leaves their appearance to the surface lookup below. The composer retains static Gaussians, transports surviving actors, and removes eligible actor primitives. Transport changes centers and covariances while retaining color; removing geometry does not erase appearance associated with a continuing track. Appendix A gives the geometric tests, capacity handling, and motion rules.

3.3 Persistent Surface Appearance

A surface can require different rendering primitives as viewpoints and visibility change. We therefore attach persistent appearance to local surface patches. Static queries use world coordinates. For actor k, the canonical query is $q = T_{k,t}^{-1}\mathbf{x}$, where $T_{k,t}$ is its supplied object-to-world transform. Queries also retain semantic and track identity; actors without a valid track and transform are excluded from persistent storage. Correct attachment requires geometry transport to agree with the same object frame.

Compatible neighboring occupancy rays form triangular patches, with point patches for isolated samples. A query selects the earliest-created compatible patch and reads its appearance by interpolation:

$$r_{\mathcal{A}}(q) = \sum_v \beta_v(q)\mathbf{c}_v, \qquad \beta_v(q) \geq 0, \quad \sum_v \beta_v(q) = 1. \tag{3}$$

The sum runs over the selected patch's vertices; $\mathbf{c}_v$ are stored colors and β_v are barycentric weights. Nearby primitives can therefore share a continuous appearance within a patch, with each channel bounded by its stored vertex values. For a candidate eligible for persistent storage, we assign

$$\mathbf{c}_{tj} = \begin{cases} r_{\mathcal{A}_{\text{in}}}(q_{tj}), & \text{if a compatible patch exists,} \\ \widetilde{\mathbf{c}}_{tj}, & \text{otherwise.} \end{cases} \tag{4}$$

This lookup precedes composition. Only final acceptance by the composer permits new atlas values to be stored, and only where compatible memory is absent. Rejected candidates write nothing; inherited patch values remain unchanged. Newly sampled appearance is stored only at accepted locations without a compatible memory entry. A missing read can reflect incomplete patch coverage or incorrect correspondence as well as an unseen surface. Each accepted Gaussian stores its assigned color, so subsequent rendering needs neither an atlas lookup nor a generative pass.

Preservation under extension. Consider a query q with a valid atlas lookup and fixed canonical coordinates, semantic identity, and track identity. Atlas updates append patches without changing existing values or their lookup priority. Consequently, across successful updates, $r_{\mathcal{A}_s}(q) = r_{\mathcal{A}_t}(q)$ for $s > t$ under unchanged lookup rules. The earlier compatible patch remains selected, and its vertices, colors, and interpolation weights are unchanged. Additional Gaussians can therefore improve support for a represented surface without introducing independent appearance assignments there. This is a property of the stored appearance; rendered pixels can still change with visibility and blending. The atlas stores generated, view-independent appearance, including residual lighting, rather than intrinsic reflectance.

3.4 Learning from Future Observations

Preserving an assignment also preserves its errors. An appearance prediction should therefore be evaluated where it will be reused, beyond the view in which it was created. Our *future read supervision* makes this later use part of training. From a generated training block, we select a time t at which appearance is stored and a later observation time s. We call t the write time and s the observation time. At s, we query the updated atlas and select visible reads from patches created at t that match semantic and track identities and assign positive interpolation weight to values first predicted at t. Copied boundary values are not treated as new predictions, and later reference RGB is used solely as the loss target.

Let $\widehat{X}_s$ be the image canvas at time s with the interpolated atlas colors inserted and X_s the reference observation. On the eligible support, the objective is

$$\mathcal{L}_{\text{future}} = d_{\text{RGB}}(\widehat{X}_s, X_s) + d_{\text{feat}}(\widehat{X}_s, X_s). \tag{5}$$

Both terms compare distributions separately for each sample and semantic class using projected quantiles (Deshpande et al., 2018); the feature term uses frozen DINOv2 features (Oquab et al., 2024). This matches the training setting: a sampled world need not reproduce the recorded scene's exact pixel colors. Surface correspondence determines which initial assignments receive supervision, while the loss compares their appearance distributions.

For eligible reads, we recover the selected initial colors from their source predictions, with values exactly equal to those stored in the atlas. Training differentiates the same interpolation used during generation, including its color bounds, through these selected colors only. All other colors and canvas pixels are detached, as are interpolation weights, geometry, patch selection, and intervening updates. The sampled video proposals are also detached, so this objective trains the appearance lifting network through its subsequent memory reads. It is a partial derivative conditional on the generated structural history; the full rollout is not differentiated. Appendix A specifies the distance and sampling details.

We combine training on paired data with training on sampled world histories. Paired training uses the video backbone's flow-matching objective and appearance, geometry, and rendering supervision;

Table 1: **Comparison of original-view video generation.** Baseline values are taken from WorldSplat (Zhu et al., 2026), Table 1, without guidance from a real first frame. MagicDrive-V2 reports FID/FVD for separate image/video tasks, and MagicDrive3D reports video-generation scores. NVS indicates whether a method supports novel-view synthesis. Methods retain their native camera configurations; scores are averaged across cameras. (*Pretrained weights are not publicly available; scores are taken from the paper.)

Method	Video backbone	NVS	$\text{FVD}_{\text{multi}} \downarrow$	$\text{FID}_{\text{multi}} \downarrow$
DriveDreamer-2 (Zhao et al., 2025)	SVD	–	105.10	25.00
MagicDrive-V2 (Gao et al., 2025)	MVDiT	–	94.84	20.91
MagicDrive3D (Gao et al., 2026)	SD-v1.5	✓	164.72	20.67
Panacea (Wen et al., 2024)	SD-2.1	–	139.00	16.96
DiVE (Jiang et al., 2024)	OpenSora-v1.1	–	94.60	–
WorldSplat* (Zhu et al., 2026)	OpenSora-v1.2	✓	<u>74.13</u>	<u>8.78</u>
Ours (completed)	**Wan2.2**	✓	**70.42**	**8.33**

Wan2.2 supplies the world appearance prior; a separate video completion module produces the final videos. – indicates that novel-view rendering is unavailable or that a metric was not reported.

boundary states used for teacher supervision are constructed from reference RGB at the first anchor only. In sampled training, initial-view appearance matching and consistency with stored appearance complement future read supervision. A term for static surfaces across views trains the scales of accepted Gaussians to improve coverage while leaving Gaussian acceptance decisions fixed. The loss weights and gradient paths for observation losses are specified in Appendix A.

3.5 Rendering and Video Completion

The saved Gaussians render directly, and their boundary renders provide the history for the next generation block. Sky and structures outside the finite occupancy volume can remain unsupported. Our optional video completion module fills unsupported regions in videos rendered from the Gaussians and applies bounded refinement in supported regions, following prior approaches that enhance rendered videos (Zhu et al., 2026). It is conditioned on raw renders, alpha, occupancy, motion, cameras, and generated style. Pixel masks define where and how the module can modify the renders: they preserve protected boundaries, limit changes within supported regions, and allow completion in eligible unsupported regions. Training targets are appearance proposals from the same world rollout. Completed images do not modify the Gaussians, the atlas, or the rendered history used for continuation. We therefore evaluate observation quality separately from coverage and persistence of the saved world. Appendix A gives the video completion architecture and compositing rule.

4 Experiments

We compare original-view and novel-view video quality and examine the effect of video completion. Our ablation study evaluates Gaussian acceptance, appearance reuse, and future read supervision using raw Gaussian renders.

4.1 Experimental Setup

Dataset and generation settings. We use nuScenes and Occ3D occupancy (Caesar et al., 2020; Tian et al., 2023), calibrated cameras, and annotated tracks. We jointly train on the three forward-facing cameras (CAM_FRONT_LEFT, CAM_FRONT, and CAM_FRONT_RIGHT), which render one shared Gaussian world. Each block contains 21 observations per camera at 448×800, aligned with four occupancy anchors. Continuation reuses the boundary anchor and adds three anchors; reference RGB is not used at inference. Evaluation uses all 150 scenes in the official nuScenes validation split, following the benchmark setting in WorldSplat (Zhu et al., 2026). Appendix A gives the implementation details.

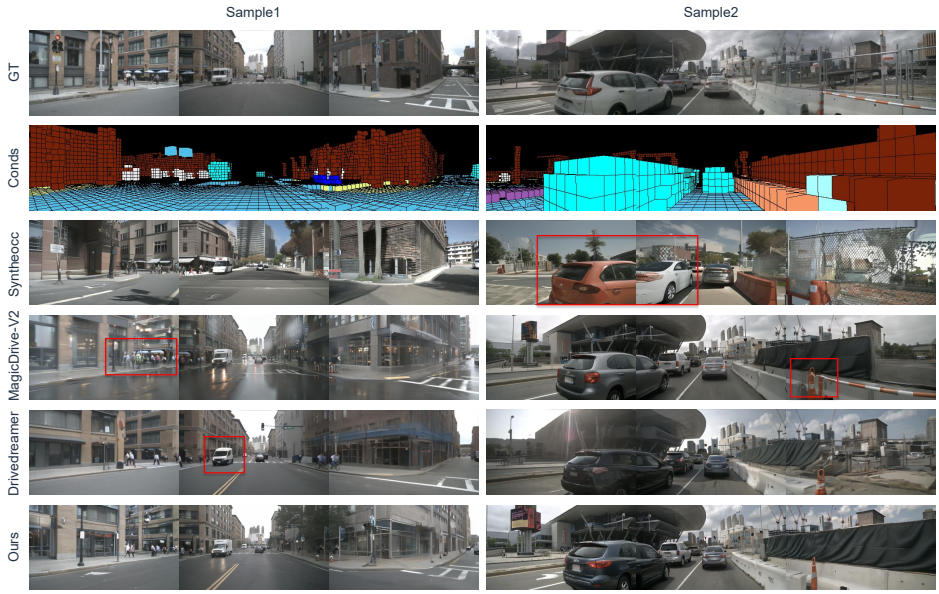


Figure 3: **Qualitative comparison of multi-view video generation.** Each scene groups front-left, front, and front-right observations at one time. Rows are GT, structural conditions, SyntheOcc, MagicDrive-V2, DriveDreamer, and ours (completed). Red boxes identify regions for local comparison. GT is shown as a reference and is not an inference input to our model. Baseline outputs are reproduced by us using each method's native conditioning inputs.

Evaluation metrics. Rendering coverage measures the fraction of queries with sufficient rendered alpha and depth agreement with occupancy. Depth error includes conflicts at queries with sufficient alpha. Cross-view RGB error (Cross RGB) and revisit RGB error use fixed correspondences defined by occupancy within the same generated world. We report FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) averaged across cameras. FID and FVD for raw renders are evaluated within OCC regions, whereas those for completed videos use full frames. Scores from these two evaluation regions are not directly comparable; holes within the fixed OCC region are retained during raw evaluation. Metric aggregation and evaluation settings are described in Appendix B.

4.2 Video Generation and Novel-View Synthesis

Original-view video generation. Table 1 compares our generated videos with the published baseline results on nuScenes. Our completed output obtains FVD 70.42 and FID 8.33, compared with the published WorldSplat scores of 74.13 and 8.78. The methods use different priors and structural inputs; this comparison evaluates their final outputs. Section 4.3 presents an ablation study of the world update mechanism.

Novel-view synthesis. Table 2 compares outputs under lateral camera shifts of ± 1, ± 2, and ± 4 m. PVG, EmerNeRF, StreetGaussian, and OmniRe reconstruct scenes from captured RGB; FreeVS* uses colored LiDAR; DiST-S (+SCC) uses video and depth (Chen et al., 2026b; Wang et al., 2025; Guo et al., 2025). These baselines retain their native inputs and published protocols. Our raw renders and completed videos are derived from the same saved world at the same camera views. Because ground-truth images are unavailable at shifted cameras, FID and FVD assess image and video distributions rather than paired reconstruction accuracy. Table 2 reports full-frame scores for our completed outputs; results for raw renders within OCC regions are reported separately in Table 4.

Qualitative comparison. Figure 3 compares our reproduced baseline outputs with our completed videos in three camera views for each of two scenes. The reference and structural conditions show the scene layout and object placement, while the highlighted regions allow local comparison of generated details.

Table 2: **Comparison of novel-view synthesis.** Baseline scores at each camera shift are taken from WorldSplat (Zhu et al., 2026), Table 2, which follows the comparison in DiST-4D (Guo et al., 2025). Methods retain their native inputs and camera configurations; our completed outputs use full frames. FreeVS* denotes the nuScenes retraining reported by DiST-4D.

		±1 m		±2 m		±4 m	
Method	Output	FID↓	FVD↓	FID↓	FVD↓	FID↓	FVD↓
PVG (Chen et al., 2026b)	Native	48.15	246.74	60.44	356.23	84.50	501.16
EmerNeRF (Yang et al., 2024)	Native	37.57	171.47	52.03	294.55	76.11	497.85
StreetGaussian (Yan et al., 2024)	Native	32.12	153.45	43.24	256.91	67.44	429.98
OmniRe (Chen et al., 2025)	Native	31.48	152.01	43.31	254.52	67.36	428.20
FreeVS* (Wang et al., 2025)	Native	51.26	431.99	62.04	497.37	77.14	556.14
DiST-S (+SCC) (Guo et al., 2025)	Native	10.12	45.14	12.97	68.80	17.57	105.29
WorldSplat (Zhu et al., 2026)	Enhanced	<u>8.25</u>	<u>40.17</u>	<u>11.26</u>	<u>47.41</u>	**13.38**	<u>64.07</u>
Ours	**Completed**	**7.75**	**39.32**	**10.03**	**44.19**	<u>13.46</u>	**62.31**

Table 3: **Ablation study.** FID/FVD use raw renders within OCC regions at ±2 m; revisit RGB error uses colors in $[0, 1]$. Gaussian counts are relative to the full model; independent appearance uses the same geometry and therefore has a count ratio of 1.00. Best and second-best values are bold and underlined, respectively, including ties.

Variant	FID↓	FVD↓	Revisit RGB↓	Relative Gaussian Count↓
w/o support check	25.38	<u>104.70</u>	<u>0.0027</u>	1.41
Independent appearance	<u>25.06</u>	109.36	0.044	**1.00**
w/o future read loss	28.71	127.62	**0.0021**	<u>1.06</u>
Full model	**24.06**	**98.45**	**0.0021**	**1.00**

4.3 Ablation Study

Table 3 evaluates the contributions of Gaussian acceptance, appearance reuse, and future read supervision after one continuation block. Removing the projected support check tests the growth of rendering primitives under shared candidate proposals. Independent appearance assignment changes only the colors of newly accepted Gaussians while replaying the full model's accepted geometry and opacity. The variant without future read supervision is trained with the same initialization procedure, budget, and remaining losses.

FID and FVD assess raw appearance within OCC regions at ±2 m; the full model uses the same evaluated outputs as Table 4. Revisit RGB measures changes in the rendered appearance of fixed static surfaces relative to an earlier state of the same world. The Gaussian count is normalized per scene to the full model, then averaged. We also evaluate coverage, depth conflicts, and the fraction of surface queries with valid rendered colors in both states (Appendix B); fewer Gaussians alone do not establish better updates.

Removing the support check increases the Gaussian count to $1.41\times$ that of the full model without improving FID or FVD. Assigning appearance independently raises revisit RGB error from 0.0021 to 0.044 despite identical geometry and opacity, supporting the role of appearance reuse in consistency. Without future read supervision, FID and FVD increase to 28.71 and 127.62, while revisit RGB error remains 0.0021. Thus, retaining an appearance assignment and learning a good assignment are distinct requirements: persistence alone does not ensure visual quality.

4.4 Revisit Consistency

Figure 4 shows raw Gaussian renders when the same regions are revisited several frames later. The curb, building, and vehicle details remain consistent across the illustrated world states, before video completion. These results are produced by our model and complement the quantitative revisit evaluation in Table 3.

Table 4: **Raw Gaussian renders and completed videos at ± 2 m.** WorldSplat values retain its published Table 3 protocol. Our two outputs use the same saved world and camera views. †Our raw FID/FVD use OCC regions; completed FID/FVD use full frames. FID/FVD across these evaluation regions are not directly comparable. OCC-IoU and Cross RGB use the same queries within OCC regions for our two outputs. – denotes unreported metrics.

Method	Output	FVD↓	FID↓	OCC-IoU (%)↑	Cross RGB↓
WorldSplat (Zhu et al., 2026)	Raw	107.58	26.73	–	–
WorldSplat (Zhu et al., 2026)	Enhanced	47.41	11.26	–	–
Ours	Raw†	98.45	24.06	**97.1**	**0.035**
Ours	**Completed**	44.19	10.03	95.3	0.050

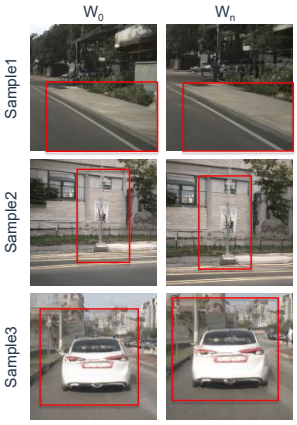


Figure 4: **Revisit consistency.** Raw renders from W_0 and W_n show consistent curb, building, and vehicle details in red boxes, without video completion.

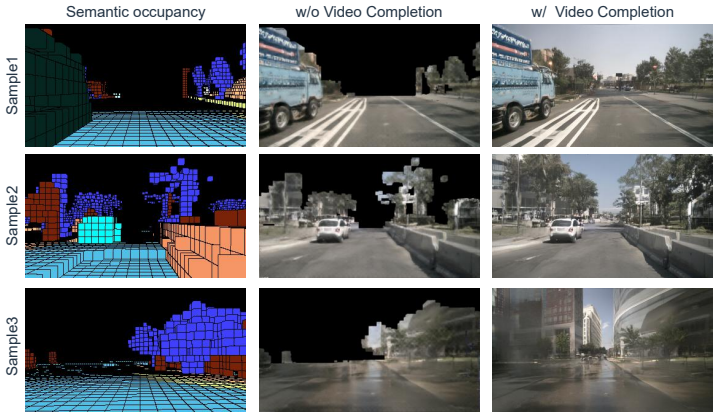


Figure 5: **Effect of video completion.** Columns show semantic occupancy, raw Gaussian renders, and completed outputs. One camera is shown for visualization. The module fills missing sky and background regions and improves visual detail while retaining the depicted road layout and object placement. The saved Gaussian world remains unchanged.

4.5 Effect of Video Completion

Figure 5 compares semantic occupancy with renders without and with video completion. The module fills missing sky and background regions while retaining the depicted road layout and vehicle placement. Table 4 reports quality within OCC regions for raw renders and over full frames for completed videos; their FID/FVD scores are not directly comparable. On the same queries within OCC regions, OCC-IoU decreases from 97.1% to 95.3%, and cross-view RGB error increases from 0.035 to 0.050 after completion. Thus, improved visual completeness comes with some loss of alignment in the output images, even though the saved Gaussian geometry and appearance remain unchanged.

5 Conclusion

OccGSGen generates persistent Gaussian worlds by separating structural updates from the assignment and storage of surface appearance. Structural evidence determines which Gaussians are accepted, the appearance atlas enables reuse, and future observations supervise initial predictions. Experiments on nuScenes assess video quality, while the ablation study shows how these components help control Gaussian growth, preserve revisit consistency, and improve appearance. Video completion fills missing image regions without changing the saved world. The method remains limited by occupancy resolution, correspondence errors, and errors in initial appearance.

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*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.pdf.

Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hongyang Li, Ya-Qin Zhang, Hangjun Ye, and Hao Zhao. DGGT: Feedforward 4D Reconstruction of Dynamic Driving Scenes using Unposed Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_DGGT_Feedforward_4D_Reconstruction_of_Dynamic_Driving_Scenes_using_Unposed_CVPR_2026_paper.html.

Yurui Chen, Chun Gu, Junzhe Jiang, Xiatian Zhu, and Li Zhang. Periodic Vibration Gaussian: Dynamic Urban Scene Reconstruction and Real-time Rendering. *International Journal of Computer Vision*, 2026b. URL https://fudan-zvg.github.io/PVG/.

Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. OmniRe: Omni Urban Scene Reconstruction. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d4938496368e1c5806138bea8837595c-Abstract-Conference.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.html.

Tianchen Deng, Xuefeng Chen, Shuang Wu, Qu Chen, Jiajun Zhu, Bo Dai, Jianfei Yang, and Hesheng Wang. GaussianDWM++: Language-Grounded 3D Gaussian Driving World Model for Unified Scene Understanding, Editing, and Multi-Modal Generation, 2026. URL https://arxiv.org/abs/2608.16234.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative Modeling Using the Sliced Wasserstein Distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Deshpande_Generative_Modeling_Using_CVPR_2018_paper.html.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street View Generation with Diverse 3D Geometry Control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/635351dc3d18116d1a44ca31592e2f9a-Abstract-Conference.html.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive-V2: High-Resolution Long Video Generation for Autonomous Driving with Adaptive Control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Gao_MagicDrive-V2_High-Resolution_Long_Video_Generation_for_Autonomous_Driving_with_Adaptive_ICCV_2025_paper.html.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. MagicDrive3D: Controllable 3D Generation for Any-View Rendering in Street Scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. URL https://arxiv.org/abs/2405.14475.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond Pixel Histories: World Models with Persistent 3D State. In *International Conference on Machine Learning*, 2026. URL https://francelico.github.io/persist.github.io/.

Jiazhe Guo, Yikang Ding, Xiwu Chen, Shuo Chen, Bohan Li, Yingshuang Zou, Xiaoyang Lyu, Feiyang Tan, Xiaojuan Qi, Zhiheng Li, and Hao Zhao. DiST-4D: Disentangled Spatiotemporal Diffusion with Metric Depth for 4D Driving Scene Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Guo_DiST-4D_Disentangled_Spatiotemporal_Diffusion_with_Metric_Depth_for_4D_Driving_ICCV_2025_paper.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, Kun Zhan, Peng Jia, and Miao Zhang. DiVE: DiT-based Video Generation with Enhanced Control. *arXiv preprint arXiv:2409.01595*, 2024. URL https://arxiv.org/abs/2409.01595.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_UniScene_Unified_Occupancy-centric_Driving_Scene_Generation_CVPR_2025_paper.html.

Leheng Li, Weichao Qiu, Yingjie Cai, Xu Yan, Qing Lian, Bingbing Liu, and Ying-Cong Chen. SyntheOcc: Synthesize Geometric-Controlled Street View Images through 3D Semantic MPIs. *arXiv preprint arXiv:2410.00337*, 2024. URL https://arxiv.org/abs/2410.00337.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. WoVoGen: World Volume-aware Diffusion for Controllable Multi-camera Driving Scene Generation. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10410_ECCV_2024_paper.php.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, and Jiahui Huang. InfiniCube: Unbounded and

Controllable Dynamic 3D Driving Scene Generation with World-Guided Video Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lu_InfiniCube_Unbounded_and_Controllable_Dynamic_3D_Driving_Scene_Generation_with_ICCV_2025_paper.html.

Wenchao Ma, Changran Liu, Sharon X. Huang, and Haomiao Jiang. Closing the Loop: Training-Free Revisit Consistency for Autoregressive Generative Rendering, 2026. URL https://arxiv.org/abs/2607.21848.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, 2020. URL https://www.matthewtancik.com/nerf.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/pdf/2304.07193.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred Neural Rendering: Image Synthesis using Neural Textures. *ACM Transactions on Graphics*, 2019. URL https://niessnerlab.org/projects/thies2019neural.html.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/cabfaeecaae7d6540ee797a66f0130b0-Abstract-Datasets_and_Benchmarks.html.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Qitai Wang, Lue Fan, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. FreeVS: Generative View Synthesis on Free Driving Trajectory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/93b4d708976a1d9b1250c400e7fda811-Abstract-Conference.html.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent Spatial Memory for Video World Models. *arXiv preprint arXiv:2606.09828*, 2026. URL https://microsoft.github.io/LatentSpatialMemory/.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards Real-world-driven World Models for Autonomous Driving. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06416.pdf.

Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wen_Panacea_Panoramic_and_Controllable_Video_Generation_for_Autonomous_Driving_CVPR_2024_paper.html.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable Memory for Video World Models, 2026. URL https://arxiv.org/abs/2608.07408.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/470629a47e2d65ce0606c40055df5d26-Abstract-Conference.html.

Tian-Xing Xu, Wenbo Hu, Yu-Kun Lai, Ying Shan, and Song-Hai Zhang. Texture-GS: Disentangling the Geometry and Texture for 3D Gaussian Splatting Editing. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03567.pdf.

Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09243.pdf.

Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, and Yue Wang. EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47fc64d05a394955b1ae2487bfef1ab0-Abstract-Conference.html.

Wangbo Yu, Kunhao Liu, Wenbo Hu, Shenghai Yuan, Chaoran Feng, Haiyang Zhou, Yukun Huang, Yiran Wang, Wang Zhao, Yingmin Luo, and Ying Shan. WorldCrafter: Consistent Video World Model with Implicit 3D-aware Memory, 2026. URL https://arxiv.org/abs/2609.24984.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. DriveDreamer-2: LLM-Enhanced World Models for Diverse Driving Video Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/33130.

Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, and Jian Yang. WorldSplat: Gaussian-Centric Feed-Forward 4D Scene Generation for Autonomous Driving. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6fa4a1def9d60b5bf6078e13c900c45c-Abstract-Conference.html.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self Gradient Forcing: Native Long Video Extrapolation. *arXiv preprint arXiv:2607.20368*, 2026. URL https://arxiv.org/abs/2607.20368.

A Implementation Details

Multi-view inputs and prediction. We jointly train on CAM_FRONT, CAM_FRONT_LEFT, and CAM_FRONT_RIGHT and construct one shared Gaussian world from the three views. Each camera contributes 21 observations at 448×800, aligned with four shared occupancy anchors through camera tokens and timestamps. Block b covers anchors $3b, \ldots, 3b + 3$; continuation retains the boundary anchor and processes three new anchors. Resampled latent controls include channels for presence and elapsed time. We adapt the frozen video backbone with LoRA and structural adapters. Boundary conditions contain rendered RGB, alpha, depth, and presence. Candidate centers may deviate by at most 0.08 m along occupancy rays; centers in ground classes 11–13 remain fixed. The limits on the appearance residual and its low frequency component are 0.08 and 0.01, respectively. Bounded affine alignment precedes replacement by compatible stored appearance.

Training objectives and gradients. For clean latent z_0, independent Gaussian noise ϵ, and $\sigma = \mathrm{clip}(\mathrm{sigmoid}(g), 10^{-5}, 1 - 10^{-5})$, $g \sim \mathcal{N}(0, 1)$,

$$z_\sigma = (1 - \sigma)z_0 + \sigma\epsilon, \qquad \mathcal{L}_{\text{flow}} = \mathbb{E}\|v_\theta(z_\sigma, \sigma, \mathcal{C}_b, H_b) - (\epsilon - z_0)\|_2^2. \tag{6}$$

The rectified flow objective (Liu et al., 2023; Team Wan et al., 2025) is augmented as $\mathcal{L}_{\text{native}} = \mathcal{L}_{\text{flow}} + 0.25\mathcal{L}_{\text{sem}} + 0.25\mathcal{L}^r_{\text{RGB}} + 0.125\mathcal{L}^r_{\text{edge}}$. Here $\mathcal{L}_{\text{sem}}$ balances semantic classes, and the superscript r denotes weighting by $(1 - \sigma)^2$. Auxiliary losses on paired data supervise appearance, geometry, rendering, and transitions. Boundary states used for teacher supervision are constructed from reference RGB at the first anchor only. Sampled training uses the current model weights, eight sampling steps, a timestep shift of 5, and a guidance scale of 1. Monitoring uses 20 steps and exponential moving average (EMA) weights when available. The implementation provides two observation loss variants: the distributional variant detaches sampling and supervises stored appearance within occupancy support, whereas the paired variant retains gradients through the final two denoising steps to train the adapters. The appearance identity and future read losses detach generated proposals. Gradients from future reads reach the selected initial colors and the appearance lifting network, while geometry, correspondence, lookup decisions, inherited appearance, and intervening states remain fixed. The feature and RGB terms of the observation loss have weights 1 and 0.25. The appearance identity and future read losses each have weight 0.5. For static scale supervision, the coverage, depth, and source RGB terms have weights 0.1, 0.05, and 0.05, respectively.

Surface lookup and actor motion. Compatible adjacent rays form triangles with edges ≤ 0.8 m and cross-product norm $> 10^{-8}$ m^2. The plane and barycentric tolerances are 0.015 m and 10^{-5}, respectively; the squared distance tolerance for point patches is 10^{-10} m^2. A spatial index with cells of width 0.10 m retrieves the earliest-created compatible patch. Inference and training use the same anchored FP32 interpolation, normalized weights, and clamping to the vertex color range. Annotated actor poses are applied once per anchor across the three views; actors without a new annotation retain their last poses, and inconsistent visible evidence is rejected. If the atlas reaches capacity, continuation terminates with an explicit failure; stored appearance is not evicted.

Future read supervision and cross-view supervision. For a training block with N occupancy anchors, write times are sampled uniformly from $t \in \{0, \ldots, N - 1\}$. For $t < N - 1$, observation times are sampled uniformly from $s \in \{t + 1, \ldots, N - 1\}$. The final anchor receives supervision from its initial view only. If no eligible future reads exist, the future loss is zero and the frequency of this case is recorded. The two distances in Eq. equation 5 use the projected quantile distance

$$d_{\mathrm{Q}}(X, Y) = \frac{1}{LK} \sum_{\ell=1}^{L} \sum_{k=0}^{K-1} \left| Q_{u_k}(\nu_\ell^\top X) - Q_{u_k}(\nu_\ell^\top Y) \right|, \tag{7}$$

where Q_u is an empirical quantile; generated and reference samples share projection directions ν_ℓ and levels u_k. We use $L = 32$, $K = 64$, $u_k = k/63$, and CPU seed $20260913 + d$ for dimension d. The RGB and feature terms are averaged over nonempty groups defined by sample and semantic class within the eligible support. The feature term is also averaged over the selected frozen layers. DINOv2 ViT-B/14-reg (Oquab et al., 2024; Darcet et al., 2024) supplies blocks 2, 5, 8, 11, projected to 256 channels on a 25×45 grid. Features are normalized across channels, while RGB uses the backbone's normalization. The loss compares distributions within each group rather than paired

pixel colors. Static scale supervision uses representative occupancy rays and detached neighbors to evaluate coverage and depth in later views, together with RGB in the source view. Any fallback to queries at earlier times is recorded. This loss updates only Gaussian scales; candidate acceptance decisions remain fixed during this supervision.

Video completion. The module in Section 3.5 uses an independent frozen STDiT2-XL/2 backbone; the implementation defaults to control adapters of rank 64 and 16 sampling steps. Its training target is the Wan appearance proposal from the same generation run, controls, and seed as the raw world; this target is not provided at inference. The conditions comprise raw RGB, alpha, depth from occupancy, semantics, motion, aligned camera poses, style, and pixel masks for the three views. Losses on velocity, temporal differences, the predicted RGB endpoint, spatial detail, and residual energy have weights 1, 0.25, 0.50, 0.50, and 0.002, respectively. The enhancement, repair, and completion regions have weights 1.50, 1.25, and 1.00; protected pixels receive no direct supervision. Four disjoint masks define regions for boundary protection, enhancement of supported interiors, repair of weak interiors and seams, and completion of unsupported regions. The low and high alpha thresholds are 0.08 and 0.70. Protected bands around physical and dynamic contours span one pixel, while seams around holes span two pixels. Repair applies to pixels supported by occupancy whose alpha is below the high threshold, together with adjacent seams. Completion is allowed outside the protected bands where both occupancy support and sufficient alpha are absent. For RGB in $[0, 1]$, the high and low frequency residual limits of the proposal are 0.25 and 0.15. The compositor uses corresponding limits of 0.20 and 0.06 for enhancement, and 0.35 and 0.12 for repair, with a 9×9 spatial low-pass filter. Protection is applied before filtering as well as during final compositing, so protected pixels retain raw RGB exactly. Completed RGB never updates the Gaussian world, atlas, or recurrent history.

B Evaluation and Comparisons

Dataset and baseline comparisons. All evaluations use the complete official nuScenes validation split of 150 scenes. Our model jointly uses the three forward-facing cameras for training and world construction. Published baseline scores retain their original inputs, camera configurations, and protocols. The original-view comparison uses the benchmark setting and published baseline scores in WorldSplat (Zhu et al., 2026), Table 1. WorldSplat uses layout and box controls, whereas OccGSGen uses fine-grained occupancy and annotated tracks; the comparison therefore concerns final systems, not isolated memory mechanisms.

Evaluation metrics and masks. Raw and completed outputs share the saved world, cameras, and timestamps. FID and FVD for our raw renders use OCC regions, whereas those for completed outputs use full frames; published baseline scores retain their reported protocols. The difference between our raw and completed scores does not isolate the effect of completion, and we do not rank published scores against our scores restricted to OCC regions. Evaluation within OCC regions requires matching masks and preprocessing for generated and reference data. FID and FVD are averaged across cameras. OCC-IoU averages class IoU between a frozen image segmenter and projected OCC semantics within a fixed mask defined by occupancy. Cross RGB measures cross-view RGB error between corresponding surfaces, using the same queries defined by occupancy for raw and completed outputs. These alignment metrics assess the output images; the geometric coverage of the saved world is unchanged by completion.

Coverage and cross-view correspondence. Occupancy fixes the query set $\mathcal{Q}$, surface correspondences, and boundary exclusions. Correspondences across views require valid projection, depth agreement, compatible semantics, and actor identity at the same physical time. Let $\mathcal{Q}_\alpha$ contain queries with sufficient alpha and valid rendered depth, and let e_d denote depth error. We define coverage and the fraction of depth conflicts as

$$C = \frac{|\{q \in \mathcal{Q}_\alpha : e_d(q) \le \tau_d\}|}{|\mathcal{Q}|}, \qquad E_{\text{conflict}} = \frac{|\{q \in \mathcal{Q}_\alpha : e_d(q) > \tau_d\}|}{|\mathcal{Q}|}.$$

Depth error includes queries whose alpha is sufficient even when their rendered depth conflicts with occupancy. RGB errors use surface queries with valid rendered colors in both views or world states and are averaged within each scene, then across scenes. RGB errors are computed only over

this shared valid set and are undefined when the set is empty. Masks and denominators are fixed independently of predicted colors or completion quality.

Ablation study. Table 3 reports measured results for four variants. The four variants share scenes, cameras, timestamps, masks, sample counts, and random seeds, and each processes an initial block followed by one continuation block. Removing the support check disables only the projected support test, retaining the other occupancy, alpha, depth, and correspondence constraints. It uses the full model's incoming world and candidate proposals; each composer maintains its own accepted history and appearance lookup, without using renders from the ablated world to generate subsequent proposals. Independent appearance assignment replays the full model's proposals, accepted geometry, opacity, and correspondences, changing only the colors of newly accepted Gaussians with compatible stored appearance. Earlier records remain fixed, and the edited colors are not fed back into proposal generation during replay. Alpha, depth, and revisit coverage therefore remain identical up to numerical tolerance, and the Gaussian count is unchanged. The variant without future read supervision is trained separately with matched architecture, initialization procedure, data, budget, and other losses; its geometry is not assumed invariant. Revisit queries lie on static OCC surfaces visible to the cameras in the initial block; this region is fixed before generation. Revisit RGB compares renders of the initial and updated worlds at fixed cameras over surface queries with valid rendered colors in both states. Coverage over the complete query set, depth conflicts, and the retained query fraction are checked to determine whether lower counts or errors come at the cost of rendering support. FID/FVD use the same evaluators restricted to OCC regions as Table 4; holes inside the fixed evaluation region remain in the evaluated images. The full-model FID/FVD in Tables 3 and 4 come from the same raw outputs and evaluation at ± 2 m. The saved Gaussian count is divided by the full-model count for each scene, then averaged across scenes. The count ratio is 1.00 for the full model by normalization and for the independent appearance variant because it replays the same geometry. Gaussian counts exclude atlas storage and do not measure GPU memory use. Capacity exhaustion and incomplete runs are recorded as failures.

C Statements

C.1 AI Use Statement

We used generative AI tools to assist with literature review, manuscript drafting and language editing, LaTeX formatting, and code inspection. The authors made all final research decisions, prepared the final figures, and take full responsibility for the accuracy of the reported results and all text, code, and visual materials in this submission.

C.2 Ethics Statement

This work studies controllable driving scene generation for simulation and autonomous driving research. It uses existing nuScenes and Occ3D resources and pretrained visual models; their use and redistribution remain subject to the respective licenses and access conditions. Generated scenes may inherit biases in road layouts, weather, and traffic participants from the source data and pretrained priors. Visually realistic images can still contain errors in geometry or motion, and synthetic driving videos may be mistaken for records of real events. Generated outputs should therefore be labeled as synthetic. The reported results do not establish suitability for vehicle control or safety certification; such use requires independent validation of physical accuracy, coverage of rare events, and performance under distribution shifts.

C.3 Reproducibility Statement

Section 3 describes the world representation, update rules, appearance atlas, training, and video completion. Section 4.1 specifies the dataset, camera inputs, generation blocks, and evaluation metrics. Appendices A and B detail the implementation, gradient paths, completion masks, evaluation regions, correspondences, metric aggregation, and ablation settings. We distinguish OCC-region metrics from full-frame metrics and identify the raw outputs shared by Tables 3 and 4. We cite published baseline scores and identify the baseline outputs we reproduced in Figure 3.