NeurIPS 2026

One World, Dual Timeline:

Decoupled Spatio-Temporal Gaussian Scene Graph for 4D Cooperative Driving Reconstruction

Yulong Chen[12*] · Xiaoyun Dong[1*] · Haoyu Zhang[13*] · Zongxian Yang[12] · Lewei Xie[12] · Jianping Wang[1] · Xinke Li[14†] · Yifan Zhang[12†] · Kai Wang[12†]

[1] City University of Hong Kong, Hong Kong, China [2] City University of Hong Kong (Dongguan), Guangdong, China [3] SLAI, Shenzhen, China

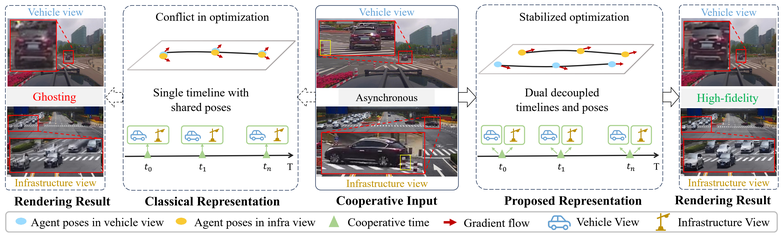
[4] Chengdu Research Institute, City University of Hong Kong, Sichuan, China * Equal contribution. † Corresponding authors.

香港城市大學 City University of Hong Kong

香港城市大學(東莞) City University of Hong Kong (Dongguan)

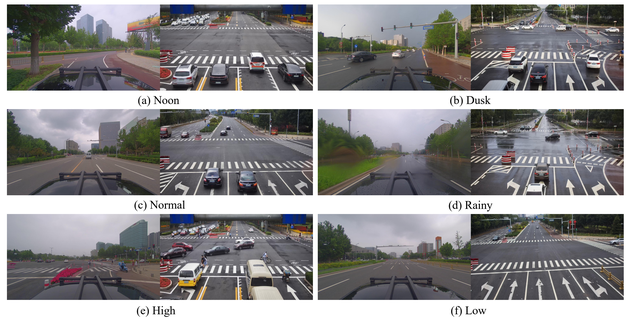
深圳河套學院 Shenzhen Loop Area Institute

1. Comparison of scene representations



Vehicle and infrastructure cameras capture the black car at different times. At 10 m/s, a 70 ms offset corresponds to 0.7 m of motion. A shared pose produces conflicting photometric gradients and ghosting.

6. V2X-Seq sequences



(a) Noon (b) Dusk (c) Normal (d) Rainy (e) High (f) Low

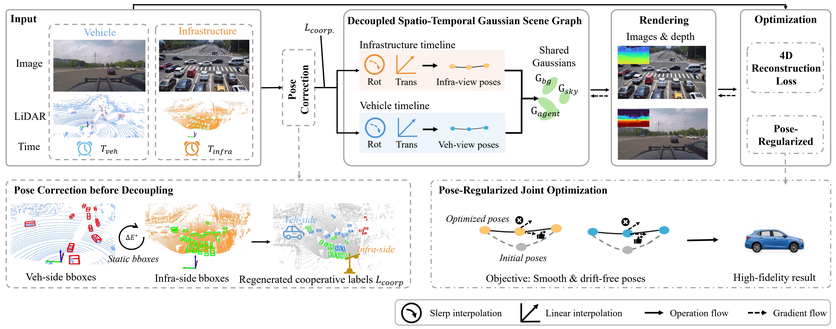
The 26 sequences span Noon, Dusk, Normal, Rainy, Low and High conditions. The median of the per-sequence median absolute offsets is 97.10 ms.

Irreducible Error under a Shared Pose

$$\mathcal{L}^{*}_{\text{single}} \geq \frac{|\Delta\tau|^2\|\boldsymbol{v}\|^2}{4}\sum_{n=1}^{N}\lambda_n > 0,$$

Under local linearized rendering, locally rigid motion and non-degenerate photometric sensitivity, the bound grows quadratically with speed and time offset.

2. Decoupled Spatio-Temporal Gaussian Scene Graph



DUST-GSG shares canonical Gaussians across sources. Each dynamic agent has separate vehicle-side and infrastructure-side pose trajectories queried at the respective capture times.

Each source updates its own pose parameters. Both sources update the canonical Gaussians.

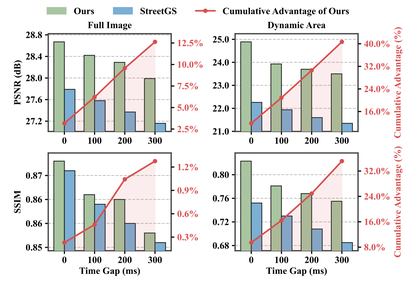
Pose Correction before Decoupling

Match co-visible static vehicles across views. Fit a 6-DoF correction to their 3D bounding-box corners and regenerate cooperative labels.

Pose-Regularized Joint Optimization

Refine Gaussian parameters and dual-timeline poses. Smooth neighboring poses and penalize drift from initialization early in training.

4. Robustness against Asynchronous Observations

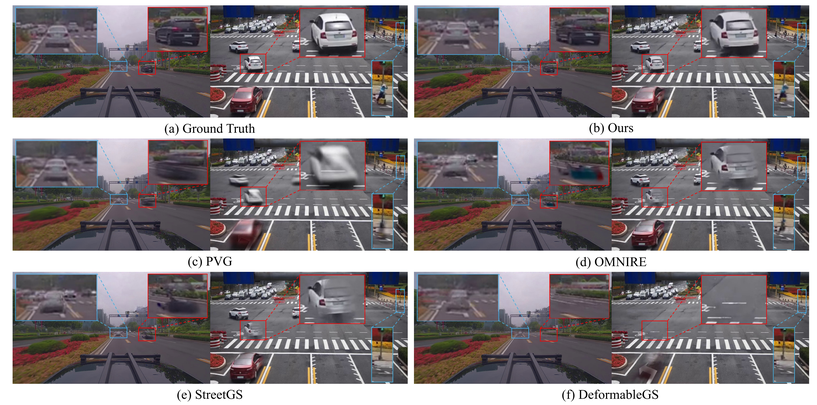


Each additional step shifts the pairing by one infrastructure frame (100 ms). DUST maintains higher Dynamic Area PSNR at all four time gaps.

The red curve shows the cumulative relative advantage over StreetGS on the right axis.

At 0 ms the original camera timestamp offsets still apply.

3. Dynamic agents under asynchronous observations

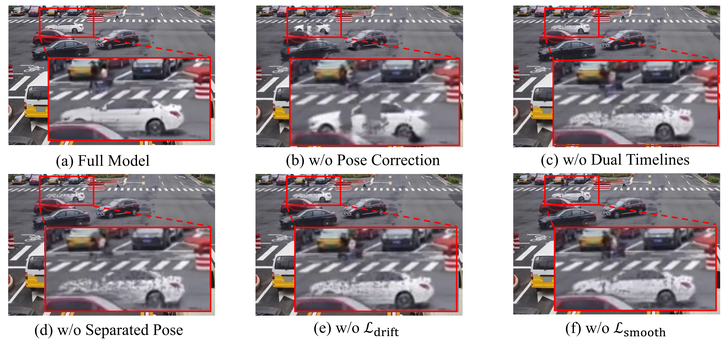


(a) Ground Truth (b) Ours (c) PVG (d) OMNIRE (e) StreetGS (f) DeformableGS

DUST reduces ghosting around moving vehicles and pedestrians in both views.

Dynamic Area · +3.20 dB PSNR over PVG, 37.7% lower FVD than OMNIRE.

5. Qualitative ablation study



(a) Full Model (b) w/o Pose Correction (c) w/o Dual Timelines (d) w/o Separated Pose (e) w/o $\mathcal{L}_{\text{drift}}$ (f) w/o $\mathcal{L}_{\text{smooth}}$

1. Full Sequence results on V2X-Seq

Full = average over all 26 sequences

Method	Scene Reconstruction							Novel View Synthesis						
	Full Image			Dynamic Area				Full Image			Dynamic Area			
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	FVD ↓	EPE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	FVD ↓	EPE ↓
3DGS (2023)	22.72	0.812	0.307	15.82	0.425	433.2	22.62	18.91	0.761	0.343	14.32	0.409	682.2	58.52
DeformGS (2024)	27.61	0.877	0.135	18.74	0.572	360.9	4.3	25.04	0.804	0.152	17.46	0.450	564.2	16.84
StreetGS (2024)	28.88	0.886	0.116	22.54	0.758	103.0	2.01	26.20	0.812	0.135	20.01	0.572	214.8	11.47
OMNIRE (2025)	**29.20**	**0.887**	**0.114**	22.90	**0.767**	**63.7**	**1.87**	**26.41**	0.813	**0.132**	**20.25**	**0.579**	**125.2**	**10.93**
CRUISE (2025)	24.47	0.826	0.187	18.57	0.561	240.9	5.74	23.63	0.801	0.199	17.68	0.488	397.6	24.49
PVG (2026)	28.97	0.879	0.157	**22.91**	0.742	131.9	2.77	26.16	**0.817**	0.171	20.09	0.562	262.3	14.24
Ours (DUST)	**29.55**	**0.889**	**0.112**	**26.11**	**0.860**	**39.7**	**1.70**	**26.82**	**0.818**	**0.130**	**21.80**	**0.631**	**97.2**	**9.94**

Best and **second best** in each column. Scene Reconstruction uses all frames. Novel View Synthesis holds out both views at every 10th cooperative timestamp.

2. Ablations on four High Crowd sequences

Variant	Full Image			Dynamic Area	
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
Full Model	**28.75**	**0.874**	**0.125**	**25.12**	**0.837**
w/o Pose Correction	27.91	0.873	0.128	23.98	0.821
w/o Dual Timelines	28.40	0.872	0.128	23.35	0.783
w/o Separated Poses	28.19	0.870	0.131	22.34	0.742
w/o $\mathcal{L}_{drift}$	28.72	0.873	0.126	24.99	0.832
w/o $\mathcal{L}_{smooth}$	28.62	0.869	0.134	24.96	0.833

Without separated poses, Dynamic Area PSNR falls from 25.12 to 22.34 dB. StreetGS with source-specific pose interpolation reaches 21.76 dB in the separate comparison.

30,000 training iterations per scene on one NVIDIA A100 GPU.