FRUC: Feedforward Dynamic Scene Reconstruction from Uncalibrated Collaborative Driving Views

Yihang Tao, Yu Guo, Zhengru Fang,* Haonan An, Yuguang Fang
Hong Kong JC STEM Lab of Smart City, City University of Hong Kong
{yihang.tommy, yu.guo, zhefang4-c, haonanan2-c}@my.cityu.edu.hk
my.Fang@cityu.edu.hk

Abstract

We present FRUC, a feedforward 3D Gaussian Splatting framework for dynamic scene reconstruction from uncalibrated collaborative driving views. Existing multi-agent reconstruction frameworks are often hindered by rigid prerequisites, demanding precise spatial calibration and slow per-scene optimization. In this paper, we rethink this task by conceptualizing a distributed multi-vehicle network as a spatio-temporally unstructured ego-centric multi-camera system, where the core challenge lies in enhancing ego-centric occluded geometry through collaboration without degrading the ego's accurately observed visible geometry, while preserving reconstruction efficiency. For efficient reconstruction, FRUC is built upon a visual grounded geometric Transformer backbone to enable one-shot, calibration-free inference from a flexible number of multi-vehicle views. To achieve non-destructive geometric supplementation under uncalibrated cross-agent misalignment, FRUC first introduces an ego-centric causal occlusion field that explicitly derives occlusion evolution as latent priors by modeling agent-wise spatio-temporal correlations. Guided by these occlusion priors, it further formulates cross-agent integration as a deterministic residual denoising process via zero-initialized injection, turning challenging cross-agent fusion into bounded residual learning for robust collaborative blind-spot completion. Through extensive evaluations on real-world V2X-Real and UrbanIng-V2X datasets, FRUC is shown to be a new state-of-the-art for the scene reconstruction of dynamic collaborative driving environments, significantly outperforming existing methods in both rendering quality and efficiency. Code is available at https://github.com/yihangtao/FRUC.git.

1 Introduction

Fast and scalable dynamic scene reconstruction forms the bedrock of modern autonomous driving, powering both on-board execution and cloud-based simulation. On-board, rapidly lifting sparse historical observations into 4D representations enables real-time what-if analysis. This allows systems to synthesize novel views and simulate potential future states for critical decision-making [25, 9, 2, 27, 12, 4]. In the cloud, efficiently transforming extensive driving logs into large-scale, dynamic digital twins provides the robust closed-loop simulators essential for training generalizable driving world models [38, 30, 13, 15, 16, 21]. To meet the efficiency demanded by these continuous sensory streams, feedforward 3D Gaussian Splatting (3DGS) has emerged as the premier single-pass solution. Consequently, recent single-vehicle pipelines have rapidly evolved from efficient surround-view static reconstruction [25] and spatial-temporal dynamic modeling [34] to unified frameworks supporting pose-free dynamic scene reconstruction directly from unposed images [39, 2].

Despite these rapid advancements, ego-centric reconstruction remains fundamentally bottlenecked by the physical line-of-sight [31, 11, 14, 5, 6, 10, 23, 24]. Severe visual occlusions in single-

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

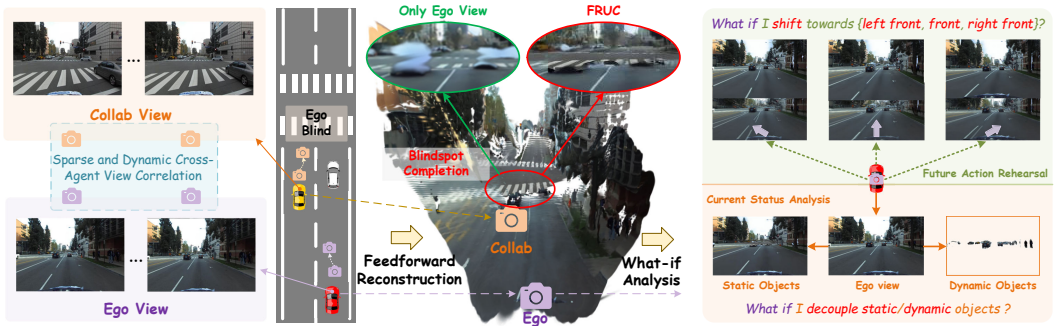


Figure 1: **Collaborative driving scene reconstruction and what-if analysis with FRUC.** By aggregating uncalibrated, sparse, and dynamic cross-agent observations, FRUC seamlessly completes the ego vehicle's blind spots via a single feedforward pass. It provides accurate occluded background recovery, leading to reliable scene editing and high-fidelity novel view synthesis.

vehicle perspectives inevitably result in incomplete scene geometry, which critically constrains what-if analysis. For instance, when removing a foreground object to synthesize novel safety-critical scenarios for closed-loop evaluation, single-vehicle representations invariably reveal unobserved "holes" in the background, severely degrading simulation fidelity. Collaborative driving systems offer a paradigm shift to shatter this physical limitation. By aggregating complementary viewpoints from distributed agents, collaborative reconstruction can effectively "see through" occlusions to establish a holistic and geometrically complete 3D environment (as illustrated in Fig. 1). Recently, methods like CRUISE [31] and V2X-Gaussians [11] pioneered this domain. However, these collaborative methods rely heavily on rigid prerequisites, assuming that multi-source agent cameras are well-calibrated and often requiring auxiliary sensor data (e.g., LiDAR point clouds) for spatial initialization. Furthermore, they depend on computationally expensive per-scene optimization. This rigid dependency makes them notoriously slow and fundamentally impractical for real-time, scalable applications.

To overcome these limitations, we re-formulate this task as *feedforward reconstruction from uncalibrated collaborative driving views*. However, extending ego-centric models to multi-agent scenarios introduces two formidable challenges due to the inherent instability of collaborative networks: ❶ *Unstructured Geometric Misalignment:* Single-vehicle multi-camera systems possess fixed relative extrinsics, allowing straightforward fusion via deterministic warping [25]. Conversely, multi-agent systems introduce uncalibrated, time-varying, and dynamically moving cameras. This lack of precise cross-agent calibration renders simple geometric projection largely ineffective for aligning these highly unstructured viewpoints. ❷ *Destructive Semantic Interference:* Single-vehicle sequences provide a highly stable spatiotemporal context with constant spatial overlap and smooth temporal transitions between consecutive frames, offering robust cues for 3D geometry inference. Introducing external cameras drastically destabilizes this pattern: spatio-temporal overlap becomes unpredictable, and cross-vehicle semantic correlations drop significantly. Forcing a network to jointly encode this highly irregular context creates a cluttered latent space, which overwhelms the network and severely corrupts the ego vehicle's previously established geometric priors, even destroying its reliable single-vehicle reconstruction capabilities (e.g., ghosting or distorted structures).

In this paper, we introduce FRUC, a unified feedforward framework designed to tackle these challenges. Rather than employing rigid geometric warping or naive feature concatenation, we conceptually re-formulate cross-agent integration as a bounded residual learning problem. Built upon a visual geometric grounded Transformer (VGGT) backbone [26], FRUC extracts ego-centric occlusion priors to guide a deterministic latent residual denoising process. This high-level design implicitly handles uncalibrated multi-agent misalignment and robustly supplements dynamic blind spots, while preserving the ego's reliable observations from destructive cross-agent semantic interference.

Our main contributions are summarized as follows:

- We propose FRUC, a unified feedforward 3DGS framework designed to seamlessly handle spatio-temporally unstructured multi-camera inputs for collaborative driving systems, achieving highly efficient one-shot inference without relying on precise multi-agent calibration.

- To achieve non-destructive geometric supplementation, we introduce an ego-centric causal occlusion field that explicitly captures occlusion evolution, coupled with a cross-agent latent residual

denoising process via prior injections. This enables robust assimilation of collaborative blind-spot information while preserving the ego vehicle's reliably observed visible geometry.

- Extensive evaluations based on the real-world V2X-Real and UrbanIng-V2X datasets demonstrate that FRUC establishes a new state-of-the-art for collaborative driving scene reconstruction, significantly outperforming existing optimization-based and single-agent feedforward methods in both rendering quality and efficiency.

2 Related Work

2.1 Feedforward 3D Reconstruction

Recent advancements in 3D Gaussian Splatting (3DGS) [13] have sparked significant interest in feedforward reconstruction methods that eliminate the need for tedious per-scene optimization by learning powerful priors from large-scale datasets. Early approaches, such as Splatter Image [22] and GS-LRM [36], demonstrate the feasibility of predicting 3DGS parameters in a single forward pass, but they primarily focus on object-centric or indoor static scenes. Subsequent works like pixelSplat [1] and MVSplat [3] leverage multi-view stereo concepts to improve geometry prediction. AnySplat [12] attempts to lift unconstrained multi-view captures into 3D Gaussians. However, these methods typically rely on dense multi-view inputs with substantial visual overlap to construct cost volumes or establish geometry priors. Consequently, they struggle to generalize to large-scale, outdoor driving environments, where onboard surround-view cameras naturally exhibit minimal overlap.

2.2 Multi-view Driving Scene Reconstruction

To address the unique demands of autonomous driving, recent efforts have shifted towards feedforward architectures tailored for large-scale, outdoor scenes. DrivingForward [25] proposes a real-time framework by jointly training a pose and a depth network to predict Gaussian primitives from flexible surround-views. STORM [34] aggregates 3D Gaussians from sparse observations using self-supervised scene flows, transforming them across multiple timesteps for dynamic reconstruction. More recently, DGGT [2] and StreetForward [35] leverage grounded transformers and cross-view attention to efficiently reconstruct dynamic driving environments directly from unposed image sequences, bypassing explicit geometric priors. Despite their impressive performance, they are limited to the singe-vehicle setting, which is inherently constrained by the ego vehicle's limited physical line-of-sight, and inevitably suffers from severe geometric incompleteness in occluded regions. Besides, these methods are tailored for fixed multi-camera view relationships within a single vehicle, limiting its applicability to unstructured and dynamic multi-vehicle multi-camera systems.

2.3 Collaborative Driving Scene Reconstruction

Collaborative driving systems aggregate distributed viewpoints to overcome the inherent physical limitations (e.g., occlusions) of single-vehicle perception. Recently, optimization-based 3DGS frameworks have been introduced to this domain. CRUISE [31] proposes a cooperative reconstruction and editing pipeline that relies heavily on dense 3D annotations to separate dynamic vehicles for novel scene synthesis. Meanwhile, V2X-Gaussians [11] introduces a V2X-tailored method that computes exact geometric ray intersections to identify overlapping areas. Despite their success, these methods share critical limitations when faced with real-world unconstrained V2X networks: they assume well-calibrated mult-agent observations, heavily rely on multi-modal inputs (e.g., LiDAR) for labeling or initialization, and operate on a computationally prohibitive per-scene optimization paradigm. In contrast, our FRUC is the first feedforward framework that elegantly bypasses these rigid constraints, implicitly learning cross-agent alignment to achieve highly efficient, single-pass dynamic scene reconstruction directly from uncalibrated collaborative driving views.

3 Methodology

We conceptualize a distributed multi-vehicle network as a spatio-temporally unstructured ego-centric multi-camera system, bypassing the need for exact camera extrinsics and hardware-level time synchronization. Given a specific spatial region, we assume a joint set of temporally ordered, uncalibrated

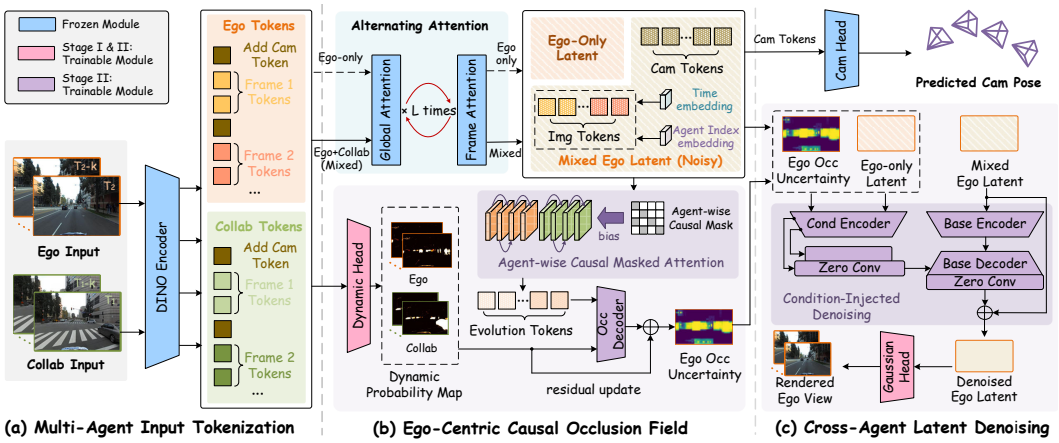


Figure 2: **The proposed** FRUC **framework.** It reconstructs collaborative driving scenes by extracting uncalibrated multi-agent context, inferring dynamic spatial priors through an ego-centric causal occlusion field, and integrating collaborative geometry via cross-agent latent residual denoising.

multi-agent input images $\{\mathbf{I}_k^\tau \mid \mathbf{I}_k^\tau \in \mathbb{R}^{H\times W\times 3}, \tau = 1,\ldots,N\}$ captured by an arbitrary agent $k \in \{e, c\}$ (where e denotes ego and c denotes collab) over N timestamps. Our objective is to reconstruct a temporally coherent, geometrically complete 4D representation in a single forward pass, as conceptually illustrated in Fig. 2.

3.1 Multi-Agent View Input Tokenization

To process the aforementioned N-timestamp multi-agent sequence, we first flatten and tokenize the inputs. For clarity, we use $N = 2$ timestamps $\tau \in \{t_0, t_1\}$ for illustration. We concatenate the raw images into a unified multi-agent sequence $\mathbf{I} = [\mathbf{I}_e^{t_0}, \mathbf{I}_c^{t_0}, \mathbf{I}_e^{t_1}, \mathbf{I}_c^{t_1}]$. Each frame is then encoded into L high-level patchified tokens $\mathbf{F}_{dino,k}^\tau$ of dimension D using a DINO encoder [17].

To process these collaborative perceptual streams without relying on rigid global coordinate systems, we formulate the distributed observations as a joint set of semantic proxies. Specifically, following the VGGT architecture [2], a learnable camera token $\mathbf{c}_{cam}$ is appended to the DINO tokens of each frame. These augmented sequences are then processed through the VGGT Alternating-Attention (AA) feature backbone, which aggregates inter-frame and intra-frame contexts to infer the implicit multi-view geometry:

$$\mathbf{F}_{m,k}^\tau, \hat{\mathbf{c}}_k^\tau = \mathcal{E}_{vggt}\Big([\mathbf{c}_{cam}, \mathbf{F}_{dino,k}^\tau]\Big), \tag{1}$$

where $\mathbf{F}_{m,k}^\tau \in \mathbb{R}^{L\times D}$ represents the aggregated deep image features absorbing rich multi-view context, and $\hat{\mathbf{c}}_k^\tau$ is the refined implicit camera token encapsulating the source view's spatial prior.

While the resulting deep features establish a powerful rendering manifold, this coarse fusion inherently creates an entangled latent space where uncalibrated collaborative perspectives interferes the ego's geometric representation. To facilitate motion-aware downstream alignment, we augment the aggregated tokens with explicit temporal and identity information. Specifically, we introduce a learnable agent-identity embedding $\mathbf{E}_{agt}(k) \in \mathbb{R}^D$ and add it to a normally distributed temporal embedding $\mathbf{E}_{tem}(\tau) \in \mathbb{R}^D$, corresponding to the discrete frame timestamp τ:

$$\mathbf{E}_{meta}(k,\tau) = \mathbf{E}_{agt}(k) + \mathbf{E}_{tem}(\tau) \in \mathbb{R}^D. \tag{2}$$

We then augment every patch token in the frame with this joint meta-data vector via broadcasting, leading to $\tilde{\mathbf{F}}_{m,k}^\tau = \mathbf{F}_{m,k}^\tau + \mathbf{E}_{meta}(k,\tau) \in \mathbb{R}^{L\times D}$. This explicit encoding of cross-agent identity and temporal causality serves as spatio-temporal anchors during the subsequent feature denoising. For simplicity, we denote the full sequence of augmented tokens as $\tilde{\mathbf{F}}_m$.

3.2 Ego-Centric Causal Occlusion Field

In complex driving scenarios, ego-vehicle blind spots are predominantly caused by the occlusion of dynamic objects like preceding vehicles or pedestrians. Rather than remaining static, these blind spots

evolve continuously alongside moving objects, occluding new areas while exposing previously hidden regions. Consequently, effectively repairing these blind spots using collaborative views requires reasoning about the entire spatio-temporal footprint of the occlusion. Motivated by this, we propose an ego-centric Causal Occlusion Field (COF) to explicitly model this temporal evolution and derive occlusion uncertainty as a structured spatial prior.

To achieve this, we first estimate a base dynamic probability map $S^{\tau}_{dyn,k}$ to identify the current spatial footprint of movable entities. Operating on the shallow, topology-preserving DINO features $\mathbf{F}^{\tau}_{dino,k}$ alongside the input images, we utilize a DPT-style [19] dynamic head Φ_{dyn}:

$$S^{\tau}_{dyn,k} = \sigma\left(\Phi_{dyn}(\mathbf{F}^{\tau}_{dino,k}, \mathbf{I}^{\tau}_{k})\right) \in [0,1]^{H\times W}, \tag{3}$$

where σ denotes the sigmoid activation. While this static prior isolates the object's current position, it lacks the kinematic context necessary to infer the surrounding occlusion dynamics. To capture these dynamics, we introduce a temporal-evolutionary occlusion inference mechanism. Operating on the flattened sequence of enhanced features $\tilde{\mathbf{Z}}_m \in \mathbb{R}^{B\times(F\cdot L)\times D}$ (where B is the batch size and F is the total number of frames), we apply an *Agent-wise Causal Masked Attention* layer. This mechanism strictly prevents information leakage across different agents or future frames, using a causal binary mask $\mathbf{M}_{cau} \in \mathbb{R}^{(F\cdot L)\times(F\cdot L)}$:

$$\mathbf{M}_{cau}[b,h,i,j] = \begin{cases} 1, & \text{if } a(i) = a(j) \text{ and } \tau(i) \geq \tau(j), \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where b and h index the batch and attention head, respectively, while i and j denote the token indices in the flattened sequence. The functions $a(\cdot)$ and $\tau(\cdot)$ map a given token index to its corresponding agent ID and frame timestamp. Intuitively, this temporal attention acts as a latent tracker. By correlating current features with their historical states within the same agent's perspective, it implicitly extracts a latent kinematic field, formalized as *motion evolution tokens* $\mathbf{Z}^{\tau}_{kin,k} \in \mathbb{R}^{L\times D}$:

$$\mathbf{Z}^{\tau}_{kin,k} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_h}} + \log \mathbf{M}_{cau}\right)\mathbf{V}, \tag{5}$$

where $\log \mathbf{M}_{cau}$ assigns $-\infty$ to the disallowed entries, strictly enforcing causality. Based on this latent kinematic field, we employ an Occlusion Decoder Φ_{occ} (a lightweight fully connected network) to model the occlusion evolution. To ensure stable optimization, we formulate this process as a motion-aware residual dilation. Instead of regressing the occlusion prior from scratch, the network explicitly starts with the base dynamic footprint $S^{\tau}_{dyn,k} \in [0,1]^{H\times W}$ and gradually learns to expand it under the guidance of the reshaped kinematic field $\mathbf{Z}^{\tau}_{kin,k} \in \mathbb{R}^{D\times H\times W}$. Specifically, the decoder predicts a spatio-temporal expansion residual $\Delta S^{\tau}_{kin,k}$:

$$\Delta S^{\tau}_{kin,k} = \Phi_{occ}(\mathbf{Z}^{\tau}_{kin,k} \oplus S^{\tau}_{dyn,k}) \in [-1,1]^{H\times W}. \tag{6}$$

This residual dictates how the object's footprint adaptively expands along its motion trajectory to cover the dynamically occluded background. We then compute the final *Occlusion Uncertainty Prior* $\mathbf{M}^{\tau}_{occ,k}$ by applying this learned residual directly to the base footprint:

$$\mathbf{M}^{\tau}_{occ,k} = \text{Clamp}\Big(S^{\tau}_{dyn,k} + \Delta S^{\tau}_{kin,k},\ 0,\ 1\Big), \tag{7}$$

where the Clamp$(\cdot)$ function restricts the resulting probability values strictly within the valid range of $[0,1]$. By employing this residual learning paradigm, the network initially anchors on the physical dynamic footprint and progressively learns the optimal spatial expansion. This process yields a continuous causal occlusion field, a structured spatial prior that preserves the coarse blind-spot location while explicitly injecting kinematic evolution. Ultimately, this field enables the model to adaptively adjust its attention around dynamically evolving occlusions based on cross-agent geometric relationships, actively fetching the most beneficial collaborative features for blind-spot completion.

3.3 Cross-Agent Latent Residual Denoising

With the established ego-centric causal occlusion field, we formulate cross-agent feature alignment as a constrained latent residual denoising process[2]. Specifically, we introduce a Cross-Agent Latent

[2]A detailed theoretical justification is provided in Appendix C.

Residual Denoising (CALRD) module to act as a spatial modulator [18]. The intuition is that while collaborative features are essential for completing blind spots, they may introduce uncalibrated geometric drift into visible regions. Therefore, CALRD utilizes the inferred occlusion prior $\mathbf{M}^{\tau}_{occ,k} \in [0,1]^{H\times W}$ to explicitly modulate feature supplementation, enabling targeted background completion without overwriting the ego vehicle's reliable local geometry.

To ensure this residual denoising process remains firmly grounded, we extract the clean ego-only features $\mathbf{F}^{\tau}_{e}$ as a pure structural reference. We feed the spatialized mixed features $\tilde{\mathbf{F}}^{\tau}_{m,k} \in \mathbb{R}^{D\times H\times W}$, the occlusion prior $\mathbf{M}^{\tau}_{occ,k}$, and the reference features $\mathbf{F}^{\tau}_{e}$ into a dual-branch convolutional architecture. The conditional branch encodes the concatenated spatial prior and reference to extract structural conditions $\mathbf{C}^{\tau}_{pri,k} \in \mathbb{R}^{D\times H\times W}$:

$$\mathbf{C}^{\tau}_{pri,k} = \mathcal{Z}_{pri}\Big(\Phi_{con}([\mathbf{M}^{\tau}_{occ,k}, \mathbf{F}^{\tau}_{e}])\Big), \tag{8}$$

where Φ_{con} is a lightweight encoder, and $\mathcal{Z}_{pri}$ is a zero-initialized 1×1 convolution. The base branch processes the mixed tokens $\tilde{\mathbf{F}}^{\tau}_{m,k}$ using a base feature encoder Φ_{fea}. The final denoised feature $\hat{\mathbf{F}}^{\tau}_{m,k}$ is obtained by predicting a residual correction and adding it to the noisy mixed tokens:

$$\hat{\mathbf{F}}^{\tau}_{m,k} = \tilde{\mathbf{F}}^{\tau}_{m,k} + \underbrace{\mathcal{Z}_{out}\Big(\Phi_{fea}(\tilde{\mathbf{F}}^{\tau}_{m,k}) + \mathbf{C}^{\tau}_{pri,k}\Big)}_{\text{Predicted Residual Correction}}, \tag{9}$$

where $\mathcal{Z}_{out}$ is another zero-initialized 1×1 convolution. This spatial modulation acts as a surgical residual denoising mechanism. Rather than directly predicting absolute target features, the network learns a precise residual correction to filter out uncalibrated collaborative noise. For unoccluded regions ($\mathbf{M}^{\tau}_{occ,k} \approx 0$), the zero-convolutions naturally suppress collaborative updates, protecting the ego vehicle's confident observations. Conversely, within dynamically shifting blind spots ($\mathbf{M}^{\tau}_{occ,k} > 0$), the residual update dynamically aggregates structural context to inpaint the missing background.

Finally, the denoised feature map $\hat{\mathbf{F}}^{\tau}_{m,k}$ is decoded by a specialized Gaussian head Φ_{gs} to predict the properties of 3D Gaussians (e.g., opacity α, color $\mathbf{c}$, scale $\mathbf{s}$, and rotation $\mathbf{q}$) directly along the camera rays. Guided by the decoupled dynamic and static streams, Φ_{gs} predicts multiple Gaussians $\mathcal{G} = \mathcal{G}_{dyn} \cup \mathcal{G}_{sta}$ distributed along each ray. This disentangles the foreground occluder from the collaboratively inpainted background. Additionally, a parallel sky head Φ_{sky} extracts environment illumination and distant sky semantics $\mathcal{I}_{sky}$ from the global context. During rasterization for novel view synthesis from an arbitrary ego viewpoint $\mathbf{P}_{novel}$, we can selectively render the full scene $\mathcal{I}_{full} = \text{Render}(\mathcal{G}, \mathbf{P}_{novel}) + \mathcal{I}_{sky}$ or seamlessly reveal the underlying static background by explicitly dropping the dynamic Gaussians $\mathcal{I}_{bg} = \text{Render}(\mathcal{G}_{sta}, \mathbf{P}_{novel}) + \mathcal{I}_{sky}$.

3.4 Model Training

Training a highly dynamic distributed network from scratch is notoriously unstable. To ensure stable convergence, we propose a two-stage progressive training curriculum.

Stage I: Single-Agent Pre-training. We first establish robust ego-centric reconstruction and kinematic priors. With the base VGGT backbone kept frozen, the Gaussian rendering head Φ_{gs}, sky head Φ_{sky}, and dynamic head Φ_{dyn} are trained using exclusively single-agent sequences. The network minimizes photometric losses $\mathcal{L}_{pho}$ for scene rendering, which calculates the discrepancy between the synthesized image $\mathcal{I}_{full}$ and the ground-truth image $\mathcal{I}_{gt}$ using a combination of L_1 and LPIPS [37] distances:

$$\mathcal{L}_{pho} = \lambda_1 \|\mathcal{I}_{full} - \mathcal{I}_{gt}\|_1 + \lambda_{lpips}\text{LPIPS}(\mathcal{I}_{full}, \mathcal{I}_{gt}). \tag{10}$$

Simultaneously, Φ_{dyn} is supervised via Binary Cross-Entropy (BCE) using pseudo ground-truth dynamic masks S^{gt}_{msk} extracted by a pre-trained SegFormer [29]:

$$\mathcal{L}_{msk} = -\frac{1}{|\Omega|}\sum_{\mathbf{u}\in\Omega}\Big[S^{gt}_{msk}(\mathbf{u})\log S^{\tau}_{dyn,k}(\mathbf{u}) + \big(1 - S^{gt}_{msk}(\mathbf{u})\big)\log\big(1 - S^{\tau}_{dyn,k}(\mathbf{u})\big)\Big], \tag{11}$$

where $\mathbf{u}$ denotes the 2D pixel coordinates and Ω represents the entire spatial image domain. The total loss for Stage I is $\mathcal{L}_{s1} = \mathcal{L}_{pho} + \lambda_{msk}\mathcal{L}_{msk}$.



Figure 3: **Qualitative comparison of NVS on V2X-Real.** The input context consists of both ego and collaborative views at $t-1$ and $t+1$, and the task is to reconstruct the ego view at time t. Here we show the results of forward-facing camera. Red boxes highlight degradations of second-best baseline.

Stage II: Cross-Agent Adaptation. Initialized with Stage I weights, we further freeze the pre-trained dynamic head Φ_{dyn}, sky head Φ_{sky}, and train the newly added `FRUC` modules alongside other trainable decoders in stage I using multi-agent sequences. To explicitly guide the residual learning, we introduce a set of auxiliary objectives $\mathcal{L}_{aux}$ (detailed in Appendix A). These losses establish a zero-sum dynamic: anchoring the occlusion prior to prevent unbounded expansion, driving background completion within occlusion boundaries, and enforcing global latent denoising to preserve the reliable ego-only baseline. Stage II combines these penalties with the base photometric loss:

$$\mathcal{L}_{s2} = \mathcal{L}_{pho} + \mathcal{L}_{aux}. \tag{12}$$

Coupled with the zero-initialization strategy, this joint objective transforms the highly non-convex multi-agent optimization into a bounded, highly stable residual learning problem.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct primary benchmark evaluations on V2X-Real [28] and evaluate generalizability on UrbanIng-V2X dataset [20]. To ensure high-quality and meaningful evaluation, we carefully pre-process the raw datasets. This involves supplementing the missing dynamic and sky masks, filtering out invalid sequences, and explicitly designating the ego vehicle, ultimately establishing two rigorous benchmarks tailored for collaborative dynamic scene reconstruction. Detailed data processing pipelines and filtering criteria are provided in Appendix A.

Implementation Details and Baselines. Our model builds upon the VGGT-1B [26] backbone and is trained on 4 NVIDIA RTX 5880 GPUs. The image resolution is set to 378×672 and the models are trained for 10 epochs. We compare our model against scene-optimized techniques like EmerNeRF [33], 3DGS [13], and V2X-Gaussians [11], as well as feed-forward methods like MVSplat [3], DrivingForward [25], STORM [34], AnySplat [12], and DGGT [2]. Except for V2X-Gaussians, all baselines are single-agent methods. For fair comparison, we adapt them by simply replacing

Table 1: **Comparison to state-of-the-art methods on the V2X-Real Dataset.** We input two frames of multi-agent views to reconstruct the intermediate ego view. **Best** in bold, second-best underlined.

Methods	Full image			Dynamic-only			Blind-Spot (NIQE↓)	Inference Time ↓
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		
Per-scene Optimization Methods								
EmerNeRF [33]	21.65	0.636	0.583	17.79	0.979	0.043	-	10min
3DGS [13]	20.19	0.616	0.592	16.85	0.976	0.035	-	14min
V2X-Gaussians [11]	22.20	0.784	0.409	23.87	**0.994**	0.011	-	21min
Generalizable Feed-forward Methods								
MVSplat [3]	16.73	0.478	0.281	15.18	0.952	0.049	-	0.31s
DrivingForward [25]	17.37	0.596	0.239	14.46	0.955	0.047	-	**0.15s**
STORM [34]	16.27	0.579	0.379	14.36	0.953	0.069	8.682	0.35s
AnySplat [12]	19.18	0.605	0.219	16.69	0.959	0.046	-	0.29s
DGGT [2]	21.85	0.804	0.129	24.76	0.992	0.129	4.301	0.54s
FRUC (Ours)	**25.89**	**0.883**	**0.076**	**25.93**	**0.994**	**0.007**	**3.949**	0.77s

Table 2: **Generalizability Comparison on the UrbanIng-V2X Dataset.** We evaluate NVS quality under both zero-shot and trained settings. **Best** in bold, and second-best underlined.

Method	Full image			Dynamic-only			Blind-Spot (NIQE↓)
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Zero-shot (Trained on V2X-Real)							
MVSplat [3]	18.58	0.463	0.380	15.15	0.966	0.056	-
DrivingForward [25]	18.09	0.676	0.335	15.66	0.970	0.037	-
STORM [34]	17.30	0.604	0.458	14.16	0.967	0.068	11.781
AnySplat [12]	21.31	0.552	0.222	17.88	0.974	0.028	-
DGGT [2]	22.23	0.724	0.286	23.71	**0.993**	0.208	7.509
FRUC (Ours)	**25.07**	**0.795**	**0.144**	**24.61**	**0.993**	**0.008**	**5.270**
Trained (Fine-tuned on UrbanIng-V2X)							
MVSplat [3]	19.24	0.467	0.345	15.86	0.969	0.039	-
DrivingForward [25]	21.80	0.754	0.258	16.71	0.974	0.028	-
STORM [34]	17.36	0.614	0.336	14.22	0.969	0.079	6.268
AnySplat [12]	22.45	0.579	0.176	19.20	0.976	**0.025**	-
DGGT [2]	25.37	0.745	0.275	24.63	**0.993**	0.285	5.728
FRUC (Ours)	**27.26**	**0.996**	**0.005**	**27.64**	0.819	0.079	**3.854**

their original single-agent sequence inputs with multi-agent sequences, while keeping their model architectures and training methods entirely unchanged. This adaptation is justified since collaborative reconstruction can be reformulated as a generalized single-agent reconstruction problem under a dynamic multi-camera setup. More details are in Appendix A.

Evaluation Metrics. To comprehensively assess reconstruction quality, we report PSNR, SSIM, and LPIPS for full image and dynamic-only settings, while using NIQE specifically to evaluate the perceptual realism of blind-spot completion. We default to the multi-frame (MF) mode for our primary benchmark comparisons, and provide additional quantitative results under the single-frame (SF) mode in Appendix B, along with detailed metric formulations.

4.2 Benchmark Comparison

Novel View Synthesis (NVS). We evaluate the rendering quality from the ego vehicle's perspective. In the default MF setting, the input context consists of 4 uncalibrated images derived from a specific ego-collab view pair across two consecutive timestamps (i.e., ego and collaborative views at $t-1$ and $t+1$) to reconstruct the target ego image at the intermediate time t. Quantitative results are averaged across all potential cross-agent view pairs that possess semantic co-visibility. Table 1 reports the comprehensive benchmark, where FRUC outperforms all baseline methods by a large margin. Notably, the degraded performance of DGGT and AnySplat indicates that naive multi-agent feature aggregation struggles with cross-agent geometric misalignment. Furthermore, despite incorporating advanced spatial reasoning modules to handle uncalibrated collaborative context, our framework maintains a highly competitive inference speed of $0.77s$, significantly faster than optimization-based paradigms. Fig. 3 further provides qualitative comparisons. While baseline methods often exhibit severe artifacts or fail to maintain the ego vehicle's correct spatial layout, our model successfully resolves the geometric interference and produces the most faithful and detailed scene reconstructions.

Generalizability. We evaluate the generalization capability of our model on the UrbanIng-V2X dataset. As shown in Table 2, we conduct evaluations under both zero-shot (direct inference using



Figure 4: **Qualitative comparison of NVS on UrbanIng-V2X.** We compare the zero-shot and trained results under MF settings. Red boxes highlight degradations of second-best baseline.

weights trained on V2X-Real) and trained (fine-tuning on UrbanIng-V2X) settings. In the zero-shot scenario, FRUC consistently outperforms feed-forward baselines, demonstrating robust cross-dataset adaptability and reliable structural priors learned from V2X-Real. When fine-tuned on the new dataset, our model further solidifies its advantage across most metrics, indicating strong capacity to fit new collaborative environments and diverse driving scenarios.

Fig. 4 further provides qualitative comparisons on UrbanIng-V2X under both settings. For *zero-shot*, FRUC transfers well from V2X-Real to UrbanIng-V2X, producing substantially clearer scene layouts and more stable object boundaries than the competing feed-forward baselines, which often suffer from severe blur, ghosting, or structural collapse. In the *trained* setting, this advantage becomes more pronounced: FRUC further improves local sharpness, lighting consistency, and geometric integrity, while maintaining reliable reconstruction of distant traffic lights, building facades, and large dynamic objects. These qualitative results corroborate the quantitative improvements, demonstrating robust cross-dataset generalization and strong adaptation after fine-tuning. More results are in Appendix B.

Blind-Spot Completion and Scene Editing. High-fidelity blind-spot completion is essential for reliable scene editing. To quantify this, we render the static scene and compute NIQE exclusively on cropped regions around the original dynamic entities. Tables 1 and 2 show that FRUC achieves the best NIQE across both datasets, demonstrating superior blind-spot completion capability. Fig. 5 further compares ego-only and collaborative reconstruction. While ego-only reconstruction reveals clear geometric holes and blurring after static rendering, FRUC leverages collaborative views to seamlessly inpaint the missing background, exposing a realistic blind-spot view and preserving reliable ego-observed regions.

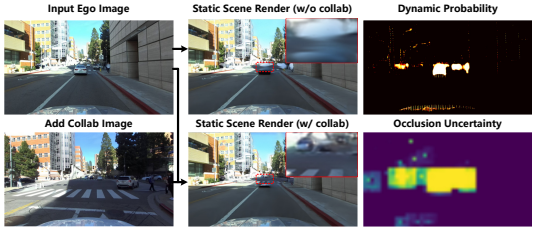


Figure 5: **Blind-spot completion on V2X-Real.**

Fig. 6 demonstrates how this capability supports object-level scene editing. Since FRUC explicitly decouples the reconstructed 3D space into static and dynamic representations, it naturally supports flexible rendering manipulations. Beyond simply rendering the global scene, our method enables fine-grained, object-level editing. As shown in the figure, by isolating specific dynamic Gaussians corresponding to a target occluder, we can selectively remove it from the scene. The framework then seamlessly fills the resulting "hole" by rendering the geometrically consistent static background, successfully achieving targeted blind-spot recovery utilizing the integrated collaborative context.

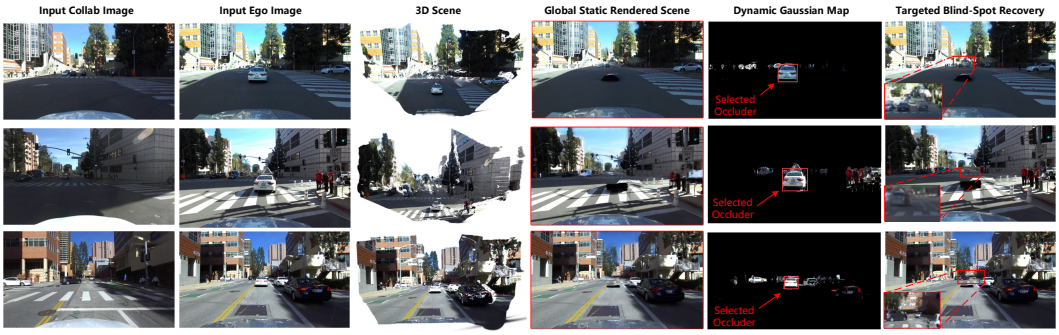


Figure 6: **More Scene Editing Results on V2X-Real.** We explicitly disentangle the 3D scene into global static and dynamic representations. By identifying specific dynamic Gaussians, our framework enables targeted blind-spot recovery utilizing collaborative context.

Table 3: **Ablation Studies on the V2X-Real Dataset. Best** in bold, and second-best underlined.

Components			Full Image			Dynamic-only			Blind-Spot
CALRD	COF	Aux.	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	(NIQE↓)
✗	✗	✓	21.94	0.799	0.138	24.71	0.985	0.098	4.181
✓	✗	✓	23.97	0.836	0.119	24.96	0.989	0.077	4.017
✓	✓	✗	22.76	0.801	0.121	24.92	0.987	0.084	4.309
✓	✓	✓	**25.89**	**0.883**	**0.076**	**25.93**	**0.994**	**0.007**	**3.949**

4.3 Ablation Studies

Effectiveness of Key Components. Table 3 show that CALRD and COF are the two core components. When both modules are removed, the performance drops most severely, indicating that naive multi-agent feature fusion cannot reliably resolve cross-agent geometric interference. When CALRD is preserved but COF is removed, the performance recovers only partially, which suggests that residual denoising alone is insufficient without a structured causal prior telling the model where collaborative completion is needed. In essence, COF provides the spatially localized, motion-aware occlusion guidance, while CALRD converts this guidance into stable ego-centric feature correction.

Effectiveness of Auxiliary Losses. Removing the auxiliary losses while keeping both CALRD and COF still produces a functional model, but the performance drops clearly, especially in blind-spot completion. This indicates that $\mathcal{L}_{\text{aux}}$ is not the source of the architectural gain itself. Instead, it serves as an optimization stabilizer that regularizes the occlusion prior, constrains the residual denoising trajectory, and encourages cooperative completion to evolve in a physically meaningful direction.

5 Conclusion

In this paper, we have introduced FRUC, a feed-forward 3DGS framework for dynamic scene reconstruction from uncalibrated collaborative driving views. To bypass the strict requirement for precise cross-agent calibration, FRUC elegantly re-formulates distributed multi-agent inputs as a spatio-temporally unstructured ego-centric multi-camera system. Specifically, we have proposed an ego-centric causal occlusion field to extract motion-aware spatial priors, and framed cross-agent feature fusion as a deterministic latent residual denoising process. This design effectively completes dynamic ego-blind spots without much corrupting reliable local observations, achieving state-of-the-art reconstruction quality and inference efficiency on V2X-Real and UrbanIng-V2X benchmarks.

Limitations and Future Work. Currently, our method may struggle in severe occlusion scenarios where weak cross-agent semantic correlation limits collaborative benefits. Besides, inaccurate dynamic probability maps can lead to failure cases, compromising the reliability of downstream scene editing. Future work will focus on enhancing the capacity of collaborative reconstruction in such severe occlusion scenarios and improving the robustness of dynamic modeling.

Acknowledgments. The research work described in this paper was conducted in the JC STEM Lab of Smart City funded by The Hong Kong Jockey Club Charities Trust under Contract 2023-0108. The work was supported in part by the Hong Kong SAR Government under the Global STEM Professorship and Research Talent Hub.

References

[1] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *CVPR*, 2024.

[2] Xiaoxue Chen, Ziyi Xiong, Yuantao Chen, Gen Li, Nan Wang, Hongcheng Luo, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Hongyang Li, Ya-Qin Zhang, and Hao Zhao. DGGT: Feedforward 4D reconstruction of dynamic driving scenes using unposed images. *arXiv preprint arXiv:2512.03004*, 2025. URL `https://arxiv.org/abs/2512.03004`.

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386, Cham, 2025.

[4] Lue Fan, Hao Zhang, Qitai Wang, Hongsheng Li, and Zhaoxiang Zhang. FreeSim: Toward free-viewpoint camera simulation in driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12004–12014, June 2025.

[5] Zhengru Fang, Senkang Hu, Haonan An, Yuang Zhang, Jingjing Wang, Hangcheng Cao, Xianhao Chen, and Yuguang Fang. PACP: Priority-Aware Collaborative Perception for Connected and Autonomous Vehicles. *IEEE Transactions on Mobile Computing*, 23(12):15003–15018, 2024.

[6] Zhengru Fang, Jingjing Wang, Yanan Ma, Yihang Tao, Yiqin Deng, Xianhao Chen, and Yuguang Fang. R-ACP: Real-time adaptive collaborative perception leveraging robust task-oriented communications. *IEEE Journal on Selected Areas in Communications*, 43(12):4215–4230, 2025. doi: 10.1109/JSAC.2025.3623179.

[7] Yu Guo, Zhengru Fang, Shengfeng He, Senkang Hu, Yihang Tao, Phone Lin, and Yuguang Fang. Universal image restoration via internalized chain-of-thought reasoning. *arXiv preprint arXiv:2606.17557*, 2026. URL `https://arxiv.org/abs/2606.17557`.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. URL `https://arxiv.org/abs/2006.11239`.

[9] Qirui Hou, Wenzhang Sun, Chang Zeng, Chunfeng Wang, Hao Li, and Jianxun Cui. DrivingScene: A multi-task online feed-forward 3D Gaussian splatting method for dynamic driving scenes. *arXiv preprint arXiv:2510.24734*, 2025. URL `https://arxiv.org/abs/2510.24734`.

[10] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: communication-efficient collaborative perception via spatial confidence maps. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4874–4886, Red Hook, NY, USA, 2022.

[11] Abhishek Dinkar Jagtap, Rui Song, Sanath Tiptur Sadashivaiah, and Andreas Festag. V2X-Gaussians: Gaussian splatting for multi-agent cooperative dynamic scene reconstruction. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 1033–1039, 2025. doi: 10.1109/IV64158.2025.11097436.

[12] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. AnySplat: Feed-forward 3D Gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023.

[14] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning Distilled Collaboration Graph for Multi-Agent Perception. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 29541–29552, 2021.

[15] Hao Lu, Tianshuo Xu, Wenzhao Zheng, Yunpeng Zhang, Wei Zhan, Dalong Du, Masayoshi Tomizuka, Kurt Keutzer, and Yingcong Chen. DrivingRecon: Large 4D Gaussian reconstruction model for autonomous driving. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 167933–167952. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f5717c76feff4f751604c0678c46627b-Paper-Conference.pdf.

[16] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024.

[17] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[19] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*, 2021. URL https://arxiv.org/abs/2103.13413.

[20] Karthikeyan Chandra Sekaran, Markus Geisler, Dominik Rößle, Adithya Mohan, Daniel Cremers, Wolfgang Utschick, Michael Botsch, Werner Huber, and Torsten Schön. UrbanIng-V2X: A large-scale multi-vehicle, multi-infrastructure dataset across multiple intersections for cooperative perception. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=iSwIkUqyqf.

[21] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4D: Efficient neural 4D decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[22] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[23] Yihang Tao, Senkang Hu, Zhengru Fang, and Yuguang Fang. Directed-CP: Directed collaborative perception for connected and autonomous vehicles via proactive attention. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7004–7010, 2025. doi: 10.1109/ICRA55743.2025.11127818.

[24] Yihang Tao, Senkang Hu, Haonan An, Zhengru Fang, Hangcheng Cao, and Yuguang Fang. Learning mutual view information graph for adaptive adversarial collaborative perception. *arXiv preprint arXiv:2602.19596*, 2026. URL https://arxiv.org/abs/2602.19596.

[25] Qijian Tian, Xin Tan, Yuan Xie, and Lizhuang Ma. DrivingForward: Feed-forward 3D Gaussian splatting for driving scene reconstruction from flexible surround-view input. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[26] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[27] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, June 2024.

[28] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2X-Real: A largs-scale dataset for vehicle-to-everything cooperative perception. In *European Conference on Computer Vision (ECCV) 2024*, pages 455–470, Berlin, Heidelberg, 2024. Springer-Verlag.

[29] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[30] Haolin Xiong, Sairisheek Muttukuru, Hanyuan Xiao, Rishi Upadhyay, Pradyumna Chari, Yajie Zhao, and Achuta Kadambi. SparseGS: Sparse view synthesis using 3D Gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, pages 1032–1041, 2025. doi: 10.1109/3DV66043.2025.00100.

[31] Haoran Xu, Saining Zhang, Peishuo Li, Baijun Ye, Xiaoxue Chen, Huan-Ang Gao, Jv Zheng, Xiaowei Song, Ziqiao Peng, Run Miao, Jinrang Jia, Yifeng Shi, Guangqi Yi, Hang Zhao, Hao Tang, Hongyang Li, Kaicheng Yu, and Hao Zhao. CRUISE: Cooperative reconstruction and editing in V2X scenarios using Gaussian splatting. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12518–12525, 2025. doi: 10.1109/IROS60139.2025.11246201.

[32] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[33] 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 *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ycv2z8TYur`.

[34] Jiawei Yang, Jiahui Huang, Boris Ivanovic, Yuxiao Chen, Yan Wang, Boyi Li, Yurong You, Apoorva Sharma, Maximilian Igl, Peter Karkus, Danfei Xu, Yue Wang, and Marco Pavone. STORM: Spatio-temporal reconstruction model for large-scale outdoor scenes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=M2NFWRPMUd`.

[35] Zhongrui Yu, Zhao Wang, Yijia Xie, Yida Wang, Xueyang Zhang, Yifei Zhan, and Kun Zhan. StreetForward: Perceiving dynamic street with feedforward causal attention. *arXiv preprint arXiv:2603.19552*, 2026. URL `https://arxiv.org/abs/2603.19552`.

[36] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. GS-LRM: Large reconstruction model for 3D Gaussian splatting. *European Conference on Computer Vision*, 2024.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[38] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. DrivingGaussian: Composite Gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024.

[39] Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. DVGT: Driving visual geometry transformer. *arXiv preprint arXiv:2512.16919*, 2025. URL `https://arxiv.org/abs/2512.16919`.

A More Implementation Details

A.1 Dataset Preparation and Utilization

Data Preparation. We unify V2X-Real [28] and UrbanIng-V2X [20] into a common OPV2V [32] format benchmark. The processed benchmark preserves the full original per-agent data layout, including RGB images, frame ids, camera parameters, ego poses, scene context files, and LiDAR files. On top of this structure, we generate sky masks and fine dynamic masks using an off-the-shelf SegFormer [29] trained on Cityscapes classes. We also establish cross-agent view-association metadata by coarsely estimating spatial overlap via LiDAR coordinates, camera poses, and Field of View (FOV) intersections, followed by manual verification. This metadata is specifically curated to filter out distant or non-overlapping ego-collaborative view pairs, ensuring the construction of a high-quality evaluation set. For UrbanIng-V2X, we additionally convert the raw multi-vehicle logs into the same file layout, remap camera ids to match the V2X-Real convention, and apply the same split protocol. Table 4 summarizes the resulting benchmark structure and clarifies the split sizes.

Data Utilization. For model inference, our pipeline strictly utilizes **only** RGB images and frame ids, which serve as the primary visual and temporal inputs, and all other signals are excluded from the model inputs. The sky masks and dynamic masks act as auxiliary supervision rather than inputs. The dynamic masks provide ground-truth supervision for the dynamic-head during training and are repurposed solely for masked metric computation during evaluation, while the sky masks assist training supervision and metric computation. To strictly preserve our fully uncalibrated testing paradigm, any global pose, LiDAR information, or view-association metadata is explicitly excluded from the model's inference inputs. Instead, the view-association metadata acts purely as an offline oracle exclusively during evaluation to construct valid collaborative benchmark pairs and determine the relevant collaborative view context for the ego vehicle. In this way, all auxiliary signals are confined to training supervision, benchmark construction, and metric computation, leaving the model to run on uncalibrated RGB images alone.

A.2 Model Training Details

During the Stage II Cross-Agent Adaptation, we explicitly supervise the model training with auxiliary losses. In real-world collaborative driving, blindly fusing multi-agent features often leads to geometric contamination and severe artifact generation due to uncalibrated errors. To mitigate this, our auxiliary losses are designed to establish an information-theoretic balance between single-agent reliability and collaborative completion. We categorize these auxiliary objectives into three functions: *Prior Stabilization*, *Cooperative Completion*, and *Ego-Manifold Regularization*.

❶ Prior Stabilization ($\mathcal{L}_{reg}$ and $\mathcal{L}_{var}$). The occlusion uncertainty prior $\mathbf{M}^{\tau}_{occ,k}$ dictates the routing of collaborative features. To prevent this prior from either expanding unbounded or collapsing to zero, we introduce two regularizers. First, the *Prior Consistency Loss* $\mathcal{L}_{reg}$ anchors the inferred occlusion mask to a dilated base dynamic footprint $\tilde{S}^{\tau}_{dyn,k} = \mathcal{D}_{\mathcal{K}}(S^{\tau}_{dyn,k})$ using an L1 penalty:

$$\mathcal{L}_{reg} = \|\mathbf{M}^{\tau}_{occ,k} - \tilde{S}^{\tau}_{dyn,k}\|_1. \tag{13}$$

Here, $\mathcal{D}_{\mathcal{K}}(\cdot)$ denotes the morphological dilation operator with a 2D square structuring element (kernel) $\mathcal{K}$ of a predefined size (e.g., 15×15 pixels). This operation expands the boundaries of the original dynamic mask $S^{\tau}_{dyn,k}$ to encompass its immediate surrounding context. This instills a conservative principle: the ego vehicle should fundamentally trust its local high-fidelity observations for static, visible regions, reserving collaborative feature aggregation strictly for dynamically occluded regions. Second, to encourage spatial diversity within the inferred mask and prevent constant-state collapse, we introduce a *Prior Dispersion Loss* $\mathcal{L}_{var}$:

$$\mathcal{L}_{var} = -\text{Var}(\mathbf{M}^{\tau}_{occ,k}), \tag{14}$$

where $\text{Var}(\cdot)$ denotes the spatial variance.

❷ Cooperative Completion ($\mathcal{L}_{coo}$). Reconstructing occluded backgrounds behind moving objects is an inherently ill-posed inverse problem if relying solely on single-agent temporal priors. To provide the explicit driving force for cross-view geometric completion, we design the *Cooperative Completion Loss* $\mathcal{L}_{coo}$. Since dynamic objects cast the most severe visual shadows, we define an active optimization target exclusively around their immediate boundaries. Specifically, we extract

Table 4: **Processed Dataset Summary.** Green checkmarks (✓) denote that the specific data modality is exposed to the model pipeline during the corresponding phase, whereas red crosses (✗) indicate it is strictly excluded. The sample counts represent the total number of distinct ego-collaborative view pairs in the dataset, reported in thousands (K).

Dataset	Split	Original Data						Added / Parsed Metadata			# Samples (K)
		RGB Images	Frame ID	Camera Params	Ego Pose	Scene Context	LiDAR Info	Sky Mask	Dynamic Mask	View Association	
V2X-Real	Train	✓	✓	✗	✗	✗	✗	✓	✓	✗	103.5
	Val	✓	✓	✗	✗	✗	✗	✓	✓	✓	29.1
UrbanIng-V2X	Train	✓	✓	✗	✗	✗	✗	✓	✓	✗	114.5
	Val	✓	✓	✗	✗	✗	✗	✓	✓	✓	44.6

the dynamic occlusion ring (or shadow boundary) $S^{\tau}_{bnd,k}$ by subtracting the base dynamic footprint from its dilated version: $S^{\tau}_{bnd,k} = \mathcal{D}_{\mathcal{K}}(S^{\tau}_{dyn,k}) \setminus S^{\tau}_{dyn,k}$. During forward rendering, we synthesize a background-only image $\mathcal{I}^{\tau}_{bg,k}$ by explicitly omitting dynamic Gaussians. By enforcing a photometric penalty between $\mathcal{I}^{\tau}_{bg,k}$ and the ground-truth image $\mathcal{I}^{\tau}_{gt,k}$ exclusively within this occlusion boundary $S^{\tau}_{bnd,k}$, we create an intense economic optimization pressure:

$$\mathcal{L}_{coo} = \frac{1}{|S^{\tau}_{bnd,k}|} \sum_{\mathbf{u} \in S^{\tau}_{bnd,k}} \|\mathcal{I}^{\tau}_{bg,k}(\mathbf{u}) - \mathcal{I}^{\tau}_{gt,k}(\mathbf{u})\|_1, \tag{15}$$

where $|S^{\tau}_{bnd,k}|$ is the total number of valid pixels in the occlusion boundary mask. This forces the network to expand $\mathbf{M}^{\tau}_{occ,k}$ to fetch and align features from collaborative agents to reconstruct these hidden geometries.

❸ **Ego-Manifold Regularization ($\mathcal{L}_{den}$).** Finally, to prevent catastrophic forgetting of the ego-only capabilities during multi-agent fine-tuning, we employ a *Ego-Reference Denoising Loss* $\mathcal{L}_{den}$. Inspired by latent denoising [8], it applies a global Mean Squared Error (MSE) objective on the latent space representations:

$$\mathcal{L}_{den} = \left\|\hat{\mathbf{F}}^{\tau}_{m,k} - \mathbf{F}^{\tau}_{e}\right\|_2^2. \tag{16}$$

Crucially, this loss is applied *globally* without spatial masking. This establishes a competitive dynamic: the global MSE acts as a strong structural prior pulling the entire feature map towards the safe ego-only baseline; meanwhile, the cooperative completion loss $\mathcal{L}_{coo}$ provides a counter-gradient, forcing the network to deviate from the ego baseline *only* in severely occluded regions where the collaborative features offer a significant rendering advantage that outweighs the MSE penalty. Additionally, to ensure ego-centric structural dominance during optimization, we employ an asymmetric loss weighting strategy that more prioritizes the ego vehicle's views (weight 1.0) over collaborative views (weight 0.1) across all photometric and perceptual penalties.

A.3 Baseline Implementation Details

Except for V2X-Gaussians [11], all comparison methods are originally designed for single-vehicle inputs. To ensure a fair comparison, we adapt them to the cooperative setting only through data reorganization, while keeping their backbone architectures, Gaussian prediction heads, rendering pipelines, and original training objectives unchanged. Concretely, each cooperative sample is converted into a unified multi-view temporal input centered on the ego agent, so that all methods receive the same cross-agent observations under a shared ego-centric reference frame.

For feed-forward multi-view methods such as AnySplat [12], MVSplat [3], and DGGT [2], the adaptation is limited to replacing the original single-vehicle data loader with a cooperative multi-view input interface. The source and target views are reorganized to match each method's expected input format, but the encoder-decoder structure, cost-volume construction, Gaussian decoding, and loss design remain exactly the same as in the original implementations. For temporal reconstruction methods such as STORM [34] and DrivingForward [25], we preserve their temporal modeling assumptions by treating the ego and collaborative observations as valid neighboring views within a short temporal window, again without modifying the core spatiotemporal architecture itself.

For scene-optimized baselines, the adaptation is performed at the scene level rather than at the network level. We export each cooperative sample into a unified scene representation that exposes synchronized multi-agent cameras and a shared ego-centric reference frame. V2X-Gaussians [11]

can then be trained on these exported scenes using its native cooperative formulation, while 3DGS [13] and EmerNeRF [33] operate on the same scene representation as conventional multi-view reconstruction baselines. In this way, the comparison focuses on each method's actual modeling capacity under identical cooperative observations, rather than on extra engineering changes to the original network design.

B More Experimental Results

B.1 Metric Definitions and Evaluation Modes

We evaluate only the ego target views during benchmarking. When the input sequence contains four frames (two timestamps from ego and collab), the script evaluates only the ego frames. Let $\mathcal{I}_{pred}$ denote the rendered image, $\mathcal{I}_{gt}$ the corresponding GT target view, and $\mathbf{M}$ a binary evaluation mask over the image plane, where masked entries are set to 1 and invalid entries are set to 0.

PSNR. PSNR measures pixel-level reconstruction fidelity through the mean squared error. For full-image evaluation, the MSE is computed over all pixels. For masked settings such as *Dynamic-only*, the binary mask is broadcast to all RGB channels, and the squared error is averaged strictly over valid masked entries. In Eq. (17), $\odot$ denotes element-wise multiplication, and $\sum \mathbf{M}$ counts the total number of valid RGB entries after mask broadcasting. Higher PSNR indicates better reconstruction quality.

$$\text{MSE}(\mathcal{I}_{pred}, \mathcal{I}_{gt}; \mathbf{M}) = \frac{\sum \left(\mathcal{I}_{pred} - \mathcal{I}_{gt}\right)^2 \odot \mathbf{M}}{\sum \mathbf{M}}, \qquad \text{PSNR} = -10 \log_{10}\big(\text{MSE}\big). \tag{17}$$

SSIM. SSIM evaluates local structural consistency instead of raw per-pixel error. Because an exact masked SSIM is difficult for patch-based metrics, the implementation adopts a region-focused approximation. Given a valid mask, the script first extracts the tight 2D bounding box of the masked region, expands it by 5 pixels on each side, and computes SSIM on the cropped patch when the crop is at least 11×11 pixels. In Eq. (18), $(x_{\min}, x_{\max}, y_{\min}, y_{\max})$ denote the horizontal and vertical bounds of the padded bounding box. Higher SSIM is better.

$$\text{SSIM} = \text{SSIM}(\mathcal{I}_{pred}[y_{\min}:y_{\max}, x_{\min}:x_{\max}], \mathcal{I}_{gt}[y_{\min}:y_{\max}, x_{\min}:x_{\max}]) \,. \tag{18}$$

This crop-based approximation is used to avoid the strong bias that zero padding would introduce in sparse masked regions.

LPIPS. LPIPS measures perceptual discrepancy in a deep feature space and is computed with an AlexNet-based LPIPS model in our script. Before evaluation, both images are linearly normalized from $[0, 1]$ to $[-1, 1]$, where $\tilde{\mathcal{I}}$ denotes the normalized image. For masked settings, the implementation extracts a 10-pixel padded bounding-box crop around the valid mask and computes LPIPS on that crop. In Eq. (19), $d_{\text{Alex}}(\cdot, \cdot)$ denotes the AlexNet-based perceptual distance used by LPIPS. Lower LPIPS indicates better perceptual similarity.

$$\tilde{\mathcal{I}} = 2\mathcal{I} - 1, \qquad \text{LPIPS} = d_{\text{Alex}}\Big(\tilde{\mathcal{I}}_{pred}, \tilde{\mathcal{I}}_{gt}\Big) \,. \tag{19}$$

Compared with direct masked zeroing, this cropped evaluation better reflects perceptual differences around dynamic objects.

NIQE. NIQE is used for *Blind-Spot Completion*, where the true background is physically invisible and thus no reference image exists after removing the dynamic occluder. In this setting, the script evaluates the background-only rendering $\mathcal{I}_{bg}$, obtained by dropping dynamic Gaussians, inside the blind-spot region. The blind-spot mask $\mathbf{M}_{bs}$ is derived from the dynamic-object mask $\mathbf{M}_{dyn}$ as

$$\mathbf{M}_{bs} = \text{Clamp}\Big(\mathcal{D}_{\mathcal{K}_{bs}}(\mathbf{M}_{dyn}) - \mathbf{M}_{dyn},\, 0,\, 1\Big), \tag{20}$$

where $\mathbf{M}_{dyn}$ marks the image region occupied by dynamic objects, $\mathcal{D}_{\mathcal{K}_{bs}}(\cdot)$ denotes morphological dilation with a kernel $\mathcal{K}_{bs}$ of size 21×21, and $\text{Clamp}(\cdot, 0, 1)$ truncates all values to the valid interval $[0, 1]$. Similar to LPIPS, NIQE is computed on a 10-pixel padded bounding-box crop around the valid blind-spot mask. In Eq. (21), the cropped region is determined by $\mathbf{M}_{bs}$ and the coordinates

Table 5: **Comparison of baselines in Single-Frame (SF) evaluation mode on V2X-Real.** The input context consists of both ego and collaborative views at t, and the task is to extrapolate the ego view at time $t+1$. **Best** in bold, and second-best underlined.

Methods	Full image			Dynamic-only			Blind-Spot (NIQE↓)
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
DrivingForward [25]	15.02	0.682	0.289	20.02	**0.986**	**0.017**	-
AnySplat [12]	19.04	0.603	0.224	16.60	0.979	0.031	-
DGGT [2]	18.02	0.585	0.226	16.57	0.978	0.209	6.021
FRUC (Ours)	**21.47**	**0.773**	**0.156**	**20.50**	0.979	0.022	**3.987**

$(x_{\min}, x_{\max}, y_{\min}, y_{\max})$ of its padded bounding box. Lower NIQE indicates that the completed region is more consistent with natural image statistics.

$$\text{NIQE} = \text{NIQE}(\mathcal{I}_{bg}[y_{\min}:y_{\max}, x_{\min}:x_{\max}])\,. \tag{21}$$

This design matches our blind-spot protocol, where only the plausibility of the completed hidden background can be assessed.

Modes. We report results under two temporal modes and three rendering settings. In *Single-frame (SF)* mode, the target view at $t+1$ is extrapolated from observations at t. In *Multi-frame (MF)* mode, the target view at t is interpolated from observations at $t-1$ and $t+1$. Spatially, *Full image* evaluates the complete rendered target view, *Dynamic-only* focuses on the image region occupied by moving objects, and *Blind-Spot Completion* focuses on the surrounding background region most strongly affected by dynamic occlusion.

B.2 More Quantitative Results

Single-frame (SF) Mode NVS Results. As discussed in the main text, we default to the multi-frame (MF) mode for our primary benchmark comparisons. For completeness, we additionally report the single-frame (SF) results for the baselines that can be fairly extended to SF evaluation. Specifically, the VGGT-based feed-forward methods, namely AnySplat, DGGT, and our FRUC, share the same predicted-pose extrapolation protocol: given the cooperative context at time t_0, the model first predicts the context camera parameters and scene representation, and the future ego target pose is then extrapolated from the predicted camera pose motion. DrivingForward is evaluated using its original native SF rendering branch under the same multi-agent input adaptation. The resulting comparison is shown in Table 5.

In SF evaluation mode, FRUC achieves the best overall performance, ranking first on all full-image metrics, Dynamic-only PSNR, and Blind-Spot NIQE. DrivingForward obtains the strongest Dynamic-only SSIM and LPIPS under its native SF branch, but its full-image fidelity remains clearly weaker, indicating that its future-view extrapolation is less stable at the scene level. By contrast, FRUC maintains consistently strong performance across both visible and occluded regions, which further supports its superior generalization and robustness in single-frame future-view prediction.

Number of Input Collaborative Views. To investigate the impact of collaborative information density, we compare DGGT and FRUC with different numbers of selected collaborative camera views on V2X-Real. In all settings, the ego input always contains two consecutive frames. $N_c = 0$ denotes the ego-only setting, where the model receives only the two ego frames. $N_c = 1$, which is also the default setting in the main paper, adds one associated collaborative camera across the same two timestamps, resulting in 2 ego images plus 2 collaborative images, i.e., 4 inputs in total. $N_c = 2$ further adds two associated collaborative cameras across the same two timestamps, resulting in 2 ego images plus 4 collaborative images, i.e., 6 inputs in total. Table 6 summarizes the results.

Compared with DGGT, FRUC better preserves the ego-only capability after full Stage II training, as reflected by its clearly stronger $N_c = 0$ performance across both full-image and dynamic-only metrics. Moreover, FRUC remains more robust as additional collaborative views are introduced: although the task becomes increasingly challenging with larger cross-agent context, our method consistently maintains higher reconstruction quality and substantially better blind-spot completion, indicating a stronger capacity to exploit varying numbers of collaborative observations without severely corrupting the ego-centric representation.

Table 6: **Effect of the Number of Input Collaborative Views on V2X-Real.** The ego input always contains two consecutive frames, while the number of collaborative inputs varies across settings.

Method	Inputs		Full image			Dynamic-only			Blind-Spot (NIQE↓)
	Ego	Collab	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
DGGT	2	0	24.80	0.838	0.179	27.48	0.995	0.105	4.938
	2	2	21.85	0.804	0.129	24.76	0.992	0.129	4.301
	2	4	20.77	0.773	0.229	23.12	0.991	0.161	4.185
FRUC (Ours)	2	0	27.02	0.893	0.006	27.93	0.995	0.005	4.985
	2	2	25.89	0.883	0.076	25.93	0.994	0.007	3.949
	2	4	23.33	0.839	0.117	24.54	0.993	0.008	3.936

C Theoretical Justification

In Section 3.3, we formulate cross-agent feature alignment as a constrained latent residual denoising process instantiated by the Cross-Agent Latent Residual Denoising (CALRD) module. In this section, we provide the formal theoretical justification for this design. We analyze the optimization bottleneck inherent in directly fine-tuning mixed latents under uncalibrated scenarios and mathematically demonstrate the convergence guarantees of our ego-conditioned residual formulation.

C.1 Optimization Bottleneck in Direct Latent Fine-Tuning

In the context of uncalibrated collaborative perception, a naive baseline directly aggregates multi-agent features to form a mixed latent $\tilde{\mathbf{F}}_m$ and subsequently fine-tunes the downstream Gaussian head Φ_{gs}. Specifically, the mixed latent can be formulated as a corrupted observation:

$$\tilde{\mathbf{F}}_m = \mathbf{F}_e + \boldsymbol{\epsilon}, \tag{22}$$

where $\mathbf{F}_e$ denotes the structurally reliable ego-centric feature and $\boldsymbol{\epsilon}$ represents the non-stationary geometric noise induced by the highly irregular and unstable cross-agent spatio-temporal context. From a statistical machine learning perspective, direct fine-tuning constitutes an empirical risk minimization (ERM) problem under severely noisy covariates:

$$\min_{\theta} \mathcal{J}(\theta) = \mathbb{E}_{(\mathbf{F}_e, \boldsymbol{\epsilon}, Y)} \left[\mathcal{L}\big(\Phi_{gs}(\mathbf{F}_e + \boldsymbol{\epsilon}), Y\big) \right], \tag{23}$$

where Y represents the target supervision signals (i.e., ground-truth images) and $\mathcal{L}$ is the rendering loss function.

To understand why optimizing this objective fails, we analyze the loss landscape via a second-order Taylor expansion around the clean ego-centric feature $\mathbf{F}_e$:

$$\mathcal{L}\big(\Phi_{gs}(\mathbf{F}_e + \boldsymbol{\epsilon}), Y\big) \approx \mathcal{L}\big(\Phi_{gs}(\mathbf{F}_e), Y\big) + \boldsymbol{\epsilon}^\top \nabla_{\mathbf{F}} \mathcal{L} + \frac{1}{2} \boldsymbol{\epsilon}^\top \mathbf{H}_{\mathbf{F}} \mathcal{L} \boldsymbol{\epsilon}, \tag{24}$$

where $\mathbf{H}_{\mathbf{F}}$ is the Hessian matrix of the loss with respect to the feature space. Because the uncalibrated cross-agent noise $\boldsymbol{\epsilon}$ exhibits high variance and lacks a consistent spatial structure, the expectation of the second-order noise penalty term $\mathbb{E}[\boldsymbol{\epsilon}^\top \mathbf{H}_{\mathbf{F}} \mathcal{L} \boldsymbol{\epsilon}]$ becomes excessively large.

To minimize the overall empirical risk $\mathcal{J}(\theta)$, the optimizer is mathematically forced to reduce this severe noise penalty. According to Lipschitz continuity, doing so requires the network Φ_{gs} to learn a highly constrained mapping with a significantly suppressed local curvature, implicitly flattening the feature manifold to accommodate $\boldsymbol{\epsilon}$. Consequently, the model struggles to disentangle the cooperative signals from the misaligned noise. In its attempt to suppress the Hessian term, the optimizer inevitably compromises the optimization of the primary ideal risk $\mathcal{L}\big(\Phi_{gs}(\mathbf{F}_e), Y\big)$, distorting the ego vehicle's reliable local geometry originally encoded in $\mathbf{F}_e$. This phenomenon manifests as *destructive semantic interference*, rendering the direct fine-tuning strategy mathematically unstable and practically infeasible.

C.2 The Philosophy of FRUC

To overcome the aforementioned optimization bottleneck, the FRUC architecture introduces a dedicated representation bottleneck, formalized as an ego-conditioned residual denoiser $\mathcal{R}_\phi$, situated

between the pre-trained feature backbone and the downstream Gaussian head Φ_{gs}. The core philosophy is to mathematically decouple the task of collaborative feature sanitization from the task of three-dimensional geometric decoding.

Instead of forcing the Gaussian head Φ_{gs} to directly map the out-of-distribution noisy latent $\tilde{\mathbf{F}}_m$ into a high-fidelity 3D Euclidean space, FRUC decomposes the original intractable empirical risk minimization into a constrained sequential optimization problem:

$$\min_{\phi,\theta} \mathbb{E}\left[\mathcal{L}\Big(\Phi_{gs}\big(\underbrace{\tilde{\mathbf{F}}_m + \mathcal{R}_\phi(\tilde{\mathbf{F}}_m, \mathbf{F}_e, \mathbf{M}_{occ})}_{\hat{\mathbf{F}}_m}\big), Y\Big)\right]. \tag{25}$$

Here, the residual mapping $\mathcal{R}_\phi$ acts as a manifold projection operator. By explicitly conditioning this projection on the uncorrupted ego reference $\mathbf{F}_e$, the network establishes a deterministic structural anchor $\mathcal{M}_{ego} = \{\mathbf{F} \mid \|\mathbf{F} - \mathbf{F}_e\|_2^2 \leq \delta\}$. The denoised latent representation $\hat{\mathbf{F}}_m$ is forcibly pulled back onto this reliable ego-centric manifold, thereby guaranteeing that the Gaussian head Φ_{gs} receives topologically coherent tokens devoid of severe geometric ambiguities.

Furthermore, the explicit derivation of the occlusion uncertainty prior $\mathbf{M}_{occ} \in [0, 1]$ injects a critical spatial inductive bias that fundamentally resolves the inherent ambiguity of cross-agent completion. This spatial modulator elegantly partitions the optimization domain into two disjoint subspaces: the reliable visible region $\Omega_{vis} = \{\mathbf{u} \mid \mathbf{M}_{occ}(\mathbf{u}) \to 0\}$ and the dynamically occluded blind spot $\Omega_{occ} = \{\mathbf{u} \mid \mathbf{M}_{occ}(\mathbf{u}) \to 1\}$. This allows the optimization of the residual $\mathcal{R}_\phi$ to be mathematically decoupled:

$$\mathcal{R}_\phi(\mathbf{u}) \to \begin{cases} -\boldsymbol{\epsilon}(\mathbf{u}), & \mathbf{u} \in \Omega_{vis} \quad \text{(Driven by } \mathcal{L}_{den}\text{)}, \\ \Delta\mathbf{F}_{coop}(\mathbf{u}), & \mathbf{u} \in \Omega_{occ} \quad \text{(Driven by } \mathcal{L}_{coo}\text{)}. \end{cases} \tag{26}$$

In unoccluded regions Ω_{vis}, the Ego-Reference Denoising Loss $\mathcal{L}_{den}$ mathematically constrains the residual mapping to approximate the negative noise $-\boldsymbol{\epsilon}$, effectively neutralizing the collaborative interference. Conversely, in dynamically occluded regions Ω_{occ}, the Cooperative Completion Loss $\mathcal{L}_{coo}$ provides a targeted photometric gradient, driving the residual to synthesize missing cooperative structures $\Delta\mathbf{F}_{coop}$.

Coupled with a strict zero-initialization strategy ($\mathcal{R}_\phi \to 0$ at $t = 0$), this formulation guarantees a well-conditioned Jacobian matrix $\nabla_\phi \hat{\mathbf{F}}_m \approx \mathbf{I}$ during early training. Consequently, the FRUC architecture transforms the intractable direct mapping problem into a bounded, spatially modulated residual optimization, providing the Gaussian decoder with a mathematically stable and geometrically disambiguated hypothesis space to achieve superior rendering fidelity.

C.3 The Effectiveness of Cross-Agent Latent Residual Denoising

The rationale and effectiveness of formulating the Cross-Agent Latent Residual Denoising (CALRD) module specifically as an ego-conditioned residual denoising process, rather than relying on conventional cross-attention fusion, can be rigorously formalized using the Information Bottleneck (IB) principle. The fundamental objective of CALRD is to extract an optimal latent representation $\hat{\mathbf{F}}_m$ that maximizes predictive power for the target view Y while minimizing the retention of irrelevant cross-agent geometric noise $\boldsymbol{\epsilon}$.

Traditional cross-attention mechanisms dynamically compute weighted combinations of multi-agent features. From an information-theoretic perspective, this operation expands the latent representation space, unintentionally maximizing the mutual information between the fused latent and the corrupted inputs. In highly uncalibrated scenarios, this expanded capacity becomes a liability, as the network is prone to memorizing spurious correlations.

In stark contrast, the CALRD paradigm explicitly acts as an information sink through a deterministic, single-step residual purification process. It is important to clarify that unlike generative Diffusion Models [7, 8], which learn to reverse a multi-step stochastic Gaussian noise process to synthesize novel data, CALRD addresses a structured geometric misalignment problem. By mathematically treating the aggregated cooperative context $\tilde{\mathbf{F}}_m = \mathbf{F}_e + \boldsymbol{\epsilon}$ as a noisy observation of the true geometric features, the CALRD module aims to optimize the Information Bottleneck Lagrangian:

$$\mathcal{L}_{IB} = -I(\hat{\mathbf{F}}_m; Y) + \beta \cdot I(\hat{\mathbf{F}}_m; \tilde{\mathbf{F}}_m), \tag{27}$$

where $I(\cdot;\cdot)$ denotes mutual information, and $\beta > 0$ is a trade-off multiplier. The first term encourages $\hat{\mathbf{F}}_m$ to retain task-relevant geometric structures for rendering Y, while the second term penalizes the complexity of the representation, forcing it to discard the non-stationary calibration noise $\boldsymbol{\epsilon}$ embedded within $\tilde{\mathbf{F}}_m$.

The structural design of the residual mapping $\mathcal{R}_\phi$ naturally enforces this compression. Because the mapping is strictly conditioned on the pure ego-reference $\mathbf{F}_e$, the mutual information can be decoupled:

$$I(\hat{\mathbf{F}}_m; \tilde{\mathbf{F}}_m) \approx I(\hat{\mathbf{F}}_m; \mathbf{F}_e) + I(\hat{\mathbf{F}}_m; \boldsymbol{\epsilon}). \tag{28}$$

The Ego-Reference Denoising Loss $\mathcal{L}_{den}$ explicitly minimizes the second component $I(\hat{\mathbf{F}}_m; \boldsymbol{\epsilon})$ by driving $\mathcal{R}_\phi(\mathbf{u}) \to -\boldsymbol{\epsilon}(\mathbf{u})$ in visible regions Ω_{vis}. Simultaneously, the Cooperative Completion Loss $\mathcal{L}_{coo}$ maximizes $I(\hat{\mathbf{F}}_m; Y)$ in the occluded blind spots Ω_{occ} by extracting necessary complementary structures $\Delta\mathbf{F}_{coop}$.

To ensure the optimization strictly evolves towards these constrained targets, CALRD follows the same dual-branch residual architecture described in Section 3.3. Specifically, the strong structural priors, namely the pure ego-reference $\mathbf{F}_e$ and the occlusion uncertainty mask $\mathbf{M}_{occ}$, are first encoded as

$$\mathbf{C}_{pri} = \mathcal{Z}_{pri}\Big(\Phi_{con}([\mathbf{M}_{occ}, \mathbf{F}_e])\Big), \tag{29}$$

and the residual correction is predicted as

$$\mathcal{R}_\phi(\tilde{\mathbf{F}}_m, \mathbf{F}_e, \mathbf{M}_{occ}) = \mathcal{Z}_{out}\Big(\Phi_{fea}(\tilde{\mathbf{F}}_m) + \mathbf{C}_{pri}\Big), \tag{30}$$

so that the final denoised feature becomes $\hat{\mathbf{F}}_m = \tilde{\mathbf{F}}_m + \mathcal{R}_\phi(\tilde{\mathbf{F}}_m, \mathbf{F}_e, \mathbf{M}_{occ})$. Here, Φ_{con} is the lightweight condition encoder, Φ_{fea} is the feature encoder, and $\mathcal{Z}_{pri}$ and $\mathcal{Z}_{out}$ are zero-initialized 1×1 convolutions. Because the weights and biases of the zero-convolutions are strictly initialized to zero, the initial state of the residual correction satisfies $\mathcal{R}_\phi \approx 0$, which implies $\hat{\mathbf{F}}_m \approx \tilde{\mathbf{F}}_m$ at the beginning of training.

This architectural design is crucial for stable conditional guidance. During the early phases of gradient descent, the zero-initialization acts as a safe exploration mechanism. The network initially preserves the raw mixed latent $\tilde{\mathbf{F}}_m$ without adding uncontrolled perturbations. As training progresses, the gradients $\nabla_\phi \mathcal{L}_{IB}$ incrementally activate the zero-convolutions, allowing the network to cautiously learn the complex non-linear interactions between the noisy latent $\tilde{\mathbf{F}}_m$ and the structural priors. This mechanism mathematically guarantees that the injection of $\mathbf{M}_{occ}$ and $\mathbf{F}_e$ acts as a deterministic, progressively strengthening regularizer, reliably guiding the residual denoising process $\mathcal{R}_\phi$ along the steepest descent path towards the ideal information bottleneck optimum, explaining the empirical superiority and training stability of CALRD over naive end-to-end multi-agent fine-tuning strategies.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction (Sec. 1) state the same core contributions that are later instantiated in Sec. 3, namely calibration-free collaborative reconstruction via the ego-centric causal occlusion field and CALRD, and these claims are supported by the quantitative and qualitative results in Sec. 4, Tables 1 and 2, and Figs. 3 and 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper explicitly discusses limitations in Sec. 5, where we state that FRUC still struggles in severe occlusion scenarios with weak cross-agent semantic correlation and can fail when dynamic probability estimation is inaccurate, and we align future work with these two specific issues.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical formulation in Sec. 3.3 is complemented by a dedicated appendix section, Appendix C, which states the optimization assumptions, introduces the information-bottleneck analysis, and provides the full derivation using numbered, cross-referenced equations.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the model design (Sec. 3), dataset processing details (Appendix A.1), baseline adaptation details (Appendix A.3), metric definitions and evaluation modes (Appendix B.1), and the complete benchmark settings and ablations in Tables 1, 2, and 3, which together specify the information needed to reproduce the reported conclusions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not provide anonymized public code or a complete reproduction package at submission time, but we plan to release the code, processed benchmark scripts, and instructions after publication so that the main results can be faithfully reproduced.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup in Sec. 4 specifies the backbone, image resolution, datasets, baselines, and evaluation protocol, while Appendix A.1, Appendix A.2, Appendix A.3, and Appendix B.1 further provide dataset preprocessing, training objectives, baseline adaptation details, and metric implementations needed to interpret the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Although we do not plot explicit error bars, the reported conclusions are supported by metrics averaged over all valid benchmark pairs under multiple settings, datasets, and ablations, including Tables 1, 2, 3, and the appendix tables for SF evaluation and collaborative-view scaling.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 4 states that our model is trained on 4 NVIDIA RTX 5880 GPUs at a resolution of 378×672, and Table 1 additionally reports the inference time of each compared method, providing the main compute information relevant to reproducing the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available driving datasets and standard benchmark baselines for scene reconstruction research, does not involve deception, human intervention, or privacy-sensitive annotation procedures, and the manuscript preserves anonymity and reports the method and evaluations in accordance with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts through safer and more complete collaborative driving reconstruction for simulation and what-if analysis in Sec. 1, while the limitations in Sec. 5 also make clear the negative consequences of failure cases in severe occlusion or inaccurate dynamic modeling for downstream scene editing and safety-critical use.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release a high-risk generative model or scraped dataset at submission time, and the proposed method is evaluated only as a research reconstruction system on existing driving benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper uses existing public datasets, backbones, and baselines with explicit citations throughout Sec. 1, Sec. 4, and Appendix A.3, but it does not itself redistribute those third-party assets or introduce new licensing terms in the submission package.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets as part of the anonymous submission; the checklist item will be updated accordingly when the planned post-publication code and benchmark release is prepared with documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or any research with human participants; all experiments are conducted on existing driving-scene datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects or crowdsourced participants, so no participant risk disclosure or IRB-style review was required for the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the proposed FRUC methodology, theoretical formulation, training pipeline, or evaluation protocol, so no declaration is required under the NeurIPS policy for core-method usage.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.