GAUSSIANCOLLAB: COMMUNICATION-EFFICIENT LATENT QUERY FUSION VIA TOKENIZED SEMANTIC GAUSSIANS FOR COLLABORATIVE OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative peception requires agents to exchange spatially grounded semantic evidence under constrained communication budgets. Existing approaches predominantly communicate dense bird's-eye-view or voxel features, which either collapse vertical geometry or incur the cost of dense 3D grids. We propose **GaussianCollab**, a collaborative perception framework that uses tokenized semantic Gaussians as the inter-agent communication medium. GaussianCollab communicates and fuses latent semantic queries anchored to explicit Gaussian geometry. This shifts Gaussian-based collaboration from attribute-only exchange to latent-query fusion, preserving explicit 3D structure for rigid transformation and geometric matching while retaining richer learned semantics. GaussianCollab performs geometry-guided matching and latent-query fusion over matched tokens, and decodes and stacks unmatched collaborator evidence. To control the resulting payload, we introduce **Semantic InConsistency (SIC)-guided selective packing** together with **Partially Embedded Query Compression (PEQC)**. Experiments on Semantic-OPV2V and Co3SOP demonstrate state-of-the-art performance on collaborative semantic occupancy prediction task. Under matched Gaussian capacity, the lowest-communication-volume operating points improve mIoU over the reproduced VOGS-L baseline by **2.95 and 3.47 points** while reducing communication volume by **7.5% and 6.9%** on Semantic-OPV2V and Co3SOP, respectively. We further evaluate robustness to token erasure and localization perturbations. These results support geometry-anchored latent Gaussian queries as an effective and communication-efficient fusion medium for vision-based collaborative occupancy prediction. Code will be made publicly available.

1 INTRODUCTION

Collaborative perception aggregates observations from multiple agents to reduce individual blind spots and extend sensing coverage (Hu et al., 2022; Yang et al., 2023). For collaborative semantic occupancy prediction specifically, under a limited communication budget, the exchanged representation must preserve both 3D geometry and task-relevant semantics while remaining alignable across coordinate frames and feasible for the communication budget.

Current multi-agent systems predominantly fuse intermediate features, yet their spatial representations remain constrained. Conventional grid-based methods either collapse height dimensions (e.g., bird's-eye view (Xu et al., 2022a)), incur the cost of dense 3D voxelization (Cao & De Charette, 2022), or lack explicit geometric primitives (e.g., tri-perspective views (Song et al., 2024)). Semantic Gaussians offer a sparse alternative with explicit 3D support, object-centric culling and lossless transformation with rigid-body primitives. VOGS-CP (Chen et al., 2025a) adapts them to collaborative semantic occupancy prediction but communicates decoded Gaussian attributes alone, discarding the learned queries associated with those primitives. GaussianCollab instead binds latent semantic queries to transformable Gaussian geometry and fuses them across agents, enabling geometry-guided spatial querying while retaining richer task-specific semantics. With the basis of achieved latent space

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

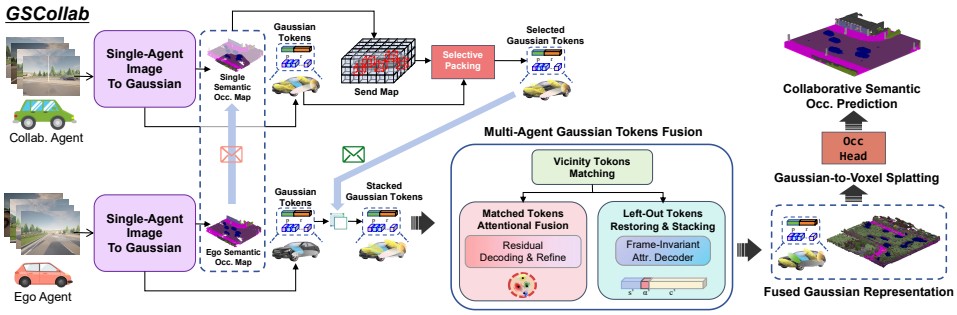


Figure 1: Overview of GaussianCollab. Tokenized semantic Gaussians are generated locally and selectively exchanged as a geometry-anchored latent communication medium for collaborative occupancy prediction.

collaborative fusion, adapting this communication medium to language, action, video, temporal prediction, or further task heads can be valuable future research directions to expand on.

We build on semantic Gaussians from previous work (Huang et al., 2024). Instead of modeling view-dependent color emissions with spherical harmonics, as commonly done in 3D Gaussian Splatting, semantic Gaussians embed task-specific latent queries into geometric primitives. We formulate this structure as **Gaussian tokens**, which couple explicit spatial anchors with learned semantic queries and serve as a finite set of communication units. Latent queries contain richer task evidence than decoded attributes, but they do not by themselves provide cross-agent spatial correspondence. Their use in collaboration therefore requires explicit geometric alignment, association of two sparse primitive sets, and different treatment of overlapping and collaborator-only evidence. Their high dimensionality also increases the raw payload.

Figure 1 summarizes **GaussianCollab**, our collaborative perception framework built upon tokenized semantic Gaussians. We redesign the single-agent pipeline to decouple pose-dependent and pose-invariant Gaussian information and compress the resulting partially embedded queries for transmission. For the multi-agent phase, we develop an SIC-guided packing strategy that identifies and broadcasts tokens from regions where collaborator predictions supplement or conflict with the ego prediction, as determined by comparing splatted semantic label maps across agents. At the receiver, explicit Gaussian geometry supplies the correspondence that latent queries lack: a spatial-vicinity graph associates incoming tokens with local tokens. Matched latent queries are fused by attention, while decoded unmatched tokens are stacked to preserve collaborator-only evidence.

Our main contributions are:

- We introduce a geometry-anchored latent-query communication medium for collaborative occupancy prediction. Each token couples transformable Gaussian geometry, which supports cross-agent correspondence, with a task-specific latent semantic query.
- We instantiate this medium with a collaboration protocol that geometrically matches sparse token sets, fuses matched latent queries, and decodes and stacks unmatched collaborator tokens. The two paths handle overlapping and collaborator-only evidence, respectively.
- We introduce SIC-guided selection and PEQC to reduce spatial and channel redundancy. Across two benchmarks, the operating points with the lowest communication volume (CV) improve mIoU while using less communication than the reproduced VOGS-L baseline.

2 Related Work

Consistency dimensions for spatial intelligence. Building general world models capable of understanding and simulating the physical world is a central goal of spatial intelligence. This goal can be examined through modal, temporal, and spatial consistency (Wei et al., 2026). Modal consistency concerns alignment of heterogeneous information (Radford et al., 2021; Liu et al., 2023; Esser et al., 2024; Tian et al., 2025). Temporal consistency concerns causal evolution (Guo et al., 2023;

Kondratyuk et al., 2023; Kong et al., 2024). Spatial consistency concerns construction and preservation of 3D geometry (Kerbl et al., 2023; Mildenhall et al., 2021; Poole et al., 2022). Multi-agent observations provide complementary geometric constraints, motivating a compact representation that can be transformed and fused across viewpoints. In this work, we address only the spatial-consistency component through vision-based collaborative semantic occupancy. Modal and temporal consistency are outside the present experimental scope and remain future extensions.

Collaborative perception. Collaborative perception often favors **intermediate fusion** (Hu et al., 2022; Yang et al., 2023) for its accuracy–bandwidth balance. The underlying representation remains a bottleneck: **BEV** (Xu et al., 2022a) compresses height, **dense voxels** (Cao & De Charette, 2022) contain substantial spatial redundancy, and **TPV** in CoHFF (Song et al., 2024) lacks explicit 3D primitives. MAC-Ego (Xu et al., 2024) studies collaborative Gaussian mapping for indoor SLAM, while VOGS-CP (Chen et al., 2025a) pioneers Gaussian-based collaborative occupancy but transmits decoded physical attributes rather than the associated latent queries. Tokenized semantic Gaussians instead bind learned semantic queries to explicit, transformable 3D geometry, providing a richer communication medium for the occupancy setting evaluated here. Whether the same carrier benefits temporal or multimodal alignment remains to be established empirically.

Semantic Gaussian splatting for semantic occupancy prediction. Standard 3D Gaussian representations parameterize view-dependent appearance for reconstruction, rendering, and generation (Kerbl et al., 2023; Jagtap et al., 2025; Chen et al., 2025b; Yang et al., 2026; Xiong et al., 2025; Gao et al., 2025a; Deng et al., 2026; Lan et al., 2025; Gao et al., 2025b). Semantic-Gaussian occupancy models instead associate semantic logits with adaptive spatial support; GSFormer (Huang et al., 2024) additionally couples each primitive with a latent query for single-agent occupancy. VOGS-CP (Chen et al., 2025a) communicates decoded semantic-Gaussian attributes for collaborative occupancy. GaussianCollab communicates geometry-anchored latent semantic queries and performs learned latent-query fusion. This richer medium increases raw payload, motivating SIC-based spatial selection and PEQC-based channel compression. We evaluate this accuracy–bandwidth trade-off for collaborative semantic occupancy.

3 Problem Formulation

3.1 General Pipeline of Collaborative Perception

We consider a multi-agent collaborative perception system with N agents, defined as $\mathcal{A} = \{a_i\}_{i=1}^{N}$. Each agent a_i observes a spatially bounded local region that may overlap with others. Collaboration is restricted to a communication range τ where observation fields intersect. For an agent a_i, the input comprises a local multi-view image set $\mathcal{I}_i$ and shared messages $\{\mathcal{M}_{j\to i}\}$ from collaborators. The vision encoder f_{vis} maps these images to an intermediate feature set $\mathcal{F}_i$. The objective is to generate a task-specific scene prediction $\widehat{\mathcal{Y}}_i$ supervised by ground-truth labels $\mathcal{Y}_i$.

The overall optimization goal of the system is to maximize perception performance $\mathcal{P}_\theta$ under a total communication budget B formulated as:

$$\max_{\theta} \sum_{i=1}^{N} \mathcal{P}_\theta\big(\widehat{\mathcal{Y}}_i, \mathcal{Y}_i\big) \quad \text{s.t.} \quad \sum_{i=1}^{N} \sum_{j\in\mathcal{N}(i)} \mathcal{C}\big(\mathcal{M}_{i\to j}\big) \le B. \tag{1}$$

Appendix A.1 gives an expanded formulation.

3.2 Spatial Representation via Semantic Gaussians

Our framework employs semantic Gaussians as the intermediate representation $\mathcal{F}_i$, enabling a tokenized exchange for perception and reasoning. A scene is represented by a set $\mathcal{G}_i = \{g_{i,k}\}_{k=1}^{K_i}$, where each primitive $g_{i,k}$ is defined by:

$$g_{i,k} = (\mathbf{p}_{i,k}, \mathbf{r}_{i,k}, \mathbf{s}_{i,k}, \alpha_{i,k}, \mathbf{c}_{i,k}, \mathbf{q}_{i,k}). \tag{2}$$

Specifically, $\mathbf{p}_{i,k} \in \mathbb{R}^3$ is the spatial center, $\mathbf{r}_{i,k}$ is a rotation quaternion, and $\mathbf{s}_{i,k} \in \mathbb{R}^3$ specifies anisotropic scale. The scalar $\alpha_{i,k}$ represents opacity, while $\mathbf{c}_{i,k}$ and $\mathbf{q}_{i,k}$ carry semantic logits and task-relevant latent content, respectively.

Physically, a primitive centered at $\mathbf{p}$ defines a 3D Gaussian distribution: $G(\mathbf{x}) = \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{p})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{p}))$, with covariance $\mathbf{\Sigma} = \mathbf{RSS}^\top \mathbf{R}^\top$. Unlike 3DGS methods optimized for rendering, our primitives serve as **task-oriented scene tokens**. They maintain sparse, adaptive, and explicit geometric support, allowing agents to align and fuse observations without transmitting dense tensors.

4 GAUSSIANCOLLAB

4.1 FROM IMAGES TO SEMANTIC GAUSSIAN REPRESENTATIONS

Given multi-view observations $\mathcal{I}$, each agent extracts multi-scale image features $\mathcal{F}_{\text{MS}} = \{\mathbf{F}^n \in \mathbb{R}^{N_v \times C \times H^n \times W^n}\}$ via a backbone f_{backbone}. These features are lifted and iteratively refined into semantic Gaussians $\mathcal{G}$ via f_{gs}:

$$\mathcal{G} = f_{\text{gs}}(\mathcal{F}_{\text{MS}}, \mathcal{G}_{\text{init}}). \tag{3}$$

We introduce three modifications to f_{gs} for tokenized collaboration; Appendix A.2 illustrates the resulting pipeline:

Decoupled anchor encoder To provide pose-aware and pose-invariant inputs for downstream multi-agent fusion, we encode the two attribute groups with independent MLPs that do not share parameters. The frame-variant attributes $(\mathbf{p}, \mathbf{r})$ and frame-invariant attributes $(\mathbf{s}, \alpha, \mathbf{c})$ are mapped to F^{var} and F^{inv}, respectively:

$$F^{\text{var}} = \text{MLP}_{\text{var}}(\mathbf{p}, \mathbf{r}), \quad F^{\text{inv}} = \text{MLP}_{\text{inv}}(\mathbf{s}, \alpha, \mathbf{c}). \tag{4}$$

The *partially embedded query* is $\mathbf{q}^{\text{part}} = \mathbf{q}^{\text{ref}} + F^{\text{inv}}$, where $\mathbf{q}^{\text{ref}}$ is the final refined Gaussian query. This separation supplies the collaborative module with a pose-invariant token component while leaving pose-dependent attributes available for explicit geometric synchronization.

Gaussian-query compression A lightweight compressor–decompressor MLP bottleneck projects $\mathbf{q}^{\text{part}}_{i,k}$ into a compact code $\tilde{\mathbf{q}}^{\text{part}}_{i,k}$ for transmission and reconstructs it as $\hat{\mathbf{q}}^{\text{part}}_{i,k}$ at the receiver. We supervise this bottleneck via a reconstruction loss $\mathcal{L}_{\text{recon}}$:

$$\mathcal{L}_{\text{recon}} = \frac{1}{K_i d_q} \sum_{k=1}^{K_i} \left\| \mathbf{q}^{\text{part}}_{i,k} - \hat{\mathbf{q}}^{\text{part}}_{i,k} \right\|_2^2, \tag{5}$$

where $\hat{\mathbf{q}}^{\text{part}}_{i,k}$ is the reconstructed partially embedded query at the receiver side and d_q is its feature dimension.

Frame-invariant attribute decoding (FIAD) The shared decoder MLP_{dec} reconstructs the raw frame-invariant anchor parameters from a decompressed query conditioned on the pose embedding $F^{\text{var}}_{i,k}$. Let $\mathbf{a}^{\text{inv}}_{i,k}$ collect the original raw scale, opacity, and semantic parameters, and let $\bar{\mathbf{a}}^{\text{inv}}_{i,k}$ denote their reconstruction. The corresponding physical Gaussian attributes are obtained through the same parameter-to-attribute mappings Φ used by the Gaussian predictor:

$$\bar{\mathbf{a}}^{\text{inv}}_{i,k} = \text{MLP}_{\text{dec}}\left(\hat{\mathbf{q}}^{\text{part}}_{i,k} + F^{\text{var}}_{i,k}\right), \qquad \{\mathbf{s}'_{i,k}, \alpha'_{i,k}, \mathbf{c}'_{i,k}\} = \Phi\left(\bar{\mathbf{a}}^{\text{inv}}_{i,k}\right). \tag{6}$$

Here, Φ applies the sigmoid-based opacity and bounded-scale mappings and the softplus semantic mapping. To supervise MLP_{dec}, the frame-invariant attribute decoding loss $\mathcal{L}_{\text{attr}}$ applies MSE in the raw anchor-parameter space, matching the implementation:

$$\mathcal{L}_{\text{attr}} = \frac{1}{K_i d_a} \sum_{k=1}^{K_i} \left\| \mathbf{a}^{\text{inv}}_{i,k} - \bar{\mathbf{a}}^{\text{inv}}_{i,k} \right\|_2^2, \tag{7}$$

where $d_a = \dim(\mathbf{a}^{\text{inv}})$ is the number of reconstructed frame-invariant anchor parameters per Gaussian. During fusion, FIAD is reused to restore only the received tokens that remain unmatched and are selected for stacking; matched tokens instead pass through attention and the shared refinement decoder.

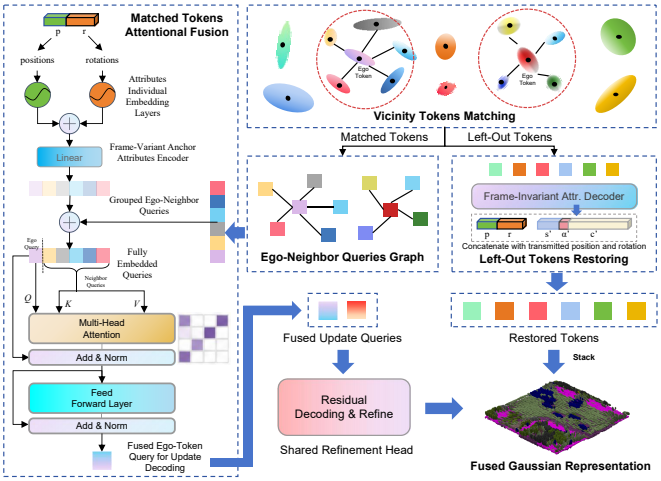


Figure 2: Multi-Agent Gaussian-Token Fusion. Spatially synchronized collaborator tokens are geometrically matched to ego tokens through a vicinity graph. Matched tokens update ego tokens through attentional aggregation, while unmatched collaborator tokens are decoded and stacked to construct a unified, spatially augmented representation.

4.2 SIC-Aware Communication

Gaussian tokenization Unlike dense grids, our framework communicates a finite set of geometry-anchored latent tokens. Each token carries a compressed real-valued feature code rather than a discrete codebook index. Each primitive $g_{i,k}$ is packed as a *Gaussian token* $t_{i,k}$:

$$t_{i,k} = \{\mathbf{p}_{i,k}, \mathbf{r}_{i,k}, \tilde{\mathbf{q}}^{\text{part}}_{i,k}\}. \tag{8}$$

To preserve the raw observation perspective, $\mathbf{p}_{i,k}$ and $\mathbf{r}_{i,k}$ are transmitted in the agent's **local coordinate system**. At the receiver, the compact code is first decompressed into $\hat{\mathbf{q}}^{\text{part}}_{j,k}$. Spatial synchronization and transformation $\mathcal{T}_{j\to i}$ then produce the *fully embedded query*:

$$\hat{\mathbf{q}}^{\text{full}}_{j,k} = \text{Embed}\big(\mathcal{T}_{j\to i}(\mathbf{p}_{j,k}, \mathbf{r}_{j,k})\big) + \hat{\mathbf{q}}^{\text{part}}_{j,k}. \tag{9}$$

The ego query is embedded analogously as $\mathbf{q}^{\text{full}}_{i,m} = \text{Embed}(\mathbf{p}_{i,m}, \mathbf{r}_{i,m}) + \mathbf{q}^{\text{part}}_{i,m}$. Appendix A.4 defines the spatial synchronization and transformation $\mathcal{T}_{j\to i}$.

Selective token packing We prioritize tokens in regions of high semantic discrepancy. Each agent splats its final single-agent Gaussian prediction, before PEQC, into a lightweight semantic occupancy map $\hat{\mathcal{Y}}_i$. Collaborators compute a semantic discrepancy mask $\mathbf{D}_{j\to i}$ by comparing the ego-provided map with the map splatted from their own pre-compression Gaussians after transformation into the ego frame:

$$\mathbf{D}_{j\to i} = \mathbb{I}\left(\text{argmax}(\hat{\mathcal{Y}}_i) \neq \text{argmax}(\hat{\mathcal{Y}}_{j\to i})\right). \tag{10}$$

The ego semantic label map required for this comparison uses the 14-state alphabet and is transmitted losslessly with four bits per voxel, packing two labels into each byte. The collaborator compares this map with its own transformed semantic label map to derive the binary SIC mask $\mathbf{D}_{j\to i}$. This mask depends on the pre-compression Gaussian predictions rather than occupancy reconstructed from decompressed queries and identifies voxels where the collaborator conflicts with or supplements the ego prediction. Only tokens located in these selected regions are packed into $\mathcal{M}_{j\to i}$. After ego-map splatting, final collaborative splatting could in principle be restricted to regions receiving new Gaussians; this regional optimization is not implemented in the current system and is left for future latency reduction.

4.3 Multi-Agent Gaussian-Token Fusion

Spatial-vicinity-based token matching As summarized in Figure 2, upon receiving and unpacking the message $\mathcal{M}_{j\to i}$, the ego agent a_i extracts a set of N_j collaborator Gaussian tokens, denoted as

$\{t_{j,n}\}_{n=1}^{N_j}$, where the index n uniquely identifies a transmitted token from collaborator a_j. Concurrently, the ego agent maintains its own set of M_i local Gaussian tokens, denoted as $\{t_{i,m}\}_{m=1}^{M_i}$, where the index m identifies an ego primitive.

To establish a geometric correspondence between the ego's local perception and the collaborators' observations, we construct a spatial-vicinity matching graph. Specifically, we compute the pairwise Euclidean distance between the spatially synchronized centers of the collaborator tokens and the ego tokens. A collaborator token $t_{j,n}$ is matched to an ego token $t_{i,m}$ if their spatial centers fall within a predefined local radius ρ. Formally, this constructs a directed bipartite graph whose edge set $\mathcal{E}_{i\leftarrow j}$ is defined as:

$$\mathcal{E}_{i\leftarrow j} = \left\{(m,n) \mid \|\mathbf{p}_{i,m} - \mathcal{T}_{j\to i}(\mathbf{p}_{j,n})\|_2 \le \rho\right\} \tag{11}$$

where $\mathbf{p}_{i,m}$ and $\mathcal{T}_{j\to i}(\mathbf{p}_{j,n})$ are the center coordinates in the ego agent's reference frame. For computational efficiency, the implementation retains at most 128 matched collaborator tokens per ego token. This radius rule accommodates variations in Gaussian density by grouping nearby collaborator queries with each ego query before attention.

Attention-based Gaussian-token aggregation To resolve spatial redundancies and integrate complementary evidence, we employ Multi-Head Attention (MHA) for query fusion. For each fully embedded ego query $\mathbf{q}_{i,m}^{\text{full}}$, let $\mathbf{Q}_{i,m}^{\text{nei}} = \{\hat{\mathbf{q}}_{j,n}^{\text{full}} \mid j \in \mathcal{N}(i), (m,n) \in \mathcal{E}_{i\leftarrow j}\}$ denote the row-stacked matrix whose rows are its sparse matched, fully embedded neighbor queries. The attentional fusion is written as:

$$\begin{aligned}\mathbf{q}'_{i,m} &= \mathrm{MHA}\big(\mathbf{q}_{i,m}^{\text{full}}, \mathbf{Q}_{i,m}^{\text{nei}}, \mathbf{Q}_{i,m}^{\text{nei}}\big) \\ &= \mathrm{Concat}_{h=1}^{H}\left[\mathrm{Softmax}\left(\frac{(\mathbf{q}_{i,m}^{\text{full}}\mathbf{W}_h^Q)(\mathbf{Q}_{i,m}^{\text{nei}}\mathbf{W}_h^K)^\top}{\sqrt{d_h}}\right)(\mathbf{Q}_{i,m}^{\text{nei}}\mathbf{W}_h^V)\right]\mathbf{W}^O.\end{aligned} \tag{12}$$

Here, the ego query supplies the attention query, while its spatially matched neighbor queries supply both the keys and values. For each ego query with at least one match, the softmax is normalized over the rows of its sparse matched-neighbor matrix; all such ego queries are processed without constructing a dense attention mask. H is the number of attention heads, d_h is the dimension of each head, and $\mathbf{W}^O$ is the learned output projection applied after concatenating the head outputs.

The expression above abstracts the attentional fusion block by its core MHA operation. The resulting attention-fused query representation $\mathbf{q}'_{i,m}$ is processed by the reused refinement module from the iterative block of the single-agent image-to-Gaussian model to predict updates and refreshed attribute values for the ego Gaussians:

$$\{\Delta\mathbf{p}_{i,m}, \mathbf{r}_{i,m}^{\text{ref}}, \mathbf{s}_{i,m}^{\text{ref}}, \alpha_{i,m}^{\text{ref}}, \mathbf{c}_{i,m}^{\text{ref}}\} = \mathrm{MLP}_{\text{refine}}(\mathbf{q}'_{i,m}) \tag{13}$$

The spatial position is updated via residual addition, whereas the rotation, scale, and opacity are directly replaced by the refined values and the semantics are updated through a semantic blending mechanism. These update rules follow VOGS-CP (Chen et al., 2025a), specifically:

$$\{\mathbf{p}_{i,m}^{\text{upd}}, \mathbf{r}_{i,m}^{\text{upd}}, \mathbf{s}_{i,m}^{\text{upd}}, \alpha_{i,m}^{\text{upd}}, \mathbf{c}_{i,m}^{\text{upd}}\} = \{\mathbf{p}_{i,m} + \Delta\mathbf{p}_{i,m}, \mathbf{r}_{i,m}^{\text{ref}}, \mathbf{s}_{i,m}^{\text{ref}}, \alpha_{i,m}^{\text{ref}}, \gamma\mathbf{c}_{i,m} + (1-\gamma)\mathbf{c}_{i,m}^{\text{ref}}\}, \tag{14}$$

where γ is a confidence weight for blending old semantics with new predicted semantics. Appendix A.3 defines γ.

By grouping these updated attributes with the fused latent features, we obtain the refined ego Gaussian primitive, denoted as $g_{i,m}^{\text{upd}} = (\mathbf{p}_{i,m}^{\text{upd}}, \mathbf{r}_{i,m}^{\text{upd}}, \mathbf{s}_{i,m}^{\text{upd}}, \alpha_{i,m}^{\text{upd}}, \mathbf{c}_{i,m}^{\text{upd}}, \mathbf{q}'_{i,m})$. By attending over geometry-matched latent Gaussian queries, this fusion performs collaborative spatial querying directly in native 3D space, allowing each ego Gaussian to integrate relevant semantic evidence from nearby collaborator Gaussians while retaining its explicit spatial anchor.

Decoding and stacking unmatched tokens To expand the perception field, we incorporate collaborator tokens that lack direct matches in the ego's set. We first recover their full physical state by decoding invariant attributes through the shared $\mathrm{MLP}_{\text{dec}}$. The restored neighbor Gaussian $g_{j,n}^{\text{rest}}$ is constructed by concatenating these decoded attributes with the spatially synchronized pose:

$$g_{j,n}^{\text{rest}} = \{\mathcal{T}_{j\to i}(\mathbf{p}_{j,n}), \mathcal{T}_{j\to i}(\mathbf{r}_{j,n}), \Phi(\mathrm{MLP}_{\text{dec}}(\hat{\mathbf{q}}_{j,n}^{\text{part}} + F_{j\to i,n}^{\text{var}}))\}, \tag{15}$$

where $F_{j\to i,n}^{\text{var}} = \mathrm{MLP}_{\text{var}}(\mathcal{T}_{j\to i}(\mathbf{p}_{j,n}), \mathcal{T}_{j\to i}(\mathbf{r}_{j,n}))$ conditions attribute recovery on the synchronized pose. Finally, the restored unmatched Gaussians are appended to the refined ego set, preserving collaborator evidence that has no local match: $\mathcal{G}_{\text{fused}} = \{g_{i,m}^{\text{upd}}\} \cup \{g_{j,n}^{\text{rest}}\}$.

Table 1: Semantic-OPV2V semantic occupancy results. All reproduced Gaussian methods use 12,800 primitives. GSCollab-112N is the highest-mIoU endpoint; GSCollab-112T is the lowest-CV endpoint. N disables SIC during both training and inference; T enables SIC during both. CoHFF values are paper-reported; the remaining values are from the matched evaluation files. Among the displayed methods, best and second-best results within each perception setting are **bolded** and underlined.

Metric / Class	Single-Agent			Collaborative Perception				
	CoHFF	GSFormer	GSFormer-M (96-dim)	CoHFF	Late Fusion	VOGS-L	GSCollab-112 (N)	GSCollab-112 (T)
IoU (↑)	38.52	67.35	**69.39**	50.46	51.71	71.54	**73.81**	72.48
mIoU (↑)	24.81	30.06	**35.30**	34.16	22.57	38.90	**43.14**	41.85
CV (MiB) (↓)	–	–	–	–	–	0.941	5.478	**0.870**
Building	**21.04**	7.65	7.21	**25.72**	4.62	4.09	7.58	7.58
Fence	20.05	16.16	**25.30**	27.83	7.95	30.35	**31.88**	30.42
Terrain	43.93	60.83	**71.20**	48.30	39.81	76.83	80.02	**80.18**
Pole	**31.66**	6.51	10.65	**42.74**	7.10	11.61	17.18	16.52
Road	55.83	79.97	**81.98**	61.77	64.81	83.85	**84.19**	82.13
Sidewalk	17.31	68.10	**69.78**	39.62	51.62	75.32	**77.18**	76.15
Vegetation	14.49	16.66	**21.58**	20.59	3.58	25.27	**26.61**	25.94
Vehicles	**58.55**	50.98	57.41	63.28	42.85	60.93	**66.39**	64.32
Wall	33.30	34.84	**36.81**	**58.27**	7.54	38.41	37.74	37.47
Guard rail	1.54	19.00	**19.60**	1.94	**39.81**	31.12	38.20	39.33
Traffic signs	0.00	0.00	**8.00**	**16.33**	1.10	12.06	14.88	13.23
Bridge	0.00	0.00	**14.07**	3.53	0.00	16.95	**35.81**	28.95

5 Experiments

5.1 Datasets and Implementation Details

We evaluate on Semantic-OPV2V (Song et al., 2024) and Co3SOP (Wu et al., 2025); Appendix A.5 describes both datasets.

In our default configuration, each agent uses $|\mathcal{G}| = 12{,}800$ Gaussian primitives and the matching radius is $\rho = 0.4$ m. The occupancy label space contains 14 states: 12 evaluated semantic classes, one empty class, and one unlabeled class excluded from metric computation. The 14-state alphabet therefore admits lossless 4-bit serialization. Table suffixes 64/80/96/112 denote PEQC output width; GSCollab-128 omits PEQC and is not a zero-compression PEQC-128 model. Further settings appear in Appendix A.6.

5.2 Evaluation Metrics

For both single- and multi-agent settings, we primarily use Mean Intersection over Union (mIoU) and IoU to evaluate per-class semantic occupancy and overall geometric distribution, respectively. All IoU, mIoU, and average precision (AP) results are reported on the 0–100 percentage scale; their absolute differences are stated in points, whereas percent signs denote relative changes. For Co3SOP, we additionally report results over visible and invisible regions; Appendix A.7 defines these regions in detail. Additionally, following previous works (Chen et al., 2025a; Song et al., 2024), we evaluate BEV segmentation by projecting 3D voxels onto the 2D BEV plane to compute 2D IoU for vehicles, roads, and general objects. Finally, collaboration efficiency is measured by the average CV (in MiB) per collaborative frame.

5.3 Quantitative Evaluation

Benchmark comparison. Table 1 reports two endpoints of the Semantic-OPV2V sweep under the same 12,800-Gaussian capacity. The mIoU-oriented GSCollab-112N obtains 73.81 IoU and 43.14 mIoU, compared with 71.54/38.90 for reproduced VOGS-L (Chen et al., 2025a). The CV-oriented GSCollab-112T obtains 72.48/41.85 at 0.870 MiB, retaining a 2.95-point mIoU gain while using 7.5% less communication than VOGS-L. The full operating-point sweep is reported in Appendix A.11, and Co3SOP results in Appendix A.8.

Table 2: Semantic-OPV2V fusion and communication controls. The VOGS-L row is the decoded-attribute baseline. The last two rows compare the two fusion contents on the GSCollab-96 backbone under the T protocol, with stacking disabled in both cases. In decoded-attribute settings, latent queries are decoded before transmission and are not sent to the receiver. GSCollab-128 omits PEQC; GSCollab-112/96 use the indicated PEQC width. N/I/T denote SIC disabled during both training and inference, an N-trained checkpoint evaluated with SIC, and SIC enabled during both. CV includes the packed ego semantic label map used to derive the SIC mask, selected features or attributes, and transmitted geometry. Among the displayed rows, best and second-best values in each metric are **bolded** and underlined.

Setting	Fused content	Stack	Protocol	IoU (↑)	mIoU (↑)	CV (MiB) (↓)
VOGS-L	Decoded attributes	–	N	71.54	38.90	0.941
GSCollab-128	Latent queries	–	N	72.88	39.71	5.113
GSCollab-128	Latent queries	✓	N	**74.11**	<u>42.83</u>	5.186
GSCollab-112	Latent queries	✓	N	<u>73.81</u>	**43.14**	5.478
GSCollab-112	Latent queries	✓	I	73.22	42.46	1.033
GSCollab-112	Latent queries	✓	T	72.48	41.85	<u>0.870</u>
GSCollab-96 backbone	Decoded attributes	–	T	71.85	38.85	**0.196**
GSCollab-96	Latent queries	–	T	72.30	39.88	0.898

Ablations and communication. Table 2 consolidates the fusion-content, stacking, and representative communication controls. In the decoded-attribute setting, the collaborator converts its latent query into explicit Gaussian attributes before transmission, so the original latent query is not available to the receiver. The matched no-stacking GSCollab-96 controls compare decoded-attribute and latent-query fusion on the same backbone and T protocol; latent-query fusion improves IoU/mIoU from 71.85/38.85 to 72.30/39.88. Matched queries contribute through latent fusion, whereas unmatched queries contribute through decoding and stacking. The no-stacking GSCollab-96 controls isolate gains from latent fusion, while enabling stacking on the GSCollab-128 backbone raises mIoU from 39.71 to 42.83. These results support fusion and stacking as complementary uses of the transmitted latent-query representation. At width 112, applying SIC only during inference (I) reduces CV from 5.478 to 1.033 MiB, while training and evaluating with SIC (T) adapts the model to SIC-guided selection and further reduces CV to 0.870 MiB. Query-width reduction alone does not guarantee a proportional reduction in CV because it imposes no entropy or rate constraint; payload also depends on SIC-selected token count and fixed geometry/map metadata. The 0.196 MiB decoded-attribute payload of the GSCollab-96 control comprises 0.158 MiB of selected Gaussian attributes and a 0.038 MiB packed ego semantic label map. The complete width-by-protocol sweep appears in Appendix A.11.

Confidence-gating control. We adapt a Where2comm-style entropy-confidence gate (Hu et al., 2022) to the same Gaussian-token stream and tune seven threshold pairs for each of five latent widths (35 inference-only runs). Even after choosing the best heuristic point separately for every width, width-matched SIC-I, which applies SIC at inference to an N-trained checkpoint, improves mIoU by 0.825–1.416 points while reducing measured communication by 22.8–43.0% in all five comparisons. Full operating points and protocol details are in Appendix A.11.

5.4 Accuracy–Communication Trade-off

Figure 3 compares mIoU against communication volume on Semantic-OPV2V. For GSCollab-112T and VOGS-L, each curve uses one checkpoint evaluated under independent erasure of received neighbor tokens; the 0% endpoint is standard inference without erasure. The CoHFF curve is reconstructed from its paper-reported feature-sparsification sweep (Song et al., 2024). Because token erasure and feature sparsification are different controls, the curves compare accuracy–payload trade-offs rather than performance under a matched corruption protocol. GSCollab-112T maintains 40.22–41.85 mIoU over 0.455–0.870 MiB, compared with 38.15–38.90 for VOGS-L over 0.471–0.941 MiB. Complete values, including all four GSCollab widths, appear in Appendix A.11, Table 10.

Additional analyses. BEV segmentation, pose and heading perturbations, and joint occupancy–detection results appear in Appendices A.9, A.9, and A.10.

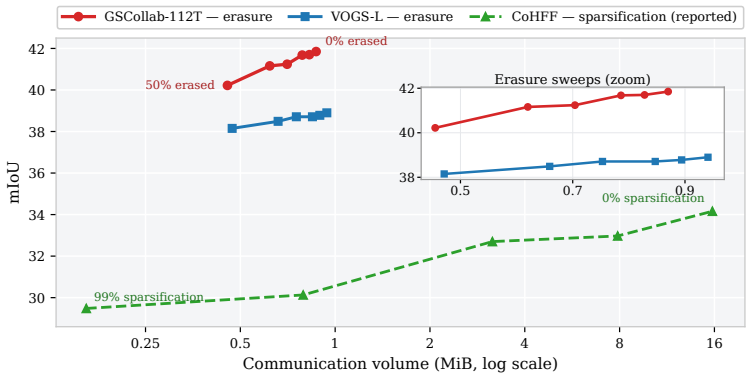


Figure 3: Semantic-OPV2V accuracy–communication trade-offs. GSCollab-112T and VOGS-L vary received-token erasure; CoHFF varies feature sparsification using paper-reported results. CoHFF's decimal-MB payloads are converted to MiB.

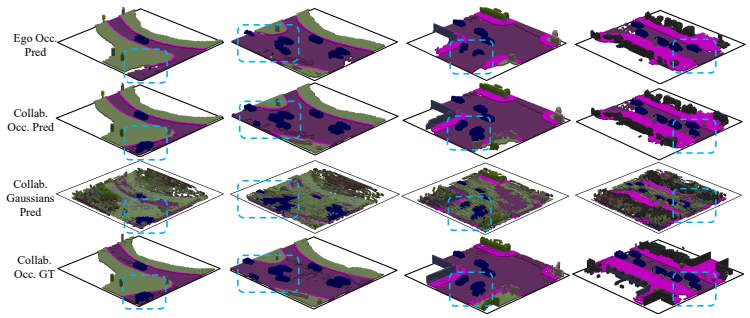


Figure 4: Semantic-OPV2V qualitative results. Each column is one collaborative frame; dashed boxes mark regions improved by collaborative Gaussian tokens.

5.5 Qualitative Evaluation

Multi-frame visualization of task results. Figure 4 shows that transmitted Gaussian tokens supplement occluded or weakly observed regions, producing collaborative predictions that more closely match the ground truth in the highlighted areas. A stepwise Co3SOP example appears in Appendix A.8, Figure 6.

6 Conclusion and Limitations

GaussianCollab uses tokenized semantic Gaussians for communication-efficient latent-query fusion in vision-based collaborative occupancy prediction. Explicit geometry supports rigid cross-agent transformation and matching, while latent queries support matched-token fusion and decoding-and-stacking of unmatched evidence. With SIC-guided selection, PEQC, and lossless 4-bit map packing, its lowest-CV operating points improve mIoU while using less communication than reproduced VOGS-L, yielding a better measured CV–mIoU frontier. Evaluation is limited to vision-based collaborative occupancy on synthetic datasets and uses simulated token erasure and localization perturbations rather than real device-to-device networking. Appendix A.13 discusses these limitations and future directions in detail.

AI use statement

Generative AI tools assisted with language editing, literature search, reference formatting, and consistency checks between the manuscript's technical descriptions, equations, tables, and implementation. The research methodology, experimental execution, and scientific conclusions remained under the authors' direction. The authors manually reviewed all AI-assisted material, verified citations and

quantitative claims against original sources and experimental records, and take full responsibility for the manuscript.

Reproducibility statement

The main paper and appendix document the method, data processing, optimization, communication accounting, protocols, robustness, and profiling. Code and execution instructions will be released.

References

Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3D semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3991–4001, 2022.

Cheng Chen, Hao Huang, and Saurabh Bagchi. Vision-only gaussian splatting for collaborative semantic occupancy prediction. *arXiv preprint arXiv:2508.10936*, 2025a.

Jialei Chen, Wuhao Xu, Sipeng He, Baoru Huang, and Dongchun Ren. I2v-gs: Infrastructure-to-vehicle view transformation with gaussian splatting for autonomous driving data generation, 2025b. URL https://arxiv.org/abs/2507.23683.

Tianchen Deng, Xuefeng Chen, Yi Chen, Qu Chen, Yuyao Xu, Lijin Yang, Le Xu, Yu Zhang, Bo Zhang, Wuxiong Huang, and Hesheng Wang. Gaussiandwm: 3D gaussian driving world model for unified scene understanding and multi-modal generation, 2026. URL https://arxiv.org/abs/2512.23180.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-First International Conference on Machine Learning*, 2024.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3D: Controllable 3D generation for any-view rendering in street scenes, 2025a. URL https://arxiv.org/abs/2405.14475.

Yuanyuan Gao, Hao Li, Jiaqi Chen, Zhengyu Zou, Zhihang Zhong, Dingwen Zhang, Xiao Sun, and Junwei Han. Citygs-x: A scalable architecture for efficient and geometrically accurate large-scale scene reconstruction, 2025b. URL https://arxiv.org/abs/2503.23044.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Anna-Maria Halacheva, Jan-Nico Zaech, Xi Wang, Danda Pani Paudel, and Luc Van Gool. Gaussianvlm: Scene-centric 3D vision-language models using language-aligned gaussian splats for embodied reasoning and beyond, 2025. URL https://arxiv.org/abs/2507.00886.

Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3D semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024.

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)*, pp. 1033–1039, 2025. doi: 10.1109/IV64158.2025.11097436.

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

Yu-Ji Kim, Dahye Lee, Jun-Seong Kim, Hyeon-Woo Nam, GeonU Kim, Yongjin Kwon, Yu-Chiang Frank Wang, Jaesung Choe, and Tae-Hyun Oh. Splatreasoner: Enhancing embodied reasoning and grounding by novel view synthesis, 2026. URL https://arxiv.org/abs/2601.13132.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud flow matching for 3D object generation, 2025. URL https://arxiv.org/abs/2411.08033.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Mohan Liu, Zhihao Gu, Xuanyu Chen, Haitian Zhang, Kaimin Mao, Yan Wu, Wei-Yun Yau, and Lin Wang. Vistavla: Geometry- and semantic-aware 3D gaussian-grounded VLA for robotic manipulation, 2026. URL https://arxiv.org/abs/2607.12356.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Yuning Peng, Haiping Wang, Yuan Liu, Yipeng Lu, Zhen Dong, and Bisheng Yang. LEGO: Leveled language gaussian splatting, 2026. URL https://arxiv.org/abs/2608.10057.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3D using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles, 2024. URL https://arxiv.org/abs/2402.07635.

Qijian Tian, Zimeng Wu, Xuhong Wang, Lizhuang Ma, and Xin Tan. Groupforward: Building referable 3D scenes via instance-grouped feed-forward gaussian splatting, 2026. URL https://arxiv.org/abs/2608.17535.

Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.

Jingxuan Wei, Siyuan Li, Yuhang Xu, Zheng Sun, Junjie Jiang, Hexuan Jin, Caijun Jia, Honghao He, Xinglong Xu, Xi bai, Chang Yu, Yumou Liu, Junnan Zhu, Xuanhe Zhou, Jintao Chen, Xiaobin Hu, Shancheng Pang, Bihui Yu, Ran He, Zhen Lei, Stan Z. Li, Conghui He, Shuicheng Yan, and Cheng Tan. The Trinity of Consistency as a Defining Principle for General World Models, February 2026.

Hanlin Wu, Pengfei Lin, Ehsan Javanmardi, Naren Bao, Bo Qian, Hao Si, and Manabu Tsukada. A synthetic benchmark for collaborative 3D semantic occupancy prediction in V2X-enabled autonomous driving. *arXiv preprint arXiv:2506.17004*, 2025. URL https://arxiv.org/abs/2506.17004.

Yajiao Xiong, Xiaoyu Zhou, Yongtao Wan, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian++: Towards realistic reconstruction and editable simulation for surrounding dynamic driving scenes, 2025. URL https://arxiv.org/abs/2508.20965.

Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv preprint arXiv:2207.02202*, 2022a.

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 International Conference on Robotics and Automation (ICRA)*, pp. 2583–2589. IEEE, 2022b. doi: 10.1109/ICRA46639.2022.9812038. URL https://arxiv.org/abs/2109.07644.

Xiaohao Xu, Feng Xue, Shibo Zhao, Yike Pan, Sebastian Scherer, and Xiaonan Huang. Mac-ego3D: Multi-agent gaussian consensus for real-time collaborative ego-motion and photorealistic 3D reconstruction, 2024. URL https://arxiv.org/abs/2412.09723.

Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems*, 36:25151–25164, 2023.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos, 2026. URL https://arxiv.org/abs/2601.00393.

Zijian Zhang, Yuqing Jiang, Qian Cheng, Xiaofan Li, Si Liu, Ding Zhao, Ping Luo, Weitao Zhou, and Haibao Yu. Gaussiandream: A feed-forward 3D gaussian world model for robotic manipulation, 2026a. URL https://arxiv.org/abs/2605.20752.

Zijian Zhang, Yuqing Jiang, Weitao Zhou, Minglei Li, Jinhao Zhang, Yao Mu, Xiaofan Li, Hao Zhao, and Haibao Yu. Gaussianwam: Distilling geometry and semantics from 3D gaussian fields into world-action models, 2026b. URL https://arxiv.org/abs/2608.24714.

A Appendix

A.1 Detailed Formulation of the General Collaborative Perception Pipeline

We consider a multi-agent collaborative perception system with N agents, defined as $\mathcal{A} = \{a_i\}_{i=1}^{N}$. Each agent a_i observes a spatially bounded local region that may overlap with others. Collaboration is restricted to a communication range τ where observation fields intersect. For an agent a_i, the input comprises a local multi-view image set $\mathcal{I}_i = \{I_{i,v}\}_{v=1}^{V}$ and shared messages $\{\mathcal{M}_{j\to i}\}$ from collaborators. Here, $I_{i,v}$ denotes the image captured from view v, and each agent has V views in total. The vision encoder f_{vis} maps these images to an intermediate feature set $\mathcal{F}_i = \{F_{i,\ell}\}_{\ell=1}^{L_i}$, where $F_{i,\ell}$ denotes one intermediate feature token or unit and L_i is the number of intermediate features possessed by agent a_i. $F_{i,\ell}$ is intentionally generic and can cover Gaussian features, BEV features, TPV features, and related structured intermediate representations. The objective is to generate a task-specific scene prediction $\widehat{\mathcal{Y}}_i$ supervised by ground-truth labels $\mathcal{Y}_i$.

As a first step, agent a_i extracts the intermediate features $\mathcal{F}_i$:

$$\mathcal{F}_i = f_{\text{vis}}(\mathcal{I}_i).$$

Collaboration involves two levels of message packing. On the unit-selection level, a function $\Psi_{j\to i}$ identifies critical feature units from a transformed set $\widetilde{\mathcal{F}}_{j\to i}$ to form the message:

$$\mathcal{M}_{j\to i} = \Psi_{j\to i}(\widetilde{\mathcal{F}}_{j\to i}), \quad \text{where} \quad \widetilde{\mathcal{F}}_{j\to i} = \{\psi_{j\to i}(F_{j,\ell})\}_{\ell=1}^{L_j}.$$

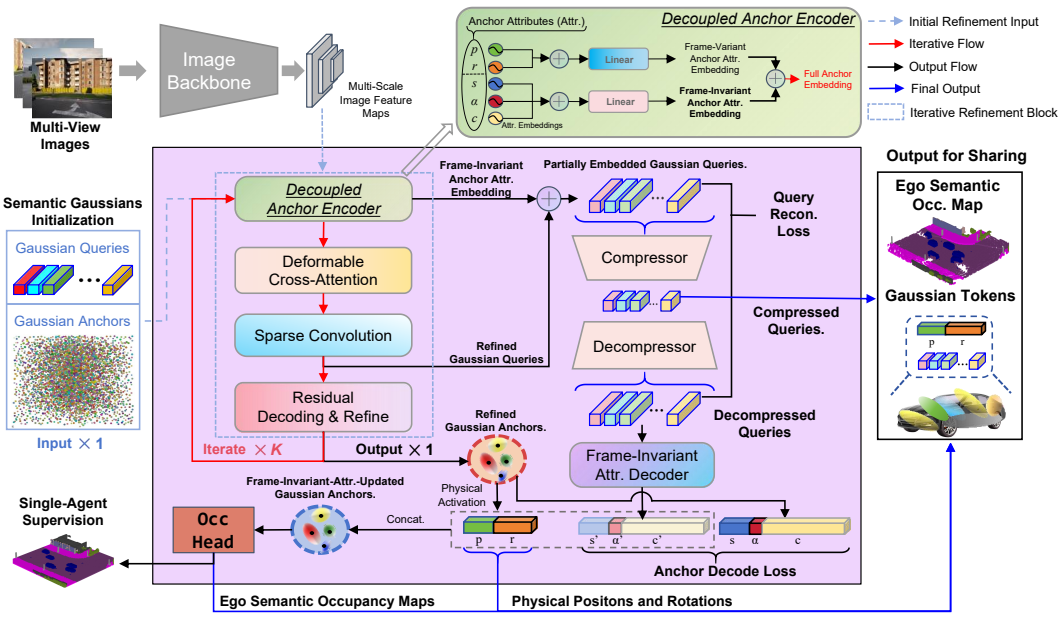


Figure 5: Single-agent pipeline. Multi-view images are lifted into Gaussian primitives, whose pose-dependent attributes and latent queries are decoupled. The geometry and compressed real-valued latent code of each primitive are then packaged as a **Gaussian Token** for inter-agent communication.

Here, $\psi_{j\to i}$ handles lower-level transformations such as attribute compression. Agent a_i fuses local features with received messages via Φ_θ:

$$\widehat{\mathcal{F}}_i = \Phi_\theta\big(\mathcal{F}_i, \{\mathcal{M}_{j\to i}\}_{j\in\mathcal{N}(i)}\big).$$

The final prediction is produced by the task head f_{task}: $\widehat{\mathcal{Y}}_i = f_{\text{task}}(\widehat{\mathcal{F}}_i)$. The system aims to maximize perception performance $\mathcal{P}_\theta$ under a total communication budget B formulated as:

$$\max_\theta \sum_{i=1}^{N} \mathcal{P}_\theta\big(\widehat{\mathcal{Y}}_i, \mathcal{Y}_i\big) \quad \text{s.t.} \quad \sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} \mathcal{C}\big(\mathcal{M}_{i\to j}\big) \le B. \tag{1}$$

A.2 Detailed Illustration of the Single-Agent Image-to-Gaussian Model

Figure 5 presents a detailed illustration of the single-agent image-to-Gaussian model GSFormer-M, modified for downstream collaboration using tokenized semantic Gaussians.

A.3 Calculation of Confidence Weight γ for Semantics Blending

The evaluated models use the `confv1` strategy. Given original semantic scores $\mathbf{c}_{i,m}$ and new predicted scores $\mathbf{c}^{\text{ref}}_{i,m}$, we first max-normalize only the original scores and then compute the blending weight in Equation 14 as

$$\bar{\mathbf{c}}_{i,m} = \frac{\mathbf{c}_{i,m}}{\max(\max(\mathbf{c}_{i,m}), \varepsilon)}, \qquad \gamma = \frac{\max(\bar{\mathbf{c}}_{i,m})}{\max(\bar{\mathbf{c}}_{i,m}) + \max(\mathbf{c}^{\text{ref}}_{i,m}) + \varepsilon}.$$

Here, $\varepsilon = 10^{-6}$ prevents division by zero. The original and refined score vectors themselves are blended in Equation 14; the normalization above is used only to obtain the original confidence.

A.4 Spatial Synchronization and Transformation through Rigid Transform

To align the observations from a collaborator agent a_j with the reference frame of the ego agent a_i, we apply a spatial synchronization operator $\mathcal{T}_{j\to i}$. This operator is modeled as a rigid body

transformation, defined by the relative rotation matrix $\mathbf{R}_{j\to i} \in SO(3)$ and the translation vector $\mathbf{t}_{j\to i} \in \mathbb{R}^3$, derived from the agents' extrinsic poses. For each received Gaussian token $t_{j,n}$, the transformation of its spatial center $\mathbf{p}_{j,n}$ is formulated as:

$$\mathcal{T}_{j\to i}(\mathbf{p}_{j,n}) = \mathbf{R}_{j\to i}\mathbf{p}_{j,n} + \mathbf{t}_{j\to i} \tag{16}$$

The orientation of the Gaussian primitive, represented by the rotation quaternion $\mathbf{r}_{j,n}$, is synchronized through quaternion multiplication to ensure consistent alignment within the ego's coordinate system:

$$\mathcal{T}_{j\to i}(\mathbf{r}_{j,n}) = \mathbf{q}_{j\to i} \otimes \mathbf{r}_{j,n} \tag{17}$$

where $\mathbf{q}_{j\to i}$ is the quaternion representation of $\mathbf{R}_{j\to i}$, and $\otimes$ denotes the Hamilton product.

The frame-invariant attributes remain unchanged by this transformation: the scale $\mathbf{s}_{j,n}$, opacity $\alpha_{j,n}$, and the reconstructed partially embedded query $\hat{\mathbf{q}}^{\text{part}}_{j,n}$. The transformed pose anchors all shared tokens in the ego coordinate system for subsequent vicinity matching and attentional fusion.

A.5 Datasets

Semantic-OPV2V. Semantic-OPV2V, introduced with CoHFF (Song et al., 2024), extends the original OPV2V collaborative-perception benchmark (Xu et al., 2022b) with semantic occupancy annotations. It provides voxel-level ground truth for 12 semantic categories mapped from the original 24 CARLA labels. The annotations are generated by majority voting over semantic LiDAR points within each voxel. Following this protocol, we use a $100 \times 100 \times 8$ grid over $x, y \in [-20, 20]$ m and $z \in [-2.3, 0.9]$ m, yielding a voxel size of 0.4 m $\times$ 0.4 m $\times$ 0.4 m.

Co3SOP. Co3SOP (Wu et al., 2025) is a high-fidelity synthetic benchmark specifically designed for V2X-enabled collaborative 3D semantic occupancy prediction. LiDAR-derived occupancy labels can contain false-empty blind spots because of sparsity and occlusion. Co3SOP instead uses a high-resolution semantic voxel sensor within CARLA to generate dense, temporally aligned semantic annotations across the sensing volume. The raw dataset provides fine-grained labels at a 0.1 m resolution across an extensive spatial grid ($1000 \times 1000 \times 70$ voxels), bounded by $x, y \in [-50, 50]$ m and $z \in [-2, 5]$ m. For our experiments, we downsample the labels to a $100 \times 100 \times 8$ resolution and crop the spatial extent to $x, y \in [-20, 20]$ m and $z \in [-2, 1.2]$ m, resulting in a 0.4 m$\times 0.4$ m$\times 0.4$ m voxel size. This configuration closely aligns with the Semantic-OPV2V setups used in prior works (e.g., CoHFF and VOGS), differing only slightly in the z-axis boundaries.

Label acquisition for invisible voxels Co3SOP and Semantic-OPV2V are derived from the same original OPV2V scenes. For this split only, we re-voxelize the single-agent Semantic-OPV2V semantic point clouds by majority voting over the Co3SOP-aligned $100 \times 100 \times 8$ grid with $z \in [-2, 1.2]$ m. We identify an invisible occupied voxel as one labeled empty in this aligned single-agent annotation but non-empty at the same voxel in the denser Co3SOP annotation. This correspondence enables reproducible IoU and mIoU evaluation over visible and invisible regions on Co3SOP.

Licenses and terms of use. Licensing information is not uniform across these releases. The official OPV2V/OpenCOOD release uses the UCLA Mobility Lab Academic Software License, which permits academic or nonprofit research use and restricts redistribution and commercial use. The official Co3SOP repository states the MIT License. The official CoHFF release provides the Semantic-OPV2V additions without stating a separate license for them; we use these additions only for academic research alongside the underlying OPV2V assets and cite both original releases.

A.6 Implementation Details

We optimize the network with AdamW (Loshchilov & Hutter, 2019) and a weight decay of 0.01. The learning rate increases linearly for 500 iterations to 2×10^{-4} and then follows a cosine decay schedule. Training has two stages: the single-agent pipeline is pretrained for 20 epochs, and the selected checkpoint initializes collaborative training.

The full multi-agent model is subsequently trained end-to-end for 30 epochs with a batch size of 1. Training uses a modified 48 GB NVIDIA RTX 4090 and consumes approximately 35 GB of GPU memory, completing in roughly 36 hours. All inference-time latency, throughput, FLOP, parameter, and peak-memory profiling is conducted separately on a standard 24 GB NVIDIA RTX 4090.

Joint occupancy–detection extension. For the preliminary joint-task experiment, a lightweight detection head reads the final fused $100 \times 100 \times 8$ semantic grid, folds the height dimension into BEV channels, and applies two 128-channel convolutional blocks before predicting classification, seven-parameter box regression, and two-bin direction outputs for two anchors per cell. The joint objective is the unit-weighted sum of the occupancy loss and the standard OpenCOOD anchor-detection loss, while auxiliary pre-fusion supervision remains occupancy-only. All joint models are trained collaboratively from scratch for 30 epochs with batch size 1, gradient accumulation over eight steps, and the same AdamW schedule. Occupancy and detection are evaluated on the Semantic-OPV2V validation set over $x, y \in [-20, 20]$ m; detection AP is reported at IoU thresholds 0.5 and 0.7.

Following GaussianFormer (Huang et al., 2024), the occupancy objective combines voxel-wise Cross-Entropy (CE) with Lovász-Softmax (Berman et al., 2018), which directly targets a tractable IoU surrogate. The collaborative fuser reuses the final Gaussian refinement decoder of the single-agent encoder, so both stages map their latent evidence into Gaussian updates through the same parameters: single-agent refinement uses evidence aggregated from local images, whereas collaborative fusion uses matched collaborator queries. The main collaborative occupancy loss backpropagates through this shared decoder, providing parameter-level supervision to the same latent-to-Gaussian mapping that produces the pre-compression single-agent Gaussians used for SIC. In both single-agent pre-training and collaborative training, an auxiliary occupancy loss $\mathcal{L}_{\text{occ}}^{\text{aux}}$ applies the same criterion to auxiliary pre-fusion predictions, including predictions reconstructed from decompressed transmitted queries. This auxiliary branch directly preserves occupancy semantics through PEQC and FIAD for subsequent fusion; it is not the source of the occupancy maps used to construct SIC masks. For branch $b \in \{\text{main}, \text{aux}\}$, the weighted occupancy loss and complete training objective are

$$\begin{aligned} \mathcal{L}_{\text{occ}}^{b} &= \lambda_{\text{CE}} \mathcal{L}_{\text{CE}}^{b} + \lambda_{\text{Lovasz}} \mathcal{L}_{\text{Lovasz}}^{b}, \quad b \in \{\text{main}, \text{aux}\}, \\ \mathcal{L} &= \mathcal{L}_{\text{occ}}^{\text{main}} + \lambda_{\text{aux}} \mathcal{L}_{\text{occ}}^{\text{aux}} \\ &\quad + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{attr}} \mathcal{L}_{\text{attr}}. \end{aligned}$$

For the evaluated configurations, the CE weight is 10, while the Lovász, auxiliary occupancy, feature-reconstruction, and frame-invariant attribute-reconstruction weights are all 1.

A.7 Details on Evaluation Metrics

We employ Intersection over Union (IoU) and mean IoU (mIoU) to quantify 3D geometric and semantic occupancy prediction performance. Overall geometric IoU treats all non-empty semantic classes as a single occupied class and is defined as $\text{IoU} = \frac{\text{TP}_{\text{occ}}}{\text{TP}_{\text{occ}}+\text{FP}_{\text{occ}}+\text{FN}_{\text{occ}}}$, where TP_{occ}, FP_{occ}, and FN_{occ} are computed after binarizing voxels as occupied or empty. The class-specific metric is defined as $\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c+\text{FP}_c+\text{FN}_c}$, where TP_c, FP_c, and FN_c respectively denote the true positive, false positive, and false negative voxel predictions for semantic class c. The mIoU is the arithmetic mean of IoU_c over all valid non-empty semantic classes. For Co3SOP, we further partition the valid evaluation voxels using the aligned labels described above. The invisible-region (Invis.) split contains dense Co3SOP voxels that are occupied but labeled empty by the single-agent Semantic-OPV2V annotation; the visible-region (Vis.) split is the complement of this mask within the valid evaluation region. The Invis. split therefore measures predictions in dense occupied regions absent from the single-agent LiDAR-derived annotation.

A.8 Extra Experiment Results on Co3SOP

Main results on Co3SOP. Table 3 reports the two endpoints of the Co3SOP sweep. The mIoU-oriented GSCollab-80N reaches 46.07 mIoU and 53.07 mIoU on invisible regions, compared with 41.19 and 46.60 for VOGS-L (Chen et al., 2025a). CV-oriented GSCollab-96T reaches 77.65 IoU and 44.66 mIoU at 0.732 MiB: 3.47 points higher mIoU and 6.9% lower CV than VOGS-L, with IoU lower by 0.09 point. The full sweep, including the 80T point that improves all three measures over VOGS-L, appears in Table 5.

Single-frame collaboration process. Figure 6 visualizes the collaboration process on Co3SOP. Neighbor agents first transform their local Gaussians into the ego frame and generate ego-frame occupancy maps. Guided by SIC masks, tokens are selected from regions where collaborator predictions supplement or conflict with the ego prediction. These selected Gaussians are then

Table 3: Co3SOP semantic occupancy results at the two sweep endpoints. GSCollab-80N is the highest-mIoU endpoint; GSCollab-96T is the lowest-CV endpoint. N disables SIC during both training and inference; T enables SIC during both. Among the displayed methods, best and second-best collaborative results are **bolded** and underlined; the better single-agent result is **bolded**.

Metric / Class	Single-Agent		Collaborative Perception			
	GSFormer	GSFormer-M (96-dim)	Late Fusion	VOGS-L	GSCollab-80 (N)	GSCollab-96 (T)
IoU (↑)	74.73	**75.49**	59.02	**77.74**	77.66	77.65
mIoU (↑)	36.30	**38.76**	30.07	41.19	**46.07**	44.66
CV (MiB) (↓)	–	–	–	0.786	4.911	**0.732**
IoU-Vis. (↑)	73.95	**74.92**	64.47	**79.18**	76.93	77.32
mIoU-Vis. (↑)	32.09	**34.30**	29.35	36.32	**40.10**	39.59
IoU-Invis. (↑)	76.28	**76.61**	49.40	75.06	**79.05**	78.27
mIoU-Invis. (↑)	42.47	**44.18**	29.45	46.60	**53.07**	50.10
Building	12.33	**13.96**	8.64	10.96	13.69	**14.28**
Fence	16.13	**18.28**	8.43	21.04	**26.78**	23.75
Terrain	71.71	**81.35**	67.74	82.44	**83.16**	82.88
Pole	9.20	**9.78**	3.87	9.88	**18.06**	13.23
Road	85.30	**87.37**	74.61	**89.34**	88.25	88.20
Sidewalk	81.95	**85.61**	66.80	**88.04**	86.27	86.49
Vegetation	43.79	**50.54**	16.09	48.41	**57.69**	57.21
Vehicles	52.37	**55.38**	45.22	62.90	**66.51**	64.91
Wall	20.11	**21.25**	4.97	21.38	**25.30**	23.87
Guard rail	**34.68**	26.64	**59.39**	46.29	47.65	47.90
Traffic signs	4.30	**8.25**	5.06	9.77	**16.33**	15.76
Bridge	3.69	**6.73**	0.00	3.88	**23.19**	17.50

fused with the ego Gaussians, expanding the ego perception range and producing a more complete collaborative occupancy prediction in the illustrated example. In this example, the selected tokens retain spatial evidence that is missing from the ego prediction.

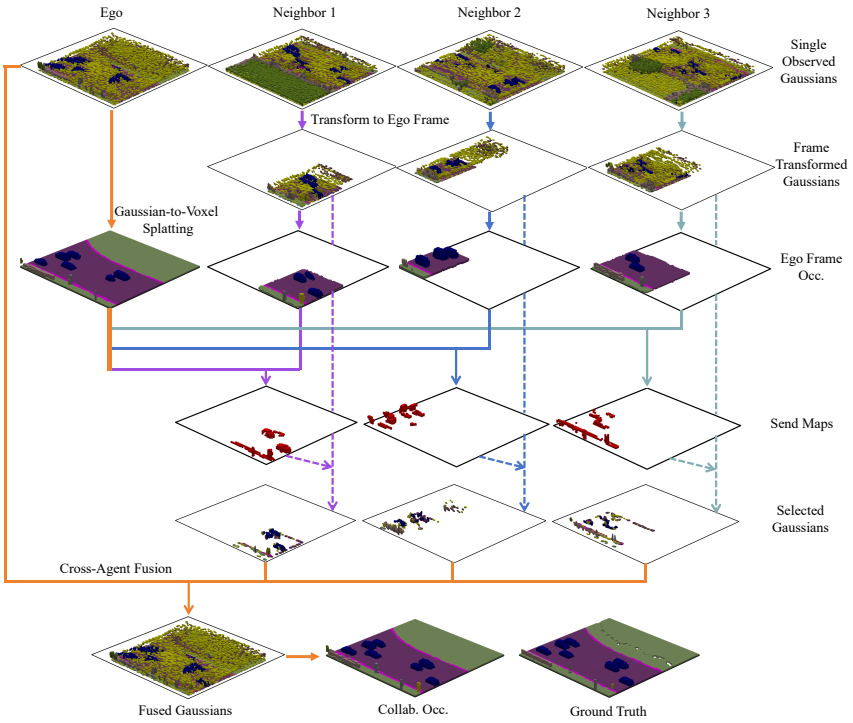


Figure 6: Single-frame collaboration process on Co3SOP. Collaborator Gaussians are transformed into the ego coordinate frame, selected using SIC masks, and fused with ego Gaussians to expand spatial coverage.

Table 4: Single-agent ablation on Co3SOP. DAE denotes the decoupled anchor encoder with independent, non-shared encoders for frame-variant and frame-invariant attributes; FIAD denotes frame-invariant attribute decoding; PEQC values denote transmitted latent width.

Method	DAE	FIAD	PEQC	IoU (↑)	mIoU (↑)
GSFormer baseline			–	74.73	36.30
GSFormer-M (w/o FIAD)	✓		–	74.94	38.05
GSFormer-M (full)	✓	✓	–	73.99	**39.84**
GSFormer-M (64-dim)	✓	✓	64-dim	**76.56**	39.17
GSFormer-M (80-dim)	✓	✓	80-dim	<u>76.29</u>	38.72
GSFormer-M (96-dim)	✓	✓	96-dim	75.49	38.76
GSFormer-M (112-dim)	✓	✓	112-dim	75.37	<u>39.41</u>

Table 5: Full Co3SOP operating-point sweep. Each populated cell is CV (MiB)/IoU/mIoU. Suffixes 64–112 denote PEQC width; 128 denotes no PEQC. N/I/T denote SIC disabled during both training and inference, an N-trained checkpoint evaluated with SIC, and SIC enabled during both training and inference, respectively. Bold cells are the endpoints used in Table 3. All CV values use lossless 4-bit packing of the ego semantic label map used to derive the SIC mask.

Method	Stack	N	I	T
VOGS-L	–	0.786/77.74/41.19	–	–
GSCollab-128	–	4.903/77.31/42.56	–	–
GSCollab-128	✓	4.460/76.18/43.93	0.929/76.53/43.10	0.909/77.17/43.44
GSCollab-64	✓	4.818/78.00/44.36	0.958/77.93/43.95	0.935/77.92/44.22
GSCollab-80	✓	**4.911/77.66/46.07**	0.767/77.65/45.23	0.760/78.00/44.84
GSCollab-96	✓	5.086/77.36/44.98	0.795/77.50/44.39	**0.732/77.65/44.66**
GSCollab-112	✓	5.013/77.13/45.37	0.814/77.32/45.40	0.754/78.01/44.08

Ablation Study of Single-Agent Model. The decoupled anchor encoder (DAE) and FIAD construct and recover the pose-decoupled embedded queries used by collaborative fusion. Because these modules implement the collaborative interface, Table 4 evaluates whether they preserve single-agent accuracy rather than measuring downstream fusion. The compressed variants remain competitive, although mIoU is non-monotonic with latent width.

Ablation study of GaussianCollab on Co3SOP. Table 5 summarizes the Co3SOP operating points across PEQC widths and SIC protocols. The 80N model attains the highest mIoU of 46.07; 80T improves IoU, mIoU, and CV over VOGS-L; and 96T reaches the lowest CV of 0.732 MiB with 44.66 mIoU.

Discussion of query dimensionality in collaborative training. In preliminary internal trials conducted before introducing a communication-side bottleneck, reducing the Gaussian-query dimension throughout the entire network weakened downstream collaborative performance. Because these exploratory controls are not presented as a formal ablation, we use this observation only as design motivation: PEQC retains the full internal query capacity and compresses partially embedded queries at the communication interface.

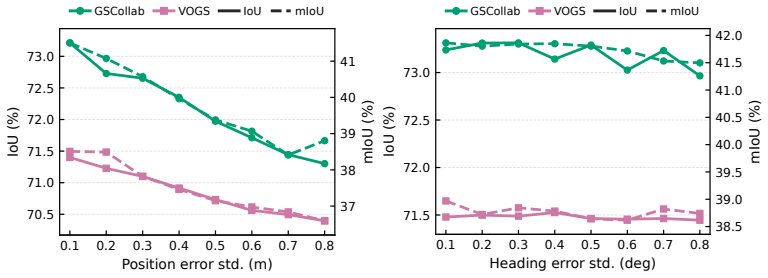
A.9 Extra Experiment Results on Semantic-OPV2V

BEV segmentation results. Table 6 reports that, using the 64-dimensional SIC-T checkpoint for both agent settings, GaussianCollab achieves the best IoUs among the compared methods for all three BEV categories. It obtains 72.76/82.47/75.49 IoU for Vehicle/Road/Others with two agents and 78.67/86.42/80.37 with up to seven agents, surpassing the corresponding VOGS-L results of 70.06/81.30/74.74 and 75.93/85.07/79.98.

Robustness to localization and heading noise. Figure 7 evaluates GSCollab and VOGS-L (Chen et al., 2025a) under simulated pose and heading perturbations on Semantic-OPV2V. These tests measure sensitivity to localization error but do not reproduce all real-world effects.

Table 6: BEV segmentation IoU on Semantic-OPV2V. GaussianCollab uses the same 64-dimensional T checkpoint (epoch 27) in both agent settings.

Method	# Agents	Vehicle	Road	Others
CoBEVT	2	46.13	52.41	-
CoHFF	2	47.40	63.36	40.27
VOGS-L	2	70.06	81.30	74.74
GSCollab-64 (T)	2	**72.76**	**82.47**	**75.49**
CoBEVT	Up to 7	60.40	63.00	-
CoHFF	Up to 7	64.44	57.28	45.89
VOGS-L	Up to 7	75.93	85.07	79.98
GSCollab-64 (T)	Up to 7	**78.67**	**86.42**	**80.37**



(a) Position Error (b) Heading Error

Figure 7: Robustness to pose (left) and heading (right) errors drawn from zero-mean Gaussian distributions with varying standard deviations. Results are obtained on Semantic-OPV2V.

A.10 Task-Head Extensibility

To test whether the fused representation can support another prediction head, we jointly train semantic-occupancy and object-detection heads on the same Gaussian tokens. These joint-task models are trained collaboratively without initialization from single-agent pretrained weights, whereas the other collaborative models use single-agent pretrained initialization. The jointly trained VOGS-L baseline obtains 71.82 IoU, 38.52 mIoU, 95.73 AP@0.5, and 91.86 AP@0.7. Across four PEQC settings and no-PEQC GSCollab-128, the joint GSCollab models obtain 72.47–73.26 IoU and 41.10–42.03 mIoU while reaching 95.76–96.16 AP@0.5 and 90.80–92.80 AP@0.7. Thus, all five GSCollab settings improve IoU, mIoU, and AP@0.5 over joint VOGS-L, while GSCollab-64, GSCollab-80, and GSCollab-112 also improve AP@0.7. Table 7 reports all settings. The detection-only baselines use the OPV2V values reported by How2comm (Yang et al., 2023) and are included for reference. These results show compatibility with two task heads; broader task and modality transfer remains untested.

A.11 Selection, Fusion-Content, and Communication Ablations

Where2comm-style confidence gating. We adapt entropy confidence to the Gaussian-token stream using $q_v = 1 - H(p_v)/\log C$ with $C = 14$ and $M_{\text{send}} = (q^{\text{ego}} < T_{\text{ego}}) \wedge (q^{\text{collab}} > T_{\text{collab}})$. We evaluate seven threshold pairs for each of five latent widths on 1,980 Semantic-OPV2V validation frames with zero pose noise and zero token erasure. These are inference-only experiments on fixed checkpoints. CV is measured in MiB per collaborative frame and includes the packed ego-to-collaborator selection map, selected latent features, and transmitted geometry. The best tested pair is $(0.98, 0.92)$ for every width; it is also the most permissive collaborator threshold tested. Across all 35 runs, CV and mIoU have Pearson correlation $+0.527$. Table 8 compares each width's best tuned heuristic point with SIC-I.

Full operating points. Table 9 reports the complete GSCollab N/I/T sweep on Semantic-OPV2V together with the no-stacking GSCollab-128N control from Table 2. Across the sweep, 112N has

Table 7: Joint occupancy and detection on Semantic-OPV2V. VOGS-L and GSCollab are trained jointly for both tasks. Suffixes 64–112 denote PEQC width; 128 denotes no PEQC. Baselines marked † are detection-only values reported by How2comm.

Model	Occ. IoU/mIoU	AP@0.5/0.7
CoBEVT†	–	81.59/67.50
Where2comm†	–	82.75/67.29
How2comm†	–	85.42/72.24
VOGS-L	71.82/38.52	95.73/91.86
GSCollab-64	73.26/41.46	96.08/92.55
GSCollab-80	72.47/41.10	**96.16/92.80**
GSCollab-96	72.97/41.20	95.76/91.05
GSCollab-112	73.16/**42.03**	96.05/92.60
GSCollab-128	72.85/41.29	95.78/90.80

Table 8: Best tuned entropy-confidence gate versus width-matched SIC-I.

Width	Heur. mIoU	SIC mIoU	Gain	Heur. volume	SIC volume	Volume red.
128	40.318	41.501	+1.183	1.576	1.105	29.9%
64	40.996	41.890	+0.894	1.659	0.946	43.0%
80	40.181	41.006	+0.825	1.639	1.115	32.0%
96	40.615	41.615	+1.000	1.495	1.154	22.8%
112	41.046	42.462	+1.416	1.525	1.033	32.3%

the highest mIoU and 112T the lowest CV; the remaining fusion-content and stacking controls are consolidated in the main-text table.

Table 9: Full GSCollab operating-point sweep on Semantic-OPV2V, including the no-stacking GSCollab-128N control. Each populated cell is CV (MiB)/IoU/mIoU. Suffixes 64–112 denote PEQC width; 128 denotes no PEQC. N/I/T denote SIC disabled during both training and inference, an N-trained checkpoint evaluated with SIC, and SIC enabled during both training and inference, respectively. Bold cells are the mIoU- and CV-oriented endpoints used in Table 1.

Method	Stack	N	I	T
VOGS-L	–	0.941/71.54/38.90	–	–
GSCollab-128	–	5.113/72.88/39.71	–	–
GSCollab-128	✓	5.186/74.11/42.83	1.105/73.15/41.50	0.942/73.47/41.92
GSCollab-64	✓	5.557/74.42/42.87	0.946/73.67/41.89	0.887/73.69/41.50
GSCollab-80	✓	5.270/73.33/41.95	1.115/72.49/41.01	0.895/72.48/41.69
GSCollab-96	✓	5.272/74.65/42.48	1.154/73.73/41.61	0.898/72.88/42.64
GSCollab-112	✓	**5.478/73.81/43.14**	1.033/73.22/42.46	**0.870/72.48/41.85**

Token-erasure robustness. The upper block of Table 10 independently erases received neighbor tokens before fusion without retransmission. Each rate is evaluated in a separate inference run with only the erasure rate changed; 0% is the standard no-erasure condition. Each entry is effective received CV/IoU/mIoU. GSCollab retains higher absolute IoU and mIoU than VOGS-L at every tested rate, although its relative degradation is larger. The lower block reproduces CoHFF's feature-sparsification sweep (Song et al., 2024). It provides a published accuracy–payload reference, not a matched token-erasure control.

A.12 Computational Overhead

Table 11 summarizes inference cost on a standard 24 GB NVIDIA RTX 4090. Relative to VOGS-L, GaussianCollab adds 0.37–0.44% parameters, 2.10–2.33% FLOPs, and 0.50–0.54% peak memory.

Table 10: Complete Semantic-OPV2V sweeps underlying Figure 3. The upper block varies independent neighbor-token erasure; all GSCollab variants use the T protocol, while VOGS-L does not use SIC. Entries are CV (MiB)/IoU/mIoU. The lower block varies CoHFF feature sparsification and reproduces its paper-reported results; decimal MB values are converted to MiB.

Independent neighbor-token erasure					
Rate	VOGS-L	GSCollab-64T	GSCollab-80T	GSCollab-96T	GSCollab-112T
0%	0.941/71.54/38.90	0.887/73.69/41.50	0.895/72.48/41.69	0.898/72.88/42.64	0.870/72.48/41.85
5%	0.894/71.56/38.78	0.844/73.65/41.46	0.853/72.45/41.89	0.855/72.88/42.51	0.828/72.47/41.70
10%	0.847/71.53/38.71	0.802/73.60/41.08	0.809/72.44/41.32	0.813/72.80/42.40	0.786/72.44/41.68
20%	0.753/71.54/38.71	0.717/73.50/40.81	0.723/72.38/41.15	0.726/72.75/42.04	0.704/72.36/41.24
30%	0.659/71.51/38.49	0.632/73.39/40.49	0.638/72.29/40.99	0.640/72.64/41.65	0.620/72.28/41.16
50%	0.471/71.40/38.15	0.463/73.08/39.97	0.466/72.05/40.49	0.468/72.33/40.93	0.455/72.02/40.22

CoHFF feature sparsification (paper-reported)					
Sparsification rate	0%	50%	80%	95%	99%
CV (MiB)	15.764	7.887	3.157	0.792	0.162
IoU	50.46	49.56	49.53	48.52	48.02
mIoU	34.16	32.97	32.70	30.13	29.48

Its measured throughput is 2.76–2.99 FPS versus 3.01 FPS, and end-to-end P95 latency spans 494.35–533.57 ms versus 501.59 ms. Neither implementation is real time, and network-transfer latency is excluded.

Table 11: Overall inference overhead. End-to-end latency is reported as P95 milliseconds.

Method	Latent query dim.	Params (M)	FLOPs/frame (TFLOPs)	End-to-end P95 (ms)	FPS	Peak GPU memory (GiB)
VOGS-L	–	73.722	0.858	501.59	3.01	2.985
GSCollab	64	73.998	0.876	498.56	2.82	3.001
GSCollab	80	74.011	0.877	533.57	2.76	3.000
GSCollab	96	74.027	0.877	494.35	2.84	3.000
GSCollab	112	74.044	0.878	496.76	2.99	3.000

Table 12 gives the module-level latency breakdown. Gaussian encoding dominates the mean end-to-end time (approximately 93% for GaussianCollab and 95% for VOGS-L). For GSCollab, we additionally derive a deployment-oriented fusion/stacking critical-path estimate per sample as $t_{\mathrm{match}} + \max(t_{\mathrm{fusion}} + t_{\mathrm{ego\ decode}}, t_{\mathrm{neighbor\ stacking}})$, assuming concurrent execution of the two independent post-matching branches. This estimate is 3.00–3.53 ms on average. The 4.32 ms VOGS-L value is its measured sequential fusion-stage latency and is included as a reference rather than as an identically scheduled critical-path estimate. Table 13 reports the individually measured GSCollab components used in the derived estimate.

End-to-end latency and its always-executed stages are profiled over 1,970 samples; neighbor-dependent stages use the 1,670 samples containing collaborator input. FLOPs, memory, and throughput use the 1,980-frame inference loop. These measurements exclude over-the-air latency.

Communication-compression observation. The operating-point sweeps show that MLP width reduction contributes substantially less communication saving than SIC and does not compound linearly with it, indicating that channel bottlenecking and semantic selection address different redundancy. This motivates future learned-codebook and explicit rate–distortion approaches rather than stronger claims about latent width alone.

A.13 Limitations and Future Work

Shared-decoder and auxiliary supervision. In early collaborative variants, the post-fusion Gaussian-update decoder did not share parameters with the single-agent refinement decoder. We observed that collaborative optimization could then allow the pre-compression single-agent Gaussians to lose reliable occupancy semantics, degrading the splatted maps from which SIC masks are

Table 12: Module latency breakdown in mean/P95 milliseconds. The GSCollab fusion/stacking critical path is the per-sample parallel-deployment estimate defined in the text; the VOGS-L entry is its measured sequential fusion-stage latency. P95 values are measured per stage and should not be summed because stages are nested or their tail events need not coincide.

(a) Encoding, selection, and overall latency.

Method	Gaussian encoding	Ego-map splatting	SIC filtering	SIC packing	End-to-end
VOGS-L	261.30/474.34	N/A	N/A	N/A	275.62/501.59
GSCollab-64	272.21/463.41	2.61/3.24	4.69/10.49	0.44/0.81	293.14/498.56
GSCollab-80	283.22/496.43	2.86/3.96	5.01/11.46	0.40/0.72	304.30/533.57
GSCollab-96	270.01/460.89	2.64/3.13	4.70/10.25	0.43/0.81	290.77/494.35
GSCollab-112	264.06/463.44	2.53/2.99	4.51/10.62	0.33/0.63	282.70/496.76

(b) Collaboration-specific stages.

Method	Neighbor transform + initial stacking	Fusion/stacking critical path	Final collaborative splatting
VOGS-L	7.63/15.79	4.32/7.59	3.00/4.37
GSCollab-64	8.12/16.70	3.53/5.12	2.67/3.41
GSCollab-80	7.73/16.07	3.52/5.26	2.79/3.73
GSCollab-96	8.05/16.59	3.43/4.96	2.65/3.52
GSCollab-112	6.73/14.21	3.00/3.81	2.50/2.98

Table 13: Individually measured GSCollab fusion/stacking components in mean/P95 milliseconds. The reference implementation profiles the ego fusion/decode and neighbor-stacking branches sequentially; the derived deployment estimate assumes that they execute concurrently after matching. The component columns are therefore not additive, and their marginal P95 values do not reconstruct the critical-path P95.

Method	Ego latent-query decode	Vicinity matching	Latent-token fusion	Neighbor stacking branch
VOGS-L	N/A	N/A	N/A	N/A
GSCollab-64	1.28/1.65	1.26/1.67	0.98/1.46	0.50/0.65
GSCollab-80	1.24/1.57	1.30/1.81	0.97/1.37	0.49/0.65
GSCollab-96	1.22/1.55	1.25/1.69	0.95/1.34	0.48/0.61
GSCollab-112	1.07/1.27	1.13/1.57	0.80/1.08	0.42/0.54

constructed. The final design therefore reuses the last single-agent Gaussian refinement decoder to project attention-fused ego queries into Gaussian updates. This parameter sharing gives single-agent refinement and multi-agent fusion a common latent-to-Gaussian update mapping: the former aggregates evidence from local image observations, whereas the latter aggregates matched queries from collaborating agents. The main collaborative occupancy objective backpropagates through the shared decoder and thereby regularizes the mapping that also produces the pre-compression Gaussian attributes used by SIC. Separately, the auxiliary single-agent occupancy loss supervises FIAD-reconstructed pre-fusion predictions, including those obtained from decompressed collaborator queries, while feature reconstruction and frame-invariant attribute reconstruction further constrain the PEQC transmission branch. Thus, shared-decoder training is directly tied to the SIC inputs, whereas the auxiliary and reconstruction losses preserve semantic information through compression for subsequent fusion. This coupling mitigates model-induced degradation of the SIC inputs, but does not eliminate intrinsic failures due to correlated errors, pose misalignment, or unstable class boundaries. Accuracy and payload also remain non-monotonic with latent width because dimensionality reduction imposes no explicit entropy or bitrate constraint; learned-codebook quantization and rate–distortion learning remain future work.

Wire format and deployment accounting. An early prototype accounted for the transmitted ego semantic label map using the integer tensor produced by `argmax`, unnecessarily inflating its payload. The revised wire format losslessly packs two 4-bit semantic labels per byte; the collaborator unpacks this map and compares it with its own prediction to derive the binary SIC mask. All updated CV results use this packed-map accounting. CV measures serialized model payloads but excludes transport headers, channel coding, retransmission, and device-to-device networking overhead. The

current implementation also performs a full final splatting pass; cached regional re-splatting remains a possible latency optimization.

Task and modality scope. Our primary evaluation concerns vision-based collaborative semantic occupancy. Every joint occupancy–detection GSCollab setting exceeds the three non-GSCollab collaborative baselines in Table 1 on both IoU and mIoU, and exceeds the three paper-reported detection references in Table 7 on both AP thresholds. The detection comparison is contextual because those values are taken from the cited paper rather than reproduced in our experiments. These results show that both heads can be jointly optimized from the same fused representation, but this two-task result does not establish transfer to language, action, video, audio, or multimodal foundation-model spaces. Existing Gaussian-based systems nevertheless motivate studying action and language: VistaVLA, GaussianWAM, and GaussianDream connect Gaussian scene representations with robotic action or world-action learning (Liu et al., 2026; Zhang et al., 2026b;a), while SplatReasoner, GaussianVLM, GroupForward, and LEGO align Gaussian scenes with language for embodied reasoning, referential grounding, or open-vocabulary understanding (Kim et al., 2026; Halacheva et al., 2025; Tian et al., 2026; Peng et al., 2026). These works are evaluated in single-agent or single-scene settings rather than cooperative multi-agent perception. Extending such capabilities to collaboration therefore remains unverified and requires a reliable multi-agent testbed together with task-specific supervision and cross-agent alignment objectives.

Data and network realism. Evaluation uses synthetic datasets with simulated independent token erasure and localization perturbations. Real deployment additionally involves asynchronous sensing, burst loss, jitter, protocol overhead, imperfect calibration, and hardware-dependent network latency. Larger real-world collaborative occupancy datasets and closed-loop networked evaluation are required to establish generalization under these conditions. SIC also has intrinsic failure modes: correlated semantic errors can yield false negatives, pose error can create false disagreement and wasted transmission, and unstable class boundaries can cause selection flicker.