SATURATION-GUIDED TEMPORAL CONTEXT ALLOCATION BEFORE 3D OCCUPANCY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scene-adaptive temporal context allocation enables 3D occupancy predictors to tailor the history length to each scene. Existing adaptive methods infer this demand from detections, predicted observability, learned features, or other predictor-generated signals. Coupling the selection rule to a particular perception architecture limits its direct transfer to other occupancy predictors, even on the same dataset. We ask whether the required history can instead be determined directly from raw camera–LiDAR observations. Our key observation is that the additional geometric coverage and visual diversity contributed by older frames saturate at scene-dependent rates. Based on this observation, we propose Composite Scene Entropy (CSE), a predictor-external proxy that quantifies this saturation over nested temporal windows. CSE combines global point-cloud entropy, local point-cloud entropy, and image-gradient entropy to characterize global geometric coverage, local structure, and visual diversity, respectively. We select the shortest window whose marginal CSE gain has saturated and feed it to an otherwise unchanged occupancy predictor. Extensive experiments on nuScenes-SurroundOcc show that one CSE calibration and its per-sample allocations can be reused across three compatible occupancy predictors; with SAGFormer, CSE improves Fixed-9 by 1.44 mIoU at a similar average window length and nearly matches Fixed-16 with 52.8% lower model-side latency. On SSCBench-KITTI-360, an independent calibration yields the same CSE parameter values as on nuScenes, and CSE allocation remains within 0.03 mIoU of Fixed-16.

1 INTRODUCTION

3D semantic occupancy prediction supports autonomous driving by reconstructing the dense geometry and semantics of the surrounding space, including regions outside the current sensor view (Cao & de Charette, 2022; Huang et al., 2023; Wei et al., 2023; Li et al., 2023a; Zhang et al., 2023). Since a single observation is inevitably incomplete under occlusion, sparse sampling, and limited sensor coverage, temporal camera–LiDAR measurements provide essential complementary evidence for a more complete scene representation (Liao et al., 2025; Leng et al., 2025).

Motivated by this need, temporal occupancy research has developed several paths for incorporating history. Recurrent state, cost volumes, feature warping, sparse association, and temporal-flow modeling aggregate historical evidence within the predictor (Liao et al., 2025; Leng et al., 2025; Chen et al., 2025; Ye et al., 2024; Yuan et al., 2025; Ouyang et al., 2025), while multimodal systems align historical LiDAR observations with camera features to strengthen geometric support (Zhang et al., 2024; Yang et al., 2025; Duan et al., 2025). These approaches primarily improve how a predefined temporal span is represented and fused. Adaptive alternatives instead vary the admitted history according to predicted observability, detected-object attributes, or learned features (Wang et al., 2026; Huang et al., 2025a; Liu et al., 2023; Li et al., 2022). These policies adapt the temporal window to scene or object conditions, but rely on signals from a specific perception pipeline, making the selection rule harder to reuse across occupancy predictors.

This coupling makes it difficult to reuse a window-selection rule across predictors, because the rule depends on signals produced by a specific perception pipeline (Wang et al., 2026; Huang et al., 2025a; Liu et al., 2023). Applying it to another occupancy predictor may require reproducing upstream computations, accessing compatible intermediate features, or retraining the allocation policy.

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.

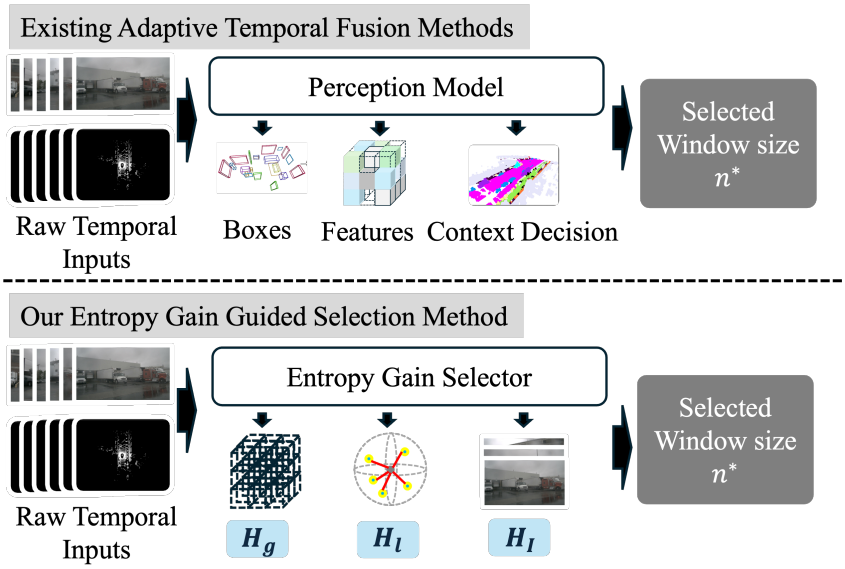


Figure 1: Temporal fusion determines how history is processed, whereas temporal context allocation determines how much history enters the predictor. We estimate the latter from the saturation of raw camera–LiDAR observations before occupancy feature extraction, then run an unchanged compatible predictor once on the selected window.

Even when two predictors process the same scene, their different internal representations can prevent them from directly sharing the same selection rule. This raises a question about where temporal allocation should take place in the pipeline. If the history length can be determined from raw observations, the decision can be made before either predictor begins feature extraction. A common criterion could then assign temporal windows to scenes and reuse those assignments across compatible predictors. We therefore ask whether scene-specific temporal demand can be estimated before occupancy inference, without using predictor features, outputs, or parameters.

To decouple temporal context allocation from predictor-specific computation, we move the selection stage ahead of occupancy inference, as illustrated in Fig. 1. Our approach builds on the observation that historical evidence exhibits scene-dependent diminishing returns: early frames often expand point cloud coverage and reveal new visual structure, whereas later frames increasingly revisit accumulated content. We quantify this progression with Composite Scene Entropy (CSE), which combines global point-cloud coverage, local geometric structure, and cumulative image-gradient diversity over nested camera–LiDAR windows. A saturation-guided rule selects the earliest window whose marginal CSE gain has saturated and supplies it to an otherwise unchanged occupancy predictor. Because CSE is computed directly from raw observations, the selector requires no detections, learned features, occupancy outputs, or predictor parameters. The resulting predictor-external interface separates how much history to admit from how a compatible predictor fuses it, enabling the same calibration and per-sample allocations to be reused across predictors on a dataset. We evaluate this principle with controlled comparisons, cross-predictor reuse, and cross-dataset calibration in camera–LiDAR occupancy prediction.

Our contributions can be summarized as follows:

- We formulate temporal context allocation as a predictor-external, sample-level stopping problem at the raw-input interface, separating how much history to admit from how a compatible occupancy predictor fuses it.

- We introduce CSE to quantify the saturation of raw geometric coverage, local structure, and visual diversity over nested temporal windows, and use a saturation-guided rule to select the earliest window at which the marginal CSE gain has saturated. This procedure requires no detections, learned features, occupancy outputs, or predictor parameters.

- Comparisons at similar average temporal budgets, together with shuffled-assignment controls, support the value of matching temporal windows to individual scenes. On nuScenes, one CSE calibration and its per-sample allocations are reused across three compatible predictors; on KITTI-360, an independent dataset-specific calibration selects the same parameter values.

2 Related Work

Temporal occupancy prediction. Dense occupancy predictors have progressed from voxel completion and lifted image features to transformer reasoning and compact Gaussian scene representations (Cao & de Charette, 2022; Huang et al., 2023; Wei et al., 2023; Li et al., 2023a; Zhang et al., 2023; Huang et al., 2024; 2025b). Temporal extensions recover occluded structure through recurrent state, cost volumes, feature warping, sparse association, temporal flow, and world-model objectives (Liao et al., 2025; Leng et al., 2025; Chen et al., 2025; Ye et al., 2024; Yuan et al., 2025; Ouyang et al., 2025; Boeder et al., 2025). Multi-modal approaches further use aligned LiDAR history to improve geometric evidence (Zhang et al., 2024; Yang et al., 2025; Duan et al., 2025). These methods principally determine how historical information is represented and fused; temporal span is predefined or implicit in predictor state.

Adaptive temporal context. Adaptive policies have been explored at different decision stages and granularities. NGOcc chooses short or long processed-feature queues per voxel using predicted observability (Wang et al., 2026). VADet supports input-level aggregation across different detector architectures, selecting the aggregation length for each detected object from its bounding box, velocity, and point density (Huang et al., 2025a). Other approaches vary sequence length or view availability to construct temporal or multi-view inputs during training (He et al., 2023; Wu et al., 2025). Camera-based BEV systems also vary temporal feature construction through recurrence or stereo matching (Li et al., 2022; Huang & Huang, 2022; Park et al., 2023; Li et al., 2023b). Our method determines a scene-level temporal window directly from raw camera–LiDAR observations before occupancy feature extraction begins. Because this decision requires no predictor-generated signals, the same calibrated rule and per-scene window assignments can be reused across compatible occupancy predictors on the same dataset.

Table 1 compares when temporal context is selected, which signals guide the decision, and the granularity of selection. These distinctions separate the ability to apply an aggregation method across architectures from the ability to select a window before running the predictor.

Table 1: Temporal context interfaces. "Pred.-derived" means that an explicit context decision uses task outputs or learned features.

Method	Decision	Signal	Granularity	Pred.-derived
ST-Occ	predictor	memory	scene/memory	–
NGOcc	predictor	observability	voxel	yes
VADet	post-detection	boxes/motion	object	yes
LEF	training	sampled length	sample	no
Ours	pre-feature	raw camera–LiDAR	scene	no

3 Problem Formulation

Each observation $V_i = (I_i, P_i, E_i)$ contains images, a LiDAR point cloud, and ego pose. At time t, candidate windows are:

$$\mathcal{V}_t^{(n)} = \{V_{t-n+1}, \ldots, V_t\}, \qquad n \in \{1, \ldots, N\}. \tag{1}$$

A policy g selects the window before inference, with sample-level quality

$$n_t^* = g(\mathcal{V}_t^{(N)}), \qquad m_t(g) = \mathrm{mIoU}_{\mathrm{pc}}\Big(f_{\theta_g}(\mathcal{V}_t^{(n_t^*)}), Y_t\Big). \tag{2}$$

Here, Y_t is ground truth and θ_g contains predictor parameters trained under g. The score $\mathrm{mIoU}_{\mathrm{pc}}$ averages over ground-truth-present classes as in Appendix E. It requires predictions and labels, which are unavailable during selection.

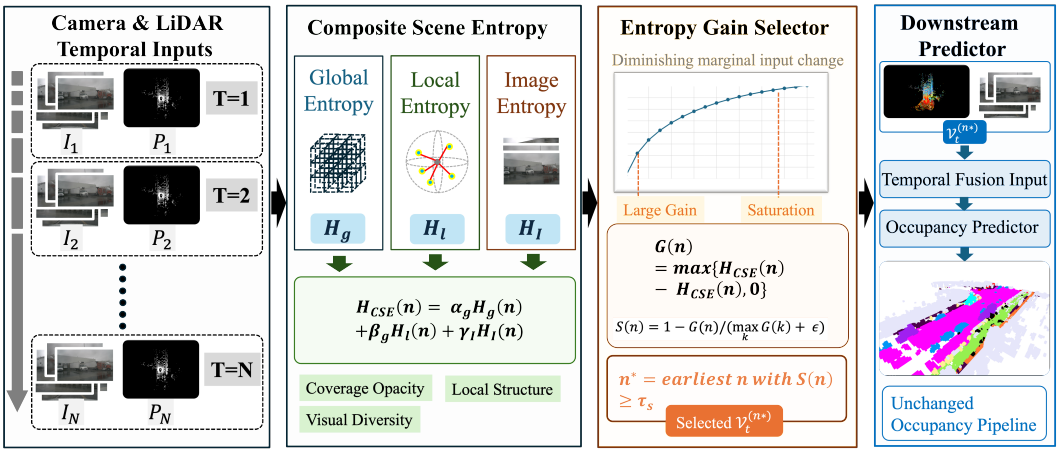


Figure 2: Raw-data CSE guides temporal allocation before an unchanged occupancy predictor.

We seek a raw-input proxy

$$q_t(n) = q(\mathcal{V}_t^{(n)}). \tag{3}$$

The proxy guides allocation from raw observations without predictor features, outputs, or parameters. Let g_q select from proxy scores and g_N always select N. Quality loss and processing cost are

$$\begin{aligned} L_t(q, g_q) &= \max\{m_t(g_N) - m_t(g_q), 0\}, \\ C_t(q, g_q) &= C_t^{\text{sel}}(q, g_q) + C_t^{\text{pred}}(g_q). \end{aligned} \tag{4}$$

The cost includes selection and prediction under a common timing boundary. We seek a proxy and allocation rule that reduce this cost while preserving task quality. We evaluate each design by

$$\overline{L}(q, g_q) = \mathbb{E}_t[L_t(q, g_q)], \qquad \overline{C}(q, g_q) = \mathbb{E}_t[C_t(q, g_q)]. \tag{5}$$

The expectation is over scenes. Composite Scene Entropy provides the proxy, and a threshold on its relative marginal gains defines the allocation rule. Calibration selects an accuracy–cost operating point.

4 Saturation-Guided Temporal Context Allocation

4.1 Composite Scene Entropy

Composite Scene Entropy (CSE) combines geometry and visual diversity. Statistics and gains omit t; $\epsilon = 10^{-6}$ is a numerical stabilizer.

Global geometric coverage. After ego-pose alignment and voxelization, $p_j^{(n)}$ is the point fraction in occupied voxel v_j, and M_n counts occupied voxels:

$$H_g(n) = -\sum_{j=1}^{M_n} p_j^{(n)} \log\left(p_j^{(n)} + \epsilon\right). \tag{6}$$

Local geometric structure. Let $\lambda_{i,1} \geq \lambda_{i,2} \geq \lambda_{i,3}$ be eigenvalues of voxel v_i's K-neighbor covariance matrix. We omit n from these window-specific eigenvalues and local responses. The differential entropy compactness index (DECI) (Barberi et al., 2023) uses

$$\rho_{i,j} = \frac{\lambda_{i,j} + \epsilon}{\sum_{r=1}^{3}(\lambda_{i,r} + \epsilon)}, \quad d_i = \left[\frac{1}{2}\log\left((2\pi e)^3 \prod_{j=1}^{3} \rho_{i,j} + 1\right) + \epsilon\right]^{-1}. \tag{7}$$

Let K_l cap the retained response count and $\mathcal{T}_n$ index the $\min(K_l, M_n)$ largest responses. Their average is

$$H_l(n) = \frac{1}{|\mathcal{T}_n|} \sum_{i \in \mathcal{T}_n} d_i. \tag{8}$$

Accumulated visual diversity. For image $I_{r,c}$ from frame r and camera c, let $C_{ab}(I_{r,c})$ count Sobel gradients in bin (a, b). We pool across N_c cameras and use $B_x \times B_y$ bins (Larkin, 2016):

$$C_{ab}^{(n)} = \sum_{r=t-n+1}^{t} \sum_{c=1}^{N_c} C_{ab}(I_{r,c}), \qquad p_{ab}^{(n)} = \frac{C_{ab}^{(n)}}{\sum_{u,v} C_{uv}^{(n)} + \epsilon},$$
$$H_I(n) = -\sum_{a=1}^{B_x} \sum_{b=1}^{B_y} p_{ab}^{(n)} \log\Big(p_{ab}^{(n)} + \epsilon\Big). \tag{9}$$

CSE combines these statistics with weights α_g, β_g, and γ_I:

$$H_p(n) = \alpha_g H_g(n) + \beta_g H_l(n), \qquad H_{\mathrm{CSE}}(n) = H_p(n) + \gamma_I H_I(n). \tag{10}$$

Thus $q_t(n) = H_{\mathrm{CSE}}(n)$. Appendices B–C specify neighborhoods, binning, and calibration.

4.2 Proxy-Constrained Window Allocation

For $N \geq 2$, selection starts at $n = 2$ with the one-frame score as its baseline. Define the increment and non-negative gain:

$$\Delta H(n) = H_{\mathrm{CSE}}(n) - H_{\mathrm{CSE}}(n-1), \qquad G(n) = \max\{\Delta H(n), 0\}. \tag{11}$$

We normalize each gain within its scene:

$$G_{\max} = \max_{2 \leq k \leq N} G(k), \qquad r(n) = \frac{G(n)}{G_{\max} + \epsilon} = 1 - S(n). \tag{12}$$

Here $r(n)$ is relative marginal gain and $S(n)$ is saturation. With $\eta = 1 - \tau_s$, the proxy constraint defines eligible windows:

$$\mathcal{F}_t(\eta) = \{n \in \{2, \ldots, N\} : r(n) \leq \eta\}. \tag{13}$$

We select the shortest eligible window and use N when none qualifies:

$$n_t^* = \begin{cases} \min_{n \in \mathcal{F}_t(\eta)} n, & \mathcal{F}_t(\eta) \neq \varnothing, \\ N, & \mathcal{F}_t(\eta) = \varnothing. \end{cases} \tag{14}$$

This defines g_q by the first crossing of $S(n) \geq \tau_s$. Increasing $\tau_s \in [0, 1]$ cannot advance selection for fixed scores. The threshold and assignments transfer across compatible predictors on the dataset.

Eligibility checks one increment, without guaranteeing task sufficiency or small later gains. Non-positive gains may trigger selection before a rebound (Appendix F). All-zero gains select $n = 2$.

Algorithm 1 evaluates all candidates for $G_{\max}$ before one predictor call. Its overhead enters Eq. (4).

Algorithm 1 Prediction-free temporal context allocation

Require: Raw windows $\{\mathcal{V}_t^{(n)}\}_{n=1}^{N}$, threshold τ_s, calibrated CSE parameters
Ensure: Selected window length n_t^*
for $n = 1$ to N **do**
  Align and voxelize points; compute $H_g(n)$ and $H_l(n)$ using Eqs. (6)–(8)
  Pool image-gradient counts and compute $H_I(n)$ using Eq. (9)
  $H_{\mathrm{CSE}}(n) \leftarrow \alpha_g H_g(n) + \beta_g H_l(n) + \gamma_I H_I(n)$
end for
$G(n) \leftarrow \max\{H_{\mathrm{CSE}}(n) - H_{\mathrm{CSE}}(n-1), 0\}$ for all $n = 2, \ldots, N$
$G_{\max} \leftarrow \max_{2 \leq k \leq N} G(k)$
for $n = 2$ to N **do**
  $r(n) \leftarrow G(n)/(G_{\max} + \epsilon)$
  if $r(n) \leq 1 - \tau_s$ **then**
    return $n_t^* = n$ {First eligible window}
  end if
end for
return $n_t^* = N$ {Full-history fallback}

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate on nuScenes-SurroundOcc (Wei et al., 2023) and SSCBench-KITTI-360 (Li et al., 2024b). The primary metric is semantic mean intersection over union (mIoU), accompanied by geometric IoU, model-side latency, memory, and average selected window. We use three camera–LiDAR occupancy predictors with compatible temporal input paths: SAGFormer (Ning et al., 2026), SDGOcc (Duan et al., 2025), and DAOcc (Yang et al., 2025). On nuScenes, SAGFormer uses 900×1600 images, whereas SDGOcc and DAOcc use their official 256×704 resolution. For each predictor, fixed and adaptive variants share the official predictor-specific batch size, architecture, training protocol, and fusion mechanism; only the temporal input policy changes. Training and inference use FP32. The CSE parameters and per-sample assignments are reused across the three nuScenes predictors without predictor-specific recalibration. For KITTI-360, we repeat the CSE calibration on the target dataset; this independent calibration selects the same parameter values as on nuScenes and is evaluated with SAGFormer.

Candidate windows contain up to 16 consecutive dataset keyframes, including the current keyframe, following the standard temporal protocol of each benchmark. If only $L < 16$ keyframes are available for a target sample, we use all available keyframes and evaluate $n \in \{1, \dots, L\}$; for $L = 1$, the selector returns $n^* = 1$. The operating threshold is $\tau_s = 0.97$. Random Window samples an integer window length uniformly from 1 to 16 for each example. Uniform Sampling uses the same number of observations as its named budget but spreads historical frames over the 16-frame span. Shuffled CSE Scores permutes scene–window assignments while retaining dynamic allocation at a comparable average budget. These controls examine the role of scene-specific assignment alongside window length, temporal span, and dynamic allocation. Appendix C contains training, timing, preprocessing, and baseline details.

The nuScenes-SurroundOcc protocol covers $[-50, 50] \times [-50, 50] \times [-5, 3]$ m^3 with 0.5 m voxels and 18 labels. SSCBench-KITTI-360 provides 12,865 urban-scene frames over nine sequences; we use its standard seven/one/one sequence split and frontal $51.2 \times 51.2 \times 6.4$ m^3 volume. For SAGFormer, paired fixed and adaptive runs use identical preprocessing, initialization, architecture, losses, optimizer, and schedule. Fixed-1/3/6/9/16 and Adaptive are trained independently from the same pretrained checkpoint. SDGOcc and DAOcc Fixed-9 use released checkpoints; Fixed-16 and Adaptive retain their official settings and differ only in temporal policy. Principal and class-wise SAGFormer tables use seed-42 checkpoints. Appendix H.1 reports mean $\pm$ standard deviation from complete Fixed-9, Fixed-16, and Adaptive trainings with seeds 42, 37, and 69; Table 6 uses the same complete-training protocol.

We calibrate CSE with SAGFormer on the nuScenes validation split. With $\gamma_I = 1$, sequential coarse sweeps over α_g and β_g use the candidate set $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and select 0.4 and 0.6, respectively; the two coefficients are optimized independently without a sum-to-one constraint. The stopping threshold is chosen separately as an accuracy–cost operating point. This nuScenes calibration is reused unchanged for SDGOcc and DAOcc. Repeating the calibration on KITTI-360 selects the same parameter values. Table 6 provides a post-selection local stability check, while Appendix C.1 reports fine-grained calibration on nuScenes-mini and inference-time parameter transfer; Appendix C documents the full parameter provenance and calibration order.

5.2 TASK QUALITY AND PREDICTOR REUSE

Tables 2 and 3 evaluate task quality under adaptive allocation and report complete class-wise results. The nuScenes comparison includes published camera, multimodal, and temporal baselines (Wei et al., 2023; Huang et al., 2025b; Zhao et al., 2026; Pavković et al., 2025; Liao et al., 2025; Chen et al., 2025; Zhang et al., 2024). The KITTI-360 comparison covers voxel, camera, LiDAR, and temporal systems (Roldão et al., 2020; Song et al., 2017; Jiang et al., 2024; Li et al., 2024a; Lee et al., 2025; Yu et al., 2024; Liu et al., 2025; Wang et al., 2025). On nuScenes, the shared allocation rule improves each predictor's nine-frame result by 1.42–1.52 mIoU. It remains within 0.04–0.22 mIoU of Fixed-16 without predictor-specific recalibration. Three complete SAGFormer trainings yield 33.35 ± 0.06, 34.83 ± 0.05, and 34.79 ± 0.05 mIoU for Fixed-9, Fixed-16, and Adaptive. The Adaptive–Fixed-9 gain exceeds the observed variation. The 0.04-point gap to Fixed-16 is of the

Table 2: Complete class-wise comparison on nuScenes-SurroundOcc. C/L denote current camera/LiDAR; CH/LH include historical inputs. Fixed-n uses n frames and +Ours uses CSE allocation. Bold and underline mark the best and second-best values within each method group.

Method	Mod.	IoU	mIoU	bus	bicycle	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other	barrier	sidewalk	terrain	manmade	vegetation
SurroundOcc	C	31.49	20.30	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
GaussianFormer2	C	30.56	20.02	20.15	12.99	27.61	30.23	11.19	15.31	12.64	9.63	13.31	22.26	39.68	23.47	25.62	23.20	12.25	20.73
GaussianFormer3D	C+L	43.30	27.10	32.70	15.80	36.10	18.60	21.70	24.10	13.00	21.30	29.00	40.60	23.70	26.90	27.30	28.20	32.60	42.30
GaussianFusionOcc	C+L	45.16	30.21	35.91	18.70	39.57	22.67	27.36	30.10	18.59	24.45	31.25	43.06	25.76	30.22	29.12	29.33	34.65	42.70
STCOcc	CH	39.21	26.03	26.01	17.69	31.32	17.71	18.31	18.01	19.32	22.08	23.79	46.92	26.77	29.63	31.64	33.66	28.45	25.15
GDFusion	CH	38.20	25.53	29.78	18.78	34.32	17.90	19.61	17.48	17.17	15.55	26.52	47.62	29.87	24.32	31.23	29.20	20.04	29.05
OccFusion	C+LH	44.66	27.30	33.70	19.60	36.20	21.70	24.80	25.30	16.30	21.80	30.00	39.50	19.90	27.10	24.90	26.50	28.90	40.40
SAGFormer-Fixed-1	C+L	41.74	28.47	29.82	21.13	33.82	**37.30**	18.59	26.25	**27.83**	18.74	22.93	31.16	39.19	23.19	27.21	27.15	31.25	39.88
SAGFormer-Fixed-3	CH+LH	46.40	30.93	33.84	23.54	38.40	23.00	27.09	27.30	24.29	24.03	31.81	41.81	24.30	32.64	29.23	30.55	37.38	45.72
SAGFormer-Fixed-9	CH+LH	52.33	33.35	33.01	23.74	39.97	24.65	28.17	26.35	25.45	24.28	32.63	47.44	30.09	33.60	33.71	35.42	43.78	51.40
SAGFormer-Fixed-16	CH+LH	**55.38**	**34.83**	34.26	23.90	41.07	25.90	28.78	26.78	25.64	25.43	33.75	**50.32**	31.03	**35.30**	**35.96**	**37.60**	**47.12**	**54.47**
SAGFormer+Ours	CH+LH	54.92	34.79	**34.60**	**24.19**	**41.43**	26.33	**29.06**	**27.06**	26.07	**25.80**	**34.02**	49.85	30.49	34.76	35.48	36.94	46.48	53.92
SDGOcc-Fixed-9	C+LH	46.21	31.70	39.70	15.60	42.40	22.20	26.90	29.70	22.70	24.20	33.20	45.90	25.10	31.20	31.80	33.50	38.90	44.50
SDGOcc-Fixed-16	C+LH	**49.42**	**33.34**	40.71	16.71	43.24	23.30	27.94	**30.86**	**24.00**	**27.19**	**35.05**	**48.04**	**29.09**	**32.94**	**32.03**	**34.13**	**40.34**	**47.87**
SDGOcc+Ours	C+LH	49.11	33.12	**40.83**	**16.88**	**43.45**	**23.53**	**28.21**	30.61	23.69	26.85	34.68	47.65	28.67	32.50	31.57	33.64	39.83	47.33
DAOcc-Fixed-9	C+LH	44.57	30.50	34.00	19.50	38.80	25.00	27.70	29.90	22.50	23.20	31.60	41.00	25.90	30.80	29.40	29.90	35.20	43.50
DAOcc-Fixed-16	C+LH	**46.88**	**32.19**	34.63	20.18	39.51	25.75	28.48	**32.81**	**24.14**	**26.09**	**33.94**	**43.86**	**28.73**	**31.26**	**31.55**	**30.27**	**38.01**	**45.83**
DAOcc+Ours	C+LH	46.33	32.02	**34.75**	**20.35**	**39.71**	**25.98**	**28.76**	32.62	23.90	25.82	33.63	43.53	28.38	30.89	31.16	29.86	37.58	45.40

Table 3: Complete class-wise comparison on SSCBench-KITTI-360. Notation follows Table 2; an independent CSE calibration on KITTI-360 selects the same parameter values as on nuScenes.

Method	Mod.	IoU	mIoU	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traffic-sign	other-struct.	other-object
LMSCNet	L	47.53	13.65	20.91	0.00	0.00	0.26	0.00	0.00	62.95	13.51	33.51	0.24	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
SSCNet	L	53.58	16.95	31.95	0.00	0.17	10.29	0.58	0.07	65.71	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.60	0.67
GaussianFormer2	C	38.37	13.90	21.08	2.55	4.21	12.41	5.73	1.59	54.12	11.04	32.31	3.34	32.01	4.98	28.94	17.33	3.57	5.48	5.88	3.54
Symphonies	CH	44.12	18.58	30.02	1.85	5.90	25.07	12.06	8.20	54.94	13.83	32.76	6.93	35.11	8.58	38.33	11.52	14.01	9.57	14.44	11.28
HTCL	CH	44.23	17.09	27.30	1.80	2.20	5.70	5.40	1.10	64.40	33.80	34.80	12.40	25.90	21.10	25.30	31.20	9.00	8.30	–	–
SOAP	CH	46.09	19.09	28.90	5.00	5.30	3.30	7.00	3.70	63.60	36.80	36.20	17.20	28.70	22.50	29.80	32.30	11.70	13.10	–	–
CGFormer	C	48.07	20.05	29.85	3.42	3.96	17.59	6.79	6.63	63.85	17.15	40.72	5.53	42.73	8.22	38.80	24.94	16.24	17.45	10.18	6.77
DISC	C	47.35	20.55	29.41	4.64	8.27	19.24	8.51	6.74	61.88	17.56	40.09	5.27	42.53	9.24	38.76	23.05	16.73	19.51	10.32	8.21
L2COcc	L	57.60	25.22	40.41	4.08	4.23	26.15	10.30	9.94	70.01	22.62	46.16	8.78	51.04	14.34	47.60	31.33	24.89	21.64	13.96	6.41
SAGFormer-Fixed-1	C+L	57.40	24.20	40.35	3.59	4.10	26.12	8.83	7.80	65.93	20.15	46.28	8.76	51.93	14.00	43.77	29.54	24.58	22.01	13.07	4.79
SAGFormer-Fixed-3	CH+LH	58.61	25.68	41.04	4.83	5.31	26.51	11.18	10.14	70.50	22.70	46.73	9.14	51.39	15.18	48.08	31.18	24.80	21.97	14.60	6.89
SAGFormer-Fixed-9	CH+LH	59.33	26.73	43.30	5.97	5.88	26.36	11.66	11.24	**71.52**	23.74	47.71	10.01	51.65	16.57	**49.82**	**34.18**	26.53	22.94	15.14	6.85
SAGFormer-Fixed-16	CH+LH	**59.92**	**27.13**	43.19	5.94	5.90	28.41	11.84	12.35	71.49	**25.10**	47.40	9.25	**52.32**	15.44	48.74	33.43	**27.16**	**24.71**	**17.07**	**8.53**
SAGFormer+Ours	CH+LH	59.72	27.10	**43.72**	**5.99**	**6.53**	**28.97**	**12.08**	**12.44**	71.38	24.99	**47.80**	**11.02**	52.28	**15.98**	49.66	32.11	25.78	23.22	15.27	8.51

same order as that variation and supports a descriptive comparison rather than statistical equivalence. On KITTI-360, independent calibration yields a 0.03 mIoU gap to Fixed-16. The class-wise tables show how these aggregate results are distributed across categories.

5.3 Scene–Window Correspondence

Table 4 tests whether the proxy criterion assigns useful windows to individual scenes. With SAGFormer on nuScenes-SurroundOcc, adaptive allocation exceeds Fixed-9 by 1.44 mIoU at a similar average temporal budget. It exceeds Random Window, Uniform Sampling, and Shuffled CSE Scores by 1.60, 0.98, and 1.08 mIoU, respectively. After rounding, Random Window has an average window length of eight frames, while the other controlled policies have nine. Shuffling retains dynamic allocation while disrupting the correspondence between scenes and their assigned windows. Its lower accuracy supports the value of scene-specific assignment at comparable average budgets.

The retrospective diagnostic evaluates sample-level sufficiency alongside dataset-level mIoU. Under the present-class criterion in Appendix E, adaptive allocation reaches the Fixed-16 tolerance on 92.64% of validation samples, compared with 55.61% for Fixed-9. These proportions measure sample-level sufficiency at a tolerance of $\delta = 0.2$ mIoU points. The harmed-sample rates are 4.33% for Adaptive and 35.62% for Fixed-9. Harm requires both a loss above δ and a window shorter than the discrete oracle choice. Thus, the harmed rate is not the complement of the sufficient-sample rate. These post-hoc diagnostics provide neither a sufficiency guarantee nor a deployable oracle.

5.4 Processing Cost and Window Allocation

We measure processing cost with selector overhead included, as defined in Eq. (4). After historical point-cloud transformation and fusion, the controller takes 101 ms and the selected SAGFormer input takes 278 ms. Their model-side total is 379 ms, compared with 803 ms for Fixed-16. Adaptive allocation saves 424 ms or 52.80% while remaining within 0.04 mIoU of Fixed-16. It takes 90 ms

Table 4: Complete accuracy–cost–sufficiency comparison with SAGFormer on nuScenes-SurroundOcc. Avg. window is the mean input window length across samples, rounded to the nearest integer. Sufficiency uses the sample-wise present-class criterion in Appendix E. Memory is peak inference process memory in MB measured with `nvidia-smi`; it includes the selector, inputs, cache, and model weights. M-FPS is model-side FPS rounded to the nearest integer.

Method	Avg. window	IoU ↑	mIoU ↑	Memory ↓	M-FPS ↑	Reach 16-frame (%) ↑	Harmed (%) ↓
Fixed-window baselines							
Fixed-1	1	41.74	28.47	3012	4	1.04	98.96
Fixed-3	3	46.40	30.93	4233	3	9.43	75.13
Fixed-6	6	49.73	32.25	4792	3	30.35	54.33
Fixed-9	9	52.33	33.35	5033	3	55.61	35.62
Fixed-16	16	**55.38**	**34.83**	7422	1	100.00	0.00
Comparable-budget controls							
Random Window	8	51.34	33.19	4982	3	54.99	42.38
Uniform Sampling 9/16	9	52.68	33.81	5134	3	59.18	34.12
Shuffled CSE Scores	9	52.55	33.71	5134	3	57.44	35.11
Adaptive selection and sufficiency oracle							
Ours-Saturation	9	54.92	34.79	5112	3	92.64	4.33
Discrete Sufficiency Oracle	12	55.31	34.81	7122	1	100.00	0.00

Table 5: Raw-data proxy study on nuScenes-SurroundOcc. Avg. is the mean selected window length across samples, rounded to the nearest integer.

Variant	Avg.	mIoU ↑	IoU ↑
w/o Global Entropy	8	34.12	54.12
w/o Local Entropy	8	34.21	54.33
w/o Image-gradient Entropy	8	34.01	54.31
Point Cloud Density	9	33.45	52.44
Non-empty Voxels	9	33.41	52.14
Global Entropy Only	10	33.75	53.32
Local Entropy Only	8	33.85	53.13
Image-gradient Entropy Only	6	32.91	51.69
Full CSE	9	**34.79**	**54.92**

longer than Fixed-9 and gains 1.44 mIoU. Directly measured end-to-end latency is 436 ms for Adaptive and 879 ms for Fixed-16. This boundary includes data I/O, transfer, point-cloud transformation, and fusion. Table 14 gives the stage-wise breakdown.

The shared threshold produces different allocations across scenes. The controller assigns 9.35% of samples to 2–3 frames, 44.44% to 4–9 frames, and 46.21% to 10–16 frames. Thus, the average budget summarizes a range of selected windows. Appendix G reports the full distribution, scene-stratified behavior, threshold sweep, memory, and timing protocol.

5.5 Raw-Data Criteria and Threshold Sensitivity

Table 5 tests the raw-data statistics used to define eligible windows. Removing any CSE component changes the selected budget and reduces accuracy under the shared threshold $\tau_s = 0.97$. Full CSE also exceeds density, occupied-voxel, and single-component proxies. Global Entropy Only uses more frames on average but remains less accurate than Full CSE. More history alone therefore does not explain the ranking. These results assess each score within the complete allocation pipeline. They support the combined criterion at the shared threshold. They do not isolate component effects at equal budgets because each score changes the window assignments. Section 5.3 separately tests scene–window correspondence with comparable-budget controls.

Table 6 checks a local 3×3 neighborhood around the calibrated weights with $\gamma_I = 1$ and $\tau_s = 0.97$. This post-selection check is separate from parameter search. Mean mIoU ranges from 34.41 to 34.79, standard deviation ranges from 0.04 to 0.07, and the average window remains eight or nine frames. The selected weights $(\alpha_g, \beta_g, \gamma_I) = (0.4, 0.6, 1.0)$ yield the highest mean of 34.79 ± 0.05 mIoU at nine frames. All tested settings lie within 0.38 mIoU. The local results show that nearby cue weights retain similar allocation budgets and task quality.

Table 6: Post-selection local sensitivity to the relative CSE cue weights on nuScenes-SurroundOcc with SAGFormer. We fix $\gamma_I = 1$ and $\tau_s = 0.97$; each cell reports mean mIoU $\pm$ standard deviation over complete end-to-end trainings with seeds 42, 37, and 69 / average selected window. Seed 42 is the checkpoint used in the principal and class-wise result. The selected configuration is bold.

$\beta_g \backslash \alpha_g$	0.2	0.4	0.6
0.4	34.44 ± 0.06 / *8*	34.55 ± 0.06 / *9*	34.69 ± 0.05 / *9*
0.6	34.41 ± 0.05 / *8*	34.79 ± 0.05 / **9**	34.65 ± 0.04 / *9*
0.8	34.59 ± 0.04 / *9*	34.56 ± 0.04 / *9*	34.48 ± 0.07 / *9*

Table 7: Local sensitivity of the saturation threshold on nuScenes-SurroundOcc with SAGFormer. Avg. is the mean selected window; memory follows the peak inference `nvidia-smi` definition in Table 4; M-FPS is model-side FPS rounded to the nearest integer.

τ_s	Avg. ↓	mIoU ↑	Memory ↓	M-FPS ↑	Status
0.92	5	33.87	4613	3	Candidate
0.97	9	34.79	5112	3	Selected
0.99	13	34.82	7053	1	Candidate

Table 7 shows how the proxy threshold changes the operating point. Raising τ_s from 0.92 to 0.97 tightens eligibility, increases the average window from five to nine frames, and raises mIoU from 33.87 to 34.79. At 0.99, the average window reaches 13 frames for only 0.03 additional mIoU. Memory rises from 5112 to 7053 MB, and model-side throughput falls from 3 to 1 FPS. We select $\tau_s = 0.97$ from this empirical accuracy–cost tradeoff. Independent calibration on KITTI-360 selects the same threshold and remaining CSE parameters.

Potentially premature crossings. The local eligibility criterion can accept a non-positive increment before a later rebound. In a post-hoc diagnostic over $27{,}430$ nuScenes training frames, non-positive CSE changes trigger 28 selections or 0.1021% of samples. Twelve cases later rebound, representing 0.0437% of all 27,430 samples. None is harmed under the retrospective criterion in Appendix E. Appendix F reports the diagnostic. Its scope is limited to non-positive-gain-triggered crossings in this training run. It does not cover positive low-gain crossings or establish behavior beyond the evaluated setting.

6 Limitations

The current study has three main scope boundaries. First, the implementation evaluates all candidate CSE scores before running the selected predictor, making it a full-candidate controller rather than a streaming early exit. Second, the present study establishes CSE as an empirical allocation criterion but does not derive a theoretical equivalence or monotone relationship between the score and occupancy-label information; Appendix F covers non-positive-gain-triggered crossings, and long ego-aligned histories may retain stale geometry from moving objects. Third, the empirical evidence is limited to compatible camera–LiDAR occupancy predictors and does not establish the same behavior for other sensor configurations or perception tasks. Progressive data-level stopping rules and learned but predictor-external proxies are natural directions for extending the current formulation.

7 Conclusion

We formulate temporal context demand as a sample-level marginal-gain problem before occupancy inference. CSE uses raw geometric coverage, local structure, and visual diversity to select scene-specific history without predictor features, outputs, or parameters. Controlled comparisons at similar average budgets support the value of scene-specific window allocation. One calibration and its allocations are reused across three predictors on nuScenes, while an independent KITTI-360 calibration yields the same parameter values. Within compatible camera–LiDAR occupancy predictors, these results support raw-evidence saturation as a predictor-external criterion for temporal context allocation.

Reproducibility Statement

Appendix C specifies datasets, temporal policies, selector parameters, training relationships, timing boundaries, and implementation details. The supplement also includes complete CSE definitions, edge cases, extra tables, and diagnostics. We plan to include anonymized code and configurations with the submission package.

AI Use Disclosure

Generative AI tools were used to assist with manuscript organization, language editing, revision of selected passages, literature search and organization, and feedback on the presentation and interpretation of existing experimental results. All AI-assisted material was reviewed and verified by the authors. The authors also verified the manuscript's technical claims, equations, citations, experimental protocols, and reported results and take full responsibility for the final content of this work.

References

Emmanuele Barberi, Filippo Cucinotta, Per-Erik Forssén, and Felice Sfravara. Deci: A differential entropy-based compactness index for point clouds analysis: Method and potential applications. *Engineering Proceedings*, 56:273, 2023. doi: 10.3390/ASEC2023-15808.

Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24943–24954, 2025.

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.

Dubing Chen, Huan Zheng, Jin Fang, Xingping Dong, Xianfei Li, Wenlong Liao, Tao He, Pai Peng, and Jianbing Shen. Rethinking temporal fusion with a unified gradient descent view for 3D semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1505–1515, 2025. doi: 10.1109/CVPR52734.2025.00148.

Zaipeng Duan, Chenxu Dang, Xuzhong Hu, Pei An, Junfeng Ding, Jie Zhan, Yunbiao Xu, and Jie Ma. SDGOCC: Semantic and depth-guided bird's-eye view transformation for 3D multimodal occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6751–6760, 2025. doi: 10.1109/CVPR52734.2025.01587.

Tong He, Pei Sun, Zhaoqi Leng, Chenxi Liu, Dragomir Anguelov, and Mingxing Tan. LEF: Late-to-early temporal fusion for LiDAR 3D object detection. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. doi: 10.1109/IROS55552.2023.10341958.

Chengjie Huang, Vahdat Abdelzad, Sean Sedwards, and Krzysztof Czarnecki. Vadet: Multi-frame lidar 3d object detection using variable aggregation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 711–720, 2025a.

Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9223–9232, 2023.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, 2024.

Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. GaussianFormer-2: Probabilistic gaussian superposition for efficient 3D occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27477–27486, 2025b. doi: 10.1109/CVPR52734.2025.02559.

Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Tianwei Lin, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20258–20267, 2024.

Kieran G. Larkin. Reflections on shannon information: In search of a natural information-entropy for images. *arXiv preprint arXiv:1609.01117*, 2016.

Hyo-Jun Lee, Yeong Jun Koh, Hanul Kim, Hyunseop Kim, Yonguk Lee, and Jinu Lee. Soap: Vision-centric 3d semantic scene completion with scene-adaptive decoder and occluded region-aware view projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17145–17154, 2025. doi: 10.1109/CVPR52734.2025.01598.

Ziyang Leng, Jiawei Yang, Wenlong Yi, and Bolei Zhou. Occupancy learning with spatiotemporal memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26569–26578, October 2025.

Bohan Li, Jiajun Deng, Wenyao Zhang, Zhujin Liang, Dalong Du, Xin Jin, and Wenjun Zeng. Hierarchical temporal context learning for camera-based semantic scene completion. In *European Conference on Computer Vision*, volume 15062 of *Lecture Notes in Computer Science*, pp. 131–148, 2024a. doi: 10.1007/978-3-031-73235-5_8.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9087–9098, 2023a.

Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. SSCBench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024b. URL https://arxiv.org/abs/2306.09001.

Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1486–1494, 2023b. doi: 10.1609/aaai.v37i2.25234.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, volume 13669 of *Lecture Notes in Computer Science*, pp. 1–18, 2022. doi: 10.1007/978-3-031-20077-9_1.

Zhimin Liao, Ping Wei, Shuaijia Chen, Haoxuan Wang, and Ziyang Ren. Stcocc: Sparse spatial-temporal cascade renovation for 3d occupancy and scene flow prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1516–1526, 2025.

Enyu Liu, En Yu, Sijia Chen, and Wenbing Tao. Disentangling instance and scene contexts for 3d semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26999–27009, 2025.

Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18534–18544, 2023.

Kanglin Ning, Yiran Zhao, Wenrui Li, Houde Quan, Qifan Li, Xingtao Wang, and Xiaopeng Fan. Learning adaptive semantic gaussian allocation for 3d occupancy, 2026. URL https://arxiv.org/abs/2607.21896. arXiv:2607.21896.

Wenzhe Ouyang, Zenglin Xu, Bin Shen, Jinghua Wang, and Yong Xu. Linkocc: 3d semantic occupancy prediction with temporal association. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1374–1384, 2025. doi: 10.1109/TCSVT.2024.3486019.

Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=H3HcEJA2Um.

Tomislav Pavković, Mohammad-Ali Nikouei Mahani, Johannes Niedermayer, and Johannes Betz. Gaussianfusionocc: A seamless sensor fusion approach for 3d occupancy prediction using 3d gaussians. *arXiv preprint arXiv:2507.18522*, 2025.

Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *International Conference on 3D Vision*, pp. 111–119, 2020.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1746–1754, 2017.

Jufei Wang, Mingwei Wei, Tianheng Qiu, Hu Wei, Huasheng Ni, and Xuan Huang. Ngocc: Normal-guided 3d occupancy prediction with observability-aware time fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 340–349, June 2026.

Ruoyu Wang, Yukai Ma, Yi Yao, Sheng Tao, Haoang Li, Zongzhi Zhu, Yong Liu, and Xingxing Zuo. L2COcc: Lightweight camera-centric semantic scene completion via distillation of LiDAR model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 716–723, 2025. doi: 10.1109/IROS60139.2025.11245916.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21729–21740, 2023.

Zhenyu Wu, Xiuwei Xu, Ziwei Wang, Chong Xia, Linqing Zhao, Jiwen Lu, and Haibin Yan. AnyView: General indoor 3D object detection with variable frames. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13333–13340, 2025. doi: 10.1109/IROS60139.2025.11246096.

Zhen Yang, Yanpeng Dong, Jiayu Wang, Heng Wang, Lichao Ma, Zijian Cui, Qi Liu, Haoran Pei, Kexin Zhang, and Chao Zhang. Daocc: 3d object detection assisted multi-sensor fusion for 3d occupancy prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. doi: 10.1109/TCSVT.2025.3610634.

Zhangchen Ye, Tao Jiang, Chenfeng Xu, Yiming Li, and Hang Zhao. CVT-Occ: Cost volume temporal fusion for 3D occupancy prediction. In *European Conference on Computer Vision*, volume 15131 of *Lecture Notes in Computer Science*, pp. 381–397, 2024. doi: 10.1007/978-3-031-73464-9_23.

Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Shanshuai Yuan, Julong Wei, Muer Tie, Xiangyun Ren, Zhongxue Gan, and Wenchao Ding. Lmpocc: 3d semantic occupancy prediction utilizing long-term memory prior from historical traversals. *arXiv preprint arXiv:2504.13596*, 2025.

Ji Zhang, Yiran Ding, and Zixin Liu. OccFusion: Depth estimation free multi-sensor fusion for 3D occupancy prediction. In *Asian Conference on Computer Vision*, pp. 232–249, 2024. doi: 10.1007/978-981-96-0972-7_14.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9433–9443, 2023.

Lingjun Zhao, Sizhe Wei, James Hays, and Lu Gan. GaussianFormer3D: Multi-modal gaussian-based semantic occupancy prediction with 3D deformable attention. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://arxiv.org/abs/2505.10685.

A Supplementary Material Overview

The supplement follows the order needed to reconstruct and then interpret the study. Appendix B completes the CSE definition, and Appendix C records the evaluation and calibration protocol. Appendix D then expands the empirical saturation evidence, while Appendices E–I provide diagnostic, allocation, cost, and qualitative analyses.

B Complete Composite Scene Entropy Definition

Historical point clouds are transformed into the current ego coordinate system and voxelized. For occupied voxel v_i, let $\lambda_{i,1} \geq \lambda_{i,2} \geq \lambda_{i,3}$ be eigenvalues of its K-neighbor covariance matrix and

$$\rho_{i,j} = \frac{\lambda_{i,j} + \epsilon}{\sum_{r=1}^{3}(\lambda_{i,r} + \epsilon)}, \quad d_i = \left[\frac{1}{2} \log \left((2\pi e)^3 \prod_{j=1}^{3} \rho_{i,j} + 1 \right) + \epsilon \right]^{-1}. \tag{15}$$

The local term is $H_l(n) = \mathrm{MeanTopK}(\{d_i\}_{i=1}^{M}, \min(K_l, M))$. For camera c at frame r, let $C_{ab}(I_{r,c})$ count Sobel-gradient pairs in two-dimensional bin (a, b). We pool unnormalized counts over the window and cameras:

$$C_{ab}^{(n)} = \sum_{r=t-n+1}^{t} \sum_{c=1}^{N_c} C_{ab}(I_{r,c}), \quad p_{ab}^{(n)} = \frac{C_{ab}^{(n)}}{\sum_{u,v} C_{uv}^{(n)} + \epsilon}, \tag{16}$$

$$H_I(n) = -\sum_{a=1}^{B_x} \sum_{b=1}^{B_y} p_{ab}^{(n)} \log(p_{ab}^{(n)} + \epsilon). \tag{17}$$

Pooling before normalization makes repeated patterns contribute progressively smaller distributional changes. If all adjacent gains are zero, the earliest eligible window is selected; if no ratio crosses τ_s, the selector returns N.

C Implementation and Evaluation Protocol

CSE input protocol. For both nuScenes and KITTI-360, CSE operates on consecutive dataset keyframes and uses at most 16 keyframes including the current one. When fewer than 16 keyframes are available, the candidate set is truncated to the available history. All image statistics used by CSE are computed after resizing the images to 704×256 pixels (width $\times$ height), independently of the input resolution used by the downstream occupancy predictor. For each dataset, the CSE voxel grid follows the spatial range and voxel size of the corresponding benchmark configuration used by the methods compared in Tables 2 and 3. This grid is fixed across compatible downstream predictors on that dataset, and points outside the configured spatial range are clipped.

SAGFormer is trained independently for evaluated fixed windows. It uses 900×1600 images on nuScenes and 376×1408 images on KITTI-360; SDGOcc and DAOcc retain their official 256×704 resolution. Each predictor uses its official batch size, and all training and inference are performed in FP32. For SAGFormer, ResNet-50 uses RayDN/ImageNet initialization on nuScenes/KITTI-360; training uses AdamW with 0.01 weight decay, a maximum learning rate of 2×10^{-4}, a one-cycle schedule, and 20/30 epochs. Preprocessing applies photometric distortion and random image flipping without additional point-cloud augmentation. Training uses an Intel i9-14900KF and four RTX 4090 GPUs with PyTorch 2.5.0, CUDA 12.1, and cuDNN 9.0.

SDGOcc and DAOcc use their predictor-specific official protocols. Within each predictor, Fixed-16 and Adaptive share model design and optimization; only temporal input assignment changes. The same CSE parameters and assignments are reused across three nuScenes predictors. Random Window samples an integer window length uniformly from 1 to 16 during both training and inference. Uniform Sampling-9 uses the current frame plus eight historical positions spread over positions 2–16, with the same policy in training and inference.

Parameter provenance and calibration. Candidate CSE configurations are assessed by validation mIoU using SAGFormer trained on the official nuScenes split. With $\gamma_I = 1$, we first fix $\beta_g = 1$ and evaluate $\alpha_g \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, selecting 0.4; we then fix $\alpha_g = 0.4$ and evaluate the same five values for β_g, selecting 0.6. These two coarse sweeps contain ten executed evaluations and nine distinct configurations because $(0.4, 1.0)$ is evaluated in both stages. The coefficients are searched independently: no constraint enforces $\alpha_g + \beta_g = 1$, and their selected sum is an empirical outcome of the search. The threshold is selected separately as an accuracy–cost operating point: moving τ_s from 0.97 to 0.99 adds four frames for only 0.03 mIoU. The resulting nuScenes calibration is reused for SDGOcc and DAOcc. On KITTI-360, a separate weight comparison on the validation split, similar to Table 6, again favors $(\alpha_g, \beta_g) = (0.4, 0.6)$. The KITTI-360 predictor is trained using assignments generated with these weights, initially selected on nuScenes. We set $K = 64$, $K_l = 10$, and the 32×32 image histogram following their source formulations (Barberi et al., 2023; Larkin, 2016); $\epsilon = 10^{-6}$ is fixed heuristically. Table 6 is a post-selection local-robustness check, not evidence of global optimality or search-order invariance. Appendix C.1 refines the coarse nuScenes solution on the official nuScenes-mini training split by searching $\alpha_g \in \{0.36, 0.38, 0.40, 0.42, 0.44, 0.46\}$ and $\beta_g \in \{0.56, 0.58, 0.60, 0.62, 0.64, 0.66\}$.

Shuffled CSE is also trained as an independent SAGFormer variant from the same pretrained initialization as the other controlled policies. It retains the same architecture, preprocessing, losses, optimizer, batch size, training schedule, and evaluator; its only designed change is the temporal context assigned to each scene. The scene–window assignments are shuffled before being supplied to the predictor, thereby breaking the correspondence between a scene's CSE-derived demand and its allocated temporal window while retaining dynamic allocation and a comparable average temporal budget. The shuffled assignment policy is applied consistently during both training and inference. Thus, the Shuffled CSE result is produced by a model trained under the same type of shuffled temporal-input policy used at evaluation, rather than by applying an inference-time perturbation to the Adaptive checkpoint. This matched training–inference construction makes Shuffled CSE a control for the scene–window correspondence instead of a test of robustness to temporal-policy distribution shift.

Runtime is reported at two measurement boundaries on nuScenes using one RTX 4090, batch size 1, and FP32 under PyTorch 2.5.0, CUDA 12.1, and cuDNN 9.0. After 10 warm-up iterations, we measure 27,430 iterations with CUDA synchronization. For each target sample, historical point clouds are transformed, aligned, and fused once, then cached for reuse across the nested candidate windows. Table 14 reports this one-time preparation separately as "PC align. + fusion"; the "CSE selector" column reports the subsequent entropy computation and stopping rule. Together, these two stages form the complete CSE selection path. Model-side latency begins after point-cloud preparation and is the sum of the reported CSE-selector and occupancy-predictor columns; the predictor processes only the selected window. The directly measured end-to-end latency includes data I/O and transfer, the complete CSE selection path, occupancy prediction, and other measured pipeline overhead, while excluding sensor acquisition. It is not reconstructed by summing independently recorded stage averages. Model-side FPS is derived from selector-plus-prediction latency and rounded to the nearest integer. Memory is peak inference process memory observed with `nvidia-smi` and includes the selector, inputs, cache, and model weights.

The planned anonymous artifact will include CSE computation, temporal assignments, configurations, and scripts for reproducing reported tables. Repository history, paths, metadata, and documentation will be anonymized before submission.

C.1 Fine-Grained Mini-Split Calibration and Transfer

Table 8 refines the coarse nuScenes solution $(\alpha_g, \beta_g) = (0.4, 0.6)$ on the official nuScenes-mini training split. Searching the denser neighborhoods shown in the table selects $(\alpha_g, \beta_g, \gamma_I, \tau_s) = (0.42, 0.58, 1.00, 0.97)$. As in the coarse search, α_g and β_g are not constrained to sum to one; the selected pair does so only as a result of the measured optimum.

To limit additional training cost, the nuScenes transfer evaluation retains the SAGFormer checkpoint trained with assignments from the original validation-calibrated CSE and replaces only the CSE parameters used to allocate windows at inference. The mini-calibrated parameters yield 54.91 IoU and 34.69 mIoU on nuScenes validation with an average of nine frames, within 0.01 IoU and 0.10

Table 8: Fine-grained CSE calibration on nuScenes-mini and inference-time parameter transfer with fixed SAGFormer checkpoints. Both checkpoints are trained on their respective training splits using CSE weights $(0.4, 0.6)$. Evaluation replaces the inference-time weights with $(0.42, 0.58)$ from nuScenes-mini. Target recalibration refers to additional target-side tuning of these transferred inference parameters.

Calibration split	α_g candidates	β_g candidates	Selected $(\alpha_g, \beta_g, \gamma_I, \tau_s)$
nuScenes-mini train	$\{0.36, 0.38, 0.40, 0.42, 0.44, 0.46\}$	$\{0.56, 0.58, 0.60, 0.62, 0.64, 0.66\}$	$(0.42, 0.58, 1.00, 0.97)$

Evaluation split	Checkpoint training split / policy	Inference CSE parameter source	Target recalib.	Avg. frames	IoU ↑	mIoU ↑
nuScenes validation	nuScenes train / validation-calibrated CSE	nuScenes-mini train	No	9	54.91	34.69
SSCBench-KITTI-360	KITTI-360 train / CSE weights $(0.4, 0.6)$	nuScenes-mini train (transferred)	No	9	59.62	27.08

mIoU of the original nuScenes operating point. This comparison therefore measures sensitivity to the inference-time calibration source under a fixed checkpoint, rather than complete calibration isolation during predictor training.

For the cross-dataset transfer, we retain the KITTI-360 SAGFormer checkpoint trained using CSE weights $(0.4, 0.6)$ and replace its inference-time weights with $(0.42, 0.58)$ from nuScenes-mini, without further KITTI-360-side tuning. This transferred configuration reaches 59.62 IoU and 27.08 mIoU with an average of nine frames, within 0.10 IoU and 0.02 mIoU of the independently calibrated KITTI-360 operating point. The comparison measures sensitivity to transferred inference-time weights with the target-trained predictor held fixed.

D Extended Motivation and Temporal Saturation

Table 9: Dataset-averaged CSE trajectory over nested windows on nuScenes-SurroundOcc.

n	H_g	H_l	H_I	H_{CSE}	$G(n)$	$S(n)$
1	7.86	1.92	3.88	8.18	–	–
2	8.24	2.66	4.74	9.63	1.45	0.000
3	8.47	2.88	5.12	10.23	0.60	0.586
4	8.64	3.02	5.43	10.70	0.47	0.676
5	8.77	3.12	5.64	11.02	0.32	0.779
6	8.87	3.20	5.77	11.24	0.22	0.848
7	8.96	3.26	5.86	11.40	0.16	0.890
8	9.04	3.32	5.91	11.52	0.12	0.917
9	9.11	3.37	5.95	11.61	0.09	0.938
10	9.17	3.40	5.98	11.68	0.07	0.952
11	9.22	3.43	6.00	11.75	0.07	0.952
12	9.26	3.46	6.02	11.80	0.05	0.966
13	9.31	3.47	6.04	11.84	0.04	0.972
14	9.33	3.49	6.05	11.87	0.03	0.979
15	9.36	3.51	6.05	11.90	0.03	0.979
16	9.39	3.52	6.06	11.92	0.02	0.986

The aggregate sequence grows rapidly through the first six frames and then flattens. The largest adjacent increase occurs when the second frame is added, after which $G(n)$ decreases from 1.45 to 0.02 by the 16-frame window. Correspondingly, the aggregate saturation ratio first exceeds the selected threshold of 0.97 at 13 frames. This dataset-level crossing is not the controller's average decision: selection is performed per sample, and scenes can saturate earlier or later than the aggregate trajectory.

Across fixed windows, composite entropy correlates with mIoU (Pearson 0.945, Spearman 1.000), and non-negative CSE gain correlates with mIoU gain (Pearson 0.854, Spearman 0.999). These associations provide a descriptive consistency check that the score and task metric follow compatible diminishing-return trends as history grows. The comparable-budget policies and shuffled assignments in Table 4 separately test whether the sample-specific allocation improves predictive performance beyond the shared ordering by window length.

E Discrete Sufficiency Diagnostic

For validation sample i, let $\mathcal{C}_i$ contain classes present in ground truth. At candidate n,

$$m_i(n) = \frac{100}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \frac{|\widehat{Y}_{i,n,c} \cap Y_{i,c}|}{|\widehat{Y}_{i,n,c} \cup Y_{i,c}|}. \tag{18}$$

The retrospective oracle selects the shortest $n \in \{1, 3, 6, 9, 16\}$ satisfying $m_i(n) \geq m_i(16) - 0.2$. A sample is harmed when its window is shorter than the oracle choice and its score is more than 0.2 points below Fixed-16. This post-hoc diagnostic is neither deployable nor a theoretical upper bound.

Under this definition, the adaptive policy is sufficient for 92.64% of validation samples and harms 4.33%, whereas Fixed-9 is sufficient for 55.61% and harms 35.62%. The contrast complements the dataset-level mIoU comparison by showing that the adaptive gain is distributed across samples rather than arising only from aggregate class averaging. Because both rates depend on the chosen 0.2 tolerance and on post-hoc predictions from the evaluated candidates, they characterize the reported operating point rather than establish tolerance-invariant sufficiency.

F Diagnostic of Non-positive-gain-triggered Crossings

The positive-part operation maps every non-positive adjacent CSE change to $G(n) = 0$ and hence $S(n) = 1$. Such a change therefore satisfies the stopping threshold and can determine the selected window if no earlier candidate has already crossed it. We examine how frequently this event occurs in the $27{,}430$ frames of the nuScenes-SurroundOcc training run. For the harm diagnostic, we apply the same present-class score and 0.2-point tolerance defined in Appendix E. This post-hoc analysis is not used to select the threshold or modify the stopping rule.

Table 10: Post-hoc diagnostic of non-positive-gain-triggered crossings on the nuScenes-SurroundOcc training set.

Diagnostic	Count (ratio)
Selection triggered by a non-positive CSE gain	28 (0.1021%)
Triggered selection followed by a positive rebound	12 (0.0437%)
Triggered selection harmed under the retrospective criterion	0 (0.0000%)

Non-positive-gain-triggered selection is therefore rare in the evaluated training run. Although 12 cases exhibit a later positive rebound, none of the 28 triggered selections produces an observed harm under the adopted retrospective criterion. This result addresses only the special case in which a non-positive adjacent gain directly determines the selected window. It does not evaluate positive but small gains that cross the threshold, nor does it establish robustness on validation data, other datasets, different sensor configurations, or streaming acquisition conditions.

G Extended Allocation and Robustness Analyses

The allocation spans the full eligible range rather than collapsing to one fixed budget. Short selections of 2–3 frames account for 9.35% of samples, 44.44% fall in the intermediate 4–9 frame range, and 46.21% retain 10–16 frames. Thus, the average nine-frame budget combines genuinely short and long decisions instead of reproducing Fixed-9 for most scenes. The large fraction of long selections also shows that the controller does not treat temporal reduction as the objective when marginal evidence continues to accumulate.

Table 11: Distribution of selected windows.

Selected window	Ratio
2–3	9.35%
4–6	19.28%
7–9	25.16%
10–12	32.47%
13–16	13.74%

Higher-entropy samples receive longer histories: low-, medium-, and high-entropy groups average 5, 9, and 13 frames. Their adaptive mIoU values are 36.01, 35.12, and 33.24, compared with Fixed-16 values of 36.28, 35.09, and 33.12. The adaptive values remain within 0.27 mIoU of Fixed-16 across the three strata while the mean allocation increases with the proxy. Because grouping and selection use the same CSE-derived signal, this analysis serves as an internal-consistency diagnostic of the allocation rule; the comparable-budget and shuffled-assignment controls provide the independent performance tests.

G.1 Frozen-Predictor Inference-Time Policy Switching

Table 12: Inference-time temporal-policy switching on nuScenes-SurroundOcc using one frozen SAGFormer checkpoint trained with Fixed-9 inputs. Avg. frames is the mean input window length across samples, rounded to the nearest integer.

Inference policy	Checkpoint	Avg. frames	IoU ↑	mIoU ↑
Fixed-9	Fixed-9, frozen	9	52.33	33.35
Random Window	Fixed-9, frozen	8	51.89	32.95
Shuffled CSE Scores	Fixed-9, frozen	9	52.67	33.69
CSE	Fixed-9, frozen	9	54.92	34.75
Fixed-16 input	Fixed-9, frozen	16	55.16	34.81

Table 12 isolates inference-time allocation by holding the Fixed-9-trained SAGFormer weights constant. CSE achieves 54.92 IoU and 34.75 mIoU. At comparable average temporal budgets, it outperforms Fixed-9 by 2.59 IoU and 1.40 mIoU and Shuffled CSE Scores by 2.25 IoU and 1.06 mIoU. It also exceeds Random Window by 3.03 IoU and 1.80 mIoU. Shuffling retains variable window lengths while disrupting their correspondence with individual scenes. With the predictor held fixed, the resulting accuracy gap supports the value of CSE-guided scene–window assignments. Fixed-16 yields the highest frozen-checkpoint result, with 55.16 IoU and 34.81 mIoU, but exceeds CSE by only 0.24 IoU and 0.06 mIoU while using 16 instead of nine frames on average. Because all policies share the same frozen predictor weights, these results further show that scene-adaptive inference itself provides a substantial benefit without policy-specific retraining.

H Extended Cost and Ablation Interpretation

H.1 Variation across complete training runs

Table 13: Training-run variation for the principal SAGFormer policies on nuScenes-SurroundOcc. Each entry reports mean ± standard deviation over three complete end-to-end trainings with seeds 42, 37, and 69. The principal and class-wise tables retain the seed-42 checkpoints.

Policy	Avg. frames	IoU ↑	mIoU ↑
Fixed-9	9	52.33 ± 0.11	33.35 ± 0.06
Fixed-16	16	55.38 ± 0.13	34.83 ± 0.05
Adaptive	9	54.92 ± 0.11	34.79 ± 0.05

The repeated trainings preserve the seed-42 ordering. Relative to Fixed-9, Adaptive gains 2.59 IoU and 1.44 mIoU at a similar average frame count; both mean separations exceed the corresponding

individual run-to-run standard deviations by more than an order of magnitude. The Adaptive–Fixed-16 gaps are 0.46 IoU and 0.04 mIoU while Adaptive uses nine rather than 16 frames on average. These descriptive statistics quantify training variation but do not constitute a paired significance, equivalence, or non-inferiority test because per-seed paired differences are not reported.

H.2 Model-Side and End-to-End Inference Cost

Table 14: Stage-wise, model-side, and end-to-end SAGFormer inference cost on nuScenes-SurroundOcc using one RTX 4090, batch size 1, FP32, and 900×1600 images. Latency is measured in milliseconds over 27,430 CUDA-synchronized iterations after 10 warm-up iterations. Model-side total comprises selector and predictor latency. End-to-end totals are measured directly and include residual pipeline overhead, so they need not equal the arithmetic sum of the displayed stages. Peak memory includes the selector, inputs, cache, and model weights.

Policy	Avg. frames	Input preparation (ms)		Model-side latency (ms)			E2E total (ms)	Peak memory (MB)
		Load + transfer	PC align. + fusion	CSE selector	Predictor	Total		
Fixed-9	9	16	21	–	289	289	335	5033
Adaptive (CSE)	9	26	22	101	278	379	436	5112
Fixed-16	16	26	34	–	803	803	879	7422

Table 14 reports both model-side and full-pipeline inference costs under a common measurement protocol. Relative to Fixed-16, Adaptive reduces model-side latency from 803 to 379 ms, saving 424 ms (52.80%), and reduces end-to-end latency from 879 to 436 ms, saving 443 ms (50.40%). Peak memory decreases from 7422 to 5112 MB, corresponding to a reduction of 2310 MB (31.12%). The predictor stage requires only 278 ms because SAGFormer processes the selected temporal window, while the predictor-external selector adds 101 ms and accounts for 26.65% of the adaptive model-side latency.

Fixed-9 remains the lowest-cost operating point, requiring 289 ms model-side latency, 335 ms end-to-end latency, and 5033 MB peak memory. At a comparable average temporal budget, Adaptive adds 90 ms model-side latency, 101 ms end-to-end latency, and 79 MB memory, while improving IoU by 2.59 points and mIoU by 1.44 points. Adaptive therefore forms an intermediate accuracy–cost operating point between Fixed-9 and Fixed-16: it approaches the accuracy of Fixed-16 with substantially lower latency and memory, while exchanging additional controller computation for higher accuracy than Fixed-9. Since the end-to-end totals are measured directly, they also include pipeline overhead outside the separately reported stages and are not defined as their arithmetic sum.

I Qualitative Allocation Behavior

Figure 3 provides a geometric view of diminishing returns. Across the displayed sequences, the first historical additions fill previously sparse regions and reveal scene boundaries that are weakly observed in the shortest window. As the window grows, new points increasingly densify structures that are already visible, while the change in overall spatial coverage becomes smaller. The examples visualize the raw-evidence pattern targeted by CSE; they do not by themselves establish occupancy accuracy or a universal saturation rate.

Figure 4 complements this progression with scene-specific selections. Some inputs become structurally stable after only a few historical observations and therefore receive short windows; others continue to gain visible coverage and retain longer histories. The selected examples thus illustrate that allocation follows the rate of marginal evidence accumulation rather than applying one temporal depth to every scene. This behavior is consistent with the broad window distribution and entropy-stratified averages in Appendix G, but the examples are descriptive and should not be interpreted as an independent generalization test.

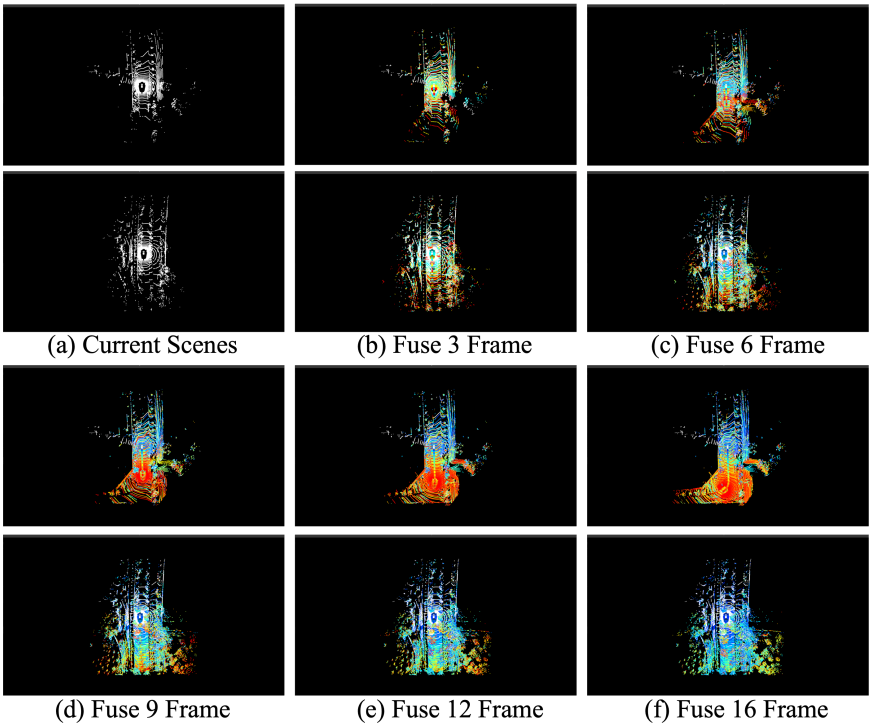


Figure 3: Early history expands point-cloud coverage, while later frames increasingly repeat observed geometry. This visualization motivates saturation but is not independent effectiveness evidence.

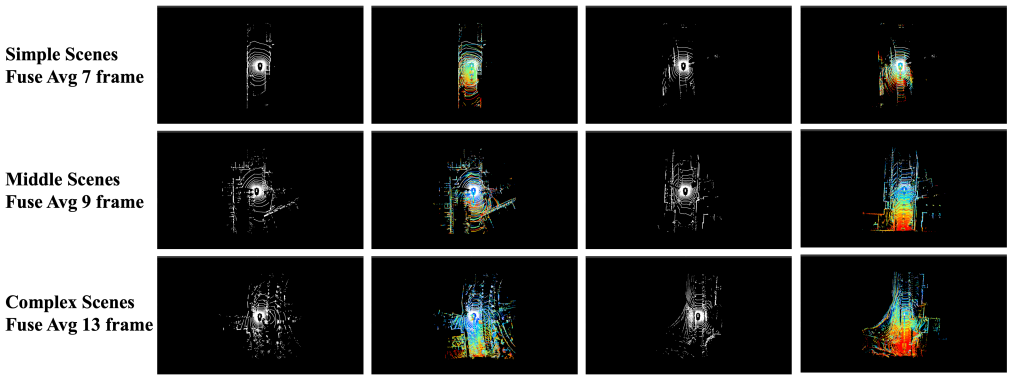


Figure 4: Examples of adaptive allocation. Earlier saturation yields shorter histories, while other scenes retain longer context. The examples illustrate controller behavior rather than establish generalization.