SPIKEOCC: TOWARDS RESILIENT MULTIMODAL 3D OCCUPANCY PERCEPTION VIA SPIKING EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal 3D occupancy prediction underpins safe autonomous driving through comprehensive scene understanding. However, existing methods have largely pursued higher accuracy through sophisticated fusion, leaving resilience to missing sensors underexplored. In this paper, we investigate spiking neural networks (SNNs) as a means of building more resilient multimodal perception. Still, the discrete nature of spiking neurons makes optimization challenging. To overcome this difficulty, SpikeOcc is introduced with two complementary components, a Spiking Mixture-of-Experts (SMoE) and Sensor Availability Learning (SAL). Specifically, SMoE organizes RGB, LiDAR, and joint fusion into specialized pathways, enabling full-width unimodal representations while retaining the original multimodal features via a bounded residual. Beyond sensor dropout, SAL employs a staged curriculum with explicit routing supervision, progressively improving robustness under severe sensing conditions. Remarkably, SAL also yields notable gains for conventional models, highlighting its broader potential for resilient multimodal learning. Extensive experiments demonstrate that SpikeOcc maintains competitive full-sensor accuracy while markedly raising the performance floor under sensor outages. SpikeOcc illustrates the promise of neuromorphic computing, pointing toward more reliable and energy-efficient autonomous driving.

1 INTRODUCTION

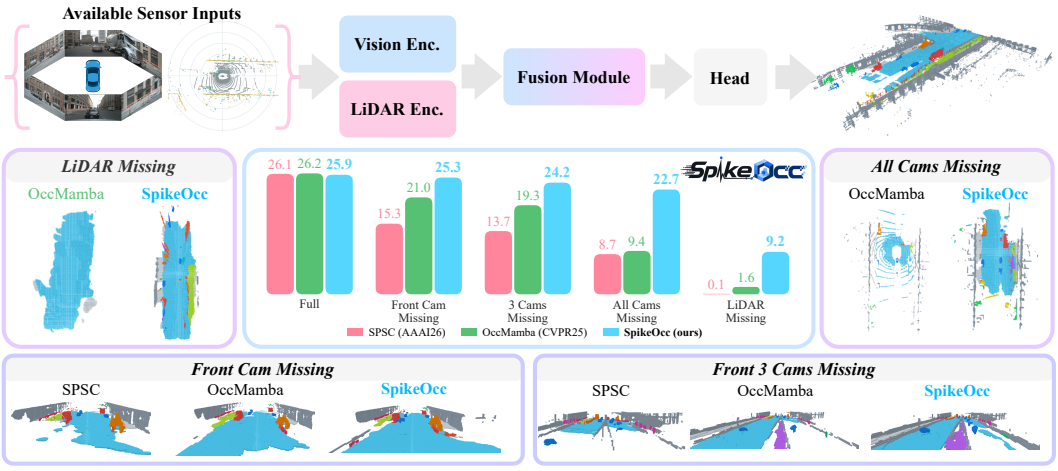
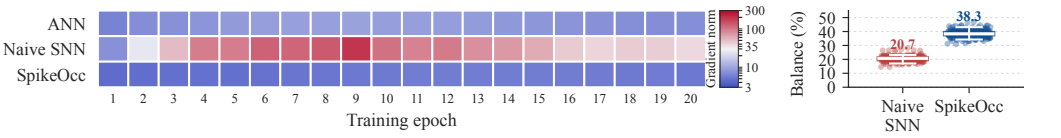


Figure 1: **Full-sensor accuracy does not guarantee resilience to sensor failures.** Existing MMOMs degrade sharply under modality outages, whereas SpikeOcc remains more resilient.

The landscape of 3D occupancy prediction has undergone a paradigm shift from single-sensor perception to tightly coupled multi-sensor fusion (Tian et al., 2023; Wang et al., 2024a; 2023b). Existing multimodal occupancy models (MMOMs) largely share a common pipeline, with major architectural advances concentrated in the fusion module. Yet, behind these advances lies an implicit but stringent assumption: All sensing modalities remain fully available at inference time.

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.



(a) Epoch-wise gradient instability **(b) Cross-modal firing balance**

Figure 2: **Naive spiking fusion exhibits instability.** Firing balance is the lower-rate modality's normalized share and reaches 50% for equal rates.

In practice, real-world sensing rarely guarantees such a fixed and complete configuration. Cameras may become unavailable because of communication dropouts or hardware malfunctions, and LiDAR faces the same risks. Recognizing this issue, BEV-based models, such as MetaBEV (Ge et al., 2023), have attracted increasing attention. Subsequent studies extend this awareness through unified fusion (Wang et al., 2024b), expert routing (Park et al., 2025), and masked-view training (Chen et al., 2024).

However, due to its complex scene structures, sensor-level failures remain largely underexplored in 3D occupancy. This leaves a critical question unanswered: **how does multimodal occupancy behave when part of its sensing stack becomes unavailable?** Fig. 1 reveals a clear failure pattern. A strong MMOM already suffers when the front camera is removed. Once an entire modality disappears, the degradation becomes catastrophic. We refer to this disproportionate collapse as modality outage fragility, raising concerns about the reliability of on-board perception.

Spiking neural networks (SNNs) have recently emerged for autonomous driving (Zhu et al., 2024), with encouraging robustness against corrupted observations (Kundu et al., 2021; Ding et al., 2024). These works motivate us to investigate whether spiking representations can also benefit structured 3D occupancy under changing sensor availability *(See Appendix A)*. Our initial results suggest that the answer is far from straightforward. In Fig. 2(a), directly converting a strong MMOM into its spiking counterpart causes severe gradient explosion during training. Camera and LiDAR features follow different representation statistics, and spiking conversion further translates this discrepancy into heterogeneous firing dynamics *(See Appendix B)*. Worse still, routing both modalities through a shared fusion pathway entangles these dynamics and makes optimization harder in Fig. 2(b).

Motivated by these observations, we introduce SpikeOcc, the first spike-driven MMOM for 3D occupancy prediction. SpikeOcc combines two complementary designs. First, we develop a Spiking Mixture-of-Experts (SMoE) that decomposes multimodal fusion into RGB, LiDAR, and joint experts. Each modality is processed along a dedicated spiking pathway, while the joint expert retains cross-modal interaction. This specialization improves naive spiking fusion and preserves fallback capability under sensor outages. Second, we introduce Sensor Availability Learning (SAL), a curriculum that progressively teaches the model to handle diverse sensor states instead of a fixed configuration. Interestingly, equipping conventional MMOMs with SAL further unlocks their resilience, delivering a second leap in performance. Our contributions are fourfold.

- We uncover modality outage fragility in MMOMs, revealing a previously overlooked gap between full-sensor accuracy and resilience under sensor failures.
- We introduce SpikeOcc, the first spike-driven MMOM for 3D occupancy prediction. Our study establishes spiking computation as a viable representation paradigm and identifies the key difficulty in transferring conventional fusion into the spiking regime.
- We develop SMoE and SAL from architectural and learning perspectives. SMoE specializes heterogeneous representations through spiking experts, while SAL explicitly learns across diverse sensor-availability states.
- Extensive experiments show that SpikeOcc performs on par with counterparts under complete sensing, while effectively raising the performance floor under modality outages.

2 Related Work

3D Semantic Occupancy Prediction. Early efforts, such as LMSCNet (Roldão et al., 2020) and JS3C-Net (Yan et al., 2021), were built primarily on LiDAR, whose sparse observations and sensing cost motivated a shift toward vision-centric occupancy perception. MonoScene (Cao & de Charette,

2022) and SurroundOcc (Wei et al., 2023) established representative monocular and surround-view formulations, followed by stronger architectures such as TPVFormer (Huang et al., 2023), OccFormer (Zhang et al., 2023), and VoxFormer (Li et al., 2023a), richer rendering strategies such as UniOcc (Pan et al., 2023) and RenderOcc (Pan et al., 2024b), and more compact representations such as COTR (Ma et al., 2024), STCOcc (Liao et al., 2025), and Dr.Occ (Zhu et al., 2026). More recently, resilience to sensor failures has emerged as an important concern in deployable perception (Orr et al., 2021; Almalioglu et al., 2022). UniBEV (Wang et al., 2024b), MetaBEV (Ge et al., 2023), and M-BEV (Chen et al., 2024) target missing or degraded observations in BEV-oriented perception, while FlexMap (Wang et al., 2026), Uni-Map (Hao et al., 2026), and SafeMap (Hao et al., 2025) extend related ideas to map construction. M2-Occ (Lin et al., 2026) specifically investigates 3D semantic occupancy prediction under incomplete multi-camera observations. Despite this progress, resilience across heterogeneous sensor availability remains underexplored in dense semantic occupancy. In this work, we take a pioneering step in utilizing SNNs for resilient occupancy learning, aiming to bridge this gap and advance multimodal occupancy inference.

Spiking Neural Networks. As the third generation of neural networks, spiking neural networks (SNNs) have shown remarkable progress across a wide range of tasks, achieving performance comparable to that of ANNs (Hu et al., 2023). Their initial applications span image classification (Yao et al., 2023; 2025), object detection (Luo et al., 2024; Li et al., 2025b; Qu et al., 2025), semantic segmentation (Lei et al., 2025), video classification (Zou et al., 2025), reinforcement learning (Xu et al., 2025; 2026), temporal prediction (Dong et al., 2026), object tracking (Zhang et al., 2026; Shan et al., 2026; Sun et al., 2026b), and event-driven recognition (Liu et al., 2021; 2025). In parallel, SNNs have officially entered the era of 3D vision (Sun et al., 2025; 2026a;c;d), with representative works including SpikePEFT (Guo et al., 2026b), Spiking Discrepancy Transformer (Lu et al., 2026), SPM (Wu et al., 2025), and 3DSMT (Zhou et al., 2026). However, so far, there has been no exciting baseline in the 3D occupancy domain. This study is dedicated to investigating the potential contributions of SNNs to this task, further advancing neuromorphic visual perception.

3 METHOD

3.1 PROBLEM FORMULATION

We consider multimodal 3D semantic occupancy prediction under incomplete sensing. Given N_c synchronized camera images $\mathcal{I} = \{I^n\}_{n=1}^{N_c}$ and an aggregated LiDAR point cloud $\mathcal{P}$, the goal is to predict the occupancy state and semantic label of each voxel in a predefined 3D space Ω. Let $\mathcal{Y} = \{0, \dots, K-1\}$ denote the label set, where 0 represents free space. The prediction is $\hat{\mathbf{Y}} \in \mathcal{Y}^{|\Omega|}$. To describe sensor availability, we introduce a binary view mask $\mathbf{m} \in \{0,1\}^{N_c}$ and a LiDAR availability indicator $m_l \in \{0,1\}$. The indicator and ratio are defined as

$$m_c = \mathbb{I}\left[\sum_{n=1}^{N_c} m_n > 0\right], \qquad r_c = \frac{1}{N_c}\sum_{n=1}^{N_c} m_n. \tag{1}$$

The resulting availability descriptor is $\mathbf{a} = (\mathbf{m}, m_c, m_l, r_c)$. A sensor configuration $s \in \mathcal{S}$ determines a masking transformation T_s, yielding the available observations $\mathcal{X}_s = T_s(\mathcal{I}, \mathcal{P})$. We require $m_c + m_l \geq 1$, so at least one modality remains available.

The objective is to learn a single predictor

$$f_\Theta : (\mathcal{X}_s, \mathbf{a}_s) \longmapsto \hat{\mathbf{Y}}_s, \qquad s \in \mathcal{S}, \tag{2}$$

that maintains accurate occupancy prediction across different sensing conditions without retraining a separate model for each configuration. Unlike conventional training (Li et al., 2025a), our learning objective must account for the risks associated with different sensing conditions.

$$R_s(\Theta) = \mathbb{E}_{(\mathcal{X},\mathbf{Y})\sim\mathcal{D}}\left[\mathcal{L}_{\text{occ}}\big(f_\Theta(T_s(\mathcal{X}), \mathbf{a}_s), \mathbf{Y}\big)\right]. \tag{3}$$

Sensor Availability Learning (Sec. 3.3) will later control how these risks are sampled during training.

3.2 SPIKING MIXTURE-OF-EXPERTS (SMOE)

Width-first Decomposition. Fig. 3 illustrates the proposed SpikeOcc architecture. The camera and LiDAR encoders first produce spatially aligned voxel features $\mathbf{C}_0$ and $\mathbf{L}_0$. After availability masking

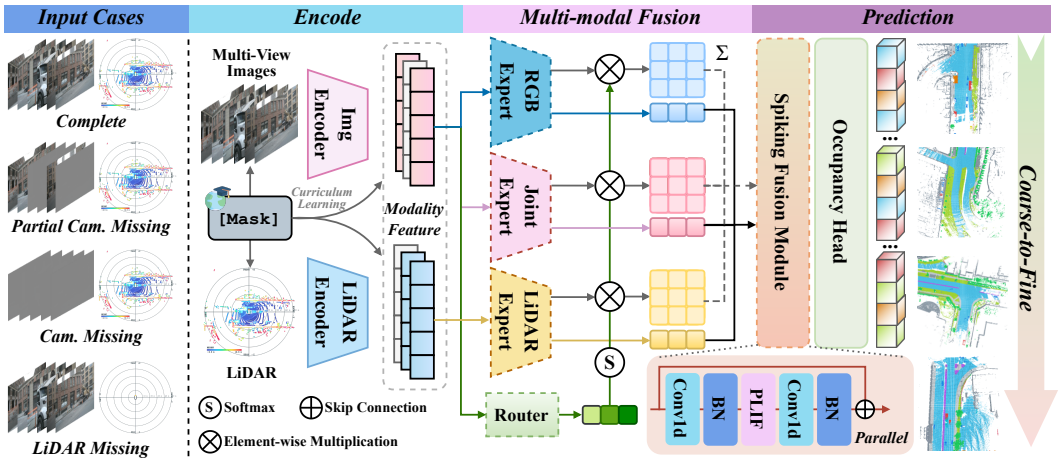


Figure 3: **Overview of SpikeOcc.** Masked camera and LiDAR inputs are encoded and processed by specialized experts. Router-weighted features pass through the spiking fusion module and occupancy head, while SAL progressively introduces missing-sensor cases during training.

and projection, we obtain

$$\begin{aligned} \widetilde{\mathbf{C}}_0 &= m_c \mathbf{C}_0, \quad \mathbf{C} = \phi_c(\widetilde{\mathbf{C}}_0), \\ \widetilde{\mathbf{L}}_0 &= m_l \mathbf{L}_0, \quad \mathbf{L} = \phi_l(\widetilde{\mathbf{L}}_0), \end{aligned} \tag{4}$$

where ϕ_c and ϕ_l are independent $1 \times 1 \times 1$ convolutional projections. Both projected features have width d, and the fusion interface has width $D = 2d$.

Rather than simply deepening a shared fusion network (Zhou et al., 2024), SMoE develops three parallel experts: an RGB expert E_c, a LiDAR expert E_l, and a joint expert E_j. Each produces a D-channel representation. This design addresses two potential failure modes. Half-width unimodal outputs would reduce the input capacity of the occupancy predictor when one modality disappears (Zagoruyko & Komodakis, 2016), whereas an unconstrained joint transformation could overwrite the original modality features (Wang et al., 2023a). SMoE therefore equips unimodal experts with full-width outputs and retains the original feature concatenation as continuous foundation.

Spiking Residual Generation. Each expert contains an independent spiking bottleneck with the same operator structure. We use Parametric Leaky Integrate-and-Fire (PLIF) neurons (Fang et al., 2021). For an input current $\mathbf{x}^t$, the membrane dynamics are

$$\begin{aligned} \widetilde{\mathbf{u}}^t &= (1-\beta)\mathbf{u}^{t-1} + \beta \mathbf{x}^t, \quad \beta = \sigma(\theta_\beta), \\ \mathbf{s}^t &= \mathbb{H}(\widetilde{\mathbf{u}}^t - V_{\text{th}}), \qquad \mathbf{u}^t = \widetilde{\mathbf{u}}^t - V_{\text{th}} \mathbf{s}^t, \end{aligned} \tag{5}$$

where θ_β is learnable, $\mathbb{H}$ is the hard-threshold function, and the last line implements soft reset (Rueckauer et al., 2017). During backpropagation, the non-differentiable threshold is approximated by the sigmoid surrogate (Wu et al., 2018; Neftci et al., 2019)

$$\frac{\partial \mathbf{s}^t}{\partial \widetilde{\mathbf{u}}^t} \approx \alpha \sigma(\alpha z)\big(1 - \sigma(\alpha z)\big), \qquad z = \widetilde{\mathbf{u}}^t - V_{\text{th}}. \tag{6}$$

Here, we set $\alpha = 4$ throughout all experiments. The spiking bottleneck compresses the input to $D/4$ channels, extracts local 3D context, and expands the result back to D:

$$\mathbf{s}_1^t = \text{PLIF}_1(\mathcal{P}_1(\mathbf{X})), \quad \mathbf{s}_2^t = \text{PLIF}_2(\mathcal{P}_2(\mathbf{s}_1^t)), \quad \mathbf{R}^t = \mathcal{P}_3(\mathbf{s}_2^t), \quad \overline{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{R}^t. \tag{7}$$

Here, $\mathcal{P}_1$ and $\mathcal{P}_3$ are convolutional projections, and $\mathcal{P}_2$ is a depthwise projection, all followed by batch normalization. The PLIF layers generate binary spikes, and the final projection produces a continuous readout. We average T internal steps and reset membrane states between samples.

Spiking Experts. For the two unimodal experts, continuous projections preserve the original information while spiking branches learn residual corrections:

$$\mathbf{F}_c = \psi_c(\mathbf{C}) + \gamma_c \overline{\mathbf{R}}_c, \qquad \mathbf{F}_l = \psi_l(\mathbf{L}) + \gamma_l \overline{\mathbf{R}}_l. \tag{8}$$

Here ψ_c, ψ_l project the modality features to width D, and γ_c, γ_l are learnable scalars initialized to 0.01. Thus, both experts initially remain close to stable continuous projections while learning spiking corrections. The joint expert instead preserves the original masked concatenation $\mathbf{B} = [\widetilde{\mathbf{C}}_0; \widetilde{\mathbf{L}}_0]$ and learns a bounded residual:

$$\mathbf{F}_j = \mathbf{B} + \rho \overline{\mathbf{R}}_j, \qquad \rho = \rho_{\max} \sigma(\theta_\rho). \tag{9}$$

The joint spiking branch receives $[\mathbf{C}; \mathbf{L}]$. We set $\rho_{\max} = 0.2$ and initialize $\rho = 0.01$. The bounded coefficient limits the residual gain, reducing the risk of large perturbations during early optimization.

Routing. The router combines global feature statistics with explicit availability information. Let $\mathbf{g}_c = \mathrm{GAP}(\mathbf{C})$ and $\mathbf{g}_l = \mathrm{GAP}(\mathbf{L})$. $\mathrm{GAP}(\cdot)$ denotes global average pooling over the spatial dimensions. A two-layer MLP predicts three logits from

$$\mathbf{h} = [\mathbf{g}_c; \mathbf{g}_l; m_c; m_l; r_c], \qquad \mathbf{z} = \mathrm{MLP}(\mathbf{h}) \in \mathbb{R}^3. \tag{10}$$

The availability vector $\mathbf{q} = (m_c, m_l, m_c m_l)$ excludes experts that require an unavailable modality. The remaining logits are normalized as

$$w_k = \frac{q_k \exp(z_k)}{\sum_{i \in \{c,l,j\}} q_i \exp(z_i)}, \qquad \mathbf{F} = \sum_{k \in \{c,l,j\}} w_k \mathbf{F}_k. \tag{11}$$

3.3 Sensor Availability Learning

Mask Propagation. SAL trains the same model across a structured set of sensing states. A sampled state is propagated consistently through the network. For the camera branch, the view mask is applied to the lifted frustum features before voxel pooling:

$$\widetilde{\mathbf{V}}_{n,u,k} = m_n \, p(d_k \mid u, n) \mathbf{g}_{u,n},$$
$$\mathbf{C}_0(v) = \sum_{\pi(n,u,d_k)=v} \widetilde{\mathbf{V}}_{n,u,k}. \tag{12}$$

Here $p(d_k \mid u, n)$ is the predicted depth distribution, $\mathbf{g}_{u,n}$ is the image context feature, and π maps a frustum sample to its 3D voxel. The same view mask excludes unavailable cameras from depth supervision, ensuring feature construction describes the same sensing state (Philion & Fidler, 2020).

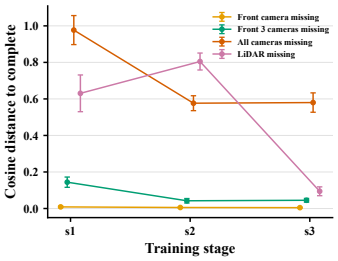


Figure 4: Representation adaptation across training stages.

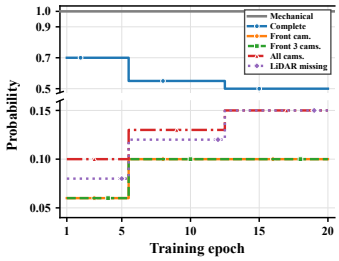


Figure 5: Curriculum learning schedule in SAL.

Curriculum Learning. Fig. 4 reveals that direct training with sensor dropout can lead to fluctuating representation distances between complete and incomplete inputs. To stabilize adaptation, we introduce a three-stage curriculum that gradually increases the exposure to sensor failures. For each sample, the sensor state is drawn from a distribution that varies with the training epoch:

$$s_b \sim \mathrm{Categorical}(\mathbf{p}_e), \qquad \mathbf{p}_e = \mathbf{p}^{(k)}, \quad e \in \mathcal{E}_k, \quad k \in \{1, 2, 3\}. \tag{13}$$

Here, e denotes the current epoch, $\mathcal{E}_k$ the epoch range of stage k, and $\mathbf{p}^{(k)}$ the corresponding sampling distribution. Fig. 5 illustrates the resulting curriculum. The curriculum progresses from complete sensing to failures. Alternative orderings are analyzed in Appendix C.

Learning Objective. To encourage the intended expert roles, we introduce a lightweight routing objective. Let y^{route} denote the target defined in *Appendix D*. We supervise the raw router logits:

$$\mathcal{L}_{\mathrm{route}} = \lambda_r \, \mathrm{CE}(\mathbf{z}, y^{\mathrm{route}}), \qquad \lambda_r = 0.05. \tag{14}$$

This auxiliary loss therefore teaches the router the responsibilities without replacing the soft mixture among valid experts. The occupancy objective remains unchanged (Li et al., 2025a). At each coarse and fine scale, it combines class-weighted cross-entropy, semantic scaling, geometric scaling (Cao & de Charette, 2022), and Lovasz-Softmax losses (Berman et al., 2018):

$$\mathcal{L}_{\mathrm{occ}} = \lambda_1 \mathcal{L}_{\mathrm{CE}} + \lambda_2 \mathcal{L}_{\mathrm{sem}} + \lambda_3 \mathcal{L}_{\mathrm{geo}} + \lambda_4 \mathcal{L}_{\mathrm{Lovasz}}. \tag{15}$$

The inherited depth loss (Li et al., 2023b) is evaluated only on valid camera views. The total training loss is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{occ}} + \mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{route}}. \tag{16}$$

Here, λ_1–λ_4 balance the contributions of the occupancy loss terms and are all empirically set to 1.

Table 1: Quantitative comparisons on the OpenOccupancy validation set. C, D, and L denote camera, depth, and LiDAR, respectively. Within each sensor setting, the **best**, second-best, and remaining results are shown in bold, underlined, and gray, respectively. For fair comparison, results marked with † are reproduced by us and averaged over 3 seeds. Same for subsequent tables.

Method	Input Modality	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive surf.	other_flat	sidewalk	terrain	manmade	vegetation
All Sensors Available																			
MonoScene (Cao & de Charette, 2022)	C	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
JS3C-Net (Yan et al., 2021)	L	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
3DSketch (Chen et al., 2020)	C&D	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
M-CONet (Wang et al., 2023b)	C&L	29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ (Pan et al., 2024a)	C&L	30.6	21.9	26.5	16.8	22.3	27.0	10.1	20.9	20.7	14.5	16.4	21.6	36.9	23.5	5.5	23.7	20.5	23.5
OccMamba (Li et al., 2025a)	C&L	35.7	**26.2**	30.2	20.5	**26.5**	29.5	**18.8**	**26.0**	23.7	19.9	**20.6**	**25.4**	38.4	26.5	27.0	26.6	**28.9**	30.5
RIOcc (Fan et al., 2025)	C&L	35.4	25.9	30.2	19.8	25.8	28.7	18.3	24.8	**31.8**	**21.8**	20.5	24.9	37.2	24.5	25.5	24.9	27.0	28.8
SPSC (Guo et al., 2026a)	C&L	35.2	26.1	**31.4**	**20.6**	25.1	**30.6**	15.7	25.9	27.1	18.8	19.7	25.2	38.4	26.9	26.8	26.5	28.5	30.4
SpikeOcc (ours)	C&L	**35.8**	25.9	30.8	19.4	25.4	30.2	13.7	25.3	26.5	17.7	19.5	25.0	**39.9**	**27.1**	**27.5**	**27.0**	28.6	**30.8**
Front Camera Missing																			
OccMamba† (Li et al., 2025a)	C&L	30.9	21.0	26.5	11.5	20.6	26.0	12.1	18.8	23.7	12.5	15.0	19.5	35.2	20.6	23.7	22.9	24.1	23.3
SPSC† (Guo et al., 2026a)	C&L	22.3	15.3	15.5	9.0	16.1	22.2	8.0	10.7	20.8	12.0	11.7	17.9	16.4	10.1	12.0	13.3	24.1	25.2
SpikeOcc (ours)	C&L	**36.2**	**25.3**	**30.1**	**17.7**	**24.2**	**29.5**	**15.1**	**21.5**	**25.9**	**17.6**	**19.6**	**24.5**	**39.7**	**25.9**	**27.0**	**26.5**	**28.5**	**30.9**
Front-3 Cameras Missing																			
OccMamba† (Li et al., 2025a)	C&L	30.4	19.3	25.1	8.8	19.1	23.8	10.0	14.6	20.9	10.6	13.7	16.8	34.8	19.9	23.0	21.9	23.7	22.6
SPSC† (Guo et al., 2026a)	C&L	21.9	13.7	13.6	6.2	13.4	18.9	6.0	8.2	18.7	9.9	10.0	15.6	16.2	9.6	11.6	12.3	23.8	24.9
SpikeOcc (ours)	C&L	**36.3**	**24.2**	**29.0**	**14.3**	**23.1**	**28.9**	**14.3**	**18.6**	**24.4**	**15.9**	**18.5**	**23.3**	**39.7**	**25.3**	**26.6**	**25.9**	**28.4**	**30.8**
All Cameras Missing																			
OccMamba† (Li et al., 2025a)	C&L	18.8	9.4	3.9	1.3	10.0	9.4	6.4	3.3	16.8	3.3	6.3	5.6	17.7	10.0	10.5	8.6	19.6	18.2
SPSC† (Guo et al., 2026a)	C&L	21.8	8.7	6.9	1.2	4.2	5.5	0.4	2.8	13.8	0.9	4.6	8.7	15.6	7.0	10.4	9.4	23.3	24.2
SpikeOcc (ours)	C&L	**36.8**	**22.7**	**27.3**	**10.1**	**21.2**	**27.6**	**12.1**	**15.4**	**22.0**	**13.2**	**17.4**	**21.1**	**40.3**	**24.0**	**26.1**	**25.4**	**28.5**	**31.1**
LiDAR Missing																			
OccMamba† (Li et al., 2025a)	C&L	7.8	1.6	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.5	3.2	4.7	3.6	0.0	0.0
SPSC† (Guo et al., 2026a)	C&L	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.4	0.4	0.0	0.0	0.0
SpikeOcc (ours)	C&L	**19.7**	**9.2**	**7.9**	**4.5**	**8.4**	**12.7**	**2.2**	**6.3**	**7.7**	**5.1**	**0.8**	**7.9**	**30.6**	**15.8**	**16.1**	**14.6**	**1.9**	**5.4**

4 EXPERIMENTS

4.1 DATASET AND METRICS

Datasets. We evaluate SpikeOcc on OpenOccupancy (Wang et al., 2023b) and SemanticKITTI (Behley et al., 2019). OpenOccupancy extends nuScenes with dense semantic occupancy annotations, containing 28,130 training frames and 6,019 validation frames with 17 semantic categories. Its evaluation space covers $[-51.2, 51.2]$ m along the X and Y axes and $[-3, 5]$ m along the Z axis, with a voxel resolution of 0.2 m. SemanticKITTI provides large-scale LiDAR sequences for semantic scene completion with 19 semantic categories. Following the official protocol, sequences 00–07 and 09–10 are used for training, sequence 08 for validation, and sequences 11–21 for testing. The scene is voxelized into a $256 \times 256 \times 32$ grid at a resolution of 0.2 m.

Metrics. We report Intersection over Union (IoU) for geometric occupancy and mean Intersection over Union (mIoU) for semantic occupancy. IoU measures occupied-versus-empty prediction quality, while mIoU averages the class-wise IoU over all semantic categories (Song et al., 2017). Additional metric definitions are provided in *Appendix E*.

4.2 IMPLEMENTATION DETAILS

Architecture and Training. Our implementation is built upon MMDetection3D (MMDetection3D Contributors, 2020). For the camera branch, we adopt an ImageNet-pretrained ResNet-50 (He et al., 2016) as the image backbone and resize the input images to 256×704. For the LiDAR branch, we aggregate 10 LiDAR sweeps and encode the resulting point clouds with a voxel-based encoder (Yan et al., 2018). The feature dimension is set to 384 throughout our experiments. Our spiking modules use both LIF (Gerstner & Kistler, 2002) and PLIF (Fang et al., 2021) neurons. PLIF is used in default. We initialize $\tau = 2$, equivalently $\beta = 0.5$, and use $V_{\text{th}} = 0.5$. We train all models using AdamW with a weight decay of 0.01 and an initial learning rate of 3×10^{-4}. Training lasts for 20 epochs with a 1,500-iteration warm-up, and gradient clipping is applied to stabilize optimization. All experiments are conducted on eight NVIDIA A800 GPUs with a global batch size of 8. The training run takes approximately 46 hours. The inference protocol is detailed in *Appendix F*.

Table 2: Performance on the SemanticKITTI test set. Within each sensor setting, the **best**, second-best, and remaining results are shown in bold, underlined, and gray, respectively.

Method	Input Modality	IoU	mIoU	road	sidewalk	parking	other ground	building	car	truck	bicycle	motorcycle	other vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic sign
All Sensors Available																						
SurroundOcc (Wei et al., 2023)	C	34.7	11.9	56.9	28.3	30.2	6.8	15.2	20.6	1.4	1.6	1.2	4.4	14.9	3.4	19.3	1.4	2.0	0.1	11.3	3.9	2.4
OccFormer (Zhang et al., 2023)	C	34.5	12.3	55.9	30.3	**31.5**	6.5	15.7	21.6	1.2	1.5	1.7	3.2	16.8	3.9	21.3	2.2	1.1	0.2	11.9	3.8	3.7
RenderOcc (Pan et al., 2024b)	C	N/A	12.8	57.2	28.4	16.1	0.9	18.2	24.9	6.0	0.4	0.3	3.7	26.2	4.9	3.6	1.9	**3.1**	0.0	9.1	6.2	3.4
M-CONet (Wang et al., 2023b)	C&L	N/A	20.4	60.6	**36.1**	29.0	**13.0**	38.4	33.8	4.7	3.0	2.2	5.9	41.5	**20.5**	35.1	0.8	2.3	**0.6**	**26.0**	18.7	15.7
OccGen† (Wang et al., 2024a)	C&L	59.9	21.1	63.1	33.0	17.3	0.3	38.1	44.1	17.4	3.8	2.9	7.9	41.4	**20.5**	46.6	4.1	2.1	0.0	13.8	28.6	**16.2**
OccMamba† (Li et al., 2025a)	C&L	59.6	**21.6**	**65.1**	35.1	16.5	0.9	38.3	44.6	21.3	**4.4**	**3.5**	8.8	41.4	19.0	45.6	**4.5**	2.0	0.0	14.0	**29.3**	16.0
SPSC† (Guo et al., 2026a)	C&L	59.4	21.3	63.4	34.5	14.5	0.2	38.0	43.7	22.2	4.2	1.6	**13.5**	40.5	19.3	46.9	3.4	1.7	0.0	13.1	28.3	15.7
SpikeOcc (ours)	C&L	**60.3**	21.5	63.1	34.4	14.7	1.1	**39.0**	**45.6**	**24.8**	2.0	2.2	9.2	**42.0**	18.8	**47.1**	4.1	2.6	0.0	13.4	**29.3**	14.5
All Cameras Missing																						
OccGen† (Wang et al., 2024a)	C&L	59.1	19.8	55.0	30.1	12.1	0.8	37.0	41.4	18.0	**5.0**	**4.2**	8.3	41.6	18.4	43.0	**4.5**	2.0	0.0	13.1	26.1	15.5
OccMamba† (Li et al., 2025a)	C&L	59.7	20.9	62.6	32.6	**16.4**	0.3	37.9	43.7	16.4	3.7	2.8	7.5	41.3	**20.2**	46.3	4.1	2.1	0.0	**13.7**	28.5	**16.0**
SPSC† (Guo et al., 2026a)	C&L	58.9	20.7	**63.9**	34.1	15.4	0.2	37.0	43.4	17.8	3.8	1.5	**11.3**	39.1	18.6	**47.3**	3.3	1.8	0.0	12.1	28.1	15.5
SpikeOcc (ours)	C&L	**60.4**	**21.4**	62.9	**34.4**	14.7	**1.1**	**39.0**	**45.6**	**24.0**	2.1	2.3	9.3	**42.2**	18.7	47.2	4.1	**2.5**	0.0	13.5	**29.3**	14.5
LiDAR Missing																						
OccGen† (Wang et al., 2024a)	C&L	4.7	0.9	8.4	6.8	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.6	0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.0
OccMamba† (Li et al., 2025a)	C&L	9.3	1.7	15.4	5.6	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.3	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0
SPSC† (Guo et al., 2026a)	C&L	11.1	1.5	15.0	5.2	1.0	0.0	0.2	2.1	0.0	**0.1**	0.0	0.1	1.4	0.0	3.1	0.0	0.0	0.0	0.2	0.0	0.0
SpikeOcc (ours)	C&L	**21.8**	**4.1**	**29.1**	**13.8**	**5.0**	**0.3**	**0.7**	**7.7**	0.0	0.0	0.0	**0.4**	**4.3**	**0.1**	**14.5**	**0.1**	0.0	0.0	**0.9**	**0.6**	**0.6**

Sensor Configurations. To evaluate resilience under varying sensor availability, we consider five sensor configurations on OpenOccupancy and three configurations on SemanticKITTI. Unless otherwise specified, all methods are evaluated under identical sensor configurations, while training follows their respective default protocols.

4.3 Comparison with State-of-the-Art Methods

OpenOccupancy. Table 1 reports the results on OpenOccupancy with v0.0 annotations under five sensor configurations. As no standard protocol exists for sensor outages in multimodal occupancy, we reproduce the competing methods under identical sensor configurations. With all sensors available, SpikeOcc remains competitive with strong ANN-based baselines (Li et al., 2025a; Guo et al., 2026a), indicating that resilience is achieved without sacrificing standard occupancy performance. As sensor availability decreases, its advantage becomes progressively more evident, particularly under complete camera or LiDAR outages. The geometry trade-off is analyzed in *Appendix G*.

SemanticKITTI. Table 2 shows a consistent trend on SemanticKITTI. In this setting, only two camera views are available, making the model more dependent on LiDAR than on OpenOccupancy. SpikeOcc remains competitive under complete sensing and largely preserves its performance when all cameras are unavailable. More notably, its advantage becomes substantially larger under LiDAR outage, where existing multimodal models degrade severely (Wang et al., 2024a).

4.4 Ablation Study

Generalization to Unseen Sensor Configurations. Instead of memorizing a finite set of predefined failure patterns, a resilient occupancy model should generalize to unseen sensor configurations. We therefore evaluate all models on held-out sensor combinations, as shown in Fig. 6. SpikeOcc maintains consistently strong performance across these unseen configurations, while SAL also improves the generalization of the ANN-based baseline (Li et al., 2025a). Separately, we compare SAL with conventional Sensor Dropout (Liu et al., 2017), which randomly removes sensor inputs during training to reduce dependence on any specific sensor subset. Beyond random sensor dropout, SAL turns sensor failures into a structured curriculum, yielding stronger generalization to unseen configurations.

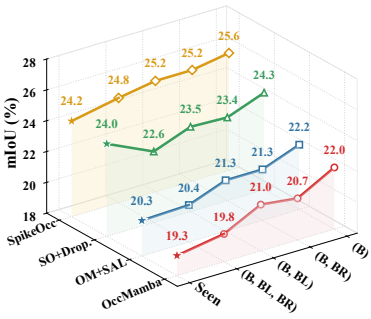


Figure 6: Generalization to unseen sensor configurations. B: Back, BL: Back left, BR: Back right.

Table 3: **Ablation on SAL and MoE under different sensor settings.** G and P denote GELU and PLIF respectively. M. means Missing.

Method	SAL	MoE	All Sensors		All Cam. M.		LiDAR M.	
			IoU	mIoU	IoU	mIoU	IoU	mIoU
OccMamba	○	–	35.7	26.2	18.8	9.4	7.8	1.6
	●	–	36.2	26.4	29.3	16.1	19.2	8.5
	○	G	30.1	21.2	19.7	9.2	10.7	2.9
	●	G	33.3	23.4	29.9	13.7	16.3	6.8
Naive SNN	○	–	35.9	16.7	21.8	8.7	0.3	0.0
	●	–	31.2	22.4	29.3	16.1	18.8	8.9
	○	P	35.2	24.7	31.5	15.6	17.4	8.1
	●	P	**35.8**	**25.9**	**36.8**	**22.7**	**19.7**	**9.2**

Table 4: **Parameter-matched comparison.** Capacity scaling versus expert specialization. All models are trained *without SAL*.

Method	Params	Full	All Cam. M.	LiDAR M.
OccMamba-128	64.6M	25.2	8.6	1.1
OccMamba-384	109.2M	26.2	9.4	1.6
OccMamba-512-ext	154.8M	26.4	9.5	2.9
Naive SNN	109.1M	16.7	8.7	0.0
Dense SNN-S	111.1M	16.1	8.1	0.1
Dense SNN-M	141.8M	18.0	8.5	1.6
Dense SNN-L	158.1M	18.5	8.9	3.7
SMoE	112.3M	**24.7**	**15.6**	**8.1**

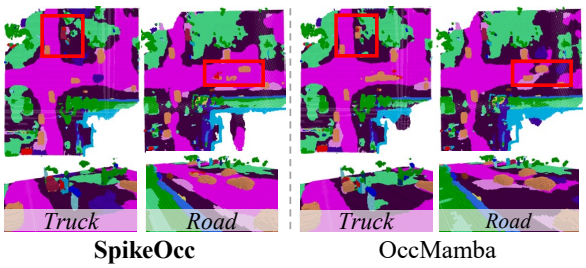


Figure 7: **Qualitative comparison of occupancy predictions.** SpikeOcc better captures small objects and preserves the geometry of major scene structures.

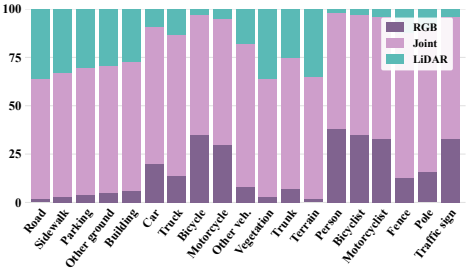


Figure 8: Class-conditioned aggregation of scene-level routing weights on SemanticKITTI under complete sensing.

Ablation on SAL and MoE. The variants differ only in the neuron operator, with identical architectures and training settings. Table 3 reveals two complementary effects. SAL consistently improves resilience across both ANN and SNN backbones, indicating that its benefit is not tied to a specific architecture. In contrast, while the ANN-based MoE brings limited benefit, the PLIF-based (Fang et al., 2021) MoE substantially closes the semantic gap of the naive SNN across all sensor conditions. Combining SAL with spiking MoE yields the strongest overall resilience, while largely preserving the full-sensor performance. More SAL-equipped baselines are reported in *Appendix H*.

Capacity Scaling. To isolate the effect of expert specialization, we construct parameter-matched dense SNNs by increasing network depth while keeping the feature dimension and operators unchanged (Eigen et al., 2014; Csordás et al., 2024). Table 4 shows simply scaling the dense architecture brings only modest improvements in mIoU, and even the Dense SNN-L remains clearly behind SMoE across all sensor settings. The gap suggests that expert specialization, rather than mere capacity scaling, is key to handling heterogeneous sensing dynamics.

Expert Specialization. After training a spiking MoE, it is natural to study the usefulness of the learned router, in the light of several pitfalls (Riquelme et al., 2021). To this end, we analyze expert routing on SemanticKITTI and visualize the resulting occupancy predictions in Fig. 7 and different aggregated routing profiles in Fig. 8. Expert weights remain shared across all voxels within each scene. The joint expert dominates most categories while RGB and LiDAR weights show distinct class-conditioned patterns. These patterns suggest that the router adjusts the relative contributions of modality-specific and joint experts according to the semantic composition of each scene.

Neuron Dynamics and Time Steps. We investigate neuron type and simulation length to understand how intrinsic spiking dynamics affect occupancy learning. As shown in Table 5, PLIF outperforms conventional IF and LIF neurons (Burkitt, 2006). A plausible explanation is that adaptive membrane dynamics better capture heterogeneous multimodal features. Table 6 further shows that additional time steps initially improve representation quality, but the gain saturates beyond $T = 4$ while the firing rate continues to increase.

Computational Efficiency. To characterize the practical cost of deployment, computational efficiency is evaluated in terms of inference latency, throughput, and theoretical energy consumption.

Table 5: Neuron type.

Neuron	τ	Learn. τ	mIoU
IF	–	–	25.2
LIF	2.0	×	25.3
PLIF	**2.0**	✓	**25.9**

Table 6: Time steps.

T	V_{th}	Firing Rate (%)	mIoU
1	0.5	7.6	24.0
2	0.5	11.1	24.5
4	**0.5**	**12.7**	**25.9**
6	0.5	13.4	25.7

Table 7: Computational efficiency.

Variant	Lat. (ms)	FPS	MACs (G)	ACs (G)	Energy (mJ)
M-CONet	1918.8	4.17	14.7	0	67.6
SpikeOcc-LIF	1862.0	4.30	1.68	1.63	9.2
SpikeOcc-PLIF	1797.3	4.45	1.68	1.86	9.4

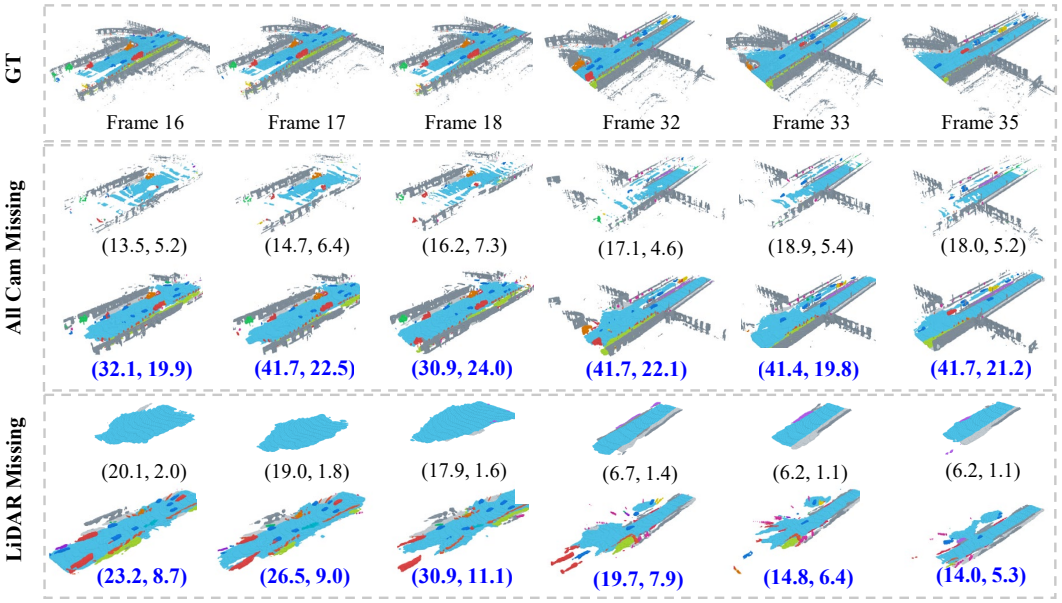


Figure 9: **More qualitative comparison with OccMamba.** SpikeOcc preserves more complete scene geometry and semantics under sensor outages. Numbers denote (IoU, mIoU).

All measurements are conducted on 8 NVIDIA A800 GPUs with batch size one under identical input settings after warm-up. For energy estimation, we follow the widely adopted 45-nm CMOS energy model (Horowitz, 2014), where a 32-bit MAC and AC operation consume 4.6 pJ and 0.9 pJ, respectively. The theoretical energy is calculated as $E_{\text{theory}} = 4.6N_{\text{MAC}} + 0.9N_{\text{AC}}$, where N_{MAC} and N_{AC} are the operation counts in billions and E_{theory} is measured in mJ. Since the pretrained camera and LiDAR encoders are shared across all variants, MACs, ACs, and theoretical energy are reported for the fusion module only. Compared with M-CONet, SpikeOcc substantially reduces fusion computation and energy while maintaining comparable latency and throughput in Table 7.

Qualitative Analysis. Figures 7 and 9 provide qualitative evidence across SemanticKITTI and OpenOccupancy. SpikeOcc better retains fine-grained objects, while its advantage becomes more pronounced under complete sensor outages, where competing predictions are often fragmented.

5 Conclusion

We present SpikeOcc, a spike-driven framework for resilient multimodal 3D occupancy prediction, targeting sensor failures that are inevitable in real-world perception systems. Rather than relying on depth scaling, a width-first spiking Mixture-of-Experts restructures multimodal fusion into specialized pathways, enabling adaptive fusion across changing sensing regimes. Complementing this, Sensor Availability Learning organizes diverse sensor states into a staged curriculum, gradually strengthening the model against increasingly incomplete observations. Extensive experiments on OpenOccupancy and SemanticKITTI demonstrate strong performance under both complete sensing and challenging modality outages. By supporting flexible inference across varying sensor availability, SpikeOcc advances occupancy prediction toward resilient deployment and offers a fresh perspective for future multimodal perception research. Limitations are discussed in *Appendix I*.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available datasets and does not involve human-subject experiments or the collection of private personal data. We are not aware of any conflicts of interest or additional ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

We provide the model formulation, training details, evaluation protocol, and additional implementation details in the main paper and appendix. All experiments follow the official dataset splits and are conducted with three random seeds. Anonymous source code, configuration files, and training scripts are provided through the OpenReview Supplementary Material.

REFERENCES

Yasin Almalioglu, Mehmet Turan, Niki Trigoni, and Andrew Markham. Deep learning-based robust positioning for all-weather autonomous driving. *Nature Machine Intelligence*, 2022.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019.

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 *CVPR*, 2018.

Anthony N. Burkitt. A review of the integrate-and-fire neuron model: I. homogeneous synaptic input. *Biological Cybernetics*, 2006.

Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, 2022.

Siran Chen, Yue Ma, Yu Qiao, and Yali Wang. M-bev: Masked bev perception for robust autonomous driving. In *AAAI*, 2024.

Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, 2020.

Róbert Csordás, Kazuki Irie, Jürgen Schmidhuber, Christopher Potts, and Christopher D. Manning. Moeut: Mixture-of-experts universal transformers. In *NeurIPS*, 2024.

Jianhao Ding, Zhiyu Pan, Yujia Liu, Zhaofei Yu, and Tiejun Huang. Robust stable spiking neural networks. In *ICML*, 2024.

Yiting Dong, Jianhao Ding, Zijie Xu, Tong Bu, Zhaofei Yu, and Tiejun Huang. Prednext: Explicit cross-view temporal prediction for unsupervised learning in spiking neural networks. In *ICLR*, 2026.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. In *ICLR Workshop*, 2014.

Baojie Fan, Xiaotian Li, Yuhan Zhou, Yuyu Jiang, Jiandong Tian, and Huijie Fan. Riocc: Efficient cross-modal fusion transformer with collaborative feature refinement for 3d semantic occupancy prediction. In *ICCV*, 2025.

Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *ICCV*, 2021.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *ICCV*, 2023.

Wulfram Gerstner and Werner M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.

Qingju Guo, Shuang Li, Binhui Xie, Jing Geng, and Wei Li. Spsc: Sparse and scalable multi-modal 3d occupancy prediction for autonomous driving. In *AAAI*, 2026a.

Zihao Guo, Jihua Zhu, Yiding Sun, Lin Chen, and Danwei Wang. Parameter-efficient fine-tuning for spiking point cloud models. *arXiv preprint arXiv:2607.29048*, 2026b.

Xiaoshuai Hao, Lingdong Kong, Rong Yin, Pengwei Wang, Jing Zhang, Yunfeng Diao, and Shu Zhao. Safemap: Robust hd map construction from incomplete observations. In *ICML*, 2025.

Xiaoshuai Hao, Yifan Yang, Mengchuan Wei, Hui Zhang, Haimei Zhao, Guang Chen, Hangjun Ye, Long Chen, and Jing Zhang. Uni-map: Unified camera-lidar perception for robust hd map construction. *npj Artificial Intelligence*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Mark Horowitz. 1.1 computing's energy problem (and what we can do about it). In *ISSCC*, 2014.

Yangfan Hu, Huajin Tang, and Gang Pan. Spiking deep residual networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, 2023.

Souvik Kundu, Massoud Pedram, and Peter A. Beerel. Hire-snn: Harnessing the inherent robustness of energy-efficient deep spiking neural networks by training with crafted input noise. In *ICCV*, 2021.

Zhenxin Lei, Man Yao, Jiakui Hu, Xinhao Luo, Yanye Lu, Bo Xu, and Guoqi Li. Spike2former: Efficient spiking transformer for high-performance image segmentation. In *AAAI*, 2025.

Heng Li, Yuenan Hou, Xiaohan Xing, Yuexin Ma, Xiao Sun, and Yanyong Zhang. Occmamba: Semantic occupancy prediction with state space models. In *CVPR*, 2025a.

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 *CVPR*, 2023a.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, 2023b.

Ziqi Li, Tao Gao, Yisheng An, Ting Chen, Jing Zhang, Yuanbo Wen, Mengkun Liu, and Qianxi Zhang. Brain-inspired spiking neural networks for energy-efficient object detection. In *CVPR*, 2025b.

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 *CVPR*, 2025.

Kaixin Lin, Kunyu Peng, Di Wen, Yufan Chen, Ruiping Liu, and Kailun Yang. M^2-Occ: Resilient 3D Semantic Occupancy Prediction for Autonomous Driving with Incomplete Camera Inputs. *arXiv preprint arXiv:2603.09737*, 2026.

Guan-Horng Liu, Avinash Siravuru, Sai Prabhakar, Manuela Veloso, and George Kantor. Learning end-to-end multimodal sensor policies for autonomous navigation. In *CoRL*, 2017.

Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, 2021.

Wenxuan Liu, Yao Deng, Kang Chen, Xian Zhong, Zhaofei Yu, and Tiejun Huang. Sota: Spike-navigated optimal transport saliency region detection in composite-bias videos. In *IJCAI*, 2025.

Yijie Lu, Zhiyi Pan, Renrui Zhang, Yanhao Jia, Ronggang Wang, and Zhaokun Zhou. Spiking discrepancy transformer for point cloud analysis. In *ICLR*, 2026.

Xinhao Luo, Man Yao, Yuhong Chou, Bo Xu, and Guoqi Li. Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection. In *ECCV*, 2024.

Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *CVPR*, 2024.

MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection, 2020.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 2019.

Itai Orr, Moshik Cohen, and Zeev Zalevsky. High-resolution radar road segmentation using weakly supervised learning. *Nature Machine Intelligence*, 2021.

Jingyi Pan, Zipeng Wang, and Lin Wang. Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction. *IEEE Robotics and Automation Letters*, 2024a.

Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023.

Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, 2024b.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *CVPR*, 2025.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020.

Jinye Qu, Zeyu Gao, Tielin Zhang, Yanfeng Lu, Huajin Tang, and Hong Qiao. Spiking neural network for ultralow-latency and high-accurate object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021.

Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, 2020.

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in Neuroscience*, 2017.

Yimeng Shan, Zhenbang Ren, Haodi Wu, Wenjie Wei, Rui-Jie Zhu, Shuai Wang, Dehao Zhang, Yichen Xiao, Jieyuan Zhang, Kexin Shi, Jingzhinan Wang, Jason K. Eshraghian, Haicheng Qu, and Malu Zhang. Sdtrack: A baseline for event-based tracking via spiking neural networks. In *CVPR*, 2026.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017.

Yiding Sun, Haozhe Cheng, Chaoyi Lu, Zhengqiao Li, Minghong Wu, Huimin Lu, and Jihua Zhu. Hyperpoint: Multimodal 3d foundation model in hyperbolic space. *Pattern Recognition*, 2025.

Yiding Sun, Chaoyi Lu, Haozhe Cheng, Jun Wang, Huimin Lu, Lin Chen, and Jihua Zhu. Curve3d: Curvature-aware masked autoencoder for self-supervised point cloud understanding. *Pattern Recognition*, 2026a.

Yiding Sun, Xiangyang Yang, Dongxu Zhang, Qirui Wang, Zijie Xu, Wenxuan Liu, Shuiwang Li, Jihua Zhu, Zhaofei Yu, and Tiejun Huang. Spikingmot: A spike-driven multi-object tracker. *arXiv preprint arXiv:2607.19875*, 2026b.

Yiding Sun, Dongxu Zhang, Jihua Zhu, Haozhe Cheng, Zhengqiao Li, Pengcheng Li, Chaowei Fang, Yonghao Dong, and Lin Chen. Tri-efficient transfer learning for point cloud videos. In *ECCV*, 2026c.

Yiding Sun, Jihua Zhu, Haozhe Cheng, Chaoyi Lu, Zhichuan Yang, Lin Chen, and Yaonan Wang. Align then adapt: Rethinking parameter-efficient transfer learning in 4d perception. *IEEE Transactions on Multimedia*, 2026d.

Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *NeurIPS*, volume 36, 2023.

Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Occgen: Generative multi-modal 3d occupancy prediction for autonomous driving. In *ECCV*, 2024a.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, 2023a.

Run Wang, Chaoyi Zhou, Amir Salarpour, Xi Liu, Zhi-Qi Cheng, Feng Luo, Mert D. Pesé, and Siyu Huang. Flexmap: Robust hd map construction under flexible camera configurations. *arXiv preprint arXiv:2601.22376*, 2026.

Shiming Wang, Holger Caesar, Liangliang Nan, and Julian F. P. Kooij. Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities. In *IV*, 2024b.

Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, 2023b.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023.

Peixi Wu, Bosong Chai, Menghua Zheng, Wei Li, Zhangchi Hu, Jie Chen, Zheyu Zhang, Hebei Li, and Xiaoyan Sun. Efficient spiking point mamba for point cloud analysis. In *ICCV*, 2025.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience*, 2018.

Zijie Xu, Tong Bu, Zecheng Hao, Jianhao Ding, and Zhaofei Yu. Proxy target: Bridging the gap between discrete spiking neural networks and continuous control. In *NeurIPS*, 2025.

Zijie Xu, Xinyu Shi, Yiting Dong, Zihan Huang, and Zhaofei Yu. Care-bn: Precise moving statistics for stabilizing spiking neural networks in reinforcement learning. In *ICLR*, 2026.

Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, 2021.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *NeurIPS*, 2023.

Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Qiuyang Zhang, Jiujun Cheng, Qichao Mao, Cong Liu, Yu Fang, Yuhong Li, Mengying Ge, and Shangce Gao. Spiketrack: A spike-driven framework for efficient visual tracking. In *CVPR*, 2026.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, 2023.

Zhaokun Zhou, Yijie Lu, Yanhao Jia, Kaiwei Che, Jun Niu, Liwei Huang, Xinyu Shi, Yuesheng Zhu, Guoqi Li, Zhaofei Yu, and Li Yuan. Spiking transformer with experts mixture. In *NeurIPS*, 2024.

Zhiming Zhou, Yong He, Qiaoyun Wu, Chaoxu Mu, and Ajmal Mian. 3dsmt: A hybrid spiking mamba-transformer for point cloud analysis. In *ICLR*, 2026.

Rui-Jie Zhu, Ziqing Wang, Leilani Gilpin, and Jason K. Eshraghian. Autonomous driving with spiking neural networks. In *NeurIPS*, 2024.

Xubo Zhu, Haoyang Zhang, Fei He, Rui Wu, Yanhu Shan, Wen Yang, and Huai Yu. Dr.occ: Depth- and region-guided 3d occupancy from surround-view cameras for autonomous driving. In *CVPR*, 2026.

Shihao Zou, Qingfeng Li, Wei Ji, Jingjing Li, Yongkui Yang, Guoqi Li, and Chao Dong. Spikevideoformer: An efficient spike-driven video transformer with hamming attention and $\mathcal{O}(t)$ complexity. In *ICML*, 2025.

A Why Spiking Representations for Incomplete Sensing?

We provide a representation-level interpretation of why discrete spiking features can be attractive under varying sensor availability. The analysis concerns the binary spike codes generated inside the spiking experts, rather than claiming that an arbitrary SNN is intrinsically robust to arbitrary sensor failures. In particular, we show that hard-threshold spiking introduces an explicit margin within which sensor-induced feature perturbations leave the discrete representation exactly unchanged.

Sensor-induced perturbation. Consider the pre-threshold membrane potential of a spiking layer at one internal time step, denoted by $\mathbf{u} \in \mathbb{R}^d$ under complete sensing. For notational simplicity, the time index is omitted; the following argument applies independently to every internal step. Let M denote a sensor-availability state. The corresponding membrane potential can always be written as

$$\mathbf{u}^M = \mathbf{u} + \mathbf{\Delta}^M, \tag{17}$$

where $\mathbf{\Delta}^M$ represents the representation perturbation induced by changing sensor availability. The binary spike codes before and after this perturbation are

$$s_i = \mathbb{H}(u_i - V_{\text{th}}), \qquad s_i^M = \mathbb{H}(u_i + \Delta_i^M - V_{\text{th}}), \tag{18}$$

where $\mathbb{H}$ is the hard-threshold function used in Eq. 5. We define the threshold margin of the i-th unit as

$$\gamma_i = |u_i - V_{\text{th}}|\,. \tag{19}$$

Proposition 1 (Threshold-margin stability). If the spike state of unit i changes under sensor state M, then the sensor-induced perturbation must be at least as large as its threshold margin:

$$s_i^M \neq s_i \quad \Longrightarrow \quad \left|\Delta_i^M\right| \geq \gamma_i. \tag{20}$$

Table 8: **Effect of curriculum ordering in SAL.** All variants use the same sensor states, while differing only in their presentation order. Results are reported in mIoU.

Schedule	Full	All Cam. M.	LiDAR M.
Static	25.4	21.5	8.5
Shuffled	25.2	21.0	8.2
Reverse	24.8	20.2	7.7
SAL	**25.9**	**22.7**	**9.2**

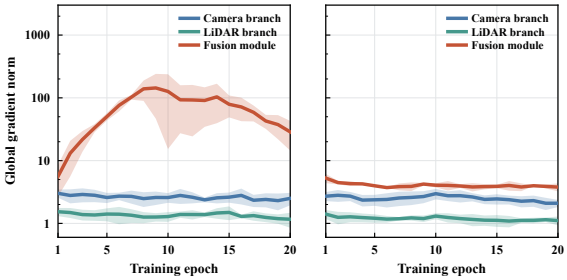


Figure 10: **Branch-wise gradient during training.**

Equivalently,

$$\mathbb{1}[s_i^M \neq s_i] \leq \mathbb{1}\left[\left|\Delta_i^M\right| \geq \gamma_i\right]. \tag{21}$$

Proof. If $u_i \geq V_{\text{th}}$ but $u_i + \Delta_i^M < V_{\text{th}}$, then $|\Delta_i^M| \geq u_i - V_{\text{th}} = \gamma_i$. Conversely, if $u_i < V_{\text{th}}$ but $u_i + \Delta_i^M \geq V_{\text{th}}$, then $|\Delta_i^M| \geq V_{\text{th}} - u_i = \gamma_i$. These are the only two cases in which the binary spike state can change, which proves Eq. 20. □

Proposition 1 gives an explicit invariance region for the forward spike representation. In particular, if $|\Delta_i^M| < \gamma_i$, the spike state is exactly preserved even though the underlying continuous membrane potential has changed. For the complete spike vector, let $d_H(\mathbf{s}, \mathbf{s}^M)$ denote the Hamming distance. Eq. 21 immediately yields

$$d_H(\mathbf{s}, \mathbf{s}^M) \leq \sum_{i=1}^{d} \mathbb{1}\left[|\Delta_i^M| \geq \gamma_i\right]. \tag{22}$$

Hence, sensor-induced feature changes do not necessarily propagate to all spiking units; only perturbations that cross their corresponding firing margins can alter the discrete code.

This property differs from a locally smooth continuous representation. For a differentiable continuous mapping $\mathbf{a} = g_{\text{cont}}(\mathbf{u})$, a sufficiently small perturbation satisfies the first-order expansion

$$g_{\text{cont}}(\mathbf{u} + \boldsymbol{\Delta}^M) - g_{\text{cont}}(\mathbf{u}) = \mathbf{J}_{g_{\text{cont}}}(\mathbf{u})\boldsymbol{\Delta}^M + o\big(\|\boldsymbol{\Delta}^M\|\big), \tag{23}$$

where $\mathbf{J}_{g_{\text{cont}}}$ denotes the local Jacobian. Whenever this Jacobian is nonzero along the perturbation direction, arbitrarily small input changes generally induce nonzero feature changes. In contrast, hard-threshold spiking provides a finite neighborhood in which the discrete forward representation remains exactly invariant.

B Diagnosing Instability in Naive Spiking Fusion

To further diagnose the instability observed in Fig. 2, we analyze the branch-wise gradient dynamics of the camera branch, LiDAR branch, and fusion module. As shown in Fig. 10, the camera and LiDAR branches of the Naive SNN remain relatively stable, while the fusion module exhibits much larger gradient magnitudes and repeated fluctuations. In contrast, SpikeOcc maintains much smoother gradient dynamics across all branches. In particular, the large oscillations in the fusion module are substantially reduced, indicating that the proposed expert decomposition alleviates interference between heterogeneous modalities and stabilizes spiking multimodal fusion.

C Alternative Curriculum Orderings

We examine whether the ordering of sensor failures matters in SAL. All variants use the same sensor states and comparable overall exposure, while only the training schedule is changed. As shown in Table 8, the default curriculum consistently outperforms alternative orderings, particularly under complete modality outages. Reversing the curriculum leads to the largest degradation, suggesting that exposing the model to severe failures too early hinders the formation of stable multimodal representations. Static and shuffled schedules perform better than the reversed order, but remain inferior to the progressive curriculum. These results indicate that SAL benefits not only from exposure to sensor failures, but also from gradually increasing their severity during training.

D Routing Target Definition

The routing target in Eq. 14 is determined by the availability of the two modalities. Let c, l, and j denote the RGB, LiDAR, and joint experts, respectively. For a sensor state characterized by m_c and m_l, we define

$$y^{\text{route}} = \begin{cases} c, & m_c = 1,\ m_l = 0, \\ l, & m_c = 0,\ m_l = 1, \\ j, & m_c = 1,\ m_l = 1. \end{cases} \tag{24}$$

Thus, LiDAR missing states are assigned to the RGB expert, while complete camera outages are assigned to the LiDAR expert. When both modalities remain available, including partial camera failures, the joint expert is used as the routing target. These cases cover all valid sensor states because $m_c + m_l \geq 1$. The target is used only by the auxiliary cross entropy loss on the raw router logits. The final representation still follows the soft routing weights in Eq. 11. Therefore, the auxiliary target encourages expert specialization without enforcing hard expert selection.

E Additional Evaluation Details

We follow the standard occupancy evaluation protocol and report geometric IoU and semantic mIoU (Wang et al., 2023b). For geometric occupancy, all non-empty semantic classes are merged into the occupied category. The IoU is computed as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{25}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative occupied voxels, respectively.

For semantic occupancy, the IoU of class k is

$$\text{IoU}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k + \text{FN}_k}, \tag{26}$$

and the mean IoU is

$$\text{mIoU} = \frac{1}{K-1} \sum_{k=1}^{K-1} \text{IoU}_k, \tag{27}$$

where class 0 denotes free space and is excluded from the semantic average. Thus, IoU measures geometric occupancy quality, while mIoU evaluates class-aware prediction over occupied voxels.

F Inference Protocol

During training, SAL samples different sensor states according to the curriculum schedule, and the auxiliary routing loss in Eq. 14 is used to encourage expert specialization. The corresponding sensor masks are propagated through the camera and LiDAR branches to construct the masked inputs.

At inference time, no curriculum sampling or routing supervision is used. The model directly receives the currently available sensor inputs together with their availability descriptor, and performs a single forward pass. The router produces the soft expert weights defined in Eq. 11, after which the fused representation is sent to the occupancy head. No retraining, test-time optimization, or repeated inference is required for different sensor configurations.

G Geometry–Semantics Trade-off under Camera Outages

An interesting trend appears as camera availability decreases in Table 1. The geometric IoU of SpikeOcc slightly increases, while the semantic mIoU gradually drops. A plausible explanation is that LiDAR provides direct and reliable geometric measurements, whereas camera features contribute richer appearance cues for semantic discrimination. Similar modality complementarity has also been observed in multimodal occupancy methods (Pan et al., 2024a). As fewer cameras remain

Table 9: **Comparison with SAL-equipped baselines on OpenOccupancy.** Values in parentheses denote the change relative to the corresponding model trained without SAL.

Method	Full		All Cam. Missing		LiDAR Missing	
	IoU	mIoU	IoU	mIoU	IoU	mIoU
OccMamba + SAL	**36.2** (+0.5)	**26.4** (+0.2)	29.3 (+10.5)	16.1 (+6.7)	19.2 (+11.4)	8.5 (+6.9)
SPSC + SAL	35.4 (+0.2)	26.2 (+0.1)	26.7 (+4.9)	13.9 (+5.2)	16.5 (+16.3)	6.0 (+5.9)
SpikeOcc	35.8 (+0.6)	25.9 (+1.2)	**36.8** (+5.3)	**22.7** (+7.1)	**19.7** (+2.3)	**9.2** (+1.1)

available, the prediction becomes more dependent on LiDAR features, which can preserve or even simplify geometric occupancy estimation but provides less semantic information. This observation suggests that geometric robustness and semantic richness are not always aligned under incomplete sensing, further motivating the preservation of modality-specific strengths in multimodal fusion.

H SAL-Equipped Baseline Comparisons

To separate the contribution of SAL from the proposed architecture, we apply the same SAL schedule and sensor masks to representative multimodal occupancy baselines. Table 9 reports the results under matched availability training. SAL consistently improves the resilience of different multimodal architectures, confirming that its benefit is not specific to SpikeOcc. Nevertheless, under the same availability training, SpikeOcc retains a clear advantage in severe sensor outages. In particular, it reaches 22.7 mIoU with all cameras missing, compared with 16.1 for OccMamba with SAL. These results indicate that SAL provides a general training benefit, while the additional robustness of SpikeOcc arises from its spiking expert architecture.

I Limitations and Future Directions

SpikeOcc focuses on a controlled setting in which sensor availability changes are known to the model. Within this scope, SMoE and SAL provide a unified predictor that avoids retraining separate models for each sensor configuration and substantially improves robustness under severe outages. Several limitations, however, remain beyond the current formulation.

First, the availability descriptor assumes that the system knows which sensors are currently valid. In practice, a sensor may produce corrupted observations before a complete failure is detected, making sensor health itself uncertain. Jointly estimating sensor reliability would require an additional fault detection or uncertainty estimation mechanism, which introduces a different learning problem from the controlled missing sensor setting studied here. Future work can therefore extend the router with learned sensor health estimation when computing expert weights.

Second, the space of possible sensor failures grows rapidly with the number of sensors. Even with six cameras and one LiDAR, there are many possible nonempty availability combinations, while real deployment further includes continuous degradations such as partial occlusion, calibration errors, and corrupted measurements. Exhaustively including these cases during training is neither practical nor desirable, since it would obscure the controlled analysis of representative failures. SAL instead uses a small set of structured states and evaluates whether the learned model generalizes beyond them. A natural extension is to replace the predefined curriculum with learned or adaptive failure sampling, allowing difficult configurations and continuous sensor degradation to be introduced according to the current model state.

Finally, SpikeOcc is designed as a single multimodal model rather than a collection of modality specific specialists. Consequently, when only one modality remains available, a dedicated unimodal model trained exclusively for that input can still retain advantages in some settings. Our objective is different. The goal is to preserve a strong performance floor without switching models or retraining when the sensing configuration changes. Future work may combine the current shared framework with stronger modality specific priors or lightweight specialist adaptation while retaining a unified inference pipeline.