SPK2IMGMOE: HETEROGENEOUS MIXTURE-OF-EXPERTS FOR HIGH-FIDELITY SPIKE-TO-IMAGE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Spike-to-image reconstruction remains challenging due to the strong heterogeneity of spiking data across space, time, and noise conditions. Most existing reconstruction methods process all features with uniform computation, which is inefficient and limits reconstruction quality. We propose Spk2ImgMoE, a Mixture-of-Experts-based framework for high-fidelity spike-to-image reconstruction. Spk2ImgMoE integrates Mixture-of-Experts at a mid-level semantic feature stage, enabling adaptive feature refinement with conditional computation. In particular, Spk2ImgMoE incorporates a Heterogeneous Mixture-of-Experts (HMoE) module that models the intrinsic heterogeneity of spiking data through structurally diverse experts with different computational capacities. In addition, a bias-driven routing mechanism introduces a learnable bias into the gating logits, serving as a dynamic load balancer that regulates expert utilization and prevents expert collapse without auxiliary loss constraints. Extensive experiments on multiple datasets show that Spk2ImgMoE improves reconstruction quality while maintaining a favorable balance between performance and computational efficiency, confirming the effectiveness of heterogeneous capacity allocation for spiking data.

1 INTRODUCTION

Spike-to-image reconstruction is fundamentally challenging due to the highly heterogeneous nature of spiking camera data (Dong et al., 2017; Zhu et al., 2019). Spike streams exhibit substantial variation across both spatial and temporal dimensions, encompassing sparse and dense regions, static backgrounds and fast-moving boundaries, as well as noise-dominated and information-rich signals (Zhao et al., 2024b; Zheng et al., 2021). This pronounced heterogeneity renders uniform feature processing inherently suboptimal, as reflected in many recent reconstruction approaches (Zhao et al., 2021; Zhang et al., 2023; Yin et al., 2025).

Most existing reconstruction methods rely on homogeneous models that apply identical computations and fixed capacity to all features (Zhao et al., 2021; Zhang et al., 2023; Jiang et al., 2024; Zhao et al., 2024b;a; Yin et al., 2025). This design leads to inefficient resource usage, over-computing on simple or static regions while providing insufficient modeling capacity for complex, dynamic, or noisy regions. Mixture-of-Experts (MoE) architectures offer a natural solution by enabling conditional computation, allowing different expert branches to contribute with input-dependent weights. As a result, MoE provides a flexible way to model the intrinsic heterogeneity of spiking data (Zamfir et al., 2024; DeepSeek-AI, 2024). However, directly applying existing MoE designs to spike-to-image reconstruction reveals two key limitations. First, experts are typically homogeneous in structure, limiting their ability to capture diverse feature characteristics. Second, expert utilization often suffers from severe load imbalance, where the gating network assigns most features to a small subset of experts (DeepSeek-AI, 2024). This imbalance results in a feedback loop during optimization, where frequently activated experts receive more gradient updates and converge faster, while under-utilized experts receive limited supervision and fail to learn meaningful representations.

To address these challenges, we propose Spk2ImgMoE, an MoE-enhanced framework for spike-to-image reconstruction that combines heterogeneous expert designs with a task-driven dynamic gating mechanism. Spk2ImgMoE introduces a novel Heterogeneous Mixture-of-Experts (HMoE) module

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.

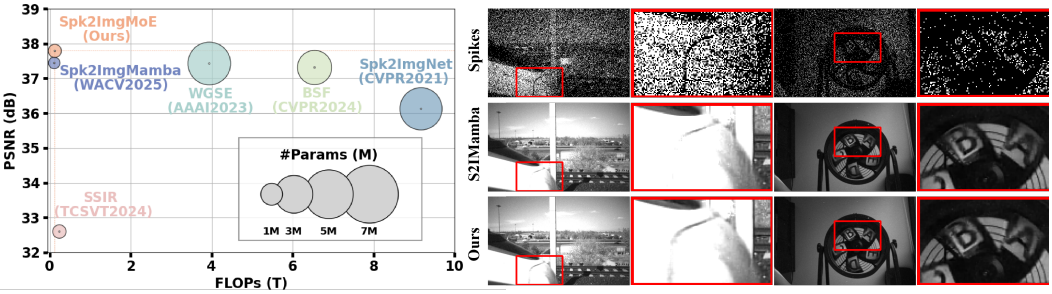


Figure 1: (Left) Comparison of PSNR, FLOPs, and model size on SREDS (Zhao et al., 2024a). Spk2ImgMoE demonstrates a favorable trade-off between reconstruction quality and computational complexity. (Right) In challenging high-speed scenarios, Spk2ImgMoE reconstructs finer details and clearer structures compared to Spk2ImgMamba (Yin et al., 2025).

at a mid-level semantic feature stage, located between channel-wise projection and local spatial aggregation, where features are semantically informative yet spatially unconstrained. This design provides experts with different computational capacities, enabling more flexible feature refinement while mitigating noise interference and spatial overfitting. To further ensure stable expert utilization and prevent expert collapse, we introduce a learnable bias into the gating network, enabling adaptive capacity allocation without auxiliary load-balancing losses. As illustrated in Fig. 1, Spk2ImgMoE reconstructs finer details and clearer structures in challenging high-speed scenarios compared to previous state-of-the-art methods, while also demonstrating a favorable trade-off between reconstruction quality and computational complexity. The main contributions are summarized as follows:

- We present Spk2ImgMoE, a Mixture-of-Experts-based framework that adapts computational capacity to the heterogeneous characteristics of spiking data for high-fidelity spike-to-image reconstruction.
- We propose a Heterogeneous Mixture-of-Experts (HMoE) module that integrates experts with different computational capacities and a bias-driven routing mechanism, allowing adaptive capacity allocation to heterogeneous spiking features while ensuring stable expert utilization without auxiliary loss constraints.
- Extensive experiments on multiple benchmark datasets show that Spk2ImgMoE achieves improved reconstruction quality while maintaining a favorable trade-off between computational efficiency and performance compared to existing state-of-the-art methods.

2 Related Work

2.1 Spike-to-Image Reconstruction

Reconstructing high-fidelity intensity images from asynchronous binary spike streams is a fundamental inverse problem in neuromorphic vision. Early approaches primarily exploited the statistical properties of photon arrivals. Representative methods, such as TFP and TFI (Zhu et al., 2019), utilized firing counts and intervals within fixed temporal windows to estimate intensity. Although computationally efficient, these methods face an inherent trade-off: long temporal windows suppress shot noise but introduce motion blur, whereas short windows preserve fast motion but produce noisy and unstable estimates. Subsequent studies incorporated short-term plasticity (Zheng et al., 2021) and retina-like mechanisms (Zhu et al., 2020) to improve temporal adaptation and refine statistical estimation. Nevertheless, their reliance on handcrafted or predefined response models limits their ability to capture complex nonlinear relationships between spike streams and scene intensity, particularly under rapid motion, low illumination, and severe noise.

Deep learning has significantly advanced spike-to-image reconstruction by learning nonlinear mappings from spike streams to scene intensity. Spk2ImgNet (Zhao et al., 2021) introduced a hierarchical CNN to extract spatio-temporal features, substantially improving reconstruction quality over analytical methods. Subsequent studies explored self-supervised learning (Chen et al., 2022) to reduce reliance on paired data and optical flow-guided reconstruction (Zhao et al., 2022) to better model

motion and recover details in dynamic scenes. Together, these methods demonstrate the effectiveness of deep networks in suppressing noise and reconstructing clear intensity images from sparse spike observations.

2.2 MIXTURE-OF-EXPERTS (MOE)

Mixture-of-Experts (MoE) scales model capacity without increasing inference cost by sparsely activating a subset of "experts" for each input token (Shazeer et al., 2017). This paradigm has been successfully adapted to vision tasks (Riquelme et al., 2021), enabling models to specialize in distinct visual patterns.

A persistent challenge in MoE training is expert load balancing. Without intervention, gating networks often collapse to using only a few experts, wasting model capacity. The standard solution involves adding an auxiliary load-balancing loss ($\mathcal{L}_{aux}$) (Fedus et al., 2022). However, in low-level vision tasks like reconstruction, this auxiliary objective can conflict with the primary pixel-wise fidelity loss, biasing the routing distribution toward artificial uniformity.

Our approach differs from standard MoE implementations in two key aspects tailored for spike reconstruction. First, instead of generic expert structures, we design heterogeneous experts to provide different computational capacities for spike features (Zamfir et al., 2024). Second, we eliminate the conflicting auxiliary loss by introducing a learnable dynamic gating bias (DeepSeek-AI, 2024). This allows expert utilization to be regulated implicitly by the reconstruction task itself, rather than by artificial distribution metrics.

3 METHODOLOGY

3.1 OVERALL ARCHITECTURE

We propose Spk2ImgMoE, a novel computation-aware framework designed for high-fidelity spike-to-image reconstruction. The input is a sparse, binary spike stream $\mathbf{S} \in \{0,1\}^{T\times H\times W}$ generated by an asynchronous integrate-and-fire mechanism (Dong et al., 2017; Zhu et al., 2019; Zhao et al., 2021). Specifically, each pixel at spatial coordinate $\mathbf{u}$ continuously accumulates incident light intensity $I(\mathbf{u}, t)$. A spike is triggered at time t_k when the temporally accumulated integral since the previous spike time t_{k-1} reaches a predefined threshold ϕ:

$$\int_{t_{k-1}}^{t_k} \alpha \cdot I(\mathbf{u}, \tau)\, d\tau = \phi, \tag{1}$$

where α denotes the photoelectric conversion efficiency.

As illustrated in Fig. 2, Spk2ImgMoE adopts a hierarchical architecture. The sparse spike input is first processed by a Dual-Stream Representation module to aggregate temporal information and generate shallow feature embeddings. These embeddings are then refined by three stacked Mixture-of-Experts (MoE) blocks, whose internal structure is shown in Fig. 2 (Bottom Left). Unlike conventional spike-to-image reconstruction backbones that apply uniform computation to all features, each MoE block starts with a Heterogeneous Mixture-of-Experts (HMoE) layer to adaptively refine features based on their heterogeneity (Zamfir et al., 2024). The resulting features are further processed by a depth-wise convolution (DW Conv) for local spatial modeling and a State Space Model (SSM) block (Gu & Dao, 2023; Yin et al., 2025) for global context modeling. A decoder then reconstructs the final intensity image.

Within the HMoE module, we introduce two key design choices. First, structurally diverse experts, namely simple experts and complex experts, are employed with different computational capacities, allowing the model to adaptively allocate resources and explicitly capture the heterogeneous characteristics of spiking data. Second, a Bias-Driven Routing Mechanism introduces a learnable bias $\mathbf{b}$ into the routing logits, which functions as a dynamic load balancer that implicitly regulates expert utilization and prevents expert collapse without relying on auxiliary loss constraints.

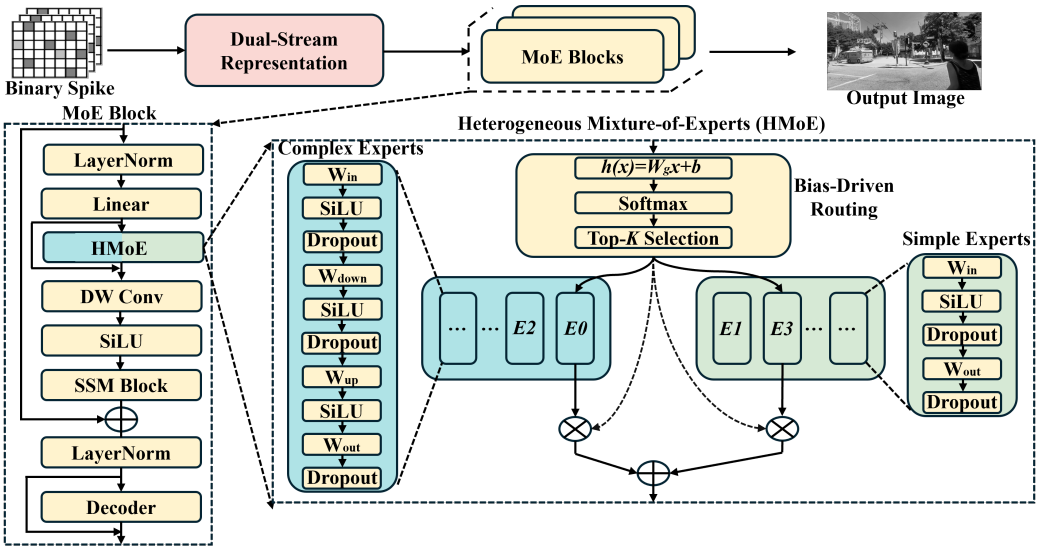


Figure 2: Overall architecture of the proposed Spk2ImgMoE framework. The input binary spike stream is first processed by a Dual-Stream Representation module to extract initial feature embeddings, which are subsequently refined by three stacked MoE Blocks. Each MoE Block incorporates a novel Heterogeneous Mixture-of-Experts (HMoE) module with two key innovations: (1) structurally diverse experts (simple and complex) with different computational capacities are employed to explicitly model the intrinsic heterogeneity of spiking data; (2) a Bias-Driven Routing Mechanism introduces a learnable bias $\mathbf{b}$ into the routing logits, serving as a dynamic load balancer that implicitly regulates expert utilization and prevents expert collapse.

3.2 Heterogeneous Mixture-of-Experts

Standard MoE architectures enable conditional computation by employing a gating network $\mathcal{G}$ to assign input-dependent weights to experts $\{E_i\}_{i=1}^{N}$ for each token $\mathbf{x} \in \mathbb{R}^D$ (Shazeer et al., 2017). The gating network produces a probability distribution $\mathbf{p}(\mathbf{x})$ over experts:

$$\mathbf{p}(\mathbf{x}) = \text{Softmax}(\mathbf{W}_g \mathbf{x}), \tag{2}$$

where $\mathbf{W}_g \in \mathbb{R}^{N \times D}$ represents the learnable routing weights. The output $\mathbf{y}$ is then obtained by the weighted aggregation of expert responses:

$$\mathbf{y} = \sum_{i=1}^{N} p_i(\mathbf{x}) \cdot E_i(\mathbf{x}). \tag{3}$$

To better align model capacity with the heterogeneous semantic complexity of spike-domain features (Zamfir et al., 2024), we propose a Heterogeneous Mixture-of-Experts (HMoE) design, as detailed in Fig. 2. The expert set $\mathcal{E} = \{E_1, \ldots, E_N\}$ is explicitly partitioned into two architecturally distinct tiers: simple experts and complex experts.

Simple Experts. Assigned to odd indices (e.g., E_1, E_3), simple experts emphasize computational efficiency. As shown in Fig. 2, each simple expert adopts a compact MLP structure that provides a lightweight transformation path. Formally, given projection matrices $\mathbf{W}_{in}$ and $\mathbf{W}_{out}$, the mapping is defined as:

$$E_{simple}(\mathbf{x}) = \mathcal{D}\Big(\mathbf{W}_{out} \cdot \mathcal{D}\big(\sigma(\mathbf{W}_{in}\mathbf{x})\big)\Big), \tag{4}$$

where $\sigma(\cdot)$ denotes the SiLU activation function and $\mathcal{D}(\cdot)$ denotes Dropout regularization.

Complex Experts. Assigned to even indices (e.g., E_0, E_2), complex experts are designed to model intricate dependencies (Zamfir et al., 2024). As depicted in Fig. 2, these experts employ a bottleneck architecture to increase representational depth. Specifically, the input feature is first projected into a compact latent space via $\mathbf{W}_{down}$ to capture condensed semantic cues:

$$\mathbf{z} = \sigma\Big(\mathbf{W}_{down} \cdot \mathcal{D}\big(\sigma(\mathbf{W}_{in}\mathbf{x})\big)\Big). \tag{5}$$

and subsequently reconstructed back to the original dimensionality through $\mathbf{W}_{up}$:

$$E_{complex}(\mathbf{x}) = \mathcal{D}\Big(\mathbf{W}_{out} \cdot \sigma\big(\mathbf{W}_{up} \cdot \mathcal{D}(\mathbf{z})\big)\Big). \tag{6}$$

The $\mathbf{W}_{down} \rightarrow \mathbf{W}_{up}$ bottleneck increases non-linear modeling capacity, providing a complementary higher-capacity transformation path.

3.3 Task-Driven Dynamic Gating

Conventional MoE gating mechanisms typically rely on an auxiliary load-balancing loss ($\mathcal{L}_{aux}$) to mitigate expert collapse, which can conflict with the primary reconstruction objective (Fedus et al., 2022). To avoid this conflict, we introduce a bias-driven routing mechanism that achieves load balancing without relying on an auxiliary loss.

Bias-Driven Routing Mechanism: As shown in Fig. 2, we introduce a learnable bias $\mathbf{b}$ directly into the routing computation. This mechanism promotes balanced expert utilization without introducing conflicting optimization objectives (DeepSeek-AI, 2024). Specifically, for an input token $\mathbf{x}$, the routing logits $\mathbf{h} \in \mathbb{R}^N$ are computed as:

$$\mathbf{h}(\mathbf{x}) = \mathbf{W}_g\mathbf{x} + \mathbf{b}. \tag{7}$$

The logits are then normalized via a Softmax operation to obtain the routing probability for the i-th expert:

$$p_i(\mathbf{x}) = \frac{\exp(h_i(\mathbf{x}))}{\sum_{j=1}^{N} \exp(h_j(\mathbf{x}))}. \tag{8}$$

The learnable bias $\mathbf{b}$ acts as a dynamic compensator during the training stage. When an expert E_i is consistently under-utilized, gradient updates naturally increase its corresponding bias term b_i, thereby elevating its routing probability. This self-regulating behavior reactivates neglected experts and achieves effective load balancing solely through the primary reconstruction objective, without requiring explicit auxiliary constraints.

The final output $\mathbf{y}$ is the weighted sum of the Top-K activated experts. Let $\mathcal{T} = \text{TopK}(\mathbf{p}(\mathbf{x}), K)$ denote the set of activated expert indices. The output aggregation is formulated as:

$$\mathbf{y} = \sum_{i \in \mathcal{T}} \left(\frac{p_i(\mathbf{x})}{\sum_{j \in \mathcal{T}} p_j(\mathbf{x})} \right) \cdot E_i(\mathbf{x}). \tag{9}$$

3.4 Loss Function

We employ a compound objective function that combines pixel-wise fidelity with perceptual quality. Crucially, unlike standard MoE training schemes that rely on an auxiliary loss ($\mathcal{L}_{aux}$) to enforce load balancing, our approach leverages the implicit self-regulating capability of the learnable bias $\mathbf{b}$ to prevent expert collapse. This design eliminates the need for $\mathcal{L}_{aux}$, enabling the optimization to focus strictly on image reconstruction:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{per}, \tag{10}$$

where λ is a hyperparameter that balances the two terms.

For the reconstruction term $\mathcal{L}_{rec}$, we calculate the ℓ_1 distance between the reconstructed image $\mathbf{I}_{rec}$ and the ground truth $\mathbf{I}_{gt}$ to ensure low-frequency signal fidelity:

$$\mathcal{L}_{rec} = \|\mathbf{I}_{rec} - \mathbf{I}_{gt}\|_1. \tag{11}$$

To further mitigate blur artifacts and enhance visual perception, we incorporate a perceptual loss $\mathcal{L}_{per}$:

$$\mathcal{L}_{per} = \sum_{l \in \mathcal{L}} \|\phi_l(\mathbf{I}_{rec}) - \phi_l(\mathbf{I}_{gt})\|_2, \tag{12}$$

where $\phi_l(\cdot)$ extracts feature maps from the l-th layer of a pre-trained feature extractor. This term compels the network to recover high-frequency structural details that align with human visual systems.

Table 1: Quantitative comparison against SOTA methods on the synthetic SREDS dataset and Real-captured datasets. We report the number of parameters and computational complexity (FLOPs). Our approach achieves an excellent balance between computational efficiency and reconstruction performance. **Bold** indicates the best performance.

Method	Backbone	Params (M)	FLOPs (T)	SREDS Dataset			Real-captured Datasets	
				PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓
TFP (Zhu et al., 2019)	-	-	-	25.35	0.690	0.2547	21.389	82.065
TFI (Zhu et al., 2019)	-	-	-	18.50	0.638	0.2590	15.967	64.252
TFSTP (Zheng et al., 2021)	-	-	-	20.68	0.618	0.2761	13.629	54.110
SSIR (Zhao et al., 2024a)	SNN	0.382	0.238	32.61	0.919	0.0500	7.376	20.420
Spk2ImgNet (Zhao et al., 2021)	CNN	3.761	9.169	36.13	0.950	0.0294	6.952	32.386
BSF (Zhao et al., 2024b)	CNN	2.477	6.532	37.32	0.956	0.0263	9.275	32.011
WGSE (Zhang et al., 2023)	CNN	3.852	3.930	37.44	0.958	0.0241	8.349	30.229
Spk2ImgMamba (Yin et al., 2025)	Mamba	0.282	0.099	37.46	0.958	0.0239	6.733	17.973
Spk2ImgMoE(Ours)	Mamba	0.364	0.126	**37.76**	**0.960**	**0.0201**	**6.693**	**15.183**

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets. Following established protocols, we utilize the synthetic SREDS dataset (Zhao et al., 2024a) for training and quantitative evaluation. Derived from the REDS dataset (Nah et al., 2019), SREDS comprises 240 training scenes and 30 testing scenes, with each scene containing 24 consecutive frames at a resolution of 1280×720. During training, scenes are cropped into non-overlapping 96×96 patches, yielding a total of 21,840 image samples. The temporal length of the input spike stream is set to $T = 61$. To assess generalization performance in real-world scenarios, we employ three real-captured datasets: **recVidarReal2019** (Zhu et al., 2020), **momVidarReal2021** (Zheng et al., 2023), and **PKU-Spike-High-Speed** (Zhao et al., 2021). These datasets feature a resolution of 400×250 and cover diverse high-speed motion scenarios in complex indoor and outdoor environments, serving as a rigorous benchmark to evaluate the model's robustness against real-world noise and dynamic motion blur.

Implementation Details. The proposed Spk2ImgMoE is implemented in PyTorch (Paszke et al., 2019) and trained on a single NVIDIA RTX A6000 GPU. We use the Adam optimizer (Kingma & Ba, 2015) with a batch size of 15 and an initial learning rate of 5×10^{-4}. The learning rate decays by a factor of 0.7 every 25 epochs. To enhance model robustness, we apply data augmentation to both spike streams and ground-truth images, including random vertical flips and rotations ($90°, 180°, 270°$). For the MoE configuration, we set the total number of experts to $N = 4$ and the number of active experts to $k = 2$.

Evaluation Metrics. We employ a comprehensive set of metrics to evaluate reconstruction quality. For synthetic data with ground truth, we report Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) (Wang et al., 2004), where higher scores indicate better fidelity, as well as Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), where lower scores denote better perceptual quality. For Real-captured Datasets, we utilize no-reference metrics including Naturalness Image Quality Evaluator (NIQE) (Mittal et al., 2013) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal et al., 2012), where lower values indicate superior image quality.

4.2 COMPARISON WITH STATE-OF-THE-ART

Quantitative Results. As summarized in Table 1, Spk2ImgMoE achieves the best performance across all evaluation metrics, outperforming existing physics-based (Zhu et al., 2019), SNN-based (Zhao et al., 2024a), CNN-based (Zhao et al., 2021; 2024b), and Mamba-based (Yin et al., 2025) methods. In a direct comparison with the previous state-of-the-art Spk2ImgMamba (Yin et al., 2025) on the synthetic SREDS dataset (Zhao et al., 2024a), Spk2ImgMoE improves PSNR by 0.30 dB and SSIM by 0.002. Notably, these performance gains are achieved with negligible parameter overhead (0.082M), suggesting that the improvements stem from more effective expert capacity allocation rather than brute-force model scaling (Zamfir et al., 2024).

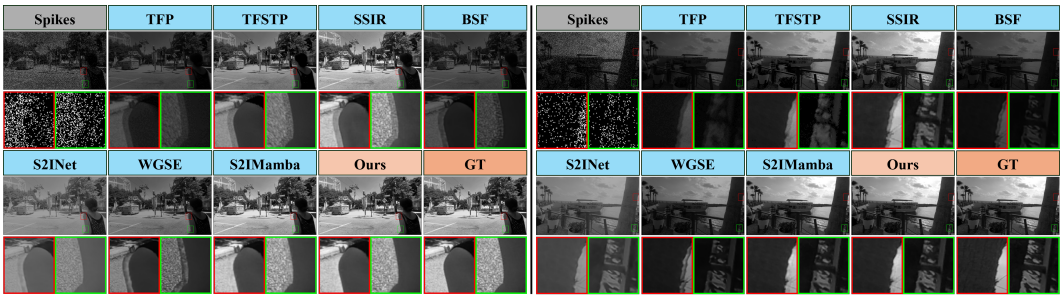


Figure 3: **Visual comparison of reconstruction quality.** We compare our proposed method against state-of-the-art approaches including TFP (Zhu et al., 2019), TFSTP (Zheng et al., 2021), SSIR (Zhao et al., 2024a), BSF (Zhao et al., 2024b), Spk2ImgNet (S2INet) (Zhao et al., 2021), WGSE (Zhang et al., 2023), and Spk2ImgMamba (S2IMamba) (Yin et al., 2025). The red and green boxes highlight zoomed-in regions. As observed, our method (Ours) recovers sharper structural details and clearer textures compared to the competing methods, which often suffer from noise or blurring artifacts.

Furthermore, to assess the generalization capability of Spk2ImgMoE under complex real-world conditions, we extend the evaluation to multiple real-captured datasets (Zhu et al., 2020; Zheng et al., 2021; Zhao et al., 2021), reconstructing frames from indices 140 to 230 at an interval of 10. As shown in Fig. 1 (Right), Spk2ImgMoE consistently recovers sharper details and clearer structural boundaries than the state-of-the-art Spk2ImgMamba (Yin et al., 2025) across diverse scenes. These visual observations are further supported by quantitative no-reference metrics. In this challenging zero-shot setting, Spk2ImgMoE achieves the lowest NIQE (Mittal et al., 2013) (6.69) and BRISQUE (Mittal et al., 2012) (15.18) scores among all compared methods. This result suggests that the proposed mid-level heterogeneous feature refinement helps decouple semantic content from sensor-specific noise, enabling the model to learn robust representations that generalize beyond the synthetic training domain (Zamfir et al., 2024).

Qualitative Results. Fig. 3 presents a visual comparison of reconstruction results on the synthetic SREDS dataset. As shown, existing methods, including the current state-of-the-art Spk2ImgMamba, often exhibit pronounced ghosting artifacts and motion blur, particularly in regions with fast motion or abrupt intensity variations (Zhao et al., 2024b). In contrast, Spk2ImgMoE produces noticeably sharper reconstructions with clearer structural boundaries and finer texture details, effectively suppressing these artifacts while better preserving the underlying scene dynamics.

4.3 Ablation Studies

We conduct ablation studies to assess the contribution of the task-driven gating mechanism and the heterogeneous expert design (Zamfir et al., 2024).

Impact of the MoE and Gating Bias. Table 2 evaluates the impact of MoE integration and the proposed learnable bias b. Directly adding MoE improves PSNR to 37.63 dB, but the gain remains limited without explicit load regulation, suggesting imbalanced expert utilization (Shazeer et al., 2017). Adding the bias term further improves PSNR to 37.76 dB and reduces LPIPS to 0.0201, indicating that the bias effectively stabilizes expert utilization without introducing an auxiliary balancing loss (DeepSeek-AI, 2024).

Impact of the Heterogeneous Expert Design. Table 3 compares the proposed heterogeneous experts with homogeneous alternatives (Zamfir et al., 2024). Specifically, All Simple Experts assigns a lightweight architecture to every expert, All Complex Experts uses a high-capacity architecture throughout, and Heterogeneous Experts combines experts with different computational capacities. The All Complex Experts setting increases the parameter count to 0.410M but achieves only 37.50 dB PSNR, suggesting that uniformly high capacity may introduce redundant computation or overfit noise-dominated features (Zhao et al., 2024b). The All Simple Experts setting also degrades performance to 37.53 dB, indicating that uniformly lightweight experts may lack sufficient representational flexibility. In contrast, the proposed heterogeneous design achieves the best trade-off, reaching 37.76 dB with 0.364M parameters and 0.126T FLOPs. These results demonstrate that combining experts with

Table 2: Ablation study of the effectiveness of MoE integration and the Learnable Dynamic Bias (b) on the SREDS and real-captured datasets.

Method	Params (M)	FLOPs (T)	SREDS Dataset			Real-captured Datasets	
			PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓
Baseline	0.282	0.099	37.46	0.958	0.0239	6.733	17.973
Baseline + MoE	0.364	0.126	37.63	0.960	0.0206	7.032	18.269
Baseline + MoE + b (Ours)	0.364	0.126	**37.76**	**0.960**	**0.0201**	**6.693**	**15.183**

Table 3: Ablation study of Expert Heterogeneity on the SREDS and real-captured datasets. The heterogeneous design achieves the best trade-off between reconstruction quality and computational complexity.

Expert Config	Params (M)	FLOPs (T)	SREDS Dataset			Real-captured Datasets	
			PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓
All Complex Experts	0.410	0.137	37.50	0.959	0.0218	6.819	18.923
All Simple Experts	0.340	0.120	37.53	0.959	0.0212	6.922	17.685
Heterogeneous (Ours)	0.364	0.126	**37.76**	**0.960**	**0.0201**	**6.693**	**15.183**

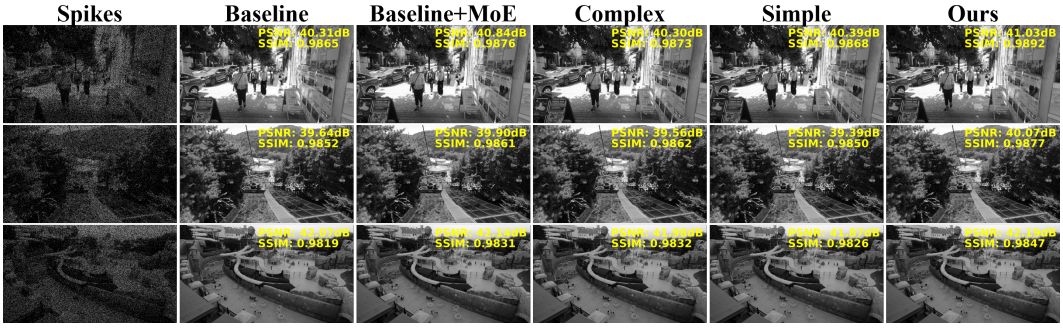


Figure 4: Visual ablation comparison on the SREDS dataset (Zhao et al., 2024a). Reconstruction quality and quantitative metrics (PSNR/SSIM) are shown for different variants. From left to right: Baseline, Baseline + MoE, All Complex Experts, All Simple Experts, and the proposed method (Ours).

different capacities enables more effective allocation of computational resources for heterogeneous spike-stream features (Zhu et al., 2019; Zhao et al., 2024b). The visual ablation results in Fig. 4 further support the quantitative trends in Tables 2 and 3. The complete model consistently achieves better PSNR, SSIM, and LPIPS scores, indicating that both the bias-driven routing mechanism and the heterogeneous expert design contribute to measurable reconstruction improvements.

Impact of the Expert Configuration and MoE Placement. We further investigate the expert configuration and the insertion position of the MoE module in Table 4. As shown in Table 4 (Left), which compares different combinations of the total number of experts N and the number of activated experts k, all MoE variants improve over the baseline (reported in the right panel). The proposed configuration with $N = 4$ experts and $k = 2$ activated experts achieves the best PSNR, SSIM, and LPIPS under a comparable computational budget. Increasing the number of experts to $N = 6$ or $N = 8$ does not bring further gains, suggesting that excessive expert capacity may introduce redundant routing rather than more effective capacity usage.

Table 4 (Right) compares different insertion positions of the MoE module. Placing the MoE at the middle feature stage outperforms both early- and late-stage placement. This result validates our choice to perform expert routing after semantic projection and before local spatial aggregation, where the features contain sufficient semantic information while retaining fine-grained spatial details.

Training Stability Across Random Seeds. To assess training stability, we compare the loss curves obtained with different random seeds in Fig. 5. Both runs exhibit consistent convergence behavior, with the training loss decreasing rapidly during the early stage and gradually stabilizing as training proceeds. Despite minor optimization fluctuations, the two curves follow similar trajectories and

Table 4: Ablation studies on the SREDS dataset. (Left) Effect of different expert configurations (N and k). (Right) Effect of the MoE insertion position.

Method	Params (M)	FLOPs (T)	SREDS Dataset PSNR ↑	SSIM ↑	LPIPS ↓
$N = 4, k = 1$	0.354	0.124	37.61	0.959	0.0235
$N = 6, k = 2$	0.364	0.126	37.63	0.960	0.0202
$N = 8, k = 2$	0.364	0.126	37.59	0.959	0.0208
$N = 4, k = 2$ **(Ours)**	0.364	0.126	**37.76**	**0.960**	**0.0201**

Method	Params (M)	FLOPs (T)	SREDS Dataset PSNR ↑	SSIM ↑	LPIPS ↓
Baseline	0.282	0.099	37.46	0.958	0.0239
Early MoE	0.372	0.128	37.54	0.960	0.0209
Late MoE	0.372	0.128	37.58	0.959	0.0211
Middle MoE (Ours)	0.364	0.126	**37.76**	**0.960**	**0.0201**

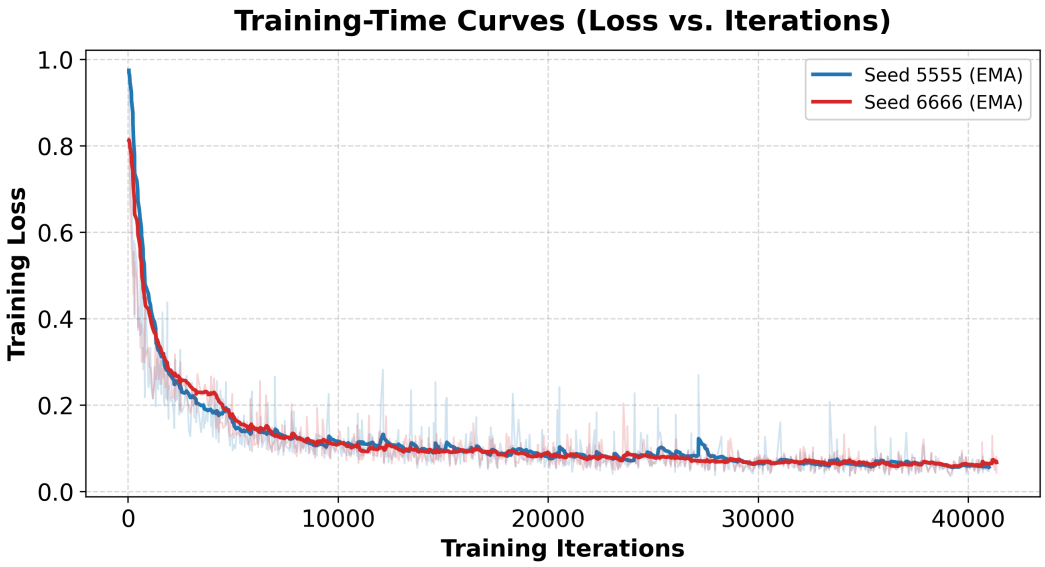


Figure 5: Training stability across different random seeds. The EMA loss curves for Seed 5555 and 6666 exhibit similar convergence trends and reach comparable final loss values.

converge to comparable final loss values. These results indicate that Spk2ImgMoE is robust to random initialization and maintains stable optimization across different seeds.

5 Conclusion

In this paper, we propose Spk2ImgMoE, pioneering the integration of the Mixture-of-Experts (MoE) paradigm into spike camera image reconstruction. To effectively handle the intrinsic data heterogeneity, we constructed a computation-aware framework driven by a Heterogeneous MoE (HMoE) mechanism. By dynamically routing features to heterogeneous experts via an auxiliary-loss-free gating strategy, our model achieves adaptive resource allocation for spike-stream reconstruction. Extensive experiments validate that this MoE-centric design significantly outperforms existing methods, establishing a new state-of-the-art benchmark.

In future work, we aim to further exploit the scalability of MoE architectures for semantic-aware spike-to-image reconstruction, while exploring hardware-friendly implementations that enable efficient deployment on neuromorphic and edge computing platforms.

AI use statement

Generative AI tools were used for English language editing and manuscript preparation assistance. No generative AI tools were used for research design, methodology development, experiments, data analysis, code development, result interpretation, or formulation of scientific claims. All AI-assisted edits and preparation work were carefully reviewed by the authors, who take full responsibility for the final content of this work.

Reproducibility statement

To support reproducibility, we describe the main model design, experimental settings, datasets, and evaluation metrics in the paper. The source code will be made publicly available upon acceptance.

References

Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 2859–2866, 2022.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. In *Proceedings of the Data Compression Conference (DCC)*, pp. 437–437, 2017.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research (JMLR)*, 23 (1):5232–5270, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Liangyan Jiang, Chuang Zhu, and Yanxu Chen. Swinsf: Image reconstruction from spatial-temporal spike streams. *arXiv preprint arXiv:2407.15708*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Anish Mittal, Anush Krishna Moorthy, and Alan C Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing (TIP)*, 21(12):4695–4708, 2012.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters (SPL)*, 20(3):209–212, 2013.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 0–0, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano, Lerrel Pinto, Neil Houlsby, et al. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8583–8595, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4): 600–612, 2004.

Jiaoyang Yin, Bin Fan, Boxin Shi, Chao Xu, and Tiejun Huang. Spk2imgmamba: Spiking camera image reconstruction with multi-scale state space models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 9342–9352, 2025.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. *arXiv preprint arXiv:2412.16434*, 2024.

Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 137–147, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11996–12005, 2021.

Rui Zhao, Ruiqin Xiong, Jing Zhao, Zhaofei Yu, Xiaopeng Fan, and Tiejun Huang. Learning optical flow from continuous spike streams. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:7905–7920, 2022.

Rui Zhao, Ruiqin Xiong, Jian Zhang, Zhaofei Yu, Shuyuan Zhu, Lei Ma, and Tiejun Huang. Spike camera image reconstruction using deep spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 34(6):5207–5212, 2024a.

Rui Zhao, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Zhaofei Yu, and Tiejun Huang. Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24955–24965, 2024b.

Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6358–6367, 2021.

Yajing Zheng, Jiyuan Zhang, Rui Zhao, Jianhao Ding, Shiyan Chen, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. Spikecv: Open a continuous computer vision era. *arXiv preprint arXiv:2303.11684*, 2023.

Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1432–1437, 2019.

Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1438–1446, 2020.