DINOv3 Perceptual Loss with Token and Correlation Matching for Spike Image Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Spike image reconstruction aims to transform asynchronous spike streams into visually meaningful images, yet it remains a challenging problem in neuromorphic vision. Existing methods typically rely on pixel-wise reconstruction losses, which often fail to capture perceptual similarity and thus produce overly smooth or visually unnatural results. To address this issue, we propose a DINOv3 perceptual loss with token and correlation matching for spike image reconstruction. Built upon DINOv3, a self-supervised vision foundation model, the proposed loss exploits both global contextual representations and dense patch-level features to provide more effective perceptual supervision. Specifically, it integrates three complementary components within a unified DINOv3 feature space: class-token matching to preserve global semantic consistency, patch-token matching to retain fine-grained details, and Gram-based correlation matching to enforce structural coherence. Compared with conventional CNN-based perceptual losses, the proposed method benefits from the global self-attention mechanism of Vision Transformers and avoids the supervised category bias introduced by classification-based pretraining, leading to better transferability and stronger generalization. The proposed loss is used only during training and introduces no additional inference overhead. Extensive experiments show that it consistently improves reconstruction performance across CNN-, Mamba-, and Transformer-based architectures. In particular, SwinSF equipped with the proposed loss achieves state-of-the-art results on the Spike-REDS and Spike-X4K benchmarks.

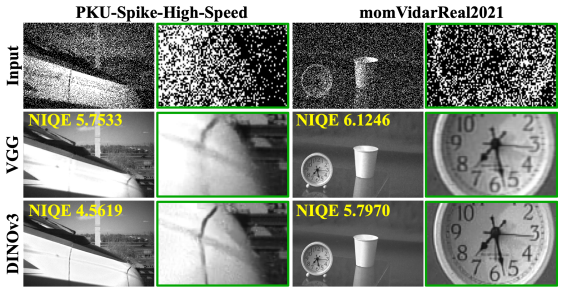
1 Introduction

Spike cameras asynchronously record per-pixel brightness changes with high temporal resolution and low power consumption, offering clear advantages over conventional frame-based cameras (Huang et al. (2023); Maqueda et al. (2018)). These properties enable reliable capture of fast motion and operation under challenging lighting conditions. However, raw spike data are not directly interpretable by humans or standard vision algorithms, making spike image reconstruction a critical step for recovering perceptually meaningful images from asynchronous spike streams. Achieving high perceptual quality is therefore essential for both human interpretation and downstream vision tasks.

Recent advances in deep learning have significantly improved spike image reconstruction. CNN-based methods, such as Spk2ImgNet (Zhao et al. (2021)) and WGSE (Zhang et al. (2023)), effectively learn hierarchical spatio-temporal features and enhance reconstruction quality. Transformer-based models, including SpikeFormer (She & Qing (2022)) and SwinSF (Jiang et al. (2024)), leverage self-attention to model long-range dependencies and complex temporal dynamics. These learning-based approaches substantially outperform traditional model-based methods, such as TFI (Zhu et al. (2019)) and TFP (Zhu et al. (2019)), in both quantitative accuracy and perceptual quality.

Despite recent progress, most existing deep models are trained using per-pixel loss functions (e.g., ℓ_1, ℓ_2, or Smooth-L_1), which measure differences at the pixel level (Dong et al. (2025); Zhang et al. (2018)). Although efficient, such losses poorly reflect perceptual similarity. For example, small spatial shifts can produce large pixel-wise errors despite minimal visual difference, and optimization with pixel-level losses often leads to overly smooth reconstructions lacking fine structural details

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.



Quantitative Comparison (NIQE ↓)

Method	PKU-Spike	momVidar
w/o perceptual loss	5.5697	8.9631
w/ L_{VGG}	5.4528	8.0250
w/ L_{DINOv3}	**5.3185**	**7.7023**

Figure 1: Qualitative and quantitative comparison of SwinSF models with VGG (Johnson et al. (2016)) and DINOv3 perceptual losses on PKU-Spike-High-Speed and momVidarReal2021.

(Ledig et al. (2017); Luo et al. (2023)). To alleviate this issue, perceptual losses based on deep feature representations have been introduced, typically using pretrained CNNs such as VGG (Johnson et al. (2016)). However, VGG-based perceptual losses are constrained by local convolutional receptive fields, limiting their ability to capture global context. In addition, because VGG is trained in a supervised manner for ImageNet classification, its features are biased toward category-specific semantics and often generalize poorly to unseen categories.

In this paper, we propose a DINOv3 perceptual loss for spike image reconstruction. Built upon the self-supervised Vision Transformer DINOv3 (Siméoni et al. (2025)), the proposed loss captures both global semantics and local details within a unified feature space. Unlike VGG-based perceptual losses, DINOv3 provides global receptive fields from early layers, enabling better modeling of holistic scene structures and fine-grained textures. As shown in Figure 1, the proposed DINOv3 perceptual loss produces more natural reconstructions with lower NIQE scores compared to the VGG-based perceptual loss, which inadequately models local details and structural coherence.

Specifically, we leverage both class tokens, which capture global semantic information, and patch tokens, which encode local details, from DINOv3 to construct a multi-component perceptual loss with three terms: (1) a global semantic alignment loss, (2) a local patch similarity loss, and (3) a Gram-based structural consistency loss for preserving structural and texture coherence. This design enables perceptual supervision that is both semantically meaningful and spatially detailed. Moreover, the self-supervised representations of DINOv3 generalize well to unseen domains, improving perceptual robustness. Finally, the Gram-based term further enhances structural consistency, reduces sensitivity to local mismatches, and alleviates visual artifacts in the reconstructed images. The contributions of this work are summarized as follows:

- We identify a fundamental limitation in existing spike image reconstruction methods, where pixel-wise and CNN-based perceptual losses inadequately model fine-grained local details and structural coherence, leading to suboptimal perceptual quality. To address this, we propose a spike image reconstruction framework that incorporates a DINOv3-based perceptual loss, leveraging its rich patch-level representations and global context modeling to provide more effective and holistic perceptual supervision.
- We design a multi-component perceptual objective that jointly enforces global semantic alignment, fine-grained local detail preservation, and inter-patch structural consistency via class-token, patch-token, and Gram-based correlation constraints, leading to more perceptually faithful, structurally coherent, and visually realistic reconstructions without introducing additional inference cost.
- The proposed framework demonstrates strong generalization across CNN-, Mamba-, and Transformer-based spike reconstruction models, consistently improving both reconstruction accuracy and perceptual fidelity across multiple benchmarks, and achieving state-of-the-art performance on Spike-REDS and Spike-X4K.

2 Related Work

Spike Image Reconstruction. Spike image reconstruction aims to recover clear intensity images from asynchronous spike streams. Earlier model-driven methods, such as TFP and TFI (Zhu et al.

(2019)), rely on the physics of spike generation and do not need large datasets, but they struggle with noise and fast motion, often producing blurry results. Recent deep learning approaches learn the mapping from spike data to images. CNN-based models like Spk2ImgNet (Zhao et al. (2021)), WGSE (Zhang et al. (2023)), and RSIR (Zhu et al. (2023)) focus on local features, while transformer models such as SwinSF (Jiang et al. (2024)) and SpikeFormer (She & Qing (2022)) capture global context. The newer Mamba-based model Spk2ImgMamba (Yin et al. (2025)) improves efficiency and temporal modeling. However, most existing methods still use pixel-wise loss functions (e.g., ℓ_1 or ℓ_2), which fail to capture perceptual information. This often leads to smooth images that lack fine details, motivating the use of perceptual losses for more realistic reconstruction.

Perceptual Loss. Traditional pixel-wise loss functions often fail to capture perceptual fidelity in image generation and reconstruction tasks. They tend to produce overly smooth results, causing images to lose fine details, textures, and natural variations, which misaligns with human visual perception (Ledig et al. (2017); Luo et al. (2023)). To overcome this issue, perceptual loss was introduced to measure image differences in a learned feature space rather than the raw pixel space. The seminal work by Johnson et al. (Johnson et al. (2016)) first demonstrated that feature activations extracted from a pretrained deep convolutional network can serve as an effective perceptual similarity metric. Using intermediate features of the VGG network (Simonyan & Zisserman (2014)), they formulated losses that encourage generated images to match the semantic and stylistic representations of target images, significantly improving results in style transfer and super-resolution. Zhang et al. (Zhang et al. (2018)) later explained why deep features are perceptually meaningful. They built a large dataset of human similarity judgments and showed that distances in deep feature space align closely with human perception. Pihlgren et al. (Pihlgren et al. (2023)) further analyzed perceptual loss designs across different architectures and domains, highlighting the effects of layer choice and normalization. More recently, Lin et al. (Lin & Yang (2023)) extended perceptual loss to diffusion models by introducing a self-perceptual loss, where the model itself serves as the feature extractor. Perceptual loss has evolved from a heuristic alternative to pixel-level objectives into a principled framework built on learned feature representations. Its variants continue to play a key role in enhancing visual quality across various image reconstruction tasks.

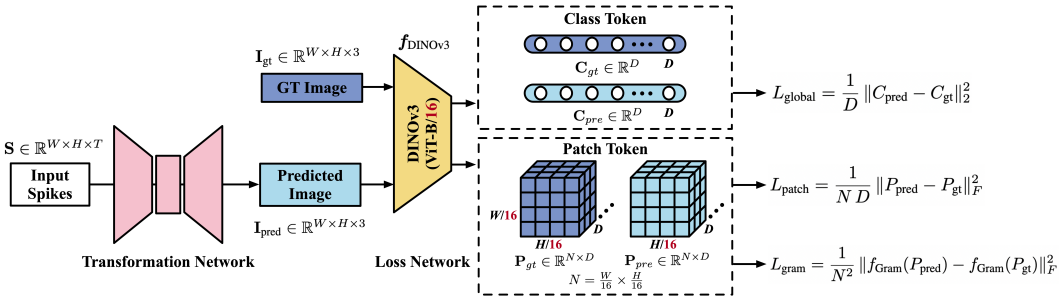


Figure 2: Overview of the proposed spike image reconstruction framework, which consists of a transformation network and a loss network. The transformation network converts spike data into reconstructed images, while the loss network incorporates the DINOv3-based perceptual loss composed of three components: a global loss L_{global} that aligns high-level semantic representations, a patch loss L_{patch} that preserves fine-grained spatial details, and a Gram loss L_{gram} that enforces structural consistency by modeling inter-patch relationships.

DINO Family Models. The pioneering work DINO, proposed by Caron et al. (Caron et al. (2021)), introduced a self-distillation framework that works well with Vision Transformers (ViTs) (Dosovitskiy (2020); Touvron et al. (2021)). DINO showed that self-supervised ViTs can naturally discover semantic parts such as heads, bodies, and objects without any labels, making them useful for segmentation and object discovery. However, scaling early DINO models to larger sizes caused instability and weaker local patch representations due to overfitting to global features. DINOv2 (Oquab et al. (2023)) improved the framework to larger datasets and model sizes, emphasizing data curation and robust training. It produces strong general-purpose visual features that transfer well across tasks, often outperforming supervised models. More recently, DINOv3 (Siméoni et al. (2025)) further advanced the framework through refined optimization and Gram anchoring to preserve dense patch consistency during long training. It achieves excellent performance in dense prediction and transfer learning, establishing it as a powerful and versatile visual foundation model.

The DINO family has also inspired broader applications. Researchers have studied the locality and interpretability of DINO features (Vanyan et al. (2023)) and applied them to biomedical imaging (Doron et al. (2023)) and remote sensing (Wang et al. (2022)), showing strong robustness and generalization. Overall, the DINO framework proves that self-distillation with transformers can learn rich semantic and dense patch-level features, enabling general visual understanding without supervision.

3 METHOD

3.1 OVERVIEW

Figure 2 illustrates the proposed spike-to-image framework, which consists of two main components: a spike-to-image transformation network and a loss network. The transformation network, implemented with SwinSF (Jiang et al. (2024)), converts spike data $\mathbf{S} \in \mathbb{R}^{W\times H\times T}$ into reconstructed images $\mathbf{I}_{\text{pred}} \in \mathbb{R}^{W\times H}$. The loss network defines the DINOv3 perceptual loss, which consists of three components based on DINOv3 features that measure the difference between the reconstructed image and the ground truth $\mathbf{I}_{\text{gt}}$.

Specifically, DINOv3 is used as a fixed loss network to extract meaningful representations from both the predicted and ground-truth images. It provides two types of feature tokens: class tokens, representing global semantics, and patch tokens, capturing local details. For each spike input, we compute the differences between the corresponding class and patch tokens from $\mathbf{I}_{\text{pred}}$ and $\mathbf{I}_{\text{gt}}$. The resulting DINOv3 perceptual loss consists of three components: a global loss that aligns high-level semantics, a patch loss that matches local details, and a Gram loss that preserves structural consistency by aligning inter-patch relationships.

3.2 DINO PERCEPTUAL LOSS

Inputs to DINOv3 Model. Each predicted image $\mathbf{I}_{\text{pred}}$ and ground-truth image $\mathbf{I}_{\text{gt}}$ is first resized to a uniform spatial resolution of $W \times H \times 3$. We use the pretrained ViT-B/16 backbone from DINOv3 as a frozen feature extractor. With patch size 16×16, each image yields $N = \frac{W}{16} \times \frac{H}{16}$ non-overlapping patches, each projected into a feature embedding of dimension $D = 768$.

DINOv3 Features. Given the frozen DINOv3 backbone f_{DINOv3}, the model extracts both global and local feature representations from the predicted and ground-truth images:

$$(C_{\text{pred}},\, P_{\text{pred}}) = f_{\text{DINOv3}}(I_{\text{pred}}), \qquad (C_{\text{gt}},\, P_{\text{gt}}) = f_{\text{DINOv3}}(I_{\text{gt}}). \tag{1}$$

Here, $C_{\text{pred}}, C_{\text{gt}} \in \mathbb{R}^D$ denote the global class tokens summarizing holistic image content, and $P_{\text{pred}}, P_{\text{gt}} \in \mathbb{R}^{N\times D}$ represent the patch-level embeddings capturing local spatial details, where $N = \frac{W}{16} \times \frac{H}{16}$ is the total number of patches.

Loss Terms. The DINOv3 perceptual loss consists of three components, each computed using the mean-squared error between corresponding feature representations:

$$L_{\text{global}} = \frac{1}{D}\,\|C_{\text{pred}} - C_{\text{gt}}\|_2^2,\; L_{\text{patch}} = \frac{1}{ND}\,\|P_{\text{pred}} - P_{\text{gt}}\|_F^2,\; L_{\text{gram}} = \frac{1}{N^2}\,\|f_{\text{Gram}}(P_{\text{pred}}) - f_{\text{Gram}}(P_{\text{gt}})\|_F^2. \tag{2}$$

where $f_{\text{Gram}}(\cdot)$ denotes the Gram matrix function (described later), $\|\cdot\|_2$ represents the Euclidean norm for vectors, and $\|\cdot\|_F$ denotes the Frobenius norm for matrices.

Among the three loss terms, the global loss L_{global} focuses on aligning high-level semantic representations, ensuring that the reconstructed image conveys the same overall meaning as the ground truth. The patch loss L_{patch} preserves fine-grained spatial details by enforcing local similarity between corresponding patch features. The Gram loss L_{gram}, described later, complements these objectives by promoting structural consistency through inter-patch relationships, thereby capturing the spatial dependencies among image regions. The overall DINO perceptual loss is formulated as a weighted combination of the three components:

$$L_{\text{DINOv3}} \;=\; \lambda_{\text{global}}\, L_{\text{global}} \;+\; \lambda_{\text{patch}}\, L_{\text{patch}} \;+\; \lambda_{\text{gram}}\, L_{\text{gram}} \tag{3}$$

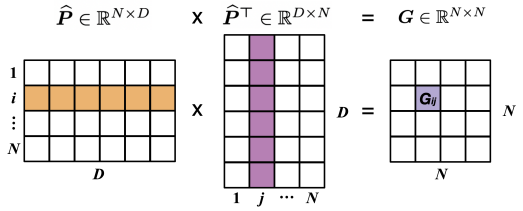


Figure 3: Visualization of the Gram matrix function $f_{\text{Gram}}(\cdot)$, which computes the Gram matrix $\boldsymbol{G} \in \mathbb{R}^{N\times N}$ from the normalized patch tokens $\widehat{\boldsymbol{P}} \in \mathbb{R}^{N\times D}$. Each element $\boldsymbol{G}_{ij}$ encodes the structural relationship between patch i and patch j.

where λ_{global}, λ_{patch}, and λ_{gram} are the hyperparameters controlling the relative importance of global alignment, local patch similarity, and structural consistency. Together, these losses establish a multi-level perceptual alignment that jointly preserves the semantic content and structural organization of the reconstructed image.

3.3 Gram Loss Term

The Gram loss L_{gram} complements the patch-based loss L_{patch} by focusing on structural and perceptual consistency rather than direct feature matching. While the patch loss enforces one-to-one similarity between predicted and reference patch embeddings, it ignores the spatial relationships among patches that are essential for perceptual coherence. As a result, reconstructed images may be patch-wise accurate but appear less natural and lack structural coherence.

To address this limitation, we introduce the Gram loss term that compares pairwise relationships among patch embeddings instead of their absolute spatial positions. By relying on patch correlations rather than spatial coordinates, the Gram loss promotes perceptual and structural consistency that more closely aligns with human sensitivity to texture and spatial organization. To compute the Gram loss, we first perform row-wise normalization on patch features, followed by the construction of Gram matrices.

Row-Wise Normalization. To quantify structural consistency between the predicted and reference patch tokens $P_{\text{pred}}, P_{\text{gt}} \in \mathbb{R}^{N\times D}$, we first apply row-wise ℓ_2-normalization to each patch embedding:

$$\widehat{P}_{\text{pred}}(i,:) = \frac{P_{\text{pred}}(i,:)}{\|P_{\text{pred}}(i,:)\|_2}, \quad \widehat{P}_{\text{gt}}(i,:) = \frac{P_{\text{gt}}(i,:)}{\|P_{\text{gt}}(i,:)\|_2}. \tag{4}$$

This normalization scales all patch features to unit length, ensuring similarity computations reflect cosine similarity rather than magnitude differences. Consequently, the resulting Gram matrices capture structural relationships among patches without being influenced by feature intensity.

Gram Matrix Function $f_{\text{Gram}}(\cdot)$. The Gram matrix function transforms local patch tokens into a Gram matrix, a relational representation that captures their interactions. This matrix is computed based on the cosine-similarity structure, which measures pairwise similarities between all patch embeddings. The Gram matrices are defined as:

$$G_{\text{pred}} = f_{\text{Gram}}(P_{\text{pred}}) = \widehat{P}_{\text{pred}}\,\widehat{P}_{\text{pred}}^{\top}, \quad G_{\text{gt}} = f_{\text{Gram}}(P_{\text{gt}}) = \widehat{P}_{\text{gt}}\,\widehat{P}_{\text{gt}}^{\top}. \tag{5}$$

where $G_{\text{pred}}, G_{\text{gt}} \in \mathbb{R}^{N\times N}$ are the Gram matrix computed from the predicted and ground truth patch tokens. Each element G_{ij} represents the cosine similarity between the ith and j^{th} patch features, encoding pairwise dependencies that describe the structural relationships among local regions. The Gram matrix thus reflects the global texture and structural layout of the image.

Gram Loss. The final Gram loss is defined as the difference between the predicted and ground-truth Gram matrices:

$$L_{\text{gram}} = \frac{1}{N^2}\,\|G_{\text{pred}} - G_{\text{gt}}\|_F^2. \tag{6}$$

By measuring differences in structural relationships, the Gram loss helps preserve texture, structure, and perceptual coherence in reconstructed images, making the results more visually consistent and perceptually aligned.

Table 1: Quantitative comparison of spike image reconstruction methods on Spike-REDS.

Method	Param. (M)	FLOPs (G)	PSNR↑	SSIM↑	LPIPS↓
TFP	–	–	22.37	0.5801	0.3035
TFI	–	–	24.94	0.7150	0.3716
TFSTP	–	–	22.37	0.7300	0.3254
WGSE	3.81	415.26	38.88	0.9774	0.0212
Spk2ImgNet	3.76	1000.5	38.44	0.9767	0.0229
SSIR	0.382	23.8	16.32	0.8451	0.1234
Spk2ImgMamba	0.282	9.9	15.15	0.8463	0.1248
SpikeFormer	7.58	67.88	37.18	0.9738	0.0580
SwinSF	1.8	415.82	39.34	0.9806	0.0191
SwinSF+L_{DINOv3}	1.8	415.82	**39.39**	**0.9810**	**0.0181**

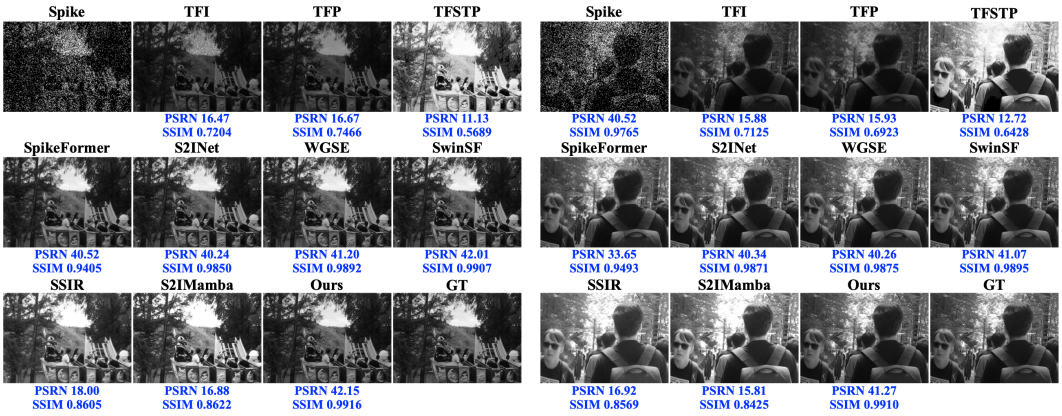


Figure 4: Quantitative comparison of different comparison methods on Spike-REDs. The PSNR and SSIM scores are shown in blue text.

4 Experiments and Results

4.1 Experimental Settings

Training Settings. For the 250×400 resolution setting, models are trained on Spike-REDs (Zhao et al. (2021)) and evaluated on Spike-REDs (Zhao et al. (2021)), PKU-Spike-High-Speed (Zhu et al. (2019)), and momVidarReal2021 (Zheng et al. (2023)). For the 1000×1000 resolution setting, models are both trained and tested on spike-X4K (Jiang et al. (2024)). spike-REDs and spike-X4K provide paired spike streams and ground-truth images, while PKU-Spike-High-Speed and momVidarReal2021 contain only spike streams and serve as challenging benchmarks for evaluating generalization under high-speed motion. The complete framework, SwinSF+L_{DINOv3}, is implemented in PyTorch. Training is conducted on a workstation equipped with an NVIDIA RTX 4090 GPU and an Intel(R) Xeon(R) Platinum 8352V CPU@2.10GHz. A batch size of 16 is used, the model begins training with a stable learning rate of 1e-4, and the StepLR scheduler halves it every 300 epochs, allowing finer updates and smooth convergence throughout all 900 training epochs.

Comparison Models and Metrics. We evaluate our approach against a variety of spike image reconstruction methods, including traditional model-based approaches (e.g., TFP (Zhu et al. (2019)), TFI (Zhu et al. (2019)), TFSTP (Zheng et al. (2021))), CNN-based frameworks (e.g., Spk2ImgNet (Zhao et al. (2021)), WGSE (Zhang et al. (2023))), SNN-based models (e.g., SSIR (Zhao et al. (2023))), Transformer-based approaches (e.g., SpikeFormer (She & Qing (2022)), SwinSF (Jiang et al. (2024))), and the recent Mamba-based model (e.g., Spk2ImgMamba (Yin et al. (2025))). For the comparison with SwinSF, we directly use the official released model weights at the official website [1], reports its results on spike-REDs, spike-X4K, PKU-High Speed, and momVidarReal2021, without retraining it. For the paired spike-REDs dataset, we evaluate reconstruction quality using three standard full-reference metrics: PSNR, SSIM, and LPIPS (Zhang et al. (2018)). For the unpaired PKU-Spike-High-Speed and momVidarReal2021 datasets, we report the no-reference metric NIQE. Additionally, we provide a comprehensive comparison of model efficiency by reporting model size (in millions of parameters) and computational cost (in GFLOPs).

[1] https://github.com/bupt-ai-cz/SwinSF

Table 2: Quantitative performance comparison on the Spike-X4K dataset.

Method	TFI	TFP	SpikeFormer	Spk2ImgNet	WGSE	SwinSF	SwinSF+L_{DINOv3}
PSNR (↑)	15.84	15.87	36.84	37.95	38.19	39.61	**40.58**
SSIM (↑)	0.6420	0.7064	0.9474	0.9519	0.9524	0.9682	**0.9695**

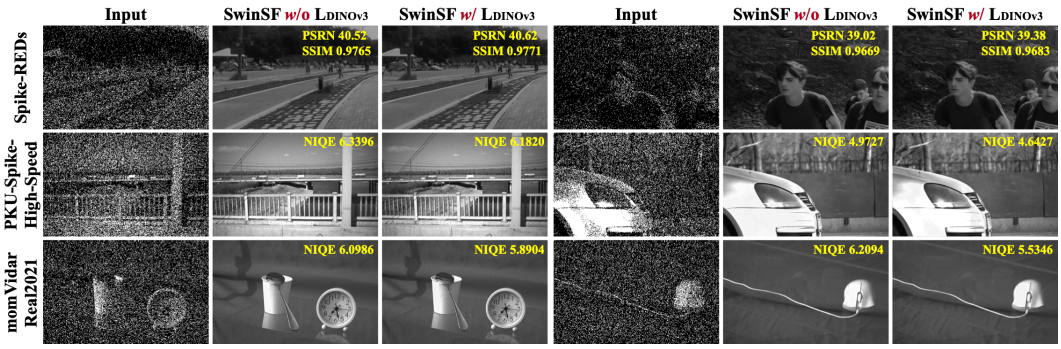


Figure 5: Visual comparison of the SwinSF model with and without L_{DINOv3}.

4.2 Experiments

Comparison with SOAT. Table 1 reports the quantitative comparison of spike image reconstruction methods on the spike-REDS dataset. Incorporating the proposed DINOv3 perceptual loss consistently improves SwinSF across all full-reference metrics (PSNR, SSIM, and LPIPS), establishing new state-of-the-art results. The gains in SSIM reflects superior preservation of spatial structure and perceptual similarity, while the reduction in LPIPS indicates closer alignment with human visual perception and enhanced naturalness of reconstructed images. Together, these results confirm that the proposed perceptual loss effectively guides the network to capture both global structures and fine local textures, leading to reconstructions that are perceptually realistic and structurally consistent. In contrast, the traditional methods TFP, TFI, and TFSTP perform poorly due to their limited ability to model complex temporal dependencies and spatial correlations in spike data. It is also noteworthy that the Mamba-based Spk2ImgMamba and SNN-based SSIR models show degraded performance on Spike-REDS, mainly producing overexposed images with unrealistic illumination and distorted contrast, as illustrated in Figure. 4. Table 2 presents the quantitative comparison of various spike image reconstruction methods on the Spike-X4K dataset. With the addition of the proposed DINOv3 perceptual loss, SwinSF+L_{DINOv3} achieves new state-of-the-art performance. All these results demonstrate the effectiveness of the proposed DINOv3 perceptual loss.

Generalization across Architectures. To investigate the generalization capability of the proposed DINOv3 perceptual loss across heterogeneous model architectures, we conduct a comprehensive evaluation on CNN-, Mamba-, and Transformer-based spike-to-image reconstruction frameworks, including Spk2ImgNet (S2I), Spk2ImgMamba (S2IMamba), and SwinSF, as summarized in Table 3. Across all three architectures, incorporating $\mathcal{L}_{\text{DINOv3}}$ consistently improves reconstruction performance in terms of PSNR, SSIM, and LPIPS. These results demonstrate that DINOv3 effectively complements pixel-wise objectives by enhancing structural fidelity and perceptual consistency across different backbone designs.

DINOv3 vs. VGG Perceptual Loss. To evaluate the effectiveness of the proposed DINOv3 perceptual loss, we compare three variants of the SwinSF model: one trained with the DINOv3 perceptual loss, one trained with the conventional VGG-based perceptual loss (Johnson et al. (2016)), and a baseline model trained without any perceptual loss. As shown in Figure. 1, the SwinSF model using L_{DINOv3} consistently outperforms both the baseline and the L_{VGG} variant across all benchmark datasets. In particular, the SwinSF+L_{DINOv3} framework achieves lower NIQE scores, reflecting superior perceptual quality and more natural-looking reconstructions (see Figure. 1). Figure. 5 shows a visual comparison of the SwinSF model with and without L_{DINOv3}. The SwinSF model with L_{DINOv3} performs better than the one without, as the perceptual supervision introduced by L_{DINOv3} enables the network to capture both global semantics and local structural details that are often overlooked

Table 3: Generalization of DINOv3 perceptual loss across different architectures on Spike-REDS.

S2I				S2IMamba				SwinSF			
Setting	PSNR↑	SSIM↑	LPIPS↓	Setting	PSNR↑	SSIM↑	LPIPS↓	Setting	PSNR↑	SSIM↑	LPIPS↓
Baseline	38.44	0.9767	0.0229	Baseline	15.15	0.8463	0.1248	Baseline	39.34	0.9806	0.0191
$+L_{\text{DINOv3}}$	38.72	0.9780	0.0216	$+L_{\text{DINOv3}}$	17.60	0.8982	0.1065	$+L_{\text{DINOv3}}$	39.39	0.9810	0.0181

Table 4: Ablation study of different loss configurations on PKU-Spike-High-Speed and momVidar-Real2021 using NIQE. Here, L_g, L_p, and L_r denote the global, patch, and Gram losses, respectively.

Loss Config.	L_g	L_p	L_r	$L_g + L_p$	$L_p + L_r$	$L_g + L_r$	$L_g + L_p + L_r$
PKU-Spike-High-Speed	5.5537	5.5356	5.5237	5.4968	5.4193	5.4369	**5.3185**
momVidarReal2021	8.6390	8.3580	8.2090	8.1027	7.9390	7.9566	**7.7023**

by pixel-level losses. Compared to L_{VGG}, L_{DINOv3} performs better because (1) DINOv3 provides stronger and more general representations, while VGG mainly focuses on local patterns, and DINOv3 captures both global and local information through self-attention, and (2) it includes a Gram loss that models relationships between patches, preserving texture and structural consistency. Together, these factors lead to reconstructions that appear more natural and perceptually realistic.

Impact of L_{global}, L_{patch}, and L_{gram}. To better understand the contribution of each component in DINOv3 perceptual loss, we perform an ablation study on its three subterms: the global semantic alignment loss (L_{global}), the local patch similarity loss (L_{patch}), and the Gram-based structural consistency loss (L_{gram}). As summarized in Table 4, among the individual components, L_{gram} achieves the best performance. By modeling pairwise correlations between patches, L_{gram} helps the reconstructed images maintain realistic structural consistency. When two components are combined, performance further improves due to the complementary nature of global semantic alignment and local detail preservation. Between the two, L_{patch} contributes slightly more than L_{global}, as it enforces similarity between local patch features, allowing the model to better preserve fine-grained textures and spatial details that are often lost with global-only supervision. The best results are obtained when all three components are applied jointly, yielding the lowest NIQE scores of 5.3185 and 7.7023 on the PKU-Spike-High-Speed and momVidarReal2021 datasets, respectively. These findings demonstrate that integrating global, local, and structural constraints enables the network to produce reconstructions that are perceptually better, more detailed, and visually natural across diverse scenes.

Impact of λ_{global}, λ_{patch}, and λ_{gram} To analyze the sensitivity of the proposed DINOv3 perceptual loss to its weighting parameters, we perform an ablation study on the three hyperparameters λ_{global}, λ_{patch}, and λ_{gram}, which control the relative contributions of the global, patch, and Gram-based loss terms, respectively. As illustrated in Figure. 6, the performance varies moderately across different weighting combinations, indicating that the proposed perceptual loss remains stable over a broad range of parameter settings. With fixed λ_{global} and λ_{patch}, increasing λ_{gram} from 0.5 to 2.0 consistently improves performance on both datasets, as the Gram-based term enhances fine structural consistency. Beyond $\lambda_{\text{gram}} = 2.0$, the improvement saturates, suggesting that an excessively strong Gram constraint may dominate the optimization and limit global coherence. Similarly, with fixed λ_{global} and λ_{gram}, increasing λ_{patch} from 0.5 to 2.0 leads to gradual improvement, as stronger local supervision helps retain detailed spatial information and texture consistency. However, overly large λ_{patch} values provide diminishing returns, indicating a trade-off between local detail preservation and global smoothness. In contrast, λ_{global} shows a weaker influence on perceptual quality compared to the other two components. A smaller value ($\lambda_{\text{global}} = 0.5$) slightly improves NIQE scores by balancing global semantic alignment with local feature fidelity. Overall, the best performance is achieved with $\lambda_{\text{global}} = 1.0$, $\lambda_{\text{patch}} = 1.0$, and $\lambda_{\text{gram}} = 2.0$, yielding the lowest NIQE scores on two datasets. These results demonstrate that properly balancing the three perceptual terms allows the model to achieve an optimal compromise between global structure preservation and local detail accuracy.

Impact of Image Resolution.

The resolution of the input image to the DINOv3 model influences the results, as higher resolutions provide richer global and patch-level representations. To examine this effect, we evaluate the

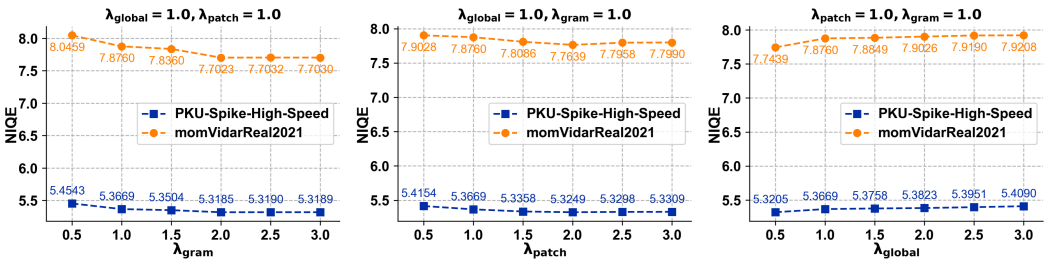
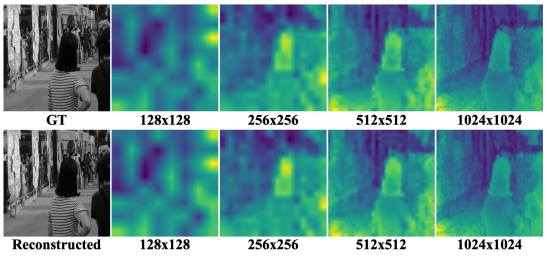


Figure 6: Impact of hyperparameter λ_{global}, λ_{patch}, and λ_{gram}.

Size	PKU		momVidar	
	L_{gram}	L_{DINOv3}	L_{gram}	L_{DINOv3}
128×128	5.5530	5.4692	8.7980	8.3956
256×256	5.5362	5.3508	8.3059	7.9036
512×512	5.5237	**5.3185**	8.2090	**7.7023**
1024×1024	**5.5229**	5.3187	**8.2058**	7.7028

(a) Quantitative comparison



(b) DINOv3 patch-token visualization

Figure 7: Effect of input resolution. (a) Quantitative comparison on PKU-Spike-High-Speed and momVidarReal2021. (b) DINOv3 patch-token heatmaps of ground-truth (top) and reconstructed images (bottom) at different resolutions.

reconstruction performance of SwinSF using L_{gram} and L_{DINOv3} across different input image sizes, ranging from 128×128 to 1024×1024. The results are summarized in Figure. 7.

Specifically, as the resolution increases from 128×128 to 512×512, the NIQE score decreases steadily, indicating that higher-resolution inputs allow DINOv3 to extract richer feature representations, leading to perceptually sharper and more natural reconstructions. Beyond 512×512, the performance saturates, suggesting that the perceptual benefit stabilizes once sufficient spatial information is available for global-context modeling. Figure. 7 presents the heatmap generated from the patch token of image with different resolution. Large resolution show clearer object boundaries and details, but also consume more memory consumption, we set the resolution as 512×512 to achieve a balance between performance and memory consumption.

5 Conclusion

In this work, we propose a DINOv3-based perceptual loss for spike image reconstruction to overcome the limitations of pixel-wise and CNN-based perceptual objectives. By leveraging the large vision foundation model DINOv3, the proposed perceptual loss jointly captures global semantics and local structures within a unified framework. A Gram-based term further improves structural consistency, resulting in sharper and more natural reconstructions. Extensive experiments show that integrating the proposed loss into SwinSF consistently improves performance across datasets and evaluation metrics. Beyond spike image reconstruction, the proposed DINOv3 perceptual loss is inherently model-agnostic and task-independent, making it readily applicable to a broad range of low-level vision tasks. It can be naturally extended to problems such as image super-resolution, denoising, and deblurring, where its ability to capture both global semantics and fine-grained local details can enhance perceptual quality, texture consistency, and structural realism. Furthermore, its strong representation capability facilitates robust generalization across diverse data domains and sensing modalities, supporting cross-domain reconstruction and real-world applications. Exploring its adaptation to different reconstruction settings may further reveal its generalization potential.

AI USE STATEMENT

Generative AI tools were used only for English grammar correction and language polishing. No generative AI tools were used for research design, methodology development, experiments, data analysis, code development, result interpretation, or scientific claims. All AI-assisted edits 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

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Guanglu Dong, Xiangyu Liao, Mingyang Li, Guihuan Guo, and Chao Ren. Exploring semantic feature discrimination for perceptual image super-resolution and opinion-unaware no-reference image quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28176–28187, 2025.

Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 25:110–119, 2023.

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

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Shanchuan Lin and Xiao Yang. Diffusion model with perceptual loss. *arXiv preprint arXiv:2401.00110*, 2023.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1680–1691, 2023.

Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5419–5427, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gustav Grund Pihlgren, Konstantina Nikolaidou, Prakash Chandra Chhipa, Nosheen Abid, Rajkumar Saini, Fredrik Sandin, and Marcus Liwicki. A systematic performance analysis of deep perceptual loss networks: Breaking transfer learning conventions. *arXiv preprint arXiv:2302.04032*, 2023.

Chen She and Laiyun Qing. Spikeformer: Image reconstruction from the sequence of spike camera based on transformer. In *Proceedings of the 2022 5th International Conference on Image and Graphics Processing*, pp. 72–78, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers. *arXiv preprint arXiv:2401.00463*, 2023.

Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 139–142. IEEE, 2022.

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

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*, volume 37, 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*, 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*, pp. 11996–12005, 2021.

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*, 34(6):5207–5212, 2023.

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*, 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 *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1432–1437. IEEE, 2019.

Lin Zhu, Yunlong Zheng, Mengyue Geng, Lizhi Wang, and Hua Huang. Recurrent spike-based image restoration under general illumination. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8251–8260, 2023.