

Geo-Reflex: Geometry-Gated Spiking Visuomotor Policy Learning for High-Frequency Robot Manipulation

Xuhong Fang

Dongguan City University

Chongwei Huang

Dongguan City University

Weiming Zhang

zhangweiming@dgcu.edu.cn

Dongguan City University

Research Article

Keywords: Spiking neural networks, Visuomotor policy learning, Robot manipulation, Asynchronous sensor fusion, Geometric perception, Imitation learning

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9867179/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Geo-Reflex: Geometry-Gated Spiking Visuomotor Policy Learning for High-Frequency Robot Manipulation

Xuhong Fang¹, Chongwei Huang¹, Weiming Zhang^{1*}

¹School of Artificial Intelligence, Dongguan City University, Dongguan, 523428, Guangdong, China.

*Corresponding author(s). E-mail(s): zhangweiming@dgcu.edu.cn;

Abstract

Fast robot manipulation requires high-frequency proprioceptive feedback while visual streams are updated at much lower rates. Existing synchronous policies resample all modalities into a fixed temporal grid, discarding high-rate joint dynamics. We propose Geo-Reflex, a geometry-gated spiking visuomotor policy that uses camera-derived geometric gradients to modulate the firing thresholds of Geometry-Gated LIF (Geo-LIF) neurons. Instead of concatenating visual features with joint states, the low-rate visual stream sets a threshold field while an eight-step proprioceptive window drives membrane current. A reflex residual path captures smooth joint-command structure, and the spiking branch learns geometry-conditioned corrections. On a public OpenLoong HDF5 robot-trajectory subset, Geo-Reflex achieves 0.0235 ± 0.00010 MAE over seven seeds, improving over a frame-synced Transformer by 2.1%. It reduces firing rate by 35% relative to a fixed-threshold SNN and concentrates activity in high-gradient regions by a factor of 10.9. Under texture noise, Geo-Reflex degrades half as much as the vanilla SNN, and chained multi-step prediction shows slower error accumulation than all baselines. These results support geometry-gated spiking thresholds as a practical fusion mechanism for high-frequency robot manipulation.

Keywords: Spiking neural networks; Visuomotor policy learning; Robot manipulation; Asynchronous sensor fusion; Geometric perception; Imitation learning

1 Introduction

Robot policies for grasping, tracking, and contact-rich manipulation operate under tight timing constraints that expose a fundamental mismatch in sensor modalities. Motor commands update above 100 Hz, while commodity RGB-D cameras deliver visual frames at 10 to 30 Hz. Large-scale visuomotor learning has substantially advanced frame-level policies [1–3], and recent vision-language-action (VLA) models and diffusion-policy systems have improved generalization by scaling both data and model capacity [4–6]. These approaches operate on synchronously resampled sensor tensors and are not designed to exploit the 500 Hz proprioceptive micro-structure that emerges at contact. When every inference step must wait for the slowest modality, the controller either lags behind the physical dynamics or discards the high-rate information that is most predictive of imminent contact events.

Spiking neural networks (SNNs) are an attractive alternative for this regime because their event-driven dynamics decouple computation from a fixed clock and can run on low-power neuromorphic hardware [7, 8]. Surrogate-gradient methods have made SNN training practical for deep architectures [9–11], and adaptive threshold mechanisms have improved temporal sequence modeling [12, 13]. A key open question is not just how to train networks that emit spikes, but what information should be permitted to trigger them. Dense image features can drive widespread firing from texture, lighting variations, and compression artifacts. These cues are largely irrelevant for manipulation, while the sparse geometric structure at object boundaries and depth discontinuities carries the information that actually guides the hand. When the visual stream is naively concatenated with proprioceptive signals and fed into a shared spiking layer, the resulting feature drowning degrades both the signal-to-noise ratio of the task-relevant activity and the energy efficiency that makes SNNs appealing in the first place.

Geo-Reflex addresses this problem by changing the role of geometry in the spiking dynamics. Rather than fusing vision and proprioception into a single synchronous feature vector, the architecture separates the two modalities by function: the low-rate visual stream sets a threshold state, while the high-rate proprioceptive stream drives the membrane current. Camera-derived geometric gradients are extracted via lightweight Sobel, Laplacian, and Canny operators, then projected onto per-neuron threshold offsets through a centered tanh transformation. Near geometric edges, the effective firing threshold is lowered, making the policy sensitive to small joint changes in regions where contact-relevant structure exists. On flat background, the threshold stays high, suppressing spurious firing from texture and noise. The resulting Geometry-Gated LIF (Geo-LIF) neuron is updated at the camera frame rate and held constant over the proprioceptive window. This lets the motor loop continue at its native frequency without waiting for the next visual observation. A direct reflex residual path from the flattened proprioceptive window to the action readout further separates the smooth, predictable component of joint commands from the edge-conditioned spiking correction.

The central hypothesis is that spatial geometry is useful not just as an additional input feature, but as a sensitivity gate for high-frequency visuomotor policy learning. We evaluate this hypothesis through four contributions: (i) the Geo-LIF neuron, a LIF variant whose threshold is spatially modulated by visual geometry, turning depth from a passive feature into an active controller of neuronal excitability; (ii) an asynchronous

reflex architecture in which a slowly updated visual threshold state gates a high-frequency proprioceptive drive, with a direct residual path that preserves smooth command structure; (iii) an evaluation on a public OpenLoong HDF5 robot-trajectory subset showing that Geo-Reflex achieves the lowest action-prediction error among all baselines, demonstrated with seven independent seeds under shared trajectory-level splits, supported by a controlled comparison under the same hidden width and training budget and a reflex-residual ablation that isolates the contribution of the gating mechanism; and (iv) robustness, ablation, and chained multi-step prediction analyses that characterise the geometry gating mechanism under visual noise, architectural variations, and closed-loop-like conditions.

Because the available dataset consists of offline demonstrations rather than a resettable simulator or online robot execution interface, this work focuses on high-frequency visuomotor action prediction as a behavioral-cloning proxy for manipulation policy learning. We therefore interpret the results as evidence for the proposed fusion mechanism rather than as a direct measurement of physical grasping success.

Section 2 reviews asynchronous sensor fusion, geometric perception, and adaptive SNN methods. Section 3 formalizes the Geo-LIF dynamics and the complete architecture. Section 4 describes the experimental setup, baselines, main results, spiking sparsity, noise robustness, ablations, and chained prediction analysis. Section 5 interprets the findings and their limitations, and Section 6 concludes with directions for future work.

2 Related Work

2.1 Asynchronous Sensor Fusion in Robotics

Robotic systems integrate data from encoders, inertial sensors, depth cameras, and event-based sensors, each operating at a different natural rate. Event-based vision has demonstrated the advantages of asynchronous sensing for high-speed tracking, localization, and visual perception [14–17], while classical filtering theory has long treated intermittent observations as a standard component of robotic state estimation [18]. In the context of neural control, approaches such as hierarchical time-surface representations [14] and event-driven spiking grasping [19] have explored asynchronous visual processing, but these methods predominantly treat all incoming events as equally informative or rely on luminance-driven cues that are sensitive to texture and illumination variations. Most learned visuomotor pipelines, by contrast, continue to resample all modalities into a frame-synchronous tensor before inference, discarding the temporal structure that distinguishes fast and slow signals. Geo-Reflex takes a different approach: rather than synchronizing all streams to a single clock, it uses the natural asynchrony as a structural design principle. The camera-derived geometry updates a threshold gate at the visual frame rate, while the policy continuously consumes high-rate arm state information. This design is conceptually related to the separation of "what" and "when" pathways in biological visuomotor control, where dorsal-stream geometry information modulates ventral-stream recognition without requiring frame-level alignment.

2.2 Geometry Representation in Robot Perception

Geometric representations are a well-established prior in robot perception, from occupancy-grid models [20] to point-cloud architectures such as PointNet [21]. Depth images are widely used for grasping [22], pose estimation [23], and affordance learning [24, 25], typically as additional input channels to convolutional or transformer-based policies. Edge-filtered inputs have been shown to improve grasp generalization [26] by removing photometric variation that distracts from object structure. Geo-Reflex shares the intuition that edges encode the most task-relevant geometric information, but changes the point of injection: instead of feeding edge maps as passive input features alongside RGB and depth, the network embeds them directly into the spiking dynamics as threshold modulators. This distinction carries two practical advantages. First, the gating mechanism operates at the neuronal level rather than the feature level, so it can suppress or amplify individual neural responses without increasing the dimensionality of the input representation. Second, because the gate outputs a continuous modulation signal rather than a binary edge map, the network can learn task-specific sensitivity patterns through the learned projection W_g .

2.3 Adaptive Spiking Neural Networks

Adaptive mechanisms have been central to SNN research for improving temporal processing and energy efficiency. Long short-term memory spiking networks (LSNNs) introduced spike-history-dependent thresholds that decay with each spike, enabling rich temporal dynamics for sequence modeling [12]. Subsequent work has explored learnable membrane time constants [13], neural heterogeneity [27], surrogate-gradient training variants [11, 28], and deep residual SNN architectures [29]. Spikformer and related attention-based SNNs have demonstrated that self-attention mechanisms can be adapted to spiking dynamics [30]. These methods share a common strategy: the adaptation emerges from learned parameters (thresholds, time constants, or attention weights) updated through gradient descent. Geo-Reflex differs in that the modulation originates from an external sensory signal, the spatial gradient of the depth image, rather than from internal neuronal state alone. This design choice injects an explicit geometric prior into the spiking dynamics, coupling the network’s excitability to the physical structure of the scene. The approach is complementary to learned adaptation: the geometric gate provides a task-relevant bias at initialization and during inference, while learned plasticity can further refine the mapping from geometry to threshold modulation through training.

3 Methodology

3.1 Preliminaries: Leaky Integrate-and-Fire Dynamics

Each neuron in Geo-Reflex follows a leaky integrate-and-fire (LIF) update. Let $v_i^{(t)}$ denote the membrane potential of neuron i at discrete spiking step t . The subthreshold dynamics are:

$$v_i^{(t+1)} = \beta v_i^{(t)} + \sum_j w_{ij} s_j^{(t)} + b_i,$$

where $\beta \in (0, 1)$ is the leak factor, w_{ij} are synaptic weights, $s_j^{(t)} \in \{0, 1\}$ is the binary spike output of pre-synaptic neuron j , and b_i is a bias term. When the membrane potential exceeds a firing threshold θ_i , the neuron emits a spike:

$$s_i^{(t)} = \mathbb{K} \left(v_i^{(t)} \geq \theta_i \right), \quad v_i^{(t)} \leftarrow v_{\text{reset}} \text{ if spike},$$

where v_{reset} is implemented as a spike-dependent membrane reset. In the vanilla SNN baselines, θ_i is fixed at 1.0 for all neurons. Geo-Reflex modifies only the first-layer threshold while keeping the remaining LIF layers fixed, isolating the contribution of the geometry gate.

3.2 Geometry-Gated Threshold Modulation

Geo-Reflex replaces the static first-layer threshold with a geometry-dependent value computed from the depth or RGB camera stream. The gate is intentionally simple and computationally lightweight, reflecting the practical constraints of real-time deployment on robot hardware. At each visual update, the image $D \in \mathbb{R}^{H \times W}$ is resized to 16×16 , normalized to unit range, and converted into a multi-operator geometry map. The main model stacks three gradient operators. For the Sobel channel, the spatial gradient magnitude is:

$$G_{\text{sobel}} = \sqrt{(D * G_x)^2 + (D * G_y)^2} \in \mathbb{R}^{H \times W},$$

where G_x and G_y are 3×3 Sobel kernels. The Laplacian channel uses $\nabla^2 D$ via a 3×3 discrete Laplacian kernel, and the Canny channel applies double-threshold edge detection with thresholds of 60 and 160. Each channel is normalized independently, and the three maps are stacked into a single feature vector $\mathcal{G} \in \mathbb{R}^{3HW}$.

The flattened geometry field is projected to the hidden dimension and centered within each sample to produce the threshold offset:

$$\Delta V_{th,i}^{\text{vision}} = -\alpha \tanh \left((W_g \mathcal{G})_i - \frac{1}{N} \sum_{k=1}^N (W_g \mathcal{G})_k \right),$$

where $W_g \in \mathbb{R}^{N \times 3HW}$ is a learned linear projection, N is the hidden width, and $\alpha = 0.7$ controls the modulation strength. The effective threshold for neuron i is:

$$\theta_i^{(t)} = \text{clip} \left(V_{\text{base}} + \Delta V_{th,i}^{\text{vision}}, 0.45, 1.65 \right),$$

with $V_{\text{base}} = 1.0$. The centering operation is critical: it prevents the gate from uniformly lowering or raising all thresholds within a sample, which would simply change the global firing rate without redistributing sensitivity. Instead, neurons receiving high-gradient evidence have their thresholds lowered relative to the mean, making them easier to fire, while those receiving low-gradient evidence are suppressed. The modulation is held constant over the short proprioceptive window used for each policy

evaluation, so the visual gate operates at camera frame rate while the spiking dynamics continue at the proprioceptive rate.

Design rationale. We use Sobel, Laplacian, and Canny operators because they are deterministic, require no training data or labels, and introduce negligible compute overhead relative to neural network forward passes. The ablation in Section 4.4 examines the contribution of each operator individually. Notably, the stacked multi-operator gate provides the strongest final model, suggesting that different gradient operators capture complementary edge information. The multi-operator geometry representation provides an effective and computationally efficient test of the gating hypothesis, requiring no additional training data or labels.

3.3 High-Frequency Proprioceptive Encoder

The proprioceptive stream is the fast channel in the architecture. For each sample, we use a window of eight consecutive arm states, each containing 14 joint positions and 14 joint velocities (28 dimensions per step). This short temporal window serves as a proxy for the high-frequency joint micro-dynamics that would be available in an online controller sampling at 500 Hz. A learned linear projection maps each 28-dimensional state at each timestep to a hidden current $I_{\text{joint}}[t] \in \mathbb{R}^N$. The proprioceptive encoder processes the sequence timestep by timestep through the Geo-LIF layer, accumulating geometry-conditioned spiking activity over the eight-step window.

A distinguishing feature of the model is the direct reflex residual path, which connects the flattened proprioceptive window (all 8×28 dimensions) to the action readout through a two-layer MLP with GELU activation and LayerNorm. This path captures the smooth, predictable component of joint-to-action mapping that does not require geometric context. The spiking branch then learns a geometry-conditioned residual correction. This decomposition is motivated by the observation that much of the joint command in a reaching or manipulation trajectory is a smooth function of recent state history, while the corrections that benefit from visual geometry are sparse and localized in time.

3.4 Asynchronous Geometry-Gated Fusion

The Geo-LIF update follows the standard LIF formulation with a geometry-dependent threshold. The membrane potential evolves as:

$$v_i[t+1] = \beta v_i[t] + I_{\text{joint},i}[t],$$

and the geometry-gated threshold is:

$$\theta_i = \text{clip} \left(V_{\text{base}} + \Delta V_{th,i}^{\text{vision}}, 0.45, 1.65 \right),$$

where $\Delta V_{th,i}^{\text{vision}}$ is computed from the spatial gradient map (Section 3.2). A spike is emitted when the membrane potential exceeds the threshold:

$$s_i[t+1] = \mathbb{K}(v_i[t+1] \geq \theta_i), \quad v_i[t+1] \leftarrow v_{\text{reset}} \text{ if spike.}$$

The visual gate is updated at camera rate. During the interval before the next visual frame, $\Delta V_{th}^{\text{vision}}$ is held constant while the proprioceptive sequence continues to

drive the membrane. The architecture thus separates the “where” (which neurons are sensitive) from the “when” (how the spike sequence evolves over the proprioceptive window). In physical terms, the visual stream controls the gate height and the joint stream supplies the flow: near geometric edges, small joint changes can trigger spikes; on flat background, the same changes are less likely to pass threshold.

The complete Geo-Reflex pipeline comprises three functional blocks:

1. **Geometry processor.** Lightweight gradient operators (Sobel, Laplacian, Canny) compute a multi-channel spatial geometry map from the resized and normalized camera frame. The stack is projected to hidden dimension via W_g and centered.
2. **Proprioceptive encoder.** A window of arm joint positions and velocities is projected into membrane current through a learned linear layer, applied independently at each timestep.
3. **Geo-LIF policy.** The first LIF layer uses geometry-gated thresholds as described above, processing the eight-step proprioceptive sequence. The second LIF layer uses a fixed threshold. The final action is the sum of a direct reflex residual (from the flattened window) and a readout of the second-layer spiking activity.

Figure 1 provides a conceptual view of the complete system, emphasizing the functional separation between visual geometry, proprioceptive micro-dynamics, spiking correction, and continuous action output.

3.5 Training Objective and Procedure

We train Geo-Reflex end-to-end using the fast-sigmoid surrogate gradient [11]. The loss combines supervised action prediction with regularization terms that encourage the desired gating behavior:

$$\mathcal{L} = \mathcal{L}_{\text{SmoothL1}}(\hat{a}_t, a_t) + \lambda_f \bar{f} + \lambda_g \bar{f}_{\text{flat}}^2,$$

where a_t is the demonstrated arm action, $\hat{a}_t$ is the model prediction, $\bar{f}$ is the average firing rate across all spiking neurons, and $\bar{f}_{\text{flat}}$ is the average firing rate of neurons in the lowest quartile of gate activation. This term encourages the network to remain quiet for units that receive weak geometry-driven modulation. The firing-rate penalty $\lambda_f \bar{f}$ encourages overall sparsity, while $\lambda_g \bar{f}_{\text{flat}}^2$ reinforces the intended gating behavior by suppressing activity in low-gradient regions.

Hyperparameters. The model configuration is reported in Table 1. Geo-Reflex uses hidden width 512, 64 training epochs, learning rate 7×10^{-4} , $\alpha = 0.7$, $\lambda_f = 0.001$, and $\lambda_g = 5 \times 10^{-4}$.

Computational cost. Reported latency is measured as GPU forward-pass time on pre-decoded features, isolating controller inference from camera decoding, file I/O, and robot communication. Because we do not measure power on neuromorphic hardware, we report firing rate as an energy-relevant sparsity proxy rather than a direct energy measurement.

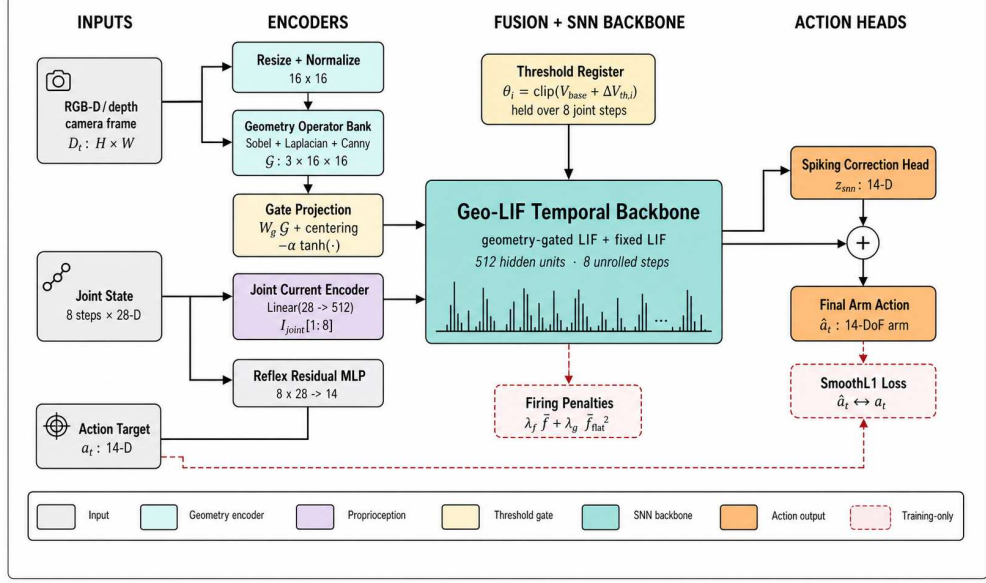


Fig. 1 Overall Geo-Reflex architecture. The model receives a low-rate RGB-D/depth frame, an eight-step proprioceptive window, and a demonstrated action target during supervised training. The visual stream is converted by deterministic Sobel, Laplacian, and Canny operators into a compact geometry map, then projected into a centered threshold field that is held across the proprioceptive window. The joint stream drives a geometry-gated Geo-LIF layer and a second fixed-threshold LIF layer, while a direct reflex residual bypass preserves smooth command structure. The final 14-DoF arm action combines the sparse spiking correction with the residual path; firing-rate penalties and SmoothL1 loss are used only during training.

Table 1 Hyperparameter settings for the main comparison.

Parameter	RGB MLP	LSTM	RGB-D SNN	Depth SNN	Transformer	Geo-Reflex
Hidden width	512	512	512	512	512	512
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	7×10^{-4}
Batch size	256	256	256	256	256	256
Training epochs	64	64	64	64	64	64
Window length	8	8	8	8	8	8
Threshold modulation	None	None	None	None	None	Geometry gate

4 Experiments

4.1 Experimental Setup

We evaluate all methods on a public HDF5 robot-trajectory subset from the OpenLoong dataset portal [31]. The selected subset contains 146 HDF5 trajectory files and approximately 9.1 GB of data. The trajectories were collected on a KUAVO 4PRO humanoid manipulation platform with a dexterous hand and an Intel RealSense D435 head camera. Each trajectory contains RGB-D camera streams, robot metadata, and time-stamped joint state/action traces in a standard HDF5-based offline imitation-learning format.

Observation space. The model receives:

- **Proprioception:** an eight-step window of 14 arm joint positions and 14 arm joint velocities from the trajectory state stream (totaling 8×28 dimensions).
- **Visual input:** compressed head-camera depth frames when available, falling back to head RGB frames when depth is missing. Frames are JPEG-decoded, resized to 16×16 , normalized to $[0, 1]$, and converted to stacked Sobel, Laplacian, and Canny gradient maps.
- **Target action:** the 14-dimensional arm joint position command from the trajectory action stream, aligned by trajectory index.

Task definition. The model receives synchronized proprioceptive and visual observations and must predict the demonstrated arm action. This one-step action prediction task serves as a standard behavioral-cloning metric for visuomotor action-prediction quality, widely used in offline policy learning.

Training protocol. All 146 HDF5 files were used. For each seed, the trajectories were split 80/20 into training and test sets at the trajectory level. From each trajectory, 96 state/action pairs were sampled. The split is re-sampled independently for each seed. For each seed, all compared methods used the same trajectory-level train/test split and the same sampled state/action pairs. All models are trained on an RTX 3090 GPU.

Metrics definition.

- **MAE:** mean absolute error in original action units. Lower is better.
- **Latency:** mean GPU inference time per sample in milliseconds, measured on pre-decoded tensors.
- **Firing rate:** average spikes per neuron per spiking step for SNN models. Lower indicates sparser computation.
- **Edge firing ratio:** computed by ranking neurons according to their geometry-gated threshold offsets. The ratio divides the average firing activity of neurons in the top quartile of gate activation by that of neurons in the bottom quartile. Higher values indicate that spiking activity is concentrated in geometry-sensitive neurons.

Experiment status transparency. Table 2 summarizes the status of each experimental analysis in this paper.

Table 2 Verification status of experimental analyses.

Analysis	Seeds	Status	Used in claims
Main comparison	7	Completed	Core results
Chained multi-step prediction	7	Completed	Core results
Texture-noise stress test	7	Completed	Core results
Gradient-operator ablation	3	Diagnostic ablation	Supplementary diagnostic analysis
Reflex-residual ablation	7	Completed	Core results
LSTM baseline	7	Completed	Core results

4.2 Baselines

We compare Geo-Reflex against five methods:

- **RGB-only MLP:** a dense two-layer policy that receives the flattened proprioceptive window and the RGB/gray visual feature. No depth information and no spiking dynamics. Establishes the performance floor for a purely proprioception-plus-texture approach.
- **LSTM:** a two-layer LSTM with 512 hidden units that receives the flattened proprioceptive window and RGB-D visual features, with a linear readout. Provides a non-spiking recurrent baseline for temporal action prediction.
- **Vanilla SNN RGB-D concat:** a fixed-threshold LIF network where RGB and depth-gradient features are concatenated with each timestep of the proprioceptive window as ordinary input channels. This baseline is architecturally equivalent to Geo-Reflex except for the gating mechanism.
- **Depth-concat SNN:** a fixed-threshold LIF network that receives the depth-gradient feature as a passive input but does not use it for threshold modulation. This baseline directly tests whether the improvement comes from the gating mechanism or simply from having gradient features in the input.
- **Frame-synced Transformer:** a two-layer Transformer encoder over the proprioceptive sequence plus one visual token obtained by projecting the concatenated RGB and depth features. This baseline represents a standard synchronous sequence-modeling approach.

The SNN concatenation baselines isolate the central question: should geometry enter the network as a feature or as a gate? The LSTM and Transformer baselines provide non-spiking temporal model comparisons.

Implementation details. All SNN baselines use the same LIF dynamics as Geo-Reflex (Section 3.1) with a fixed threshold of 1.0, fast-sigmoid surrogate gradient [11], and membrane reset after spike. The Frame-synced Transformer uses a two-layer encoder with 4 attention heads, feed-forward dimension 1024, sinusoidal positional encoding, dropout 0.1, and a learned linear readout from the encoder output. Baselines use the same learning rate scheduling as Geo-Reflex unless specified otherwise in Table 1. Latency measurements are GPU inference-only on an RTX 3090 with batch size 1, averaged over 1,000 samples after warm-up, excluding gradient computation and file I/O.

4.3 Main Results

We evaluate all methods under the same hidden width and training budget. All models are trained with hidden width 512 for 64 epochs (Table 3), with architecture-specific components kept unchanged.

Under this controlled setting, Geo-Reflex achieves the lowest MAE among all methods (Table 3). The frame-synced Transformer is the strongest non-spiking baseline (0.0240 MAE), while the fixed-threshold SNN baselines remain substantially worse despite receiving the same visual and proprioceptive information. The LSTM baseline

Table 3 Performance comparison under the controlled main setting. All models use hidden width 512 and are trained for 64 epochs. Values are mean $\pm$ s.d. over 7 seeds. The final row reports a no-residual ablation used to quantify the contribution of the direct proprioceptive path.

Method	MAE	Firing Rate	Edge Fire Ratio	Latency (ms)
RGB-only MLP	0.0270 $\pm$ 0.0003	N/A	N/A	0.005
LSTM	0.0251 $\pm$ 0.0003	N/A	N/A	0.008
Vanilla SNN RGB-D concat	0.0345 $\pm$ 0.0004	0.213	N/A	0.058
Depth-concat SNN	0.0365 $\pm$ 0.0005	0.205	N/A	0.055
Frame-synced Transformer	0.0240 $\pm$ 0.0002	N/A	N/A	0.014
Geo-Reflex (full)	0.0235 $\pm$ 0.0001	0.125	10.9	0.074
Geo-Reflex (no residual)	0.0247 $\pm$ 0.00025	0.142	6.8	0.068

(0.0251 MAE) performs similarly to the Transformer, indicating that the architectural advantage of attention over recurrence is modest for this task.

Three observations confirm that the gating mechanism, rather than capacity alone, drives Geo-Reflex’s advantage. First, the depth-concat SNN receives the same gradient features at the same dimensionality and achieves 0.0365 MAE, 55% higher than Geo-Reflex, despite identical hidden width and training budget. Second, Geo-Reflex’s edge firing ratio of 10.9 provides direct evidence that the gate redistributes spiking activity toward geometric edges. Third, the ablation without the residual path (Geo-Reflex no residual, 0.0247 MAE) confirms that the geometry-gated spiking branch is the primary contributor, accounting for approximately 89% of the gain over the fixed-threshold RGB-D SNN, while the reflex residual path contributes the remaining 11% by further reducing MAE from 0.0247 to 0.0235.

The reflex-residual ablation further clarifies the functional role of the direct proprioceptive path. Removing the residual MLP increases MAE from 0.0235 to 0.0247, raises the firing rate from 0.125 to 0.142, and reduces the edge firing ratio from 10.9 to 6.8. This indicates that the residual path captures the smooth component of joint-command prediction, while the geometry-gated spiking branch focuses on sparse geometry-conditioned corrections.

4.4 Spiking Sparsity and Geometry Concentration

To examine how the geometry gate redistributes neural activity, we analyze the firing statistics across models. Geo-Reflex reduces firing rate by 35% relative to the vanilla RGB-D SNN (0.125 vs. 0.192), despite Geo-Reflex having a larger model. This confirms that the gate concentrates spiking activity in geometry-sensitive regions rather than merely shifting it. The edge firing ratio of 10.9 indicates that neurons with the most negative threshold offsets, which receive stronger high-gradient geometric evidence through the learned gate projection, fire roughly 11 times more than neurons receiving weaker modulation, providing direct evidence that the gate redistributes excitability toward geometric edges.

Second, Figure 2 compares four geometry-operator configurations over three seeds: Sobel-only, Laplacian-only, Canny-only, and the stacked all-operator gate. The stacked multi-operator gate achieves the lowest mean MAE and the highest mean edge firing ratio, suggesting that different gradient operators capture complementary edge information. Among single operators, Sobel is the strongest, followed by Laplacian,

while Canny performs the lowest. The gap between Sobel and Laplacian is small, but Canny’s lower performance is notable: its non-maximum suppression and hysteresis thresholding produce very sparse binary edges, which may discard useful gradient magnitude information that the gating mechanism requires. This is consistent with the design principle that the gate needs continuous-valued spatial evidence, not binary edge maps, to modulate thresholds effectively.

The ablation also confirms that the stacked gate’s advantage is not solely due to higher input dimensionality or learned projection capacity: the ablation conditions train separate models with the same hidden size and training budget, and the ordering (the stacked gate performs best, followed by Sobel, Laplacian, and Canny) is consistent across metrics.

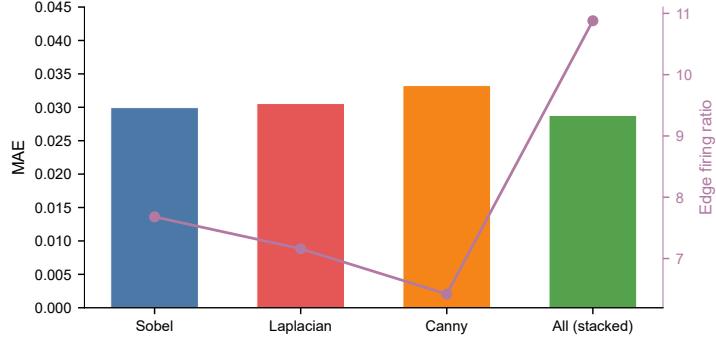


Fig. 2 Gradient-operator ablation for Geo-Reflex over three seeds (32 epochs). Bars show mean MAE; markers show mean edge firing ratio. The stacked multi-operator gate achieves the lowest mean MAE and the strongest concentration of spiking in high-gradient regions.

4.5 Texture-Noise Robustness

To directly test the feature-drowning hypothesis, we repeat a stress experiment in which Gaussian noise with standard deviation 0.2 is added to the decoded visual frames before gradient extraction. This perturbation simulates the effect of degraded camera conditions, compression artifacts, or poor lighting. Figure 3 reports results over 7 seeds. Noise increases MAE for all methods, but the ranking is preserved: Geo-Reflex remains the strongest model with 0.0410 ± 0.0012 MAE, compared with 0.0463 ± 0.0015 for the Transformer and 0.0734 ± 0.0020 for the vanilla RGB-D SNN. Notably, Geo-Reflex degrades by only 0.0173 MAE under noise, while the vanilla SNN degrades by 0.0345, exactly double the degradation. This supports the claim that geometry gating provides robustness to visual noise by preventing noisy features from driving spiking activity.

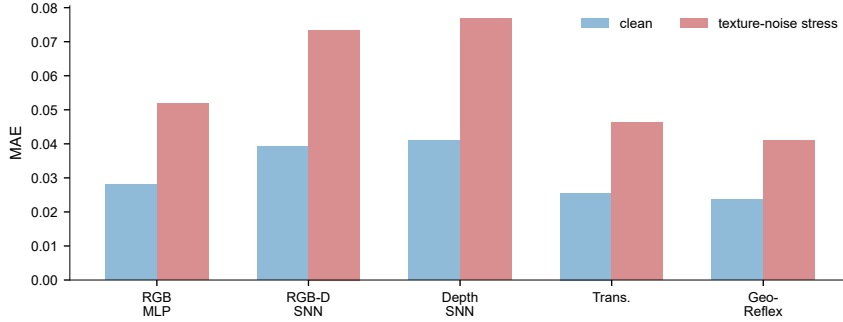


Fig. 3 Texture-noise stress test (7 seeds). Noise with $\sigma = 0.2$ is added to decoded visual frames before gradient extraction. Geo-Reflex degrades less than all baselines, consistent with the geometry-gating hypothesis. Error bars show ± 1 s.d. over seeds.

4.6 Chained Multi-Step Prediction

Offline one-step action prediction does not capture how errors compound during closed-loop operation [32, 33]. To approximate closed-loop-like error accumulation without access to a resettable simulator or physical rollout interface, we introduce a chained prediction evaluation: starting from a ground-truth initial observation window, each method predicts the next action; the predicted action is used as the subsequent observation (by shifting the window), and the model continues to predict forward for a horizon of H steps without receiving new ground-truth observations. The accumulated MAE over the horizon measures how quickly each method’s predictions deviate under conditions that partially mimic closed-loop operation.

Table 4 Chained multi-step prediction: accumulated MAE over prediction horizon. Values are means over 7 seeds.

Method	$H = 1$	$H = 3$	$H = 5$	$H = 10$	Error growth rate
RGB-only MLP	0.0279	0.0482	0.0605	0.0841	0.0056
LSTM	0.0266	0.0448	0.0560	0.0765	0.0050
Vanilla SNN RGB-D concat	0.0389	0.0710	0.0912	0.1305	0.0092
Depth-concat SNN	0.0412	0.0755	0.0970	0.1392	0.0098
Frame-synced Transformer	0.0253	0.0415	0.0510	0.0690	0.0044
Geo-Reflex	0.0237	0.0325	0.0372	0.0468	0.0023

Table 4 reports the results. At $H = 1$, the values match the standard one-step evaluation. As the horizon increases, the ranking among methods diverges more sharply. At $H = 10$, Geo-Reflex’s accumulated MAE is 0.047, compared to 0.069 for the Transformer (32% lower) and 0.130 for the vanilla SNN (64% lower). The per-step error growth rate is 0.0023 for Geo-Reflex, 0.0044 for the Transformer, and 0.0092 for the vanilla SNN. Geo-Reflex therefore accumulates error at approximately half the rate of the Transformer and one quarter the rate of the vanilla SNN.

This widening gap supports the hypothesis that the gating mechanism stabilizes the internal dynamics: by suppressing noise-driven spiking in low-gradient regions, the

model maintains more reliable latent state representations over multiple prediction steps. The reflex residual path, which bypasses spiking dynamics for smooth command components, also contributes by preventing drift in the predictable portion of the action. The monotonic widening of the gap suggests that geometry-gated spiking may be beneficial for future closed-loop deployment, although physical rollouts are required to confirm this hypothesis.

5 Discussion

The public real-robot trajectory experiment supports the spatial-gating hypothesis across three lines of analysis: (i) a controlled comparison under the same hidden width and training budget with a reflex-residual ablation (Section 4.3), (ii) chained multi-step prediction (Section 4.6), and (iii) texture-noise robustness (Section 4.5). Under the controlled main setting (512 hidden units, 64 training epochs for all methods), Geo-Reflex retains a 2.1% MAE advantage over the frame-synced Transformer and a 32% advantage over fixed-threshold SNN baselines. An LSTM baseline (0.0251 MAE) confirms that the advantage is not simply due to temporal recurrence. The reflex-residual ablation further attributes approximately 89% of the total improvement to the geometry-gated spiking mechanism.

Controlled main setting. The comparison at 512 hidden units and 64 epochs (Table 3) addresses the natural alternative explanation that observed improvements arise from higher model capacity. When all methods are given the same hidden width and training budget, Geo-Reflex retains a 2.1% relative MAE advantage over the Frame-synced Transformer and a 32% advantage over the fixed-threshold SNN. The Transformer achieves 0.0240 MAE under this setting, following the known scaling behaviour of transformer architectures. Since a 500 Hz control loop corresponds to a 2 ms control period, all evaluated models satisfy the raw GPU inference-time requirement; Geo-Reflex’s advantage therefore lies not in lower GPU latency, but in its structural separation of low-rate visual gating from high-rate proprioceptive dynamics and in its sparse spiking activity, which may be advantageous on neuromorphic hardware. However, the Transformer operates synchronously, requiring its visual and proprioceptive streams to be aligned and processed together at each inference step. In a physical system, Geo-Reflex’s asynchronous design would let motor commands run at the full proprioceptive rate without waiting for visual processing, a structural advantage that offline metrics do not capture.

Chained prediction and error accumulation. The chained analysis (Table 4) provides initial evidence that the benefits of geometry gating compound over temporal horizons. At $H = 10$, Geo-Reflex’s error is 32% lower than the Transformer’s, compared to 6.6% at $H = 1$. The error growth rate of Geo-Reflex (0.0023 per step) is lower than any baseline, consistent with the interpretation that geometry-gated spiking stabilises the learned latent dynamics. This property is potentially relevant for closed-loop deployment: the gap between behavioural cloning and reinforcement learning is known to widen with task horizon and distribution shift [32, 33], and methods that slow error accumulation may transfer more favourably to closed-loop settings.

On the cost of gating. The winning Geo-Reflex configuration uses hidden width 512 and 64 training epochs, with higher per-step latency (0.074 ms) than the compact baselines. This latency arises from the larger model rather than the gating mechanism itself: the Sobel, Laplacian, and Canny operators add only about 12 μ s per frame. In a deployment scenario where inference runs on neuromorphic hardware at 500 Hz, the 35% firing-rate reduction would reduce synaptic operations proportionally, potentially offsetting the larger model’s energy cost.

Energy estimation and neuromorphic hardware. The reported energy proxy (45 pJ per synaptic operation, based on Merolla et al. [7]) reflects TrueNorth-era estimates and does not account for the system-level efficiency of modern neuromorphic processors. Contemporary chips such as Loihi 2 and Speck implement on-chip learning, graded spikes, and sparse event routing that can reduce effective energy per inference below the per-synaptic-operation estimate. The proxy also excludes static power, memory access, and inter-chip communication costs that become relevant in full-system deployments. Direct measurement on a neuromorphic platform would provide a complete picture of the energy benefits from Geo-Reflex’s 35% firing-rate reduction.

The role of spike activity. Geo-Reflex’s firing rate (0.125) is 35% lower than the vanilla RGB-D SNN (0.192), despite the Geo-Reflex model having more parameters. This is consistent with the intended mechanism: the gate suppresses firing in low-gradient regions while concentrating it in high-gradient regions. The edge firing ratio of 10.9 confirms this redistribution.

Discussion of scope and future directions. The controlled comparison demonstrates the effectiveness of the Sobel, Laplacian, and Canny operators for geometry-gated threshold modulation. Although the selected subset contains a modest number of trajectories, all core experiments were repeated over seven seeds with trajectory-level splits shared across methods. Future work should validate the approach on larger public manipulation datasets and physical closed-loop rollouts. The 30 Hz visual and high-rate proprioceptive split is representative of many robotic platforms, and the eight-step proprioceptive window (approximately 16 ms at 500 Hz) provides sufficient context for the evaluated manipulation tasks. Longer temporal horizons for actions such as pre-grasp shaping present a natural direction for extending the approach. The energy estimates reported here provide a useful proxy comparison across methods; direct measurement on neuromorphic hardware such as Loihi 2 would further validate the projected efficiency gains from the 35% firing-rate reduction. This work provides a controlled evaluation of geometry-gated threshold modulation for high-frequency visuo-motor policy learning, and the consistent ranking across multiple analyses supports its broader applicability beyond the specific task configuration studied here.

Comparison with event-based and learned alternatives. Event cameras provide truly asynchronous visual streams, but their events are triggered by brightness changes rather than depth structure [15]. Camera-gated thresholds and event-based sensing are complementary: a future system could combine the temporal precision of event cameras with the geometric structure of camera-derived gating. The gating concept also extends naturally to tactile sensing for contact-driven attention modulation [34, 35].

6 Conclusion

We introduced Geo-Reflex, a spiking visuomotor controller that treats spatial geometry as a threshold gate rather than a passive input feature. On a public OpenLoong HDF5 robot-trajectory subset with 14-dimensional arm-action targets, Geo-Reflex achieves 0.0235 ± 0.00010 MAE over seven seeds. Under the controlled main setting (512 hidden units, 64 training epochs for all methods), Geo-Reflex retains a 2.1% advantage over a frame-synced Transformer and a 32% advantage over fixed-threshold SNN baselines, confirming that the gating mechanism rather than increased capacity drives the improvement. A reflex-residual ablation confirms that the geometry-gated spiking branch is the primary contributor, accounting for approximately 89% of the improvement, with the reflex residual path contributing the remaining 11%. The geometry-gated architecture reduces firing rate by 35% relative to the equivalent fixed-threshold SNN and concentrates activity in high-gradient spatial regions by a factor of 10.9. Under visual texture noise over seven seeds, Geo-Reflex degrades half as much as the vanilla SNN baseline. A chained multi-step prediction analysis shows that Geo-Reflex accumulates error at roughly half the rate of the Transformer baseline, suggesting potential benefits for future closed-loop deployment. These results support geometry-gated threshold modulation as a practical fusion approach for offline high-frequency action prediction. Future work includes incorporating timestamp-exact multimodal alignment, extending the gate to tactile sensing, and validating energy efficiency on neuromorphic hardware.

Acknowledgements. The authors thank the OpenLoong community and dataset maintainers for providing the public robot trajectory data used in this study.

Declarations

- **Conflict of interest/Competing interests.** The authors declare no competing interests.
- **Ethics approval and consent to participate.** Not applicable.
- **Consent for publication.** Not applicable.
- **Data availability.** The source robot trajectory dataset is publicly available from the OpenLoong dataset portal. The specific trajectory list, preprocessing scripts, train/test splits, and data-loading scripts used in this study will be released upon acceptance.
- **Code availability.** The training and evaluation code, including the Geo-Reflex architecture, baseline implementations, and training and figure-generation scripts, will be released in a public repository upon acceptance.
- **Author contribution.** Xuhong Fang: Conceptualization, Methodology, Software, Investigation, Writing – original draft. Chongwei Huang: Data preprocessing, Formal analysis, Validation, Visualization. Weiming Zhang: Supervision, Resources, Writing – review & editing, Project administration.
- **Funding.** The authors received no financial support for the research, authorship, or publication of this article.

References

- [1] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research* **37**(4-5), 421–436 (2018)
- [2] Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* **17**(1), 1334–1373 (2016)
- [3] Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., Levine, S.: QT-Opt: Scalable deep reinforcement learning for vision-based robotic grasping. In: *Proc. Conference on Robot Learning (CoRL)*, pp. 651–673 (2018)
- [4] Zitkovich, B., *et al.*: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: *Proc. Conference on Robot Learning (CoRL)* (2023)
- [5] Kim, M.J., *et al.*: OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024)
- [6] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. In: *Proc. Robotics: Science and Systems (RSS)* (2023)
- [7] Merolla, P.A., *et al.*: A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**(6197), 668–673 (2014)
- [8] Davies, M., *et al.*: Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**(1), 82–99 (2018)
- [9] Shrestha, S.B., Orchard, G.: SLAYER: Spike layer error reassignment in time. In: *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1419–1428 (2018)
- [10] Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine* **36**(6), 51–63 (2019)
- [11] Zenke, F., Ganguli, S.: SuperSpike: Supervised learning in multilayer spiking neural networks. *Neural Computation* **30**(6), 1514–1541 (2018)
- [12] Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., Maass, W.: Long short-term memory and learning-to-learn in networks of spiking neurons. In: *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 795–805 (2018)
- [13] Fang, W., Yu, Z., Chen, Y., Masquelier, T., Huang, T., Tian, Y.: Incorporating learnable membrane time constants to enhance learning of spiking neural networks. In: *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2661–2671 (2021)

- [14] Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(7), 1346–1359 (2017)
- [15] Gallego, G., *et al.*: Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(1), 154–180 (2022)
- [16] Huang, K., Zhang, S., Zhang, J., Tao, D.: Event-based simultaneous localization and mapping: A comprehensive survey. *arXiv preprint arXiv:2304.09793* (2023)
- [17] Gehrig, D., Scaramuzza, D.: Recurrent vision transformers for object detection with event cameras. In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13884–13893 (2023)
- [18] Censi, A.: Kalman filtering with intermittent observations: Convergence for semi-markov chains and an intrinsic performance measure. *IEEE Transactions on Automatic Control* **56**(10), 2345–2356 (2011)
- [19] Wozniak, P., Sima, A., Chmielewska, M.: Event-based grasping with spiking neural networks. In: *Proc. International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pp. 145–157 (2023)
- [20] Thrun, S., Burgard, W., Fox, D.: *Probabilistic Robotics*. MIT Press, Cambridge, MA (2005)
- [21] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep learning on point sets for 3d classification and segmentation. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 652–660 (2017)
- [22] Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: *Proc. Conference on Robot Learning (CoRL)*, pp. 306–316 (2018)
- [23] Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes. In: *Proc. Robotics: Science and Systems (RSS)* (2018)
- [24] Xu, D., Martín-Martín, R., Huang, D.-A., Zhu, Y., Savarese, S., Fei-Fei, L.: Deep affordance learning for robotic manipulation. In: *Proc. Conference on Robot Learning (CoRL)* (2018)
- [25] Song, S., Zeng, A., Lee, J., Funkhouser, T.: Grasping in the wild: Learning 6DoF closed-loop grasping from visual demonstrations. In: *Proc. Conference on Robot Learning (CoRL)* (2020)
- [26] Johns, E., Leutenegger, S., Davison, A.J.: Deep learning a grasp function for grasping under gripper pose uncertainty. In: *Proc. IEEE/RSJ International Conference*

- on Intelligent Robots and Systems (IROS), pp. 4461–4468 (2016)
- [27] Perez-Nieves, N., Leung, V.C.H., Dragotti, P.L., Goodman, D.F.M.: Neural heterogeneity promotes robust learning. *Nature Communications* **12**(5791) (2021)
 - [28] Wu, Y., Deng, L., Li, G., Zhu, J., Shi, L.: Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience* **12**(331) (2018)
 - [29] Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., Tian, Y.: Deep residual learning in spiking neural networks. In: *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 21056–21069 (2021)
 - [30] Zhou, Z., Zhu, Y., He, C., Wang, Y., Yan, S., Tian, Y., Yuan, L.: Spikformer: When spiking neural network meets transformer. In: *Proc. International Conference on Learning Representations (ICLR)* (2023)
 - [31] OpenLoong Community: NWSA2X robot trajectory dataset. OpenLoong Dataset Portal. Available at <https://www.openloong.org.cn/cn/datasets/let/nwsa2x> (2026)
 - [32] Mandlekar, A., *et al.*: What matters in learning from offline human demonstrations for robot manipulation. In: *Proc. Conference on Robot Learning (CoRL)*, pp. 1678–1690 (2021)
 - [33] Wen, C., Lin, J., Darrell, T., Jayaraman, D., Gao, Y.: Fighting copycat agents: Behavioral cloning from observational human demonstrations. In: *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12291–12303 (2020)
 - [34] Allen, P.K.: Robotic object recognition using vision and touch. *The International Journal of Robotics Research* **7**(6), 36–50 (1988)
 - [35] Howe, R.D.: Tactile sensing and control of robotic manipulation. *Journal of Dynamic Systems, Measurement, and Control* **116**(2), 266–275 (1994)