**ICRA2026 Late Breaking Results Poster presented at**
**2026 IEEE International Conference on Robotics and Automation (ICRA 2026)**
**June 1-5, 2026. Vienna, Austria**

# State-Space Time Surfaces for Event-Based Zero-Shot Robotic Grasping and Scene Reconstruction

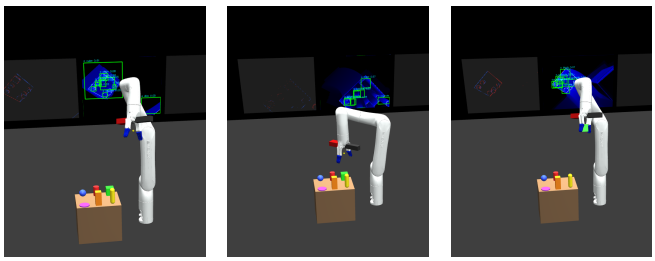
Gu Gong and David Navarro-Alarcon
The Hong Kong Polytechnic University

***Abstract*—Event cameras report per-pixel brightness changes asynchronously with microsecond latency, but their output is incompatible with vision foundation models trained on conventional images. We propose State-Space Time Surfaces (S3TS), a training-free representation that recasts exponential-decay time surfaces as a diagonal state-space model with multi-scale temporal channels and input-dependent selective decay inspired by Mamba. The resulting three-channel pseudo-RGB image is fed directly to a frozen OWLv2 detector for zero-shot, text-prompted object detection from events alone. We demonstrate two applications on a simulated 6-DOF manipulator: (i) event-only grasping with near-nadir refinement that localizes objects without any depth sensor, and (ii) dense 3D scene reconstruction via multi-view TSDF fusion with neuromorphic per-vertex surface descriptors. S3TS detects over twice as many objects as single-channel event representations and produces faithful 3D workspace meshes—all without network training or fine-tuning.**

## I. Introduction

Event cameras [1], [2] are neuromorphic sensors that report per-pixel log-intensity changes asynchronously, offering microsecond temporal resolution and high dynamic range (>120 dB). However, their sparse, asynchronous output is fundamentally incompatible with vision-language models (VLMs) [3] that expect dense RGB frames—creating a *modality gap* that prevents leveraging open-vocabulary detectors [4] for event-driven robotics.

Existing event representations—binary frames, histograms, surfaces of active events (SAE), and single-timescale time surfaces [5]—each discard critical temporal information or operate at a fixed scale, yielding single-channel images with insufficient structure for zero-shot VLM detection. We address this with three contributions: **(1)** A formal equivalence between exponential-decay time surfaces and first-order state-space models, extended to a *multi-scale diagonal SSM* with input-dependent selective decay [6], [7]—requiring no learned parameters. **(2)** The first pipeline connecting event cameras to frozen VLMs for zero-shot, text-prompted robotic grasping. **(3)** A neuromorphic temporal-geometric mesh where per-vertex S3TS descriptors encode illumination-invariant surface properties via TSDF fusion [9].



(a) Scan & detect (b) Refine & approach (c) Grasp & lift

Fig. 1: S3TS event-only grasping pipeline. In-scene screens show event stream (left) and S3TS pseudo-RGB with OWLv2 detections (right).

## II. Methodology

### A. S3TS Representation

The classical time surface at pixel $(x, y)$ is $\mathcal{T}(x, y, t) = \exp(-(t - t_{\text{last}})/\tau)$, the impulse response of a first-order LTI system $\dot{s} = -s/\tau + u$ with $A{=}-1/\tau$, $B{=}C{=}1$.

**Multi-scale extension.** We deploy $K{=}3$ parallel channels with $\tau \in \{5,\ 50,\ 5000\}$ ms, forming a diagonal state-space model:

$$\dot{\boldsymbol{s}} = \operatorname{diag}(-1/\tau_1, \ldots, -1/\tau_K)\, \boldsymbol{s} + \mathbf{1}_K\, u, \quad \boldsymbol{y} = \boldsymbol{s}, \qquad (1)$$

equivalent to a diagonal S4 model [6] with analytically prescribed eigenvalues. The fast channel captures motion/texture; medium captures edges; slow retains persistent scene structure (≈94% energy after a 0.3 s hold).

**Selective decay.** Inspired by Mamba [7], we adapt $\tau_k$ via local event rate: $\tau_k^{\text{sel}} = \tau_k \cdot \sigma(\alpha - \beta\, \hat{r})$, where $\sigma$ is the logistic sigmoid and $\hat{r}$ the spatially smoothed, normalized rate. High activity compresses $\tau$ (prevents saturation); low activity extends it (preserves faint structure)—entirely analytical and training-free.

**Pseudo-RGB encoding.** The three channels map to RGB: $\boldsymbol{I}_{\text{S3TS}} = \lfloor 255 \cdot \text{clip}([s_1, s_2, s_3]^\top, 0, 1) \rfloor$, with CLAHE [8] enhancement, yielding images compatible with frozen VLMs.

### B. Event-Only Grasping

We propose a complete event-only grasping pipeline (Fig. 1) requiring *no RGB input and no depth sensor*. **(i) Active scanning:** sinusoidal joint-space oscillation excites events at object boundaries; the arm stabilizes for detection. **(ii) Zero-shot detection:** frozen OWLv2 [4] processes the S3TS image with

TABLE I: Zero-shot detection results (OWLv2-Base, $c_{\min}$=0.05).

| Representation | Det. | Top-1 | Mean |
|---|---|---|---|
| Binary Event Frame | 3 | 0.189 | 0.103 |
| Event Histogram | 3 | 0.467 | 0.208 |
| Single-$\tau$ TS | 2 | 0.318 | 0.185 |
| SAE | 1 | 0.394 | 0.394 |
| Multi-Scale TS | 7 | 0.448 | 0.121 |
| **S3TS (Ours)** | **7** | **0.486** | **0.129** |

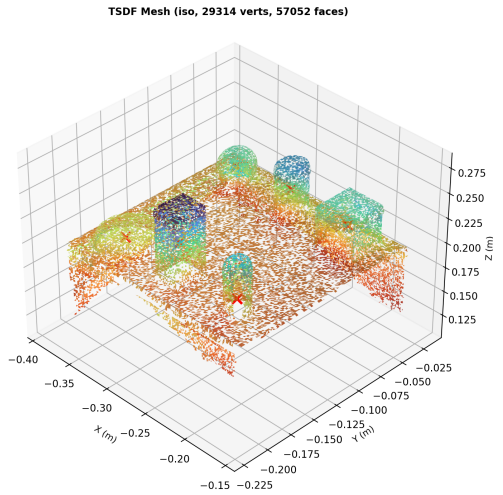



Fig. 2: Dense temporal mesh from multi-view TSDF fusion (29k vertices, 57k faces, 2 mm voxels). Per-vertex colors encode S3TS temporal descriptors across three timescales.

text queries. **(iii) Depth-free 3D localization:** detection centroid defines a viewing ray intersected with a per-object height plane. **(iv) Near-nadir refinement:** the arm repositions above the coarse estimate, re-scans and re-detects from near-vertical view, reducing XY error by 5–15 mm. **(v) Grasp execution:** orientation-constrained IK (full $6\times6$ Jacobian) enforces vertical descent with FK-verified waypoints ($\leq$3 mm lateral deviation).

### C. Neuromorphic Scene Reconstruction

We extend S3TS to a *per-surface temporal descriptor* fused into a volumetric model. The manipulator traverses 152 viewpoints, executing active scans at each. Depth and S3TS features are jointly fused into a TSDF volume [9]:

$$\boldsymbol{c}(\mathbf{p}) \leftarrow \frac{W(\mathbf{p})\,\boldsymbol{c}(\mathbf{p}) + w^{(v)}\,\boldsymbol{s}^{(v)}(\pi(\mathbf{p}))}{W(\mathbf{p}) + w^{(v)}}, \tag{2}$$

where $\boldsymbol{s}^{(v)}=[s_1, s_2, s_3]^\top$ is the S3TS state at the projected pixel. Marching Cubes [10] extracts a *temporal mesh* with per-vertex S3TS descriptors encoding illumination-invariant, multi-scale surface response (Fig. 2).

## III. Experiments and Results

Experiments use MuJoCo [11] with a 6-DOF manipulator, a simulated DAVIS event camera (via v2e [12]), and an RGBD sensor.

**Detection comparison.** Table I summarizes zero-shot detection across six representations fed to the same frozen OWLv2-Base. Single-channel representations (binary frame, single-$\tau$ TS, SAE) yield $\leq$3 detections. S3TS achieves 7 detections with the highest top-1 confidence (0.486 vs. 0.448 for non-adaptive multi-scale), confirming that selective decay improves the strongest responses.

**Grasping.** S3TS enables successful zero-shot grasps on text-specified objects (e.g., "grasp the cube") using events alone. Near-nadir refinement reduces XY error by 5–15 mm—critical for small objects ($\leq$40 mm).

**Reconstruction.** Multi-view TSDF fusion produces a dense temporal mesh (29k vertices, 57k faces, 2 mm voxels) in 74.5 s. Per-vertex S3TS descriptors encode fine texture ($s_1$), structural edges ($s_2$), and persistent geometry ($s_3$), creating a neuromorphic surface map (Fig. 2).

## IV. Conclusion

We presented S3TS, a training-free event representation grounded in state-space theory that bridges the modality gap between neuromorphic sensors and frozen vision-language models. The formal SSM equivalence enables principled multi-scale decomposition and input-dependent selective decay without learned parameters. Demonstrated on zero-shot robotic grasping and dense 3D reconstruction, S3TS provides a practical, theoretically motivated pathway for deploying event cameras in open-vocabulary manipulation systems.

## References


[1] P. Lichtsteiner, C. Posch, and T. Delbrück, "A 128×128 120 dB 15 $\mu$s latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] G. Gallego *et al.*, "Event-based vision: A survey," *IEEE Trans. PAMI*, vol. 44, no. 1, pp. 154–180, 2022.

[3] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[4] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," in *Proc. NeurIPS*, 2024.

[5] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. PAMI*, vol. 39, no. 7, pp. 1346–1359, 2017.

[6] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *Proc. ICLR*, 2022.

[7] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Proc. ICML*, 2024.

[8] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994, pp. 474–485.

[9] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. ACM SIGGRAPH*, 1996, pp. 303–312.

[10] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.

[11] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IROS*, 2012, pp. 5026–5033.

[12] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. PAMI*, vol. 43, no. 6, pp. 1964–1980, 2021.