EventPrune: Cascaded Event-Assisted Token Pruning for Efficient Embodied Spatial Reasoning

Pengtao Ma[1,†], Ziliang Zhou[1,†], Ciyu Ruan[1], Haoyang Wang[3], Kaiyuan Li[1], Zihang Gong[2], Wenhua Ding[1], Chen Gao[3,*], Jingao Xu[4], Xinlei Chen[1,*]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Harbin Institute of Technology [3]Tsinghua University [4]The University of Hong Kong

{mpt25,zhouzl25}@mails.tsinghua.edu.cn, jingaoxu@hku.hk
chgao96@gmail.com, chen.xinlei@sz.tsinghua.edu.cn

Abstract

Embodied spatial reasoning requires models to track continuous motion and precise geometric structure, but the quadratic attention cost of Transformer-based Video-LLMs makes dense visual tokens computationally expensive. Existing token pruning paradigms predominantly rely on discrete static snapshots, failing to preserve the motion and geometric cues essential for reasoning. We propose Event Cascade Pruning (ECP), to our knowledge the first training-free framework that leverages the high-frequency motion cues from event cameras as a continuous event-guided motion prior to guide token selection. ECP combines three stages: Event-Guided Temporal Sampling to anchor motion-informative keyframes, Event-guided Motion Saliency Filtering to suppress event-inactive visual tokens, and Event-Attention Ranking Fusion to calibrate spatial attention with motion-salient dynamics. With 80% visual token reduction, ECP outperforms the full-token baseline (37.62% vs. 36.31%) while achieving 1.89× inference speedup and 52% GFLOPs reduction. We further introduce ESR-Real, the first real-world RGB-event benchmark for embodied spatial reasoning, where ECP improves accuracy by 2.68 percentage points over full-token baselines. Code is available at https://github.com/Ma-Potter/EventPrune.

1 Introduction

Video-LLMs [23, 37, 6] are increasingly used for embodied spatial reasoning [49, 39, 13, 58, 52], such as indoor navigation and drone perception [53, 36, 15]. Unlike conventional video understanding, embodied scenarios involve rapid egocentric motion and continuously changing scene geometry [7, 38]. These settings require models to process video efficiently [20, 45] while preserving motion continuity and geometric structures for reliable spatial reasoning.

To mitigate the "visual token explosion" and high inference latency in these real-time settings, training-free visual token pruning has become an important optimization strategy [10, 48, 20]. However, existing RGB-only pruning signals are fragile for highly dynamic egocentric videos at both temporal and spatial levels. (1) For inter-frame pruning, existing methods filter temporal redundancy by comparing pixel differences or feature similarities [51]. Such appearance-based comparison not only introduces significant latency but also fails to capture continuous dynamic characteristics like speed and trajectory, making it unreliable for dynamic information extraction. (2) For intra-frame pruning, methods based on attention weights [10, 27, 48] or diversity metrics [3, 25] may discard

[†]Equal contribution.
[*]Corresponding authors: Xinlei Chen and Chen Gao.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

structure-critical tokens. Our analysis reveals a severe peripheral spatial bias in attention scores: attention concentrates disproportionately on image borders (Fig. 3), causing task-relevant interior regions to be under-retained. Diversity- or merging-based strategies may scatter retained tokens or merge local neighborhoods, fragmenting object boundaries and weakening geometric continuity. Consequently, existing pruning methods often fail to preserve the motion and structural cues required for spatial reasoning under high compression.

In contrast to existing software-based approaches, we identify event cameras as a natural external cue for these issues. As bio-inspired sensors that asynchronously record pixel-level brightness changes with microsecond resolution, event cameras offer unique cues for both dynamics and geometry: (1) At the inter-frame level, event streams directly record motion activity at each pixel. This provides an instant, training-free prior for anchoring causally critical transitions and filtering temporally redundant regions without expensive feature matching. (2) At the intra-frame level, events exhibit high sensitivity to edges and textures, representing the geometric structures that biased attention mechanisms neglect. These signals offer a reliable reference to calibrate attention scores and ensure the retention of tokens with true spatial saliency.

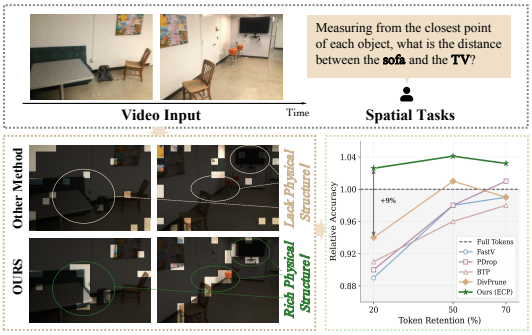

Figure 1: Token pruning for spatial reasoning. Unlike prior pruning, ECP preserves task-relevant geometric cues under high sparsity.

However, events are asynchronous external signals, fundamentally heterogeneous from the synchronous grid features inside Video-LLMs. How to deeply integrate this heterogeneous prior into the inference pipeline to genuinely guide token selection remains non-trivial. To overcome this, we propose Event Cascade Pruning (ECP), a training-free event-assisted pruning framework with three cascaded modules. Event-Guided Temporal Sampling (ETS) anchors keyframes using event activity flux and temporal changes; Event-guided Motion Saliency Filtering (EMSF) removes low-motion regions via event saliency maps; and Event-Attention Ranking Fusion (EARF) calibrates attention-based token scores through non-parametric rank alignment, preserving tokens with joint event-semantic saliency.

Extensive experiments demonstrate the effectiveness of ECP. At 20% token retention, ECP outperforms all evaluated training-free pruning baselines and surpasses the full-token baseline (37.62% vs. 36.31%), while achieving 1.89× speedup and 52% GFLOPs reduction. On our real-world ESR-Real benchmark with native event signals, ECP improves over the full-token baseline by 2.68 percentage points at 50% retention. These results show that event-assisted pruning improves efficiency while preserving, and in some cases improving, accuracy for embodied spatial reasoning.

Our main contributions are:

(1) We propose Event Cascade Pruning (ECP), to our knowledge the first training-free event-assisted pruning framework for embodied spatial reasoning, establishing a cascaded temporal–spatial–representation pruning pipeline that removes redundancy while preserving motion, geometric, and attention-derived semantic cues.

(2) We introduce ESR-Real, a real-world RGB-event benchmark for embodied spatial reasoning with over 700 QA pairs across 6 categories.

(3) We achieve the best performance among evaluated training-free pruning methods: ECP outperforms the full-token baseline at 80% token reduction with 1.89× speedup and 52% GFLOPs reduction, with consistent gains on both simulated-event and real-event benchmarks.

2 Related Work

Video-LLMs and Embodied Spatial Perception. Recent Large Vision-Language Models (LVLMs) [2, 24, 28, 5, 44, 12] integrate visual encoders with LLMs [40, 1] for unified multimodal understanding, including spatial reasoning about distance and direction [55]. Representative Video-

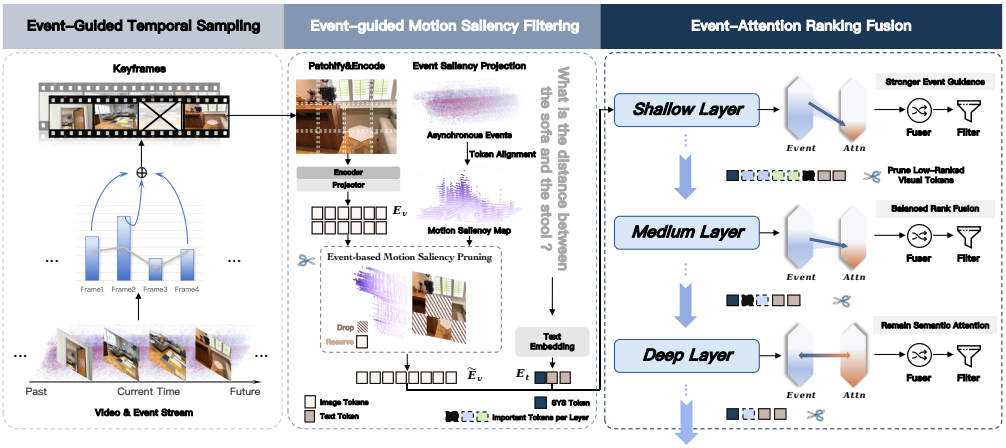

Figure 2: Overview of Event Cascade Pruning (ECP). ECP comprises three cascaded modules: ETS selects keyframes via event activity flux; EMSF maps aligned events to token saliency and filters visual tokens; EARF fuses event and attention ranks with layer-wise weights, retaining key visual tokens. Scissors indicate the removal of the remaining visual tokens.

LLMs such as LLaVA-NeXT [29], Qwen2.5-VL [6], LLaVA-OneVision [23], and LongVA [54] improve fine-grained and long-context video understanding, but their high-resolution visual tokens lead to substantial computational cost. This bottleneck is especially problematic for embodied spatial reasoning, where models must respond to rapid ego-motion and changing geometry under latency constraints.

Visual Token Pruning. To mitigate computational bottlenecks, visual token pruning has emerged as a critical optimization strategy [11, 21, 22], operating primarily across inter-frame and intra-frame dimensions. Inter-frame methods such as DTD [51] remove redundancy via pixel or feature differences, while intra-frame methods select or merge tokens using attention [10, 27, 48], diversity [3], token merging [8], or hybrid objectives [25]. However, these software-only cues are fragile for embodied spatial reasoning: appearance differences miss high-frequency motion continuity, attention scores suffer from peripheral spatial bias [47], and diversity or merging objectives can fragment local geometric structures. Consequently, existing pruning methods often discard motion- or structure-critical tokens under high compression.

Event-based Vision in VLMs. Event cameras (Dynamic Vision Sensors) asynchronously capture pixel-level brightness changes with microsecond resolution, offering intrinsic high dynamic range and motion sensitivity [17, 9, 42]. With recent commercialization by manufacturers such as Prophesee and Sony, these sensors are gaining traction in embodied AI applications including autonomous navigation, drone localization for precision landing [43], long-range UAV detection [60], and robotics. While demonstrating exceptional performance in vision tasks such as SLAM [41, 62], optical and scene flow [61, 19], event denoising and deraining [34, 35], and video reconstruction [18, 32], their integration into LVLMs remains limited. Recent event-LVLM methods such as EventCLIP [46], EventBind [59], and EventGPT [30] align event representations with vision-language embeddings or generate event-based captions, but typically require dedicated event encoders or additional training. In contrast, ECP uses events as a training-free score-level motion and structure prior for pruning visual tokens in existing RGB Video-LLMs, without parameter updates or event-embedding injection.

3 Methodology

Existing pruning methods [10, 48, 3, 25, 31] struggle to preserve both dynamics and geometry required for embodied spatial reasoning. Event cameras [17, 9] provide a lightweight external cue by recording asynchronous brightness changes that reveal motion activity and motion-induced structural cues. The key challenge lies in bridging the representational gap between asynchronous event streams and the synchronous grid features of Video-LLMs [6, 23]. We propose Event Cascade Pruning (ECP), a training-free framework shown in Figure 2, which cascades temporal–spatial–semantic

pruning: ETS selects keyframes, EMSF filters low-motion regions, and EARF fuses event saliency with attention-derived semantic scores for layer-wise token pruning.

3.1 Preliminaries

Attention-based Pruning. Token pruning reduces computational overhead by retaining the top-k tokens based on an importance score $\mathbf{s} \in \mathbb{R}^{T \times N}$. Existing methods [10, 48, 27] derive $\mathbf{s}$ from text-image attention weights:

$$\mathbf{S}_{\text{img}}^{(l)} = \frac{1}{m} \sum_{i=1}^{m} \text{Atten}^{(l)}(\mathbf{X}_I, \mathbf{X}_T^{(i)}) \tag{1}$$

where $\mathbf{X}_I$ and $\mathbf{X}_T$ denote image and text tokens, and m is the number of text tokens.

Event Camera as Motion Prior. Unlike frame-based sensors, event cameras [17] asynchronously trigger events $e_k = (\mathbf{u}_k, t_k, p_k)$ only when logarithmic intensity changes exceed a threshold:

$$|\ln I(\mathbf{u}, t) - \ln I(\mathbf{u}, t_{\text{prev}})| \geq \theta \tag{2}$$

The resulting stream $\mathcal{E}$ offers microsecond-level temporal resolution, providing rich cues about motion causality and geometric boundaries that can be leveraged for spatial reasoning.

3.2 Event-Guided Temporal Sampling

Existing sampling strategies struggle to balance coverage and efficiency: uniform sampling misses transient dynamics, while appearance-based methods incur high computational cost and are sensitive to motion blur. We propose ETS, a lightweight paradigm that leverages high-temporal-resolution event streams to anchor motion-rich moments and critical change points in scene dynamics.

Event Activity Flux. To align temporal sampling with non-uniform motion activity, we divide the event stream into fixed temporal windows $W_n = [n\Delta t, (n+1)\Delta t)$ indexed by n. Applying a spatiotemporal density filter $\Phi(\cdot)$ to mitigate sensor noise, we define the window-level activity flux as S_n. A larger S_n indicates stronger event activity within window n:

$$S_n = \sum_{k=1}^{|\mathcal{E}|} \Phi(e_k)\, \mathbb{I}[t_k \in W_n]\,. \tag{3}$$

Dual-Criteria Frame Anchoring. ETS anchors frames using two complementary cues: high event activity S_n preserves motion-rich moments, while large activity changes indicate critical change points in scene dynamics:

$$\Delta S_n = |S_n - S_{n-1}|, \quad n \geq 1. \tag{4}$$

A large ΔS_n indicates that the event activity changes sharply between adjacent windows, and is therefore treated as a critical temporal anchor. Given the budget N_{target}, ETS jointly selects high-S_n windows and top-ranked ΔS_n windows, then refines the set by pruning clustered low-activity candidates or filling large temporal gaps. The selected windows are mapped to the nearest RGB keyframes. Since both cues are computed on density-filtered window activity, ETS suppresses event jitter while maintaining temporal coverage.

3.3 Event-guided Motion Saliency Filtering

Following temporal sampling, spatial redundancy remains a bottleneck. RGB-based motion differencing requires additional frame processing and can degrade under fast motion or blur, whereas event streams provide lightweight, high-temporal-resolution cues of local brightness changes. EMSF distills the event activity aligned with each selected keyframe into a token-aligned motion prior, enabling early suppression of event-inactive visual tokens.

Token-aligned Motion Saliency. For each selected keyframe f, let $\mathcal{W}_f = [a_f, a_f + \Delta t)$ denote its aligned event window. Each visual token $i \in \{1, \ldots, N\}$ corresponds to a spatial support $\Omega_i^{(f)}$ in the image plane. Given the density-filtered event stream $\tilde{\mathcal{E}}$, we compute the event activity associated with token i as:

$$C_i^{(f)} = \sum_{e_k \in \tilde{\mathcal{E}}} \mathbb{I}\left[(x_k, y_k) \in \Omega_i^{(f)},\ t_k \in \mathcal{W}_f\right], \quad M_i^{(f)} = \mathcal{N}_f\left(C_i^{(f)}\right), \tag{5}$$

where $\mathcal{N}_f(\cdot)$ adds $\epsilon = 10^{-6}$ to each event score and divides by the maximum within keyframe f. The resulting score $M_i^{(f)} \in [0, 1]$ measures the event-guided motion saliency of visual token i.

Budgeted Motion-token Retention. With retention ratio ρ, EMSF keeps $K = \lfloor \rho N \rfloor$ visual tokens per keyframe by maximizing the retained motion saliency:

$$\mathcal{I}_{\text{EMSF}}^{(f)} = \arg \max_{\mathcal{I} \subseteq \{1,\dots,N\},\ |\mathcal{I}|=K} \sum_{i \in \mathcal{I}} M_i^{(f)}. \tag{6}$$

By reducing the number of visual tokens from N to K before subsequent attention, EMSF reduces the visual-token attention term from $O(N^2)$ to $O(K^2) = O(\rho^2 N^2)$ and provides motion-salient candidates for EARF.

3.4 Event-Attention Ranking Fusion

While EMSF suppresses event-inactive visual tokens, the remaining candidates still dominate subsequent attention and KV-cache storage [47, 26]. EARF further sparsifies these tokens at selected pruning layers $\mathcal{L}_p$ by calibrating attention-based importance with event saliency. Importantly, events are used only to compute pruning scores and masks, rather than being encoded, concatenated with RGB embeddings, or injected into hidden representations.

Empirical Motivation: Peripheral Sink. Our analysis of 21,920 frames across 28 transformer layers (Figure 3) reveals a consistent *Peripheral Sink* in Video-LLMs [6]: visual attention concentrates more on image borders than on central regions, with the peripheral-to-center attention ratio exceeding 1 across all layers ($p < 10^{-10}$). This spatial bias resembles the attention sink observed in LLMs [47], but appears over peripheral visual tokens rather than initial text tokens. The bias is layer-dependent, with the middle-layer group showing the highest mean peripheral-to-center ratio of $5.64\times$. Moreover, the layer-wise bias profiles are highly correlated across datasets ($r = 0.897$), showing similar peripheral concentration in both evaluated settings. This observation motivates event saliency as a complementary cue for token ranking. Detailed per-layer statistics are provided in Appendix A.

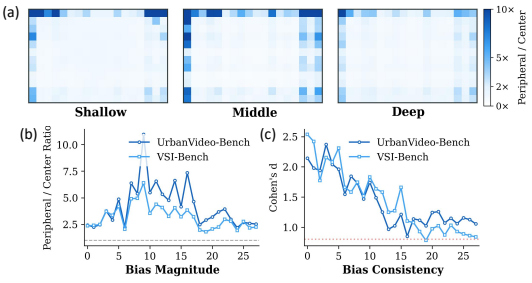

Figure 3: Peripheral Sink. (a) Attention maps reveal border-biased visual attention. (b) Peripheral-to-center ratio measures bias magnitude across layers. (c) Cohen's d measures bias reliability, with similar trends across datasets.

Fusion via Rank Projection. Starting from the EMSF-retained candidates, EARF further sparsifies visual tokens at pruning layers $\mathcal{L}_p$. For pruning layer l and keyframe f, let $\mathcal{V}^{(l,f)}$ denote the active visual-token set before pruning. EARF first derives a text-conditioned visual importance score from the preceding post-softmax multimodal self-attention map:

$$A_i^{(l,f)} = \frac{1}{|\mathcal{Q}_s|} \sum_{q \in \mathcal{Q}_s} \mathbf{A}^{(l-1)}[q, i], \quad i \in \mathcal{V}^{(l,f)}. \tag{7}$$

Here $\mathcal{Q}_s$ denotes the scoring query-token set used to read out visual importance, and $\mathbf{A}^{(l-1)}$ is the head-averaged post-softmax attention map used for scoring. The event-guided saliency $M_i^{(f)}$ is obtained from EMSF for the same visual token. Since attention scores and event saliency scores lie on different numerical scales and often exhibit long-tailed distributions (see Figure 8 in Appendix), direct arithmetic fusion can be dominated by a few extreme tokens. EARF therefore projects both signals into a shared rank space within each keyframe:

$$R_{\mathcal{V}^{(l,f)}}(\phi, i) = \frac{\text{rank}_{\mathcal{V}^{(l,f)}}(\phi_i)}{\max(|\mathcal{V}^{(l,f)}| - 1, 1)}, \quad \phi \in \{A^{(l,f)}, M^{(f)}\}. \tag{8}$$

Here $\text{rank}_{\mathcal{V}^{(l,f)}}(\phi_i)$ denotes the zero-based ascending rank of token i among active visual tokens in $\mathcal{V}^{(l,f)}$, so larger scores receive larger normalized ranks. This non-parametric projection mitigates scale mismatch while preserving ordinal token importance.

Layer-wise Rank Calibration and Pruning. EARF computes the calibrated pruning score as:

$$S_{\text{calib}}^{(l,f,i)} = (1-\gamma_l) R_{\mathcal{V}^{(l,f)}}(A^{(l,f)}, i) + \gamma_l R_{\mathcal{V}^{(l,f)}}(M^{(f)}, i). \tag{9}$$

The weight γ_l primarily follows the bias reliability measured by Cohen's d (Figure 3(c)): high-d layers receive stronger event-guided correction, while deeper low-d layers use smaller γ_l to preserve content-dependent semantic attention.

Given the layer-wise retention ratio ρ_l, EARF keeps $K_l^{(f)} = \max(1, \lfloor \rho_l |\mathcal{V}^{(l,f)}| \rfloor)$ visual tokens with the highest calibrated scores in each keyframe:

$$\mathcal{I}_{\text{EARF}}^{(l,f)} = \arg \max_{\mathcal{I} \subseteq \mathcal{V}^{(l,f)},\ |\mathcal{I}| = K_l^{(f)}} \sum_{i \in \mathcal{I}} S_{\text{calib}}^{(l,f,i)}. \tag{10}$$

Visual tokens outside $\mathcal{I}_{\text{EARF}}^{(l,f)}$ are pruned, while text and special tokens are retained. By lowering the calibrated ranks of tokens whose attention scores are inflated by peripheral bias but lack event support, EARF makes such tokens less likely to be retained, reducing the presence of event-unsupported position-biased keys and values in the downstream KV cache. Appendix C provides a mechanistic interpretation of how EARF reduces event-unsupported position-biased keys and values in subsequent attention computation.

4 Experiments

We evaluate ECP on embodied spatial reasoning benchmarks using both simulated and real event streams. Our experiments address four questions: (1) Does ECP outperform existing pruning methods? (2) What are the computational gains? (3) Is ECP effective on both simulated and real events? (4) How does each component contribute?

4.1 Experimental Setup

Benchmarks. We focus on spatial reasoning during motion in 3D physical space and evaluate ECP in two event-stream settings. For public RGB-only benchmarks, we use VSI-Bench [49] (indoor navigation) and UrbanVideo-Bench [57] (outdoor drone scenarios). We focus on 13 ego-motion-centric tasks covering perception, cognition, and reasoning/planning, such as relative distance, proximity, route planning, and landmark positioning. Since these datasets lack native events, we generate simulated events offline with vid2e [18] from each original video at its native frame rate. ETS uses the full simulated event stream for temporal scoring, while EMSF/EARF only use event windows aligned with the retained RGB keyframes. To further validate effectiveness on native event signals, we introduce ESR-Real, a synchronized RGB-event benchmark for embodied spatial reasoning with over 700 QA pairs across 6 categories (see Figure 4 for distribution). Dataset construction details are provided in Appendix E.1.

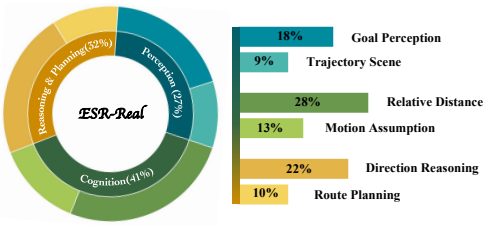

Figure 4: ESR-Real task distribution across Perception, Cognition, and Reasoning.

Baselines. We integrate ECP into Qwen2.5-VL-7B-Instruct [6] and compare with eight training-free baselines. Seven of them are RGB-only pruning methods: DTD [51] uses inter-frame differencing; FastV [10], PyramidDrop [48], and VTW [27] use attention-based pruning; DivPrune [3] uses diversity-based pruning; and BTP [25] combines attention saliency with diversity. VisionZip [50] compresses redundant visual-encoder tokens. We also include Direct Event Pruning (DEP), an event-only baseline that directly uses token-aligned event density maps as pruning saliency under the same token budgets. DEP controls for whether improvements come from event access alone. All methods are evaluated at 70%, 50%, and 20% final visual-token retention relative to the unpruned model.

Implementation. Experiments run on NVIDIA A6000 GPUs with FlashAttention-2 [14] acceleration. All methods are implemented on the same codebase for fair comparison. Following standard practice in VLM evaluation [23, 49, 6], we use greedy decoding for deterministic inference. Event

Table 1: **Accuracy comparison on embodied spatial reasoning benchmarks**. Avg.Acc is computed over all 13 selected tasks; six representative task columns are shown here and the remaining seven are reported in Appendix F. Best in bold. *Task definitions*: **SE.Pos**=Start/End Position; **Goal.Det**=Goal Detection; **Rel.Dis**=Relative Distance; **Proximity**=Proximity Estimation; **Ass.Reason**=Association Reasoning; **Landmark.Pos**=Landmark Positioning.

Method	Avg.Acc	Perception		Cognition		Reasoning & Planning	
		SE.Pos	Goal.Det	Rel.Dis	Proximity	Ass.Reason	Landmark.Pos
Original	36.31	55.10	39.79	44.51	50.00	16.90	37.41
70% Retaining							
DTD	36.36	55.10	39.44	42.63	46.88	15.96	37.76
FastV	36.28	57.14	39.08	44.51	50.00	16.90	37.41
PDrop	36.73	53.06	40.49	44.83	46.88	15.96	36.71
DivPrune	36.16	55.10	39.79	43.26	46.88	15.96	37.06
BTP	35.95	57.14	39.44	43.26	46.88	17.37	36.71
VTW	36.45	57.14	39.44	44.83	50.00	17.84	36.71
VisionZip	36.64	53.06	40.00	43.57	50.00	**19.25**	36.71
DEP	36.88	53.10	38.30	45.80	50.00	18.80	**37.80**
OURS	**37.78**	**59.18**	**40.49**	**49.22**	**56.25**	17.37	37.76
50% Retaining							
DTD	36.55	59.18	39.44	42.01	53.12	17.37	35.31
FastV	35.67	51.02	37.32	42.32	46.88	17.37	37.06
PDrop	35.67	51.02	39.79	41.69	46.88	16.43	**37.41**
DivPrune	36.77	48.98	39.79	44.51	46.88	18.31	35.31
BTP	35.09	51.02	40.49	40.13	50.00	15.96	35.66
VTW	33.84	42.86	33.45	38.87	53.12	14.08	34.97
VisionZip	36.54	53.06	41.38	41.69	53.12	**19.72**	37.06
DEP	33.28	38.80	32.80	38.20	53.10	15.00	34.60
OURS	**38.16**	**59.18**	**42.25**	**50.47**	**53.12**	18.78	37.06
20% Retaining							
DTD	35.37	51.02	37.68	39.50	50.00	16.43	35.31
FastV	32.64	28.57	29.58	33.23	46.88	14.55	31.82
PDrop	33.12	26.53	31.69	33.86	46.88	15.96	32.52
DivPrune	34.40	32.65	36.97	36.36	56.25	16.90	34.62
BTP	33.34	30.61	33.10	35.11	56.25	12.68	33.92
VTW	30.50	20.41	32.75	36.36	53.12	13.15	32.17
VisionZip	35.79	36.73	**40.34**	40.44	46.88	17.84	34.62
DEP	29.95	22.40	30.70	36.10	**59.40**	12.20	30.10
OURS	**37.62**	**55.10**	38.03	**44.20**	56.25	**20.19**	**36.36**

streams are aligned with visual tokens following the procedure described in Section 3.3. Pruning layers are determined by the layer-wise bias dynamics (Figure 3b): we intervene at stages preceding rapid bias escalation to preemptively correct attention drift before it propagates through the network. The fusion weight γ_l is modulated according to bias consistency (Figure 3c): layers with high and stable Cohen's d receive stronger event-guided motion prior correction, while layers exhibiting content-dependent variation preserve greater semantic flexibility. We report inference latency, GFLOPs, and peak memory. For RGB-only public benchmarks, offline vid2e synthesis is excluded from the reported inference cost.

4.2 Main Results

Accuracy Comparison. Table 1 shows that ECP maintains accuracy across the three retention ratios. At 70% retention, the large pool of retained tokens provides redundancy that can compensate for imperfect token selection, allowing existing methods to remain competitive. Under aggressive compression, this redundancy is reduced, making the selection of informative tokens more critical. At 20% retention, FastV and VTW drop to 32.64% and 30.50%, respectively, while ECP achieves 37.62%, exceeding the full-token baseline (36.31%). DEP obtains 33.28% at 50% retention and 29.95% at 20%, trailing ECP by 7.67 percentage points at the latter budget. ECP also exceeds VisionZip by 1.14–1.83 percentage points across the three retention ratios. The comparison with DEP, which also has access to events, supports the benefit of the proposed selection and fusion procedure beyond event access alone.

For Relative Distance, ECP achieves 44.20% at 20% retention, close to the full-token baseline (44.51%) and above the strongest pruning alternative, VisionZip (40.44%). For Start/End Position, ECP matches the full-token accuracy of 55.10%, while DEP obtains 22.40%. These comparisons suggest that calibrated event guidance preserves useful evidence for distance and position questions. Gains are smaller for Landmark Positioning, which can rely more on static scene context.

Table 2: **Accuracy comparison on ESR-Real benchmark.** ECP demonstrates consistent effectiveness on real event streams captured by physical sensors. Best in bold. *Task definitions*: **Goal.Percept**=Goal Perception; **Traj.Scene**=Trajectory Scene; **Rel.Dis**=Relative Distance; **Motion.Ass**=Motion Assumption; **Dir.Rea**=Direction Reasoning; **Route.Plan**=Route Planning.

Method	Avg.Acc	Perception		Cognition		Reasoning & Planning	
		Goal.Percept	Traj.Scene	Rel.Dis	Motion.Ass	Dir.Rea	Route.Plan
Original	51.90	63.28	31.75	69.35	39.56	30.82	63.77
70% Retaining							
DTD	51.20	**65.62**	28.57	67.34	35.16	30.82	66.67
FastV	51.48	62.50	34.92	68.34	37.36	30.82	63.77
PDrop	51.20	60.94	30.16	69.85	37.36	29.56	66.67
DivPrune	50.92	62.50	33.33	68.84	38.46	27.67	63.77
BTP	50.21	60.94	31.75	68.34	36.26	28.30	63.77
VTW	50.78	57.81	31.75	**69.85**	37.36	29.56	66.67
VisionZip	48.45	59.38	31.75	68.84	34.07	**31.45**	65.22
DEP	50.60	60.20	28.60	68.30	37.40	29.60	68.10
OURS	**52.89**	64.06	**38.10**	68.84	**39.56**	30.82	**68.12**
50% Retaining							
DTD	49.51	64.06	28.57	65.83	34.07	26.42	68.12
FastV	48.94	63.28	31.75	64.32	36.26	26.42	62.32
PDrop	50.49	63.28	33.33	65.83	37.36	29.56	63.77
DivPrune	50.21	65.62	30.16	64.82	40.66	27.67	62.32
BTP	50.21	63.28	34.92	64.32	38.46	28.30	65.22
VTW	43.58	64.84	26.98	47.24	31.87	28.30	59.42
VisionZip	49.05	60.16	33.33	65.33	37.36	31.45	66.67
DEP	41.50	64.80	27.00	42.70	27.50	27.00	59.40
OURS	**54.58**	**67.97**	**38.10**	**69.85**	**40.66**	**32.70**	**69.57**
20% Retaining							
DTD	46.40	66.41	25.40	55.78	34.07	27.04	62.32
FastV	39.77	66.41	20.63	37.19	34.07	25.79	55.07
PDrop	39.49	66.41	23.81	29.65	38.46	28.30	59.42
DivPrune	45.56	67.19	28.57	48.24	35.16	31.45	59.42
BTP	43.30	66.41	20.63	49.25	34.07	25.16	57.97
VTW	39.07	62.50	23.81	28.64	35.16	33.96	56.52
VisionZip	44.76	63.28	26.98	57.79	32.97	29.56	57.97
DEP	40.20	65.60	28.60	29.10	37.40	**34.60**	52.20
OURS	**50.07**	**67.19**	**31.75**	**60.30**	**38.46**	28.93	**69.57**

To validate effectiveness on real sensors, we evaluate on ESR-Real (Table 2). At 50% retention, ECP achieves 54.58%, surpassing the full-token baseline (51.90%) by 2.68 percentage points. Under more aggressive 20% retention, ECP remains the most robust pruning method, outperforming the second-best baseline (DTD: 46.40%) by 3.67 percentage points. The corresponding gains over VisionZip range from 4.44 to 5.53 percentage points. Additional evaluations on Video-MME and LLaVA-OneVision-1.5-8B are reported in Appendices I and J.

Efficiency Analysis. Table 3 reports online inference cost with event streams available either natively or precomputed. At 20% retention, ECP achieves 1.11s latency (1.89Œ speedup), 141.9 GFLOPs (52% reduction), and 16637 MB memory. Unlike DTDs pixel-wise differencing or DivPrunes $O(N^2)$ pairwise distances, ECPs sparse event accumulation provides motion-aware guidance with low overhead. For RGB-only public benchmarks, offline vid2e generation is used only to construct the simulated-event evaluation setting and is excluded from the reported inference cost.

Table 3: Efficiency comparison on the Urbanvideo-Bench (20% retention ratio).

Method	Ratio	Time (s)	GFLOPs ↓	Mem (MB)
Full Tokens	-	2.10	294.6	18848
DTD	20%	1.80	185.9	16880
FastV	20%	1.16	173.9	17068
VTW	20%	1.18	189.8	17200
PDrop	20%	1.24	187.5	17169
DivPrune	20%	1.25	187.1	17169
BTP	20%	1.55	187.5	17169
Ours	20%	**1.11**	**141.9**	**16637**

Qualitative Analysis. Figure 5 shows retained tokens under high compression for locating the sofa when standing by the TV and facing the chair. Attention-based pruning concentrates tokens near image borders, while diversity-based pruning leaves fragmented object boundaries (yellow circles); both incorrectly predict "Back right". ECP uses event-guided motion cues to retain more continuous chair contours and better coverage of the sofa (green circles), correctly predicting "Front Left". This example illustrates how preserving task-relevant geometry can support spatial reasoning with fewer tokens.

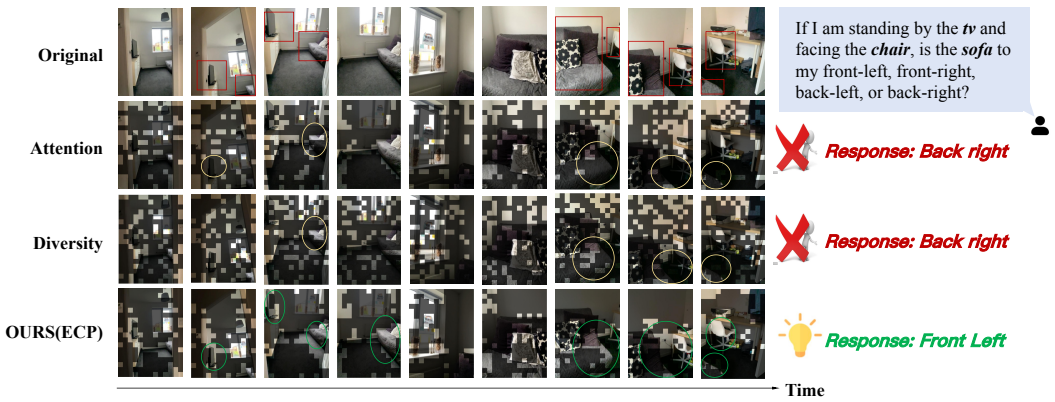

Figure 5: **Qualitative comparison.** ECP retains physical structures even at high token compression.

4.3 Ablation Studies

We conduct ablation studies to validate each component and design choice, examining three aspects: module contribution, fusion weight strategy, and layer selection.

Component Contribution. Table 4 reports component ablations averaged over the three retention ratios. ETS mainly improves efficiency, reducing latency from 2.08s to 1.58s while slightly improving accuracy by 0.21 percentage points. EMSF and EARF improve accuracy by 0.91 and 0.84 percentage points individually. Adding ETS to EMSF or EARF further improves accuracy, while EMSF+EARF without ETS is less effective, suggesting that temporal anchoring is important for coherent event-attention fusion. The full cascade achieves the best measured trade-off, with 37.85% accuracy and 1.16s latency. Appendix G provides task-category breakdowns.

Table 4: Component ablation.

ETS	EMSF	EARF	Lat.(s)	Acc.(%)
-	-	-	2.08	36.31
✓	-	-	1.58	36.52
-	✓	-	1.83	37.22
-	-	✓	1.53	37.15
✓	✓	-	1.47	37.51
✓	-	✓	1.17	37.75
-	✓	✓	1.59	37.00
✓	✓	✓	**1.16**	**37.85**

Fusion Weight Analysis. Table 5 examines the layer-wise fusion weight γ_l with pruning layers fixed. Three trends emerge: (1) Increasing the uniform event weight improves accuracy from 34.18% to 37.47% up to $\gamma = 0.7$, while event-only rank fusion ($\gamma = 1.0$) slightly drops to 37.36%, indicating that semantic attention remains complementary. (2) The shallow-heavy Decay pattern ($0.8 \to 0.5 \to 0.3$) outperforms Growth ($0.3 \to 0.5 \to 0.8$) by 1.05 percentage points. (3) The bias-guided schedule $(0.8, 0.6, 0.5)$ achieves the best accuracy of 37.62%. This aligns with Figure 3(c): high-d shallow layers receive stronger event-guided calibration, whereas lower-d deeper layers use smaller γ_l to preserve content-dependent semantic attention.

Table 5: Layer-wise fusion-weight ablation at 20% retention (S/M/D: shallow/mid/deep).

Strategy	γ_S	γ_M	γ_D	Acc.(%)	Analysis
Uniform	0.0	0.0	0.0	34.18	Baseline (Attn)
	0.3	0.3	0.3	35.66	Weak Fusion
	0.5	0.5	0.5	36.21	Medium Fusion
	0.7	**0.7**	**0.7**	**37.47**	**Strong Fusion**
	1.0	1.0	1.0	37.36	Pure Event
Peak ($\wedge$)	0.2	0.6	0.2	34.60	Mid-focus
Valley ($\vee$)	**0.6**	0.2	0.6	36.32	Ends-focus
Growth ($\nearrow$)	0.3	0.5	0.8	35.99	Suboptimal
Decay ($\searrow$)	**0.8**	0.5	0.3	37.04	**Supports Hyp.**
Best	**0.8**	**0.6**	**0.5**	**37.62**	**Best Balance**

Layer Selection Strategy. Table 6 compares bias-guided and uniformly spaced pruning layers with identical ECP modules and final token budgets. Uniform spacing is slightly better at 70% retention (38.10 vs. 37.78), where token redundancy can mask suboptimal layer placement. From 70% to 20%, accuracy drops 3.24 percentage points for uniform spacing but only 0.16 for bias-guided placement. Bias-guided placement performs better at both 50% and 20%, supporting early intervention to limit peripheral-bias propagation through the residual stream and KV cache when redundancy is limited.

Table 6: Layer selection ablation.

Ratio	Variant	Avg.	Urban	VSI
Full	Baseline	36.31	40.86	32.03
70%	Uniform	**38.10**	**42.91**	33.13
70%	**Ours**	37.78	42.49	**33.35**
50%	Uniform	37.09	42.56	29.88
50%	**Ours**	**38.16**	**42.85**	**33.76**
20%	Uniform	34.86	39.99	29.77
20%	**Ours**	**37.62**	**42.78**	**32.76**

5 Conclusion and Limitations

We presented Event Cascade Pruning (ECP), a training-free event-assisted token pruning framework for embodied spatial reasoning. ECP combines event-guided frame selection, spatial filtering, and attention-rank calibration to retain motion and geometric evidence under compression. At 20% token retention, ECP surpasses the full-token baseline (37.62% vs. 36.31%) with $1.89\times$ speedup and 52% GFLOPs reduction. We also introduced ESR-Real, a synchronized RGB-event benchmark for embodied spatial reasoning, where ECP improves accuracy by 2.68 percentage points at 50% retention. Evaluations with simulated and native events support the effectiveness of event-assisted token selection. More broadly, this work illustrates how event sensing can guide efficient visual computation in embodied Video-LLMs.

Limitations. ECP assumes synchronized RGB-event inputs; it targets platforms where an event camera is paired with RGB sensing, rather than serving as an RGB-only online acceleration method. Although event sensors remain less ubiquitous than RGB cameras, commercialization by Sony and Prophesee makes this setting increasingly feasible. More broadly, ECP illustrates sensor-assisted pruning and motivates future external-prior-guided or RGB-only distilled pruning schemes.

Acknowledgments and Disclosure of Funding

This paper was supported by the Natural Science Foundation of China under Grant 62371269, Shenzhen Low-Altitude Airspace Strategic Program Portfolio (Grant No. Z25306110), Shenzhen Science and Technology Program ZDCY202517012, Tsinghua University-Toyota Motor Corporation Joint Research Center for AI Technology of Automated Vehicle, Meituan Academy of Robotics Shenzhen.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025.

[4] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

[7] Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. Eagle: Egocentric aggregated language-video engine. *arXiv preprint arXiv:2409.17523*, 2024.

[8] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[9] Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermuller, and Yezhou Yang. Recent event camera innovations: A survey. In *European conference on computer vision*, pages 342–376. Springer, 2024.

[10] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

[11] Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2061–2070, 2023.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[13] Ronghao Dang, Yuqian Yuan, Yunxuan Mao, Kehan Li, Jiangpin Liu, Zhikai Wang, Xin Li, Fan Wang, and Deli Zhao. Rynnec: Bringing mllms into embodied world. *arXiv preprint arXiv:2508.14160*, 2025.

[14] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

[15] Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6342–6352. IEEE, 2025.

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[17] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[18] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3586–3595, 2020.

[19] Ruishan Guo, Ciyu Ruan, Haoyang Wang, Zihang Gong, Jingao Xu, and Xinlei Chen. x^2-fusion: Cross-modality and cross-dimension flow estimation in event edge space. *arXiv preprint arXiv:2603.16671*, 2026.

[20] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19959–19973, 2025.

[21] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 13358–13376, 2023.

[22] Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models–from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[25] Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. Balanced token pruning: Accelerating vision language models beyond local optimization. *arXiv preprint arXiv:2505.22038*, 2025.

[26] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

[27] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342, 2025.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[30] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29139–29149, 2025.

[31] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43 (6):1964–1980, 2019.

[33] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3857–3866, 2019.

[34] Ciyu Ruan, Zihang Gong, Ruishan Guo, Jingao Xu, and Xinlei Chen. Edmamba: Rethinking efficient event denoising with spatiotemporal decoupled ssms. *arXiv preprint arXiv:2505.05391*, 2025.

[35] Ciyu Ruan, Ruishan Guo, Zihang Gong, Jingao Xu, Wenhan Yang, and Xinlei Chen. Pre-mamba: A 4d state space model for ultra-high-frequent event camera deraining. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9169–9180. IEEE, 2025.

[36] Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, Mohamed Elnoor, Anuj Zore, Brian Ichter, Fei Xia, Jie Tan, Wenhao Yu, and Dinesh Manocha. Convoi: Context-aware navigation using vision language models in outdoor and indoor environments. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13837–13844. IEEE, 2024.

[37] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[38] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11101–11122, 2024.

[39] Jacob Thompson, Emiliano Garcia-Lopez, and Yonatan Bisk. Rem: Evaluating llm embodied spatial reasoning through multi-frame trajectories. *arXiv preprint arXiv:2512.00736*, 2025.

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: open foundation and fine-tuned chat models. arxiv. *arXiv preprint arXiv:2307.09288*, 10, 2023.

[41] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018.

[42] Haoyang Wang, Ruishan Guo, Pengtao Ma, Ciyu Ruan, Xinyu Luo, Wenhua Ding, Tianyang Zhong, Jingao Xu, Yunhao Liu, and Xinlei Chen. Event camera meets mobile embodied perception: abstraction, algorithm, acceleration, application. *ACM Computing Surveys*, 58(8):1–41, 2026.

[43] Haoyang Wang, Jingao Xu, Xinyu Luo, Ting Zhang, Xuecheng Chen, Ruiyang Duan, Yunhao Liu, Jianfeng Zheng, Weijie Hong, Xiaoqiang Ji, et al. mme-loc: Facilitating accurate drone landing with ultra-high-frequency localization. *IEEE Transactions on Mobile Computing*, 2026.

[44] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[45] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947, 2024.

[46] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023.

[47] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. *URL https://arxiv. org/abs/2309.17453*, page 1, 2023.

[48] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

[49] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[50] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19792–19802. IEEE, 2025.

[51] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10807–10816, 2025.

[52] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025.

[53] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024.

[54] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[55] Weichen Zhang, Zile Zhou, Xin Zeng, Liu Xuchen, Jianjie Fang, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3d-vqa: A benchmark for embodied spatial concept reasoning with multimodal large language model in open space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12784–12791, 2025.

[56] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[57] Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, et al. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32400–32423, 2025.

[58] Baining Zhao, Ziyou Wang, Jianjie Fang, Chen Gao, Fanhang Man, Jinqiang Cui, Xin Wang, Xinlei Chen, Yong Li, and Wenwu Zhu. Embodied-r: Collaborative framework for activating embodied spatial reasoning in foundation models via reinforcement learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11071–11080, 2025.

[59] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *European Conference on Computer Vision*, pages 477–494. Springer, 2024.

[60] Zhiting Zhou, Xingchen Liu, Xinglin Yu, Jiashen Chen, Haoyang Wang, Jingao Xu, Yunhao Liu, and Xinlei Chen. Eventradar: Long-range visual uav discovery through spatiotemporal event sensing. *arXiv preprint arXiv:2606.11285*, 2026.

[61] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018.

[62] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 989–997, 2019.

A Spatial Attention Bias Analysis

This appendix provides statistical analysis of the spatial attention bias in vision-language models. We analyze 21,920 frames across all 28 transformer layers from UrbanVideo-Bench (472 videos, 15,104 frames) and VSI-Bench (213 videos, 6,816 frames).

A.1 Spatial Region Definition

All attention patterns are standardized onto a 12×18 spatial grid (216 tokens). As illustrated in Figure 6, we partition this grid into three regions using a 15% margin threshold: Corner (8 tokens, 3.7%), Edge (68 tokens, 31.5%), and Center (140 tokens, 64.8%). The Peripheral region (Corner $\cup$ Edge, 76 tokens, 35.2%) serves as the primary metric for quantifying spatial attention bias.

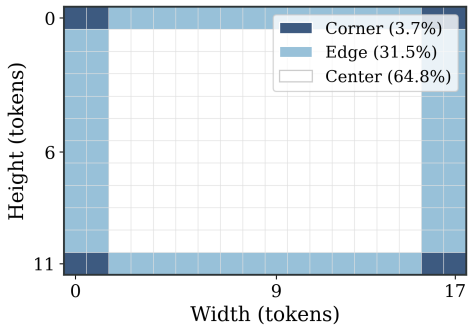

Figure 6: Spatial region partitioning. Peripheral = Corner (3.7%) + Edge (31.5%); Center (64.8%) serves as baseline.

A.2 Statistical Analysis

Table 7 provides complete per-layer statistics; Table 8 summarizes bias characteristics by layer group.

Table 7: Complete layer-wise statistics ($n = 21{,}920$). t: one-sample t-statistic against $\mu_0 = 1.0$; all layers satisfy $p < 10^{-10}$.

L	Corner	Edge	Periph.	d	t	L	Corner	Edge	Periph.	d	t
0	7.70	1.80	2.41	2.24	331	14	13.70	4.90	5.82	0.99	146
1	7.00	1.80	2.31	2.09	310	15	9.00	3.30	3.87	1.23	182
2	7.60	1.90	2.48	1.88	279	16	12.60	5.50	6.25	0.80	118
3	12.30	2.70	3.75	**2.30**	340	17	9.70	3.50	4.20	1.08	160
4	9.20	2.30	3.05	2.01	297	18	4.00	2.10	2.33	1.04	154
5	16.80	3.20	4.66	1.99	295	19	3.80	2.40	2.57	0.91	134
6	6.40	1.70	2.18	1.56	231	20	5.00	2.60	2.84	1.10	163
7	20.30	4.20	5.92	1.72	254	21	8.10	2.70	3.23	1.11	165
8	19.40	3.60	5.26	1.73	256	22	10.20	2.90	3.64	1.00	148
9	**47.70**	5.00	**9.53**	1.35	200	23	7.30	2.40	2.92	1.11	164
10	18.90	3.20	4.89	1.59	235	24	4.40	1.80	2.11	1.01	150
11	18.00	4.50	5.89	1.41	209	25	6.30	2.30	2.72	1.04	154
12	12.40	4.10	4.95	1.26	187	26	5.70	2.10	2.47	1.04	153
13	8.00	3.90	4.31	0.95	141	27	5.90	2.00	2.44	0.98	145

Key findings from Table 7: (1) All 28 layers show significant peripheral bias ($p < 10^{-10}$). (2) Layer 9 exhibits peak ratio ($9.53\times$); Layer 3 achieves highest d (2.30). (3) Mean ratio across layers is $3.89\times$. (4) Cross-dataset correlation $r = 0.897$ ($p = 1.05 \times 10^{-10}$) shows similar layer-wise profiles across the evaluated datasets.

Table 8: Summary statistics by layer group, computed by averaging unrounded per-layer values.

Group	Layers	Periph.	Corner	Edge	d
Shallow	0–7	3.35×	10.92×	2.45×	1.97
Middle	8–16	5.64×	17.75×	4.21×	1.26
Deep	17–27	2.86×	6.42×	2.44×	1.04

A.3 Effect Size Interpretation

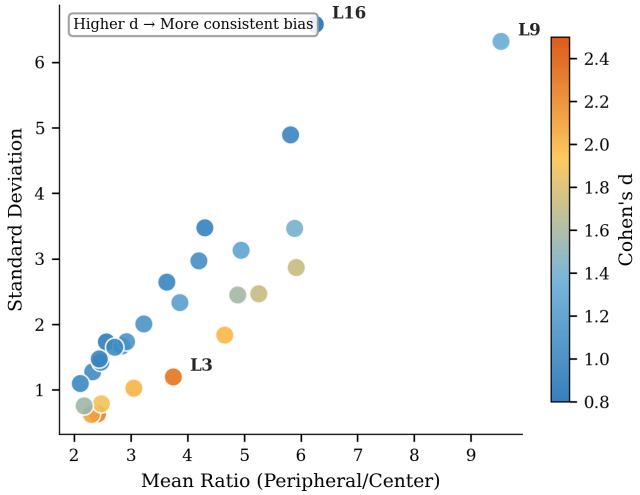

Figure 7: Bias magnitude vs. consistency. Cohen's $d = (\mu - 1)/\sigma$ penalizes variance: Layer 9 has peak ratio (9.53×) but moderate d (1.35); Layer 3 achieves highest d (2.30) with lower ratio (3.75×).

Figure 7 shows why high bias magnitude does not imply high effect size. Layers with high Cohen's d provide more reliable pruning decisions due to consistent bias patterns.

A.4 Design Implications

Pruning Layer Selection. Effective pruning layers should satisfy two criteria: (1) intervene before bias escalation to prevent error propagation, and (2) exhibit high Cohen's d for reliable decisions. The ablation study empirically validates this strategy.

Fusion Weight γ_l. Shallow layers exhibit the highest consistency ($d = 1.97$), making event-based calibration most reliable there. We adopt a decay schedule where γ_l decreases from shallow to deep layers. The ablation study empirically validates this strategy.

B Long-tailed Distribution Analysis

This appendix analyzes the distributional properties of attention scores and event density, which motivates our rank-based fusion strategy in EARF (Section 3.4).

As shown in Figure 8, both attention scores and event density exhibit severe long-tailed distributions: a small fraction of tokens dominate the score mass while the majority cluster near zero. This distributional skewness poses a fundamental challenge for multimodal fusion—direct arithmetic combination (e.g., weighted sum) would be dominated by outliers from either modality, failing to capture the relative importance across the full token population. Our rank-based projection normalizes these skewed distributions into uniform ranks, enabling fair comparison and robust fusion.

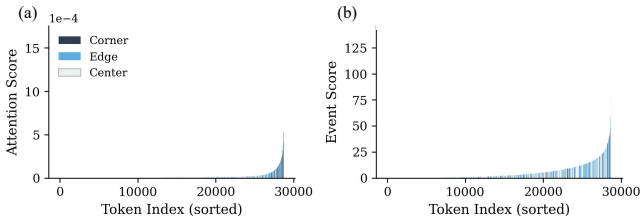

Figure 8: Long-tailed distribution of attention scores and event density. Both modalities exhibit severe skewness, motivating rank-based fusion.

C Mechanistic Analysis of KV-Cache Purification

This appendix provides an analytical interpretation of how event-unsupported peripheral sink tokens can degrade downstream key–value states in Video-LLMs, and how EARF mitigates this issue through event-assisted rank calibration. We use "KV cache" broadly to refer to the active keys and values retained for subsequent attention computation.

C.1 Problem: Peripheral Sink Tokens in the KV Cache

For clarity, we focus on one pruning layer l. EARF performs score calibration and top-K pruning within each keyframe, but the retained tokens are then merged into the active sequence for subsequent multimodal self-attention. Let $\mathcal{T}^{(l)}$ denote the always-retained text and special tokens, and let $\mathcal{V}^{(l)} = \bigcup_f \mathcal{V}^{(l,f)}$ denote all active visual tokens before pruning, where $\mathcal{V}^{(l,f)}$ is the active visual-token set of keyframe f. The active key–value set is:

$$\mathcal{K}^{(l)} = \mathcal{T}^{(l)} \cup \mathcal{V}^{(l)}. \tag{11}$$

For a query vector q, attention over the global active set $\mathcal{K}^{(l)}$ produces:

$$z_q^{(l)} = \sum_{i \in \mathcal{K}^{(l)}} \alpha_{q,i}^{(l)} v_i^{(l)}, \quad \alpha_{q,i}^{(l)} = \frac{\exp\left(q^\top k_i^{(l)} / \sqrt{d_k}\right)}{\sum_{j \in \mathcal{K}^{(l)}} \exp\left(q^\top k_j^{(l)} / \sqrt{d_k}\right)}, \tag{12}$$

where d_k denotes the key dimension. Pruning changes only the visual-token subset of $\mathcal{K}^{(l)}$, while text and special tokens are always kept.

As demonstrated in Appendix A, visual attention in the evaluated Video-LLM exhibits a consistent peripheral sink: border and corner tokens receive inflated attention scores relative to central tokens. Let $\mathcal{P}^{(l,f)} \subseteq \mathcal{V}^{(l,f)}$ denote event-unsupported peripheral visual tokens in keyframe f, and let $\mathcal{P}^{(l)} = \bigcup_f \mathcal{P}^{(l,f)}$. Define $\mathcal{S}^{(l)} = \mathcal{K}^{(l)} \setminus \mathcal{P}^{(l)}$ as the remaining active tokens. The attention output can be decomposed as:

$$z_q^{(l)} = \sum_{i \in \mathcal{S}^{(l)}} \alpha_{q,i}^{(l)} v_i^{(l)} + \sum_{j \in \mathcal{P}^{(l)}} \alpha_{q,j}^{(l)} v_j^{(l)}. \tag{13}$$

The first sum gives the contribution of the remaining active tokens, and the second that of event-unsupported peripheral tokens. When attention-based pruning preferentially retains tokens in $\mathcal{P}^{(l)}$, these tokens occupy the limited visual-token budget and continue contributing keys and values to later attention computation. Under high compression, each such token also displaces a potentially structure-relevant visual token, weakening the retained visual context.

C.2 Mechanism: Cascading Degradation Across Layers

The degradation can propagate through the network. At pruning layer l, if the retained visual-token set contains many event-unsupported peripheral tokens, their values contribute to the layer output and their representations are forwarded through residual connections and feed-forward blocks. Subsequent attention layers then operate on this biased retained set. If later pruning decisions are again based mainly on attention scores, the same peripheral sink tokens can be repeatedly favored, causing the error to compound.

This mechanism is most harmful at aggressive retention ratios, where the token budget is too small to mask suboptimal token placement. This is consistent with the empirical results in Table 1: attention-based baselines such as FastV and PyramidDrop degrade substantially at 20% retention, whereas ECP maintains robust accuracy by incorporating event-guided calibration.

C.3 Solution: EARF as a Purification Mechanism

EARF mitigates this issue by using event saliency only at the score-and-mask level, without encoding events as additional hidden states. For pruning layer l and keyframe f, EARF uses the text-conditioned attention score $A_i^{(l,f)}$ from the preceding head-averaged post-softmax multimodal self-attention map, following the main-text definition, and the EMSF event saliency $M_i^{(f)}$ for the same visual token. Although attention is computed over the global active sequence, EARF calibrates and ranks visual tokens within each keyframe set $\mathcal{V}^{(l,f)}$, matching the implementation. To avoid scale mismatch, both signals are projected into a shared per-keyframe rank space:

$$R_{\mathcal{V}^{(l,f)}}(\phi, i) = \frac{\mathrm{rank}_{\mathcal{V}^{(l,f)}}(\phi_i)}{\max(|\mathcal{V}^{(l,f)}| - 1, 1)}, \quad \phi \in \{A^{(l,f)}, M^{(f)}\}. \tag{14}$$

The calibrated EARF score is:

$$S_{\mathrm{calib}}^{(l,f,i)} = (1 - \gamma_l) R_{\mathcal{V}^{(l,f)}}(A^{(l,f)}, i) + \gamma_l R_{\mathcal{V}^{(l,f)}}(M^{(f)}, i). \tag{15}$$

Consider an event-unsupported peripheral sink token j with high attention rank but weak event support:

$$R_{\mathcal{V}^{(l,f)}}(A^{(l,f)}, j) \approx 1, \quad R_{\mathcal{V}^{(l,f)}}(M^{(f)}, j) \approx 0.$$

Attention-only pruning would assign this token a near-maximal score. In EARF, however, its calibrated score becomes:

$$S_{\mathrm{calib}}^{(l,f,j)} \approx 1 - \gamma_l. \tag{16}$$

By contrast, an event-supported token u with comparable attention rank and high event rank obtains:

$$S_{\mathrm{calib}}^{(l,f,u)} \approx 1.$$

More generally, the score gap between an event-supported token u and an event-unsupported token j in the same keyframe is:

$$S_{\mathrm{calib}}^{(l,f,u)} - S_{\mathrm{calib}}^{(l,f,j)} = (1 - \gamma_l)\Delta R_A + \gamma_l \Delta R_M, \tag{17}$$

where $\Delta R_A = R_{\mathcal{V}^{(l,f)}}(A^{(l,f)}, u) - R_{\mathcal{V}^{(l,f)}}(A^{(l,f)}, j)$ and $\Delta R_M = R_{\mathcal{V}^{(l,f)}}(M^{(f)}, u) - R_{\mathcal{V}^{(l,f)}}(M^{(f)}, j)$. When the two tokens have comparable attention ranks and their event ranks differ substantially, the margin is dominated by the event term and is approximately γ_l. Importantly, EARF does not simply penalize all peripheral tokens: a peripheral token with strong event support can still be retained. The mechanism specifically targets tokens whose attention rank is inflated but whose event rank is low.

Given the layer-wise retention budget $K_l^{(f)}$, EARF keeps the visual-token set $\mathcal{I}_{\mathrm{EARF}}^{(l,f)}$ with the highest calibrated scores in each keyframe. After per-keyframe pruning, the retained visual tokens are merged across keyframes, yielding:

$$\widetilde{\mathcal{V}}^{(l)} = \bigcup_f \mathcal{I}_{\mathrm{EARF}}^{(l,f)}, \quad \widetilde{\mathcal{K}}^{(l)} = \mathcal{T}^{(l)} \cup \widetilde{\mathcal{V}}^{(l)}. \tag{18}$$

For any pruned visual token $j \notin \widetilde{\mathcal{V}}^{(l)}$, its key and value are no longer forwarded to subsequent layers. The attention output is therefore computed over a purified active set:

$$\widetilde{z}_q^{(l)} = \sum_{i \in \widetilde{\mathcal{K}}^{(l)}} \widetilde{\alpha}_{q,i}^{(l)} v_i^{(l)}, \quad \widetilde{\alpha}_{q,i}^{(l)} = \frac{\exp\left(q^\top k_i^{(l)} / \sqrt{d_k}\right)}{\sum_{r \in \widetilde{\mathcal{K}}^{(l)}} \exp\left(q^\top k_r^{(l)} / \sqrt{d_k}\right)}. \tag{19}$$

This removes event-unsupported position-biased keys and values from downstream attention, increasing the relative proportion of tokens that carry semantic, motion, or geometric support.

C.4 Connection to Attention Sink Literature

Prior work on efficient LLMs identified that certain tokens in autoregressive models can act as "attention sinks", accumulating attention mass regardless of semantic relevance. Our analysis reveals an analogous spatial phenomenon in vision-language models: peripheral visual tokens can serve as two-dimensional attention sinks due to positional biases.

The key difference is that event streams provide an external motion/structure cue for distinguishing structure-supported tokens from event-unsupported peripheral artifacts. Unlike text-only settings where sink tokens must be detected from internal attention statistics alone, EARF leverages token-aligned event saliency as a training-free prior to calibrate pruning decisions. This external prior reduces the downstream presence of event-unsupported position-biased keys and values while preserving semantic attention through the layer-wise weight γ_l.

D Simulated Event Generation Protocol

For RGB-only public benchmarks, we generate simulated events offline with vid2e from each original source video at its native frame rate. If a video is recorded at 30 fps, vid2e is run on the full 30-fps sequence and synthesizes events between adjacent original frames by triggering events when interpolated logarithmic brightness changes exceed the contrast thresholds. We use the standard setting $c_p = 0.2$, $c_n = 0.2$, and $t_{\text{ref}} = 0$, where c_p and c_n are the positive and negative contrast thresholds and t_{ref} is the refractory period. This offline synthesis is used only to construct paired RGB-event inputs for benchmarks without native events. For ESR-Real, events are captured directly by the Prophesee EVK4 hardware and do not require vid2e simulation.

E ESR-Real Dataset

To comprehensively evaluate the embodied cognition capabilities of models during motion, we designed a structured benchmark consisting of three primary dimensions: *Perception, Cognition*, and *Reasoning & Planning*. Each cognitive ability is manifested through specific sub-tasks. We provided handcrafted question prototypes for each task to ensure the evaluation captures the model's spatial reasoning rather than linguistic patterns. The task set logic is as follows:

- **Perception:** This level evaluates the model's foundational ability to recognize environmental goals (Goal Perception) and maintain visual continuity during camera movement (Trajectory Scene).
- **Cognition:** Moving beyond raw vision, these tasks measure the model's capacity to internalize spatial metrics, such as calculating relative proximity (Relative Distance) and predicting metric changes during dynamic actions (Motion Assumption).
- **Reasoning & Planning:** This is the highest level of embodied intelligence, requiring the model to resolve orientation-based logic (Direction Reasoning) and generate strategic action sequences based on current orientation and target goals (Route Planning).

E.1 Hardware Setup

ESR-Real is captured using a synchronized RGB-event camera rig consisting of two sensors rigidly mounted on a custom handheld platform:

Event Camera: Prophesee EVK4 HD, equipped with the Sony IMX636ES event-based vision sensor. Key specifications include: 1280×720 pixel resolution, $4.86\,\mu$m pixel pitch, $>$120 dB dynamic range, microsecond-level temporal resolution (equivalent to $>$10,000 fps), and low-light sensitivity down to 0.08 lux. The asynchronous pixel-level readout enables motion capture without motion blur, making it ideal for embodied spatial reasoning under ego-motion.

RGB Camera: Hikvision MV-CS050-10UC industrial camera, featuring the Sony IMX264 global shutter CMOS sensor. Specifications: 2448×2048 pixel resolution (5 MP), $3.45\,\mu$m pixel pitch, up to 60 fps frame rate, and USB 3.0 interface. The global shutter eliminates rolling shutter artifacts, ensuring geometric consistency with the event stream.

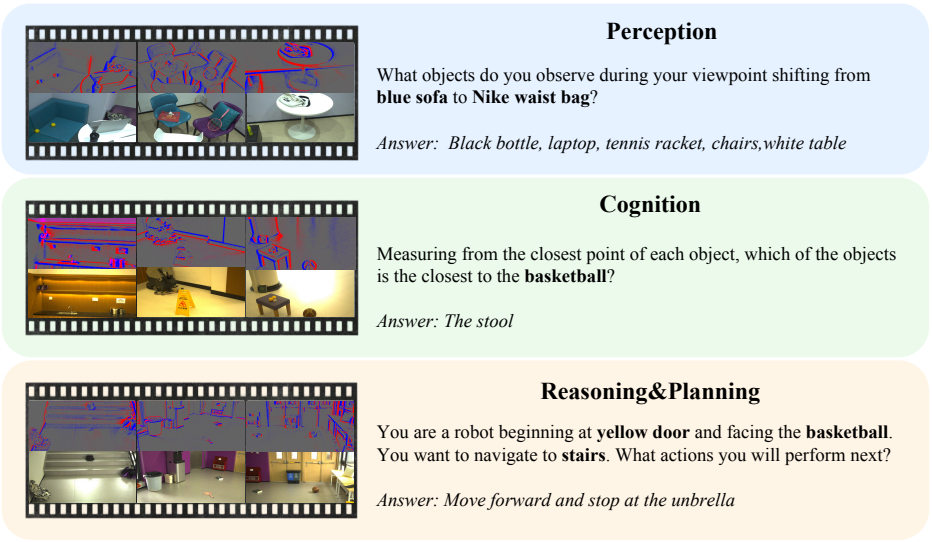

Figure 9: ESR-Real dataset overview

Table 9: **ESR-Real task set overview.**

Perception	
Goal Perception	At the initial position, you are asked to navigate to the {target_object}. When you reach the last frame of the video, is the destination within view at your current location?
Trajectory Scene	What objects do you observe during your viewpoint shifting along {trajectory_reference}
Cognition	
Relative Distance	Measuring from the closest point of each object, which of these objects is the closest to the {target_object}?
Motion Assumption	How does the distance between you and the {target_object} change during your {action_description}?
Reasoning & Planning	
Direction Reasoning	You're a robot standing by the center of {position_object} and facing the center of {orient_object}, is the center of {query_object} to my left, right, or back? (Back defined as $\geq$ 135 degrees turn)
Route Planning	You are a robot beginning at {position_object} and facing the {orient_object}. You want to navigate to {end_object}. What actions you will perform next?

Both cameras are rigidly mounted on a 3D-printed bracket with a baseline of approximately 5 cm, enabling synchronized capture of the same scene from nearly co-located viewpoints.

E.2 Stereo Calibration and Alignment

To project RGB frames and event streams into a common coordinate system, we perform stereo calibration. The calibration procedure consists of three stages:

Camera Calibration: We employ Zhang's flexible calibration method [56] for both intrinsic and extrinsic parameter estimation. For the RGB camera, we capture 20–30 images of a 4×7 checkerboard pattern (square size: 80 mm) at various orientations and distances. Corner detection is performed using sub-pixel refinement, and the camera parameters are estimated via nonlinear optimization. For the event camera, we use Prophesee's Metavision SDK calibration pipeline with a blinking checker-

board pattern displayed on a monitor. The blinking pattern generates consistent events at pattern boundaries, enabling reliable corner detection directly from the event stream.

Stereo Calibration: To estimate the extrinsic parameters between the RGB and event cameras, both sensors simultaneously capture the same 4×7 checkerboard pattern. Since event cameras output asynchronous event streams rather than intensity images, we first reconstruct intensity frames from the event stream using E2VID [33]. Using the detected corners from both the RGB frames and E2VID-reconstructed event frames, we estimate the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ that define the rigid transformation between the two camera coordinate systems.

Event-RGB Alignment: Using the calibrated stereo parameters, we warp RGB frames into the event camera's coordinate system. Event streams remain in their native resolution (1280×720 pixels), while RGB frames are projected onto the event image plane to achieve pixel-level alignment.

E.3 Data Collection Protocol

Data collection was conducted across diverse indoor environments including offices, laboratories, corridors, and common areas. Each recording session involves handheld traversal with natural ego-motion (walking, turning, approaching objects), with sequence durations ranging from 50.1 to 153.3 seconds. RGB video and continuous event streams are recorded simultaneously, followed by post-capture annotation of spatial reasoning QA pairs by the internal research team.

The final ESR-Real benchmark comprises over 700 QA pairs across 6 task categories, covering perception, cognition, and reasoning capabilities essential for embodied AI systems.

E.4 MCQ Generation

This section details our automated annotation pipeline, designed to scale the production of high-quality spatial reasoning benchmarks. The objective is to leverage the robust visual perception and logical reasoning capabilities of advanced Large Multimodal Models (LMMs) to transform raw egocentric videos into structured, logic-grounded datasets.

We have designed a multi-stage Chain-of-Thought (CoT) framework tailored for embodied spatial understanding, consisting of the following two primary stages:

Stage 1: Translating visual information into a comprehensive topological and spatial metadata map.

Object Segmentation and Disambiguation: The model identifies the 20 most salient and identifiable objects within the video. To prevent ambiguity, a strict "Unique Naming" protocol is enforced. If multiple objects of the same category exist (e.g., two monitors), they are assigned unique identifiers based on objective visual attributes (e.g., "monitor by the window") while avoiding unquantifiable subjective adjectives (e.g., "nice", "big").

Fine-grained Spatial Mapping: The system generates over 25 detailed spatial descriptions covering relative orientations, distances, and overall scene layout.

Proximity Identification: All object pairs within a 1-meter radius are explicitly mapped as "proximity pairs" to serve as constraints for subsequent logic generation.

Stage 2: Logic-driven QA Synthesis–Populate complex reasoning templates with the structured metadata.

Spatial Reasoning Logic Constraints: To ensure that questions require high-level reasoning rather than simple local associations, we implement a "Non-Proximity Constraint." For instance, objects selected for Route Planning and Scene Recall must not be proximity pairs, forcing the model to reason over extended trajectories.

Automated Template Population: The system systematically varies reference frames (robot position and orientation) to generate diverse perspectives. Rigorous filters are applied to ensure that the correct options (ground truth) are derived directly from the Phase 1 metadata and that distractors are logically distinct.

By utilizing Gemini-3Pro at this stage, we ensure that the generated questions are grounded in a deep understanding of the 3D environment, providing a challenging and high-fidelity benchmark for embodied AI agents.

E.5 Human Refinement

The generated MCQs may contain invalid questions, ambiguous options, incorrect answers, and various other issues that require further human refinement. These issues stem from two main sources: a) Even the most advanced Video-LLMs lack the ability to fully understand fine-grained embodied movements within confined indoor spaces. b) The indoor scenarios are densely populated and spatially complex, making dynamic video comprehension challenging. We approach the human refinement process from four aspects:

a) **Invalid/ambiguous questions:** For example, in Goal Perception tasks, an instruction to "navigate to the chair" leads to ambiguity when multiple chairs are present in a living room. The navigation target should be clarified to "navigate to the wooden armchair near the window" to ensure precise target identification.

b) **Object hallucination:** This refers to the presence of furniture or indoor elements in the correct option that were never actually captured during the Trajectory Scene (e.g., mentioning a lamp that was occluded or absent).

c) **Spatial and metric inaccuracies:** Descriptions involving Direction Reasoning or Relative Distance are often incorrect or imprecise due to the constantly changing camera perspective during motion.

d) **Choices with insufficient differentiation or logic errors:** In Motion Assumption and Route Planning tasks, one scenario is that distractors are too similar to the correct action sequence. Another is that the ground truth failed to represent the optimal strategic action based on the agent's current orientation.

The entire refinement process required over 200 person-hours.

Statistical reliability. To account for correlations among QA pairs from the same recording, we compare ECP and the full-token baseline using 10,000 paired cluster-bootstrap resamples over recordings. At 50% token retention, the micro-accuracy gain is 2.68 percentage points (95% CI: $[+0.43, +4.87]$); the macro-accuracy gain, with equal weighting of the six tasks, is 3.39 points (95% CI: $[+1.05, +5.59]$). Leave-one-recording-out gains range from 2.24 to 3.11 points, remaining positive for every omitted recording.

F Remaining Results of Effectiveness Experiment

In the main text, we report six representative task columns from UrbanVideoBench and VSI-Bench due to space. This appendix reports the remaining seven task columns; together with Table 1, the results cover all 13 selected tasks used to compute Avg.Acc. Table 10 shows that ECP is particularly stable under aggressive compression. When the retention rate drops to 20%, VTW decreases to 30.50%, whereas ECP maintains 37.62%, close to its 70% retention result. This trend suggests that event-assisted pruning helps preserve motion- and geometry-relevant evidence when the token budget becomes sparse. The remaining task columns show the following patterns:

a) Perception: ECP remains strong on Object Recall and Scene Recall. At 20% retention, it achieves 73.68% in Object Recall, higher than the full-token baseline's 65.79%, suggesting that the retained tokens can still cover key environmental entities.

b) Cognition: ECP maintains the highest Cognitive Map accuracy of 54.36% at 20% retention, indicating stronger layout consistency under sparse token budgets.

c) Reasoning & Planning: For High-Level Planning, ECP is consistently strong across pruning ratios, achieving 50.00% accuracy at 70% retention and 52.12% at 20% retention. At 70% retention, it clearly outperforms BTP (44.49%) and DivPrune (42.80%), while other baselines show larger variation across retention ratios.

Table 10: Accuracy comparison of remaining tasks on embodied spatial reasoning benchmarks. *Task definitions*: **Obj.Recall**=Object Recall, **Scene.Recall**=Scene Recall, **Abs.Dis**=Absolute Distance Estimation; **Cog.Map**=Cognitive Map Construction; **Rel.Dir**=Relative Direction Understanding; **Route.Plan**=Route Planning; **HL.Plan**=High-Level Planning.

Method	Avg.Acc	Perception		Cognition		Reasoning & Planning		
		Obj.Recall	Scene.Recall	Abs.Dis	Cog.Map	Rel.Dir	Route.Plan	HL.Plan
Original	36.31	65.79	69.70	23.05	51.43	34.74	29.53	45.23
70% Retaining								
BTP	35.95	65.79	69.70	23.31	51.45	34.14	28.19	44.49
DivPrune	36.16	65.79	69.70	23.53	52.28	35.14	30.87	42.80
FastV	36.28	65.79	69.70	23.47	51.87	34.54	28.86	44.49
PDrop	36.73	65.79	69.70	23.85	**53.53**	35.54	30.87	44.49
DTD	36.36	63.16	69.70	24.01	51.45	35.74	29.53	46.61
VTW	36.45	65.79	69.70	23.97	51.45	35.14	30.20	42.80
VisionZip	36.64	**68.42**	69.70	23.15	53.47	36.34	30.87	42.32
DEP	36.88	68.40	69.70	24.28	53.50	**36.55**	28.90	41.50
OURS	**37.78**	65.79	**69.70**	**24.32**	52.70	33.94	**31.54**	**50.00**
50% Retaining								
BTP	35.09	65.79	69.70	22.20	49.38	34.14	30.20	43.64
DivPrune	36.77	63.16	69.70	24.05	53.53	**36.14**	29.53	46.61
FastV	35.67	65.79	66.67	22.14	50.21	34.74	30.20	47.46
PDrop	35.67	65.79	69.70	22.16	50.62	34.74	28.86	45.76
DTD	36.55	63.16	69.70	24.17	53.11	35.14	29.53	47.03
VTW	33.84	68.42	**72.73**	22.44	49.38	34.74	30.87	38.98
VisionZip	36.54	68.42	69.70	23.41	51.84	35.74	30.87	43.15
DEP	33.28	**71.10**	72.70	21.18	49.00	35.14	30.20	36.50
OURS	**38.16**	63.16	69.70	**24.50**	**54.36**	34.14	**32.21**	**48.73**
20% Retaining								
BTP	33.34	65.79	66.67	19.95	51.87	34.74	30.20	47.88
DivPrune	34.40	60.53	66.67	20.33	49.79	35.74	27.52	51.27
FastV	32.64	57.89	63.64	22.79	47.72	32.73	**34.23**	49.58
PDrop	33.12	63.16	66.67	21.50	53.53	32.53	31.54	48.31
DTD	35.37	65.79	72.73	22.58	52.28	33.53	29.53	47.88
VTW	30.50	42.11	54.55	12.14	47.72	35.54	30.20	41.95
VisionZip	35.79	63.16	69.70	22.52	52.24	36.15	31.54	48.96
DEP	29.95	42.10	51.50	11.11	44.10	**37.35**	30.90	42.30
OURS	**37.62**	**73.68**	**72.73**	**23.05**	**54.36**	36.95	30.20	**52.12**

G Additional Ablation Study of Component Analysis

To provide a more granular view of each component, we conduct an ablation study on the combined validation set of UrbanVideo-Bench and VSI-Bench. We analyze Event-Guided Temporal Sampling (ETS), Event-guided Motion Saliency Filtering (EMSF), and Event-Attention Ranking Fusion (EARF) under token retention rates of 20%, 50%, and 70%, and report averaged accuracy. The results are summarized in Table 11.

G.1 Component-wise Contribution Analysis

ETS: The Spatiotemporal Anchor for Cognition. ETS provides temporal anchoring for spatial cognition. As shown in Table 11, ETS alone achieves the highest Cognition accuracy among the tested variants (44.71%). Adding ETS to EMSF increases Cognition accuracy from 43.11% to 44.65%, while adding it to EARF increases accuracy from 42.99% to 44.25%. These results suggest that anchoring motion-informative keyframes preserves coherent temporal context for downstream dynamic scene understanding, including mapping and distance estimation.

EMSF: The High-SNR Filter for Reasoning. EMSF alone achieves the highest Reasoning & Planning accuracy (28.31%), slightly exceeding full ECP (28.17%). This suggests that EMSF acts as a denoising mechanism: filtering background redundancies in event-inactive regions can improve the signal-to-noise ratio (SNR) of motion-relevant evidence, helping the LLM focus on dynamic interactions. However, relying solely on EMSF may remove background context needed for Cognition, where its accuracy (43.11%) is lower than that of full ECP (44.06%).

Table 11: **Additional ablation study of component analysis.** We perform a granular breakdown of accuracy metrics mapped to the validation baseline. The best results are **bolded** and second-best are underlined. Full ECP achieves the highest overall accuracy and Perception score while maintaining the lowest latency among multi-component variants.

Variant	Components			Latency (s) ↓	Overall Acc (%) ↑	Task Categories Acc (%)		
	ETS	EMSF	EARF			Perception	Cognition	R&P
ETS	✓	-	-	1.58	36.52	48.06	**44.71**	26.67
EMSF	-	✓	-	1.83	37.22	49.20	43.11	**28.31**
EARF	-	-	✓	1.53	37.15	50.25	42.99	28.18
ETS+EMSF	✓	✓	-	1.47	37.51	49.84	44.65	27.76
ETS+EARF	✓	-	✓	1.17	37.75	50.71	44.25	28.01
EMSF+EARF	-	✓	✓	1.59	37.00	48.66	43.68	27.93
Full ECP (Ours)	✓	✓	✓	**1.16**	**37.85**	**50.74**	44.06	28.17

EARF: The Semantic Booster for Perception. EARF contributes to fine-grained visual recognition. Adding EARF to ETS+EMSF increases Perception accuracy from 49.84% to 50.74%, the highest score among the tested variants and a gain of 0.90 percentage points. This suggests that rank-based fusion combines the high temporal resolution of event cues with the semantic richness of RGB representations, helping preserve fine-grained evidence for Perception tasks.

G.2 Overall Accuracy and Latency

Among the tested variants, full ECP achieves the highest overall accuracy (37.85%) and the lowest latency (1.16 s). Compared with ETS+EARF, adding EMSF improves overall accuracy by 0.10 percentage points, with similar latency (1.16 vs. 1.17 s). Full ECP also achieves the highest Perception score (50.74%), although ETS alone and EMSF alone obtain higher scores in Cognition and Reasoning & Planning, respectively. These results support the complementary roles of temporal selection, spatial filtering, and rank fusion.

H Causal Event-Guided Temporal Sampling

ETS selects RGB frames using event activity and its temporal change. The implementation used in the main experiments computes normalization statistics and selects frames globally from the complete sequence. We refer to this implementation as Offline ETS. Here, we describe Causal ETS (Online), which replaces global normalization and global frame selection with historical statistics and a bounded online buffer. EMSF and EARF remain unchanged.

H.1 Prefix-only scoring and frame selection

For the n-th event window, let $\mathrm{EventVoxel}_n$ denote its event representation. We compute the activity and its change from the preceding window as

$$S_n = \sum \mathrm{EventVoxel}_n, \qquad \Delta S_n = |S_n - S_{n-1}|. \tag{20}$$

The current frame receives the priority

$$P_n = \tfrac{1}{2} z_{<n}(S_n) + \tfrac{1}{2} z_{<n}(\Delta S_n) + \mathbf{1}[\max\{z_{<n}(S_n), z_{<n}(\Delta S_n)\} \geq 1], \tag{21}$$

where $z_{<n}$ uses exponential moving average (EMA) statistics maintained separately for activity and activity change over windows $1, \ldots, n-1$. We initialize $\Delta S_1 = 0$ and use zero standardized scores until two historical observations are available. The indicator gives a unit priority bonus when either standardized score reaches the threshold. The current observation is incorporated into the statistics only after the current selection decision.

Candidate RGB frames follow a fixed temporal schedule that is independent of the final video duration. The selector maintains a buffer of at most 64 RGB frames and their priorities. Each arriving candidate is inserted into this buffer. On overflow, the selector finds the temporally closest pair and removes its lower-priority member, subject to protecting the first and most recent frames of the observed prefix. If a member of the pair is protected, the other member is removed. This rule balances

event informativeness and temporal coverage without accessing future frames. When a question arrives, the buffered frames are passed to EMSF and EARF.

H.2 Causality verification and results

Both normalization and buffer updates depend only on the observed prefix. The implementation uses no full-video maxima, quantiles, or future windows. Prefix-invariance tests verify that changing a future continuation leaves the candidate schedule and decisions within the common prefix unchanged. These checks establish causality of the temporal selector; they do not measure the end-to-end throughput of the Video-LLM.

Table 12: Comparison of Offline ETS and Causal ETS (Online), with the subsequent EMSF and EARF modules unchanged. Scores are reported in percent.

	UrbanVideo-Bench + VSI-Bench		ESR-Real	
Retention	Offline ETS	Causal ETS	Offline ETS	Causal ETS
70%	37.78	37.76	52.89	52.75
50%	38.16	37.96	54.58	54.44
20%	37.62	37.25	50.07	49.79

Table 12 shows that Causal ETS remains within 0.37 percentage points of Offline ETS across both evaluation settings and all three retention ratios. Across 642 event-available videos, the selected frame sets have a mean Jaccard similarity of 93.10% with the controlled Offline ETS reference. These results support extending event-activity-based selection to causal streaming while keeping the downstream pruning procedure unchanged.

I General Video Understanding on Video-MME

We evaluate broader video understanding on the event-available subset of Video-MME [16]. This evaluation covers eight categories: object recognition, attribute understanding, action recognition, OCR, counting, temporal understanding, synopsis, and object reasoning. Events provide motion and structural priors for selecting RGB tokens; semantic and appearance evidence still comes from the retained RGB representations and text-conditioned attention.

Table 13: Overall results on the event-available Video-MME subset. Accuracy is reported in percent; all methods use the same evaluation subset.

Retention	Method	Accuracy
100%	Full-token baseline	72.47
70%	FastV	74.16
70%	DEP	69.66
70%	ECP	73.60
50%	FastV	74.16
50%	DEP	68.54
50%	ECP	73.60
20%	FastV	59.55
20%	DEP	61.80
20%	ECP	65.73

As shown in Table 13, ECP exceeds the full-token baseline by 1.13 percentage points at 70% and 50% retention. At 20% retention, it exceeds FastV and DEP by 6.18 and 3.93 points, respectively. FastV performs better at the two higher retention ratios, while ECP has an advantage under the more aggressive 20% budget. This pattern supports the ability of event guidance to preserve useful RGB evidence beyond spatial reasoning without implying a uniform advantage across tasks or budgets.

Table 14 provides all eight category results. At 20% retention, ECP improves over both alternatives in action recognition and object reasoning, and retains the full-token synopsis score. Advantages

Table 14: Complete category-level results on the event-available Video-MME subset. All scores are percentages. Object Rec., Action Rec., and Object Reason. denote object recognition, action recognition, and object reasoning.

Retention	Method	Total	Object Rec.	Attribute	Action Rec.	OCR	Counting	Temporal	Synopsis	Object Reason.
100%	Full-token	72.47	81.08	85.71	71.43	84.62	39.13	77.78	92.86	57.69
70%	FastV	74.16	78.38	82.14	71.43	92.31	47.83	88.89	92.86	61.54
70%	DEP	69.66	78.38	78.57	64.29	84.62	39.13	77.78	78.57	65.38
70%	ECP	73.60	81.08	78.57	71.43	84.62	47.83	77.78	92.86	65.38
50%	FastV	74.16	83.78	82.14	67.86	92.31	43.48	77.78	92.86	65.38
50%	DEP	68.54	83.78	75.00	60.71	84.62	39.13	66.67	78.57	61.54
50%	ECP	73.60	81.08	78.57	75.00	92.31	43.48	77.78	92.86	61.54
20%	FastV	59.55	56.76	50.00	60.71	92.31	39.13	77.78	85.71	53.85
20%	DEP	61.80	72.97	67.86	60.71	76.92	30.43	33.33	78.57	61.54
20%	ECP	65.73	72.97	64.29	64.29	84.62	34.78	44.44	92.86	69.23

vary by category: for example, FastV obtains higher OCR, counting, and temporal-understanding scores at this budget. In prolonged static scenes, especially for questions about color, text, or low-contrast details, an event prior can become less informative. Conversely, ego-motion can generate events along otherwise static scene contours, allowing background RGB evidence to participate in event-guided selection.

J Cross-Backbone Evaluation on LLaVA-OneVision

We evaluate ECP on LLaVA-OneVision-1.5-8B-Instruct [4] to examine whether its bias-guided calibration principle transfers beyond Qwen2.5-VL. Visual-token layouts, attention patterns, and pruning sensitivity can differ across backbones; the layer indices and fusion weights selected for one model therefore need to be adapted for another.

J.1 Layer-wise bias and backbone adaptation

We repeat the peripheral-bias analysis using the same UrbanVideo-Bench and VSI-Bench samples as for Qwen2.5-VL. Table 15 reports the Peripheral/Center ratio and Cohen's d for three depth regimes of the 36-layer OneVision backbone. The Peripheral/Center ratios are lower than those measured for Qwen2.5-VL. Within OneVision, both the ratio and Cohen's d decrease with depth, motivating stronger event calibration in shallow layers and greater reliance on attention in deeper layers. These regimes describe measured bias patterns rather than fixed causal functions of Transformer stages.

Table 15: Layer-wise peripheral attention bias in LLaVA-OneVision-1.5-8B. Stage boundaries are specific to this backbone.

Stage	Layers	Peripheral/Center	Cohen's d
Shallow	L0–9	$2.53\times$	1.78
Middle	L10–21	$2.43\times$	1.49
Deep	L22–35	$2.20\times$	1.11

For this backbone, we reproject event maps to its native visual-token layout and adapt the token budget to its 36-layer architecture. The event-fusion weights decrease with depth as $(0.7, 0.6, 0.5)$. Thus, the transferable component is the bias-guided adaptation principle; the token mapping, pruning-layer configuration, and fusion weights are adapted to the backbone.

J.2 Cross-backbone results

We compare ECP with six pruning methods (DTD, FastV, PyramidDrop, DivPrune, BTP, and VTW) at three token-retention ratios on the public spatial benchmarks and ESR-Real. Table 16 summarizes the strongest non-ECP pruning result in each setting.

Table 16: Cross-backbone results with LLaVA-OneVision-1.5-8B. Best non-ECP denotes the strongest result among the six comparison methods in each setting. Scores are percentages, and gains are percentage points.

Dataset	Retention	Best non-ECP	ECP	Gain
UrbanVideo-Bench + VSI-Bench	70%	36.43	**37.08**	+0.65
UrbanVideo-Bench + VSI-Bench	50%	35.38	**36.64**	+1.26
UrbanVideo-Bench + VSI-Bench	20%	33.55	**34.33**	+0.78
ESR-Real	70%	40.90	**41.61**	+0.71
ESR-Real	50%	42.31	**43.44**	+1.13
ESR-Real	20%	41.04	**42.74**	+1.70

As shown in Table 16, ECP exceeds the strongest comparison method by 0.65–1.70 percentage points across all six settings. These results support transferring event-assisted calibration to a backbone with a different visual-token layout, depth, and attention-bias strength.

K Broader Impacts and Responsible Release

Broader impacts and responsible use. ECP reduces the visual-token computation required by Video-LLMs for embodied spatial reasoning, which may lower latency, memory usage, and energy consumption on resource-constrained RGB-event platforms. These efficiency gains may also support more local processing and reduce reliance on cloud offloading. However, as with other egocentric perception technologies, easier deployment in privacy-sensitive environments could raise monitoring or privacy concerns, and incorrect spatial predictions could affect downstream navigation or interaction decisions. ECP should therefore be treated as an efficiency mechanism for perception rather than a complete decision-making system, and safety-relevant deployment requires task-specific validation, privacy-preserving data handling, human oversight, and compliance with applicable regulations.

Responsible release of ESR-Real. ESR-Real is collected in controlled indoor environments for spatial-reasoning evaluation, not identity recognition. Before release, we review the RGB-event data, QA annotations, and metadata to remove or blur incidental faces, personal identifiers, sensitive documents or screens, and location-revealing metadata when present. The released benchmark will not include identity, biometric, or person-tracking labels, and will be distributed with documentation and research-use terms restricting re-identification, privacy-invasive monitoring, and unvalidated safety-critical deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scoped claims: ECP as a training-free RGB-event pruning framework with three event-guided stages, the ESR-Real benchmark, and the reported accuracy/efficiency gains. These claims are supported by the method in Section 3, the main results and ablations in Section 4 and Tables 1, 2, 3, 4, 5, and 6, the ESR-Real details in Appendix E, and the assumptions and limitations in Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Section 5, including the main assumption that ECP requires synchronized RGB-event inputs and is intended for platforms with paired RGB-event sensing rather than RGB-only online acceleration. The section also notes the current practical limitation that event sensors are less ubiquitous than RGB cameras and discusses future RGB-only distilled pruning directions.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix C states the assumptions and provides the derivation/proof for the EARF rank-calibration analysis, showing how the calibrated score changes pairwise token ordering and removes pruned visual keys/values from downstream attention.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the experimental setup needed to reproduce the main results, including the base model, benchmarks, event generation/alignment procedure, pruning budgets, decoding strategy, hardware, metrics, baselines, and ablations in Section 4 and Appendices D–G. The supplementary package further provides the ECP inference implementation, configuration files, demo data, and scripts for demo and full evaluation, while external assets such as Qwen2.5-VL weights are identified with required path and schema instructions.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized access to the ECP code, configuration files, run scripts, and the ESR-Real benchmark, including annotations, RGB videos, and event directories; the supplementary README provides environment setup, demo and full-evaluation commands, and configuration files for the reported retention settings. External assets that cannot be redistributed in the package, such as Qwen2.5-VL weights are identified with required paths and usage instructions.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: ECP is training-free, so optimizer, learning-rate schedule, and training-checkpoint selection are not applicable; Section 4.1 specifies the base model, benchmarks, task subsets, baselines, retention ratios, decoding strategy, metrics, event generation setting, and implementation details. The method hyperparameters and their choices are described in Sections 3.2–3.4, Appendices D–G, and the supplementary configuration files/README for the reported retention settings.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main benchmark results are deterministic full-set evaluations under fixed model weights, fixed evaluation sets, identical prompts, and greedy decoding, so repeated-run error bars are not applicable under the stated protocol. For the statistical analysis that motivates EARF, Appendix A reports effect sizes, one-sample t-statistics with $p < 10^{-10}$, and cross-dataset correlation over $n = 21{,}920$ frames.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All reported experiments are inference-only evaluations of fixed checkpoints, so the relevant resource requirement is the online evaluation cost rather than training or fine-tuning compute; Section 4.1 specifies the NVIDIA A6000 GPU environment, FlashAttention-2 acceleration, shared implementation, and deterministic decoding protocol. Table 3 reports the measured per-sample online latency, GFLOPs, and peak GPU memory for the main efficiency comparison, with ablation latencies reported in Tables 4 and 11, and the paper explicitly states that offline vid2e event synthesis is excluded from the online inference cost.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and the research conforms to it. The work uses public benchmarks and controlled RGB-event recordings for spatial reasoning, does not perform person identification or make decisions about individuals, and the data collection protocol for ESR-Real is described in Appendix E.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix K discusses both positive and negative broader impacts: ECP may reduce latency, memory usage, and energy consumption for embodied Video-LLM inference and support more local processing, while lower-cost egocentric perception could raise privacy-sensitive deployment, monitoring, and downstream navigation-error concerns. The appendix also discusses responsible-use mitigations including privacy-preserving data handling, human oversight, task-specific validation, and compliance with applicable safety and privacy regulations.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Appendix K describes safeguards for the responsible release of ESR-Real, including review of RGB-event data, QA annotations, and metadata, removal or blurring of incidental faces, personal identifiers, sensitive documents or screens, and location-revealing metadata when present, and exclusion of identity, biometric, or person-tracking labels. The release will include documentation and research-use terms restricting re-identification, privacy-invasive monitoring, and unvalidated safety-critical deployment.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original sources for the Qwen2.5-VL backbone, HuggingFace Transformers implementation, vid2e simulator, VSI-Bench and UrbanVideo-Bench evaluation data, Gemini-assisted annotation tool, FlashAttention-2, and all pruning baselines. The supplementary code preserves the Apache-2.0 license headers for adapted Qwen2.5-VL/Transformers components, and external model weights or full benchmark assets are not redistributed without their original terms.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces ESR-Real and the ECP code as new assets. Appendix E documents the dataset hardware, calibration and alignment, data collection protocol, task taxonomy, QA generation and refinement pipeline, intended use, limitations, privacy safeguards, and release contents; the supplementary package includes README documentation, evaluation scripts, and configurations needed to inspect and run the released ECP code.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing experiments or research with human subjects. The ESR-Real QA refinement described in Appendix E.5 was conducted by the authors/internal research team as dataset curation, so there were no external participants, crowd workers, participant-facing instructions, screenshots, or compensation.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve crowdsourcing experiments or human-subject studies. ESR-Real records controlled indoor scenes for spatial reasoning and does not collect or analyze participant responses, identities, biometric attributes, or personal traits; the manual QA refinement described in Appendix E.5 was conducted by the authors/internal research team as dataset curation.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 4.1 describes the use of Qwen2.5-VL-7B-Instruct as the backbone Video-LLM, and Appendix E.4 describes the use of Gemini-3Pro for automated ESR-Real MCQ synthesis. The generated QA pairs are human-refined in Appendix E.5, and ECP itself is a training-free pruning method without backbone parameter updates.