RESET: RESIDUAL-ENERGY SUBSPACE EXPANSION VIA TOKENS FOR EFFICIENT VLM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) achieve strong visual understanding but incur substantial memory and inference costs due to long visual token sequences. While numerous token compression strategies have been explored, we revisit the problem from a representative subset selection perspective, asking which visual tokens can collectively preserve diverse information with minimal redundancy. Based on this view, we propose RESET, a training-free visual token compression framework based on residual subspace selection. For images, RESET iteratively selects tokens with the highest residual energy, favoring previously unexplained feature directions and reducing spatial redundancy. For videos, RESET constructs a compact historical reference via SVD and conditions current-frame selection on the subspace orthogonal to historical information, thereby suppressing temporal redundancy. RESET is text-independent, plug-and-play, and compatible with FlashAttention, enabling seamless integration into existing image and video inference pipelines without modifying the underlying attention mechanism. Extensive experiments across multiple VLMs and benchmarks show that RESET consistently outperforms strong state-of-the-art token compression methods, validating residual-subspace selection as a simple and effective principle for representative visual token compression. Our code is available at this link.

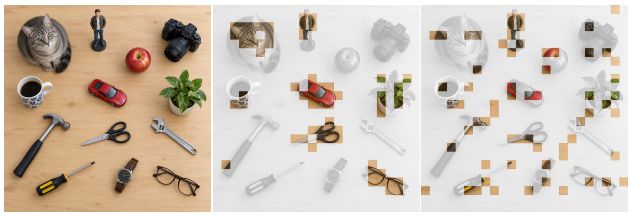
1 INTRODUCTION

Multimodal Large Language Models (MLLMs) typically convert visual inputs into sequences of visual tokens and jointly model them with textual tokens, enabling strong image and video understanding capabilities and substantially expanding the scope of generative and visual-language applications (Liu et al., 2024b; Bai et al., 2025; Chen et al., 2024c; Li et al., 2024; Guo et al., 2024). However, the growing visual sequence length has become a major bottleneck for efficient inference. A single image can already produce hundreds of visual tokens, while videos further accumulate visual representations along the temporal dimension. As the visual context grows, the cost of visual prefill and LLM attention increases rapidly, accompanied by substantial KV-cache and GPU-memory overhead. Efficient attention kernels such as FlashAttention reduce the cost of attention computation (Dao et al., 2022; Dao, 2024). However, they do not eliminate the fundamental overhead introduced by processing long visual token sequences. This has motivated a broad line of visual token compression methods, especially training-free approaches that reduce visual context before or during LLM inference.

Existing training-free methods estimate token importance or redundancy using signals such as attention, feature similarity, diversity, reconstruction error, coverage, and temporal novelty. Representative image-oriented approaches include FastV, VisionZip, DivPrune, PACT, ApET, MMTok, and VisionTrim (Chen et al., 2024a; Yang et al., 2025a; Alvar et al., 2025; Dhouib et al., 2025; Ma et al., 2026; Fang et al., 2026; Dong et al., 2026; Yu et al., 2026), while recent video-oriented methods exploit temporal similarity, hierarchical memory, surprise, or streaming-aware compression (Cho et al., 2026; Xie et al., 2026; Chen et al., 2026b; Wang et al., 2026b; Tang et al., 2026; Fu et al., 2025b). Despite their different scoring signals and compression strategies, a large class of existing methods is primarily designed to answer which tokens appear important or redundant under the current scoring context. For compression, however, the more fundamental question is: *what information remains unexplained given what has already been retained or observed?* Token value is inherently conditional. Once a representation direction has been covered by retained tokens, subsequent tokens carrying a similar direction should contribute less new information, even if their standalone scores remain high.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This issue becomes more pronounced in video, where much of the current visual content may already be explained by causal visual history. Evaluating tokens only within the current frame or the neighborhood window can therefore conflate visual saliency with genuinely novel information.



(a) Input Image (b) ApET (c) RESET

Figure 1: Qualitative token selection under the same 64-token budget on LLaVA-1.5-7B. Highlighted regions denote retained visual tokens.

Motivated by this observation, we revisit reconstruction-based visual token compression from the perspective of conditional residual information and propose RESET (**R**esidual-**E**nergy **S**ubspace **E**xpansion via **T**okens), a unified framework for reference-conditioned residual subspace preservation. RESET first removes from the current visual tokens the components already explained by a reference subspace and then performs pivoted QR (QRCP)-style residual selection in the resulting residual space. At each iteration, the token with the largest residual energy is retained, after which its feature direction is orthogonally removed from the remaining candidates. As illustrated in Fig. 1, under the same token budget, ApET tends to concentrate its selections on a subset of visual regions, whereas RESET makes more effective use of the budget to cover a broader set of semantically salient regions across the image. This QRCP-based expansion evaluates each token by its incremental contribution conditioned on both the reference and previously selected tokens (Gu & Eisenstat, 1996; Golub & Van Loan, 2013). The same selection operator applies to both images and videos, differing only in reference construction: images use a zero reference, whereas videos recursively construct a compact, parameter-free historical subspace from previously retained features to remove information already explained by causal history. The selected tokens are subsequently used according to the corresponding compression stage, through either pruning or feature aggregation. RESET is training-free, attention-free, and introduces no additional learnable parameters.

The main contributions of this work are summarized as follows.

- **Conditional compression perspective.** We identify that the value of a visual token should be assessed relative to information that has already been retained or observed, and formulate token selection in terms of its remaining unexplained information rather than its standalone importance.
- **Unified residual-subspace framework.** We propose RESET, a training-free and attention-free framework that performs reference-conditioned residual initialization followed by dynamic residual-energy subspace expansion. The same selection operator naturally supports image compression with a zero reference and video compression with a recursively constructed, parameter-free historical reference.
- **Strong empirical performance.** Extensive experiments across image and video benchmarks demonstrate that RESET achieves state-of-the-art performance under different model and compression settings, validating conditional residual information as an effective principle for both offline and streaming visual token compression.

2 Related Work

2.1 Image Token Compression

Image token compression reduces visual redundancy to improve VLM inference efficiency. FastV (Chen et al., 2024a), VisionZip (Yang et al., 2025a), and VisionTrim (Yu et al., 2026) exploit attention or feature similarity for token pruning and merging, while DivPrune (Alvar et al., 2025) and PRUNESID (Fang et al., 2026) further incorporate token diversity. MMTok (Dong et al., 2026) and FLoC (Cho et al., 2026) formulate token selection from coverage-based perspectives. Closely related to our work, ApET (Ma et al., 2026) adopts a reconstruction view, using approximation error from sampled token bases to guide compression.

2.2 Video Token Compression

Video token compression additionally exploits temporal redundancy across frames. FrameFusion (Fu et al., 2025b) combines cross-frame similarity with token importance, FLoC (Cho et al., 2026) performs representative selection within temporal blocks, and SURGE (Tang et al., 2026) retains temporally surprising tokens. For streaming scenarios, STC (Wang et al., 2026b), FluxMem (Xie et al., 2026), and StreamingTOM (Chen et al., 2026b) progressively manage long visual streams through hierarchical compression, adaptive memory, or fixed-budget causal token reduction.

2.3 Subspace Selection

Subspace selection provides a classical mathematical foundation for identifying representative elements in high-dimensional data. The column subset selection problem seeks actual matrix columns whose span accurately approximates the original matrix (Boutsidis et al., 2009; 2014). QR factorization with column pivoting greedily selects columns according to their residual contribution (Golub & Van Loan, 2013), while strong rank-revealing QR further improves rank-revealing properties through additional column interchanges (Gu & Eisenstat, 1996). These ideas provide a natural basis for transferring residual-based subset selection to visual token compression.

Building on these directions, RESET formulates visual token compression as reference-conditioned residual subspace selection. It is training-free, requires no modification to the underlying attention operator, and remains compatible with FlashAttention-based inference. Its plug-in design enables simple integration into existing image and video compression pipelines while effectively reducing visual-token redundancy.

3 Method

3.1 Problem Formulation and Method Overview

Problem formulation. We consider a video stream $\mathcal{V} = \{F_t\}_{t=1}^{N_f}$, where F_t is the frame arriving at timestep t and N_f is the number of frames. The visual encoder and projector convert each frame into

$$X_t = [x_{t,1}, \ldots, x_{t,N_t}] \in \mathbb{R}^{d \times N_t}, \tag{1}$$

where N_t is the number of visual tokens, d is the feature dimension, and $x_{t,i} \in \mathbb{R}^d$ denotes the i-th token. We seek a training-free selection operator

$$\widetilde{X}_t = \mathcal{C}(X_t; M_{t-1}, K) \in \mathbb{R}^{d \times K}, \tag{2}$$

where M_{t-1} denotes the causal visual history available before F_t, and $K < N_t$ is the target token budget. The compression uses only the current frame and previously observed visual information.

Method overview. The proposed RESET (**R**esidual-**E**nergy **S**ubspace **E**xpansion via **T**okens) preserves visual information that remains unexplained by already retained content. As illustrated in Fig. 2, RESET first constructs a compact historical subspace from previously retained tokens, then applies reference-conditioned QRCP to progressively select complementary residual directions. For video compression, grouped parallel execution reduces frame-wise compression latency by shortening the sequential dependency depth across frames.

3.2 From Reconstruction Error to Residual Selection

Motivated by the connection between reconstruction error and information preservation (Ma et al., 2026), we view token compression as a subset selection problem. For a generic token matrix $X = [x_1, \ldots, x_N] \in \mathbb{R}^{d \times N}$, we select an index set $\mathcal{S} \subseteq \{1, \ldots, N\}$ with $|\mathcal{S}| = K$:

$$\mathcal{S}^* = \arg\min_{|\mathcal{S}|=K} \left\| X - \widehat{X}_{\mathcal{S}} \right\|_F^2, \tag{3}$$

where $\widehat{X}_{\mathcal{S}}$ is the least-squares reconstruction of X from the subspace spanned by the selected columns, and $\|\cdot\|_F$ denotes the Frobenius norm.

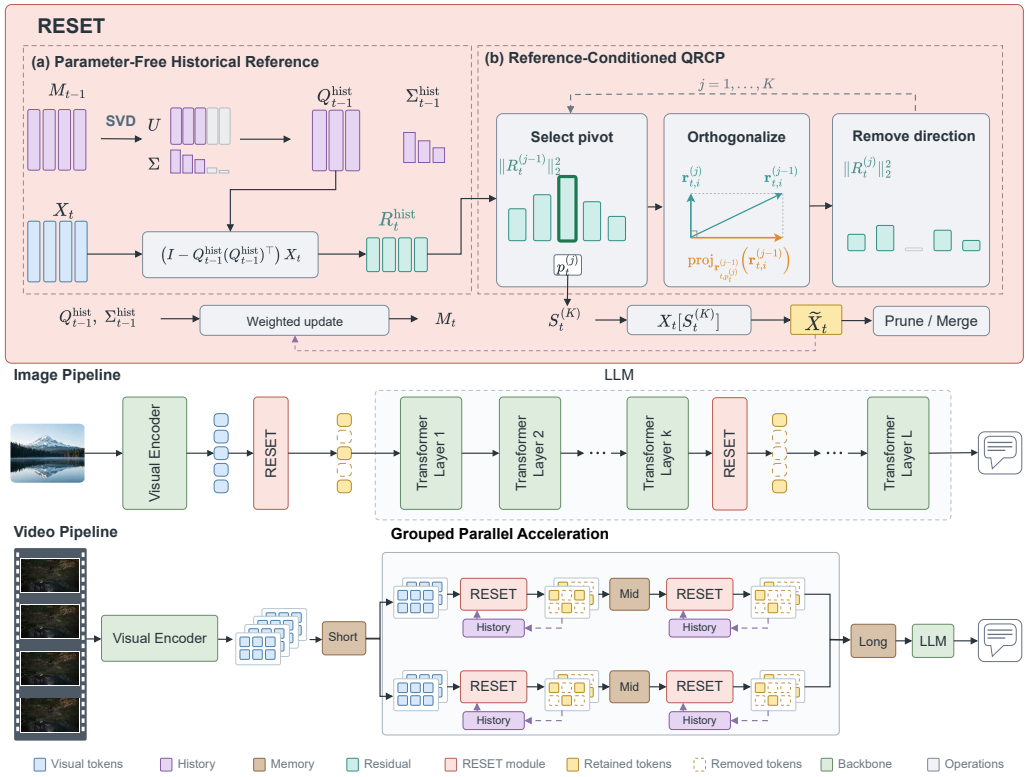


Figure 2: Overview of RESET. (a) Parameter-Free Historical Reference maintains a compact historical basis via SVD-based adaptive rank estimation. (b) Reference-Conditioned QRCP removes history-explained components and performs residual-based token selection. RESET supports both image token compression and hierarchical video memory, with grouped parallel execution reducing frame-level sequential dependency.

Defining

$$R_{\mathcal{S}} = X - \widehat{X}_{\mathcal{S}}, \qquad \mathcal{E}(\mathcal{S}) = \|R_{\mathcal{S}}\|_F^2, \tag{4}$$

minimizing reconstruction error is equivalent to minimizing the remaining residual energy. Motivated by this objective, we adopt a greedy residual-pivoting strategy: at each step, we prioritize a token with large unexplained residual energy and remove its represented direction before the next selection. This criterion provides an efficient QRCP-style approximation to the combinatorial subset-selection problem rather than directly solving Eq. 3. For video, previously retained information can additionally serve as a reference subspace, yielding the history-conditioned formulation below.

3.3 Parameter-Free Historical Reference

For each frame $t > 1$, RESET summarizes historical information using a compact representation

$$M_{t-1} \in \mathbb{R}^{d \times c_{t-1}}, \tag{5}$$

where d is the feature dimension and c_{t-1} is the number of historical columns. We compute its compact SVD as

$$M_{t-1} = U_{t-1}\Sigma_{t-1}V_{t-1}^{\top}, \tag{6}$$

where $q_{t-1} = \min(d, c_{t-1})$, $U_{t-1} \in \mathbb{R}^{d \times q_{t-1}}$, $\Sigma_{t-1} \in \mathbb{R}^{q_{t-1} \times q_{t-1}}$, and $V_{t-1} \in \mathbb{R}^{c_{t-1} \times q_{t-1}}$. Let $\sigma_{t-1,1} \geq \cdots \geq \sigma_{t-1,q_{t-1}}$ denote the singular values.

We determine the effective historical rank using a numerical-rank criterion. Let ϵ_{mach} denote machine precision and $r_{\text{cap}} = \min(d, K)$, where K is the token budget. Then

$$\tau_{t-1} = \sigma_{t-1,1} \max(d, c_{t-1})\epsilon_{\text{mach}}, \qquad r_{t-1}^{\text{hist}} = \min\left(r_{\text{cap}}, \sum_{i=1}^{q_{t-1}} \mathbf{1}[\sigma_{t-1,i} > \tau_{t-1}]\right), \tag{7}$$

where τ_{t-1} is the numerical threshold and r_{t-1}^{hist} counts the singular directions that remain distinguishable from numerical noise. The threshold scales with three factors. The leading singular value $\sigma_{t-1,1}$ adapts the threshold to the overall magnitude of the historical representation, so that the criterion remains relative to the scale of its spectrum. A larger matrix size $\max(d, c_{t-1})$ or a larger machine epsilon ϵ_{mach} raises τ_{t-1}, making the criterion more conservative and requiring a singular value to be larger in order to be retained. Consequently, larger matrices or lower numerical precision tend to discard weaker singular directions that are more difficult to distinguish from numerical noise. The cap $r_{\text{cap}} = \min(d, K)$ further restricts the historical subspace to both the feature dimension and the downstream compression capacity, avoiding an additional manually tuned rank hyperparameter.

The retained historical basis and singular values are

$$Q_{t-1}^{\text{hist}} = U_{t-1}[:, 1 : r_{t-1}^{\text{hist}}], \qquad \Sigma_{t-1}^{\text{hist}} = \Sigma_{t-1}[1 : r_{t-1}^{\text{hist}}, 1 : r_{t-1}^{\text{hist}}], \tag{8}$$

where $Q_{t-1}^{\text{hist}} \in \mathbb{R}^{d \times r_{t-1}^{\text{hist}}}$ forms an orthonormal basis of the retained historical subspace.

Given the current-frame tokens $X_t \in \mathbb{R}^{d \times N_t}$, we remove the component explained by history:

$$R_t^{\text{hist}} = \left(I_d - Q_{t-1}^{\text{hist}} (Q_{t-1}^{\text{hist}})^\top\right) X_t, \qquad \eta_t = \frac{\|R_t^{\text{hist}}\|_F^2}{\|X_t\|_F^2}, \tag{9}$$

where I_d is the identity matrix, R_t^{hist} is the history-conditioned residual, and $\eta_t \in [0, 1]$ is the unexplained energy ratio. For $t = 1$, we set $R_1^{\text{hist}} = X_1$ and $\eta_1 = 1$.

After obtaining the retained tokens $\widetilde{X}_t \in \mathbb{R}^{d \times K}$, we update the historical representation as

$$M_t = \left[\sqrt{1 - \eta_t}\, Q_{t-1}^{\text{hist}} \Sigma_{t-1}^{\text{hist}},\ \sqrt{\eta_t}\, \widetilde{X}_t\right], \tag{10}$$

where $M_t \in \mathbb{R}^{d \times c_t}$ with $c_t = r_{t-1}^{\text{hist}} + K$. Since $r_{t-1}^{\text{hist}} \le r_{\text{cap}}$, the number of historical columns remains bounded by $K + r_{\text{cap}}$ rather than growing with the video length. For $t = 1$, the historical term is absent and $M_1 = \widetilde{X}_1$. Thus, RESET recursively summarizes preceding visual context. In the serial formulation, the historical reference requires no manually specified reference-window size.

3.4 Reference-Conditioned Pivoted QR Selection

Although R_t^{hist} removes history-explained information, redundancy may remain among current-frame tokens. We therefore perform QRCP-style pivot selection in the residual space.

Let $\mathcal{S}_t^{(j)}$ be the selected token indices after j steps. We initialize

$$R_t^{(0)} = R_t^{\text{hist}}, \qquad \mathcal{S}_t^{(0)} = \varnothing. \tag{11}$$

At step $j = 1, \ldots, K$, let $\mathbf{r}_{t,i}^{(j-1)} = R_t^{(j-1)}[:, i] \in \mathbb{R}^d$ denote the residual vector of candidate token i. The next pivot is

$$p_t^{(j)} = \arg \max_{i \notin \mathcal{S}_t^{(j-1)}} \left\| \mathbf{r}_{t,i}^{(j-1)} \right\|_2^2, \tag{12}$$

and its normalized residual direction is

$$\mathbf{u}_t^{(j)} = \frac{\mathbf{r}_{t,p_t^{(j)}}^{(j-1)}}{\left\| \mathbf{r}_{t,p_t^{(j)}}^{(j-1)} \right\|_2}, \tag{13}$$

with $\mathbf{u}_t^{(j)} = \mathbf{0}$ when $\mathbf{r}_{t,p_t^{(j)}}^{(j-1)} = \mathbf{0}$. When the remaining residual energy is exhausted, RESET continues selecting until the prescribed budget K is reached, while zero-residual pivots introduce no additional residual direction. We then remove this direction from the remaining residuals and update the selected set:

$$R_t^{(j)} = \left(I_d - \mathbf{u}_t^{(j)} (\mathbf{u}_t^{(j)})^\top\right) R_t^{(j-1)}, \qquad \mathcal{S}_t^{(j)} = \mathcal{S}_t^{(j-1)} \cup \{p_t^{(j)}\}. \tag{14}$$

Because $R_t^{(0)}$ has already removed the historical subspace and each subsequent update removes previously selected directions, Eq. 12 favors tokens that contribute complementary information relative to both history and current-frame selections. After K iterations, QRCP yields the selected anchor index set $\mathcal{S}_t^{(K)}$. The residual features are used only to determine these anchor indices, while the final compressed representation is constructed in the original feature space.

3.5 GROUPED PARALLEL ACCELERATION

Historical conditioning introduces a recursive dependency across frames, since the historical state must be updated before processing the next frame within the same group. Although QRCP within each frame remains sequential, the frame-level dependency can be shortened by processing multiple independent historical groups in parallel.

Consider a buffer of T frames, partitioned into G independent groups of length $B = \lceil T/G \rceil$, with all-zero frames padded when necessary:

$$\mathcal{G}_g = \left\{ F_{(g-1)B+1}, \ldots, F_{gB} \right\}, \qquad g = 1, \ldots, G. \tag{15}$$

Each group maintains an independent historical state $M_{g,b}$, initialized with $M_{g,0} = \varnothing$, and $X_{g,b} = X_{(g-1)B+b}$ for $b = 1, \ldots, B$. We set $Q^{\text{hist}}_{g,0} = \varnothing$, yielding $R^{\text{hist}}_{g,1} = X_{g,1}$.

At each step, the historical basis is constructed only from preceding compressed frames within the same group, yielding

$$R^{\text{hist}}_{g,b} = \left(I_d - Q^{\text{hist}}_{g,b-1} (Q^{\text{hist}}_{g,b-1})^\top \right) X_{g,b}. \tag{16}$$

Since the G historical states are independent, QRCP can be applied concurrently across groups, after which each group updates its own state according to Eq. 10.

This replaces the original recursion over T frames with G parallel recursions of length $B = T/G$, reducing the frame-level sequential depth from T to B. The token budget and residual-energy pivot criterion remain unchanged, although restricting history to individual groups may produce different token selections from a single full-history group.

Instantiation for image and video compression. For image compression, we follow the integration pipeline of ApET (Ma et al., 2026) and replace its token-selection module with RESET. Since a standalone image has no temporal history, $R^{\text{hist}} = X$, and RESET simplifies to residual QRCP selection. For video compression, we retain the short–medium–long memory hierarchy of FluxMem (Xie et al., 2026). RESET determines the anchor tokens through reference-conditioned QRCP, while the subsequent pruning or anchor-based merging follows the corresponding FluxMem memory-transition setting. The resulting compressed representations are used for both downstream inference and historical-reference updates. We further use a fixed retention ratio, making the retained token count predictable from the input size and thereby facilitating memory estimation under constrained inference resources. Unless otherwise specified, the remaining settings follow the corresponding baseline configurations.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Implementation Details. We evaluate RESET on three representative VLMs. For image understanding, we consider LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). We report the main results on LLaVA-1.5-7B in the main paper, while the corresponding results on LLaVA-NeXT-7B and Qwen2.5-VL-7B are provided in the appendix. The image compression stages follow the ApET implementation (Ma et al., 2026) for consistent comparison with existing token compression methods. For video understanding, we adopt Qwen2.5-VL-7B as the base model and follow the short-medium-long memory hierarchy and the corresponding configurations of FluxMem (Xie et al., 2026). For image compression, the LLaVA-based models use a fixed number of retained tokens, whereas Qwen2.5-VL uses a fixed retention ratio. For video compression, we use the same retention ratio for the short-to-medium and medium-to-long transitions, set to 6.25% for online benchmarks and 50% for offline benchmarks. For grouped parallel acceleration, we set the buffer length T to the number of frames used by the medium-memory stage, thereby introducing no additional buffer-size hyperparameter.

Benchmarks. We evaluate image understanding performance on a comprehensive set of benchmarks, including GQA (Hudson & Manning, 2019), MMBench (MMB) (Liu et al., 2024c), MMBench-CN (MMB-CN) (Liu et al., 2024c), MME (Fu et al., 2026), POPE (Li et al., 2023), ScienceQA (SQA)

Table 1: **Image understanding performance on LLaVA-1.5-7B.** Results are reported under three average visual-token budgets across LLM layers, namely 192, 128, and 64 tokens, following the ApET evaluation protocol. "–" denotes unavailable results. Avg. denotes the average relative performance retention with respect to the vanilla upper bound. † indicates the reproduced results.

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	TextVQA	VizWiz	Avg.
Upper Bound, 576 Tokens (**100%**)									
LLaVA-1.5-7B Vanilla	61.9	64.7	58.1	1862.0	85.9	69.5	58.2	50.0	100%
Retain 192 Tokens (↓ 66.7%)									
ToMe (Bolya et al., 2022)	54.3	60.5	–	1563.0	72.4	65.2	52.1	–	88.80%
FastV (Chen et al., 2024a)	52.7	61.2	57.0	1612.0	64.8	67.3	52.5	50.8	91.06%
SparseVLM (Zhang et al., 2024b)	57.6	62.5	53.7	1721.0	83.6	**69.1**	56.1	50.5	96.08%
PDrop (Xing et al., 2025)	57.3	63.3	56.8	**1797.0**	82.3	69.0	56.5	51.1	97.38%
VisionZip (Yang et al., 2025a)	59.3	63.0	–	1783.0	85.3	68.9	57.3	–	97.64%
ApET† (Ma et al., 2026)	60.3	63.8	57.0	1793.2	86.3	68.3	55.7	**54.0**	99.11%
MMTok (Dong et al., 2026)	60.1	63.4	–	1773.9	**86.4**	68.8	**57.7**	–	98.16%
RESET (Ours)	**60.4**	**64.4**	**57.2**	1765.3	**86.4**	68.7	56.9	**54.0**	**99.45%**
Retain 128 Tokens (↓ 77.8%)									
ToMe (Bolya et al., 2022)	52.4	53.3	–	1343.0	62.8	59.6	49.1	–	80.40%
FastV (Chen et al., 2024a)	49.6	56.1	56.4	1490.0	59.6	60.2	50.6	51.3	86.18%
SparseVLM (Zhang et al., 2024b)	56.0	60.0	51.1	1696.0	80.5	67.1	54.9	51.4	93.70%
PDrop (Xing et al., 2025)	57.1	61.6	**56.6**	1761.0	82.3	68.4	56.6	51.0	96.62%
VisionZip (Yang et al., 2025a)	57.6	62.0	–	1763.0	83.2	**68.9**	56.8	–	96.19%
ApET† (Ma et al., 2026)	59.1	62.5	55.8	1763.2	86.3	68.5	54.2	54.9	98.10%
MMTok (Dong et al., 2026)	59.3	62.3	–	**1779.1**	86.3	68.8	**57.0**	–	97.50%
RESET (Ours)	**59.8**	**62.9**	56.4	1768.3	**86.8**	68.5	56.4	**55.7**	**99.22%**
Retain 64 Tokens (↓ 88.9%)									
ToMe (Bolya et al., 2022)	48.6	43.7	–	1138.0	52.5	50.0	45.3	–	69.68%
FastV (Chen et al., 2024a)	46.1	48.0	52.7	1256.0	48.0	51.1	47.8	50.8	77.49%
SparseVLM (Zhang et al., 2024b)	52.7	56.2	46.1	1505.0	75.1	62.2	51.8	50.1	87.29%
PDrop (Xing et al., 2025)	47.5	58.8	50.5	1561.0	55.9	69.2	50.6	50.7	86.42%
VisionZip (Yang et al., 2025a)	55.1	60.1	–	1690.0	77.0	69.0	55.5	–	92.82%
ApET† (Ma et al., 2026)	57.2	59.9	54.0	1679.8	84.7	67.6	53.0	55.0	95.64%
MMTok (Dong et al., 2026)	**58.3**	61.2	–	**1715.3**	85.8	**69.2**	**56.0**	–	96.07%
RESET (Ours)	57.9	**61.6**	**54.1**	1700.7	**86.6**	68.1	55.0	**57.2**	**97.61%**

(Lu et al., 2022), TextVQA (Singh et al., 2019), and VizWiz (Gurari et al., 2018). For video understanding, we consider both online streaming and offline long-video scenarios. Specifically, we adopt OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026) for online video understanding evaluation, and VideoMME (Fu et al., 2025a), MLVU (Zhou et al., 2025), and LongVideoBench (Wu et al., 2024) for offline long-video understanding evaluation.

4.2 Main Results

Image Understanding Tasks. As shown in Table 1, we evaluate RESET on eight image understanding benchmarks with LLaVA-1.5-7B under three token budgets of 192, 128, and 64. RESET achieves the highest average performance retention across all compression settings, preserving 99.45%, 99.22%, and 97.61% of the vanilla performance, respectively. Notably, since RESET follows the same overall compression framework as ApET and differs mainly in the anchor-token selection strategy, the comparison directly reflects the benefit of our residual-based selection. As the token budget decreases from 192 to 64, the performance-retention margin over ApET increases from 0.34 to 1.12 and 1.97 percentage points, respectively. This widening gap indicates that RESET becomes increasingly effective under more aggressive compression, where selecting a compact yet representative token subset is more critical.

Video Understanding Tasks. As shown in Table 2, RESET remains effective when extended from image token compression to both online and offline video understanding. Despite operating on compressed visual representations, RESET consistently surpasses the Qwen2.5-VL backbone across all five aggregate benchmarks, improving StreamingBench, OVO-Bench, VideoMME, MLVU, and LongVideoBench by 4.8, 5.5, 0.8, 5.9, and 2.1 points, respectively. Compared with existing video compression methods, RESET further achieves the best results on StreamingBench, OVO-

Table 2: **Performance comparison on online and offline video understanding benchmarks.** StreamingBench reports real-time accuracy and OVO-Bench reports overall accuracy. Video-MME is evaluated without subtitles. "–" denotes unavailable results. The best results are shown in **bold**.

Method	Size	Frames	Online Video		Offline Video					
			StreamingBench	OVO-Bench	VideoMME				MLVU	LongVideoBench
			Real-Time	Overall	Short	Medium	Long	All	M-Avg	Val
Open-source Offline MLLMs										
LongVA (Zhang et al., 2024a)	7B	128	60.0	–	61.1	50.4	46.2	52.6	56.3	–
LongVU (Shen et al., 2024)	7B	1 fps	–	48.5	–	–	–	60.6	65.4	–
LLaVA-Video (Zhang et al., 2024c)	7B	1 fps	–	53.1	–	–	–	63.3	70.8	–
LLaVA-OneVision (Li et al., 2024)	7B	32	71.1	52.9	70.1	56.4	48.8	58.4	64.7	56.5
InternVL2.5 (Chen et al., 2024b)	8B	64	–	–	–	–	–	64.2	68.9	60.0
Video Token and Memory Efficiency Methods										
Qwen2.5-VL (Bai et al., 2025)	7B	1 fps	73.9	49.8	73.8	62.4	53.8	63.3	67.9	60.7
ReKV (Di et al., 2025)	7B	0.5 fps	69.1	–	–	–	–	–	68.5	–
StreamChat (Xiong et al., 2025)	8B	1 fps	64.7	–	–	–	–	–	–	–
LiveVLM (Ning et al., 2025)	7B	0.5/0.2 fps	72.9	–	66.7	56.4	48.8	57.3	66.3	–
StreamMem Yang et al. (2025b)	7B	4.0/0.5 fps	–	–	71.5	62.4	52.3	62.1	65.9	–
SURGE Tang et al. (2026)	7B	–	–	–	–	–	–	60.9	65.7	59.4
FlashVID Fan et al. (2026)	7B	–	–	–	71.1	58.7	49.1	59.6	–	58.1
Sparsity Forcing Chen et al. (2026a)	7B	–	–	–	–	–	–	64.0	69.5	–
QDP Zhang et al. (2026)	7B	1 fps	–	–	–	–	52.6	63.6	–	61.9
StreamingTOM Chen et al. (2026b)	7B	0.5 fps	–	–	69.9	**65.2**	**54.0**	63.0	–	–
EarlyTom Wang et al. (2026a)	7B	$\leq$768	–	–	70.7	61.6	53.6	61.9	–	–
HAWK Zhu et al. (2026)	7B	–	–	–	60.1	51.0	45.6	52.2	–	–
ProtoKV Minh et al. (2026)	7B	–	77.3	54.4	–	–	–	–	–	–
FluxMem Xie et al. (2026)	7B	1 fps	76.4	53.3	**76.9**	65.1	**54.0**	**65.3**	73.1	61.1
RESET (Ours)	7B	1 fps	**78.7**	**55.3**	**76.9**	62.2	53.2	64.1	**73.8**	**62.8**

Table 3: **Ablation on the parallel group size.** Experiments are conducted on the six real-time tasks of OVO-Bench: OCR, ACR, ATR, STU, FPD, and OJR, with the medium-memory window fixed to 64 frames for all configurations, using a single NVIDIA H800 GPU.

Group Size	Parallel Groups	OVO-Bench Real-Time Accuracy (%) ↑							Compression Latency (ms) ↓	End-to-End Latency (ms) ↓
		OCR	ACR	ATR	STU	FPD	OJR	Avg.		
64	1	92.6	67.9	73.3	57.3	74.3	65.8	71.9	771.6	2097.1
32	2	92.6	67.9	73.3	57.9	74.3	66.9	72.1	453.5	1748.0
16	4	92.6	68.8	72.4	57.3	74.3	66.9	72.0	289.0	1564.7
8	8	92.6	68.8	71.6	56.7	73.3	67.4	71.7	212.6	1478.3
4	16	92.6	67.9	73.3	57.3	73.3	66.3	71.8	177.2	1445.5
2	32	92.6	67.9	74.1	59.6	73.3	67.9	72.6	158.4	1421.6

Bench, MLVU, and LongVideoBench. These consistent gains across heterogeneous video benchmarks suggest that reference-conditioned residual selection effectively removes redundant visual information while preserving task-relevant content, allowing compression to improve rather than merely retain the performance of the uncompressed backbone.

4.3 Ablation Studies

Analysis on Parallel Group Size. We study the effect of parallel group size on the Real-Time Visual Perception subset of OVO-Bench, with the medium-memory window fixed to 64 frames. As shown in Table 3, decreasing the group size introduces more parallel groups and substantially reduces compression latency, from 771.6 ms at a group size of 64 to 158.4 ms at a group size of 2. End-to-end latency is correspondingly reduced from 2097.1 ms to 1421.6 ms. Meanwhile, accuracy remains stable across all configurations, ranging from 71.7% to 72.6%, with the best average accuracy obtained at a group size of 2. These results show that grouped parallel execution effectively reduces the sequential overhead of residual selection while maintaining stable prediction accuracy across different group sizes.

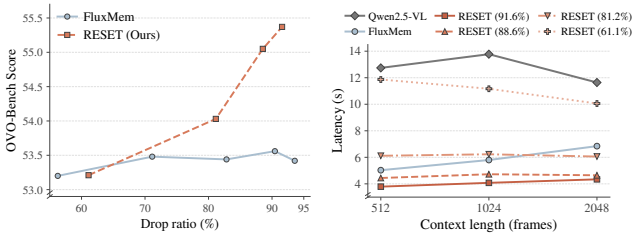
Effectiveness of Historical Reference. Unlike image compression, video compression must account for substantial temporal redundancy across frames. We therefore introduce a parameter-free historical reference to remove information already explained by previous visual content before residual-based token selection. As shown in Fig. 3, incorporating the historical reference consistently improves performance across MLVU, LongVideoBench, and VideoMME under all retention ratios.



Figure 3: **Ablation on the historical reference.** We compare RESET with and without the parameter-free historical reference under different retention ratios on three offline video benchmarks.

The gains are particularly evident on LongVideoBench and VideoMME, indicating that conditioning on historical information helps allocate the limited token budget to temporally novel content rather than repeatedly preserving redundant visual information.

Inference Efficiency and Drop-Ratio Analysis. We further evaluate RESET under different compression levels and context lengths. As shown in Fig. 4(a), RESET achieves higher OVO-Bench scores as the visual token drop ratio increases, whereas FluxMem remains relatively stable under fixed-threshold settings, suggesting that RESET is particularly effective under aggressive compression. Fig. 4(b) shows that token compression substantially reduces end-to-end latency compared with Qwen2.5-VL, while RESET further achieves lower latency than FluxMem under high drop ratios. Latency does not necessarily increase monotonically with context length because Qwen2.5-VL uses a fixed visual-pixel budget. As more frames are included, fewer pixels are allocated to each frame, reducing the per-frame spatial resolution and partially offsetting the additional temporal computation.



(a) Accuracy vs. drop ratio. (b) Latency vs. context length.

Figure 4: **Accuracy and efficiency analysis.** (a) OVO-Bench score under different visual token drop ratios. For FluxMem, fixed-threshold variants are used to obtain controlled compression levels. (b) End-to-end latency under different context lengths. The percentages in parentheses indicate the overall visual token drop ratio over the entire video, i.e., the fraction of visual tokens removed relative to the uncompressed input.

5 Conclusion

This paper presents RESET, a training-free visual token compression framework based on reference-conditioned residual subspace preservation. By removing information already explained by the reference subspace and iteratively selecting tokens with the largest residual energy, RESET measures token value by its incremental contribution rather than by standalone importance. The same residual selection operator naturally supports both image and video compression, using the zero reference for images and the parameter-free historical reference for videos. Experiments across image and video benchmarks show that RESET consistently preserves strong understanding performance under aggressive token reduction, while grouped parallel execution further reduces the computational overhead of residual selection. These results demonstrate the effectiveness of conditional residual information as a unified principle for efficient visual token compression.

Limitation. The efficiency advantage of RESET is less pronounced under mild compression, where the overhead of residual selection becomes relatively more significant. Future work may explore more lightweight selection and subspace-update strategies to further improve efficiency in this regime.

AI USE STATEMENT

In this work, we use generative AI tools for writing refinement and code editing assistance. The core research ideas, methodology design, implementation, and experimental analysis are conducted by the authors. All AI-assisted outputs are manually reviewed and verified for correctness and consistency before use. We take full responsibility for the final content of this work, including all claims and artifacts.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. All experimental settings are included in Appendix A.3. The datasets used in our experiments are publicly available.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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.

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.

Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pp. 968–977. SIAM, 2009.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

Feng Chen, Yefei He, Lequan Lin, Jing Liu, Chenhui Gou, Bohan Zhuang, and Qi Wu. Sparsity forcing: reinforcing token sparsity of mllms. In *International Conference on Learning Representations*, volume 2026, pp. 155748–155762, 2026a.

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*, pp. 19–35. Springer, 2024a.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24675–24685, 2026b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

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*, pp. 24185–24198, 2024c.

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. Floc: Facility location-based efficient visual token compression for long video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 102973–102996, 2026.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14582–14592. IEEE, 2025.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, volume 2026, pp. 48075–48099, 2026.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. Flashvid: Efficient video large language models via training-free tree-based spatiotemporal token merging. *arXiv preprint arXiv:2602.08024*, 2026.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. *arXiv preprint arXiv:2603.09480*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025a.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22654–22663. IEEE, 2025b.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

ZJ Guo, ML Li, MY Zhou, WL Peng, S Li, ZX Qian, and XP Zhang. Survey on digital watermarking technology for artificial intelligence generated content models. *J. Cybersecur*, 2(1):13–39, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Qiankun Ma, Ziyao Zhang, Haofei Wang, Zhen Song, Jie Chen, and Hairong Zheng. Apet: Approximation-error guided token compression for efficient vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26306–26316, 2026.

Le Tu Ngoc Minh, Jinyeong Lim, and Dongsu Han. Protokv: Streaming video understanding under delayed query with summary-state memory. *arXiv preprint arXiv:2606.26762*, 2026.

Zhenyu Ning, Guangda Liu, Qihao Jin, Chengwei Li, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Chong Tang, Sannara Ek, Dirk Koch, Robert Mullins, Alex Weddell, and Jagmohan Chauhan. Surge: Surprise-guided token reduction for efficient video understanding with vlms. In *International Conference on Learning Representations*, volume 2026, pp. 3849–3868, 2026.

Hesong Wang, Xin Jin, Lu Lu, Chenhaowen Li, Jian Chen, Qiang Liu, and Huan Wang. Earlytom: Early token compression completes fast video understanding. *arXiv preprint arXiv:2605.30010*, 2026a.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18523–18533, 2026b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. *arXiv preprint arXiv:2603.02096*, 2026.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603. IEEE, 2025.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *International Conference on Learning Representations*, volume 2025, pp. 69332–69351, 2025.

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)*, pp. 19792–19802. IEEE, 2025a.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025b.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *The Fourteenth International Conference on Learning Representations*, 2026.

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*, 2024a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Qihui Zhu, Tao Zhang, Yuchen Wang, Zijian Wen, Mengjie Zhang, Shuangwu Chen, Xiaobin Tan, Jian Yang, Yang Liu, Zhenhua Dong, et al. Hawk: Head importance-aware visual token pruning in multimodal models. *arXiv preprint arXiv:2604.07812*, 2026.

A Experimental Details

A.1 Baseline Details

We compare RESET with representative baselines on both image and video understanding tasks. For image understanding, we consider training-free visual token compression methods spanning token merging, attention-guided pruning, progressive token reduction, reconstruction-based selection, and coverage-based selection. For video understanding, we additionally include representative offline video MLLMs and efficient video understanding methods based on token compression, KV-cache management, memory retrieval, and sparse computation.

A.1.1 Image Understanding Baselines

ToMe. ToMe (Bolya et al., 2022) reduces the visual sequence by progressively merging similar tokens rather than directly discarding them. It employs an efficient bipartite matching strategy to identify similar token pairs and combines their representations, thereby reducing token redundancy while preserving aggregated visual information.

FastV. FastV (Chen et al., 2024a) exploits the observation that visual-token attention becomes increasingly sparse in deeper layers of vision-language models. It estimates visual-token importance from attention patterns in early LLM layers and removes low-importance visual tokens in subsequent layers, reducing the computation of deeper transformer blocks without additional training.

SparseVLM. SparseVLM (Zhang et al., 2024b) performs text-guided visual token sparsification using the internal attention patterns of the language model. It identifies visual-relevant text tokens to estimate visual-token importance, progressively removes less relevant tokens through a layer-adaptive sparsification strategy, and recycles discarded information into compact representations to alleviate information loss.

PDrop. PDrop (Xing et al., 2025) progressively reduces visual redundancy along the depth of the language model. Instead of pruning a large fraction of visual tokens at a single layer, it divides the LLM into multiple stages and drops tokens at stage boundaries according to lightweight instruction-conditioned relevance scores, resulting in a gradually decreasing visual-token sequence.

VisionZip. VisionZip (Yang et al., 2025a) performs text-agnostic visual compression before the visual representations are processed by the language model. It identifies a set of dominant tokens according to the information aggregated by the vision encoder's self-attention, while the remaining representations are merged according to feature similarity to construct contextual tokens. The dominant and contextual tokens jointly form the compressed visual representation.

ApET. ApET (Ma et al., 2026) evaluates visual-token informativeness from the perspective of linear approximation error rather than attention-based importance. It constructs a set of basis tokens to approximate the original visual representations and assigns higher importance to tokens with larger reconstruction errors, which are less well explained by the existing basis. The selected tokens and basis tokens form the retained set, while the remaining tokens are merged into their most similar retained representations.

MMTok. MMTok (Dong et al., 2026) formulates visual token selection as a multimodal maximum-coverage problem. It jointly considers text-vision coverage, which preserves information relevant to the input query, and vision–vision coverage, which maintains representative information from the original visual token set. The two objectives are incorporated into a submodular coverage function, with tokens iteratively selected according to their marginal coverage gains.

A.1.2 Video Understanding Baselines

We consider two groups of video baselines. The first group consists of representative open-source video MLLMs that directly process sampled video frames. The second group includes methods specifically designed to improve the efficiency of long or streaming video understanding through token compression, memory management, KV-cache retrieval, or sparse computation.

LongVA. LongVA (Zhang et al., 2024a) investigates long-video understanding from the perspective of context-length extension rather than explicit visual token compression. It transfers the long-context capability of the language backbone to the visual modality, enabling substantially longer sequences of video frames and visual tokens to be processed within a unified multimodal context.

LongVU. LongVU (Shen et al., 2024) introduces adaptive spatiotemporal compression for long-video understanding. It first removes temporally redundant frames based on visual similarity and then performs text-guided frame reduction and spatial token compression to retain information relevant to the input query.

LLaVA-Video. LLaVA-Video (Zhang et al., 2024c) extends the LLaVA family toward video understanding through large-scale video instruction tuning. Its training data covers diverse video-language tasks such as detailed captioning, open-ended question answering, and multiple-choice question answering, providing strong general-purpose video understanding capabilities.

LLaVA-OneVision. LLaVA-OneVision (Li et al., 2024) unifies single-image, multi-image, and video understanding within a shared multimodal architecture. Its video capability benefits from cross-scenario transfer, allowing representations learned from image and multi-image supervision to generalize to video inputs.

InternVL2.5. InternVL2.5 (Chen et al., 2024b) is a general-purpose open-source MLLM supporting images, multiple images, and videos. It improves upon InternVL2 through enhanced data quality, training strategies, and model scaling, and serves as a strong offline video understanding baseline.

Qwen2.5-VL. Qwen2.5-VL (Bai et al., 2025) serves as the uncompressed backbone of our video experiments. It supports dynamic-resolution visual encoding and explicit temporal modeling for video inputs. We report its original inference performance without additional visual token compression as the reference backbone.

ReKV. ReKV (Di et al., 2025) is a training-free framework for streaming video question answering based on KV-cache retrieval. It processes incoming frames with a bounded active context, stores previously computed video KV representations, and retrieves query-relevant historical caches when a question arrives, avoiding repeated processing of the entire video history.

StreamChat. StreamChat (Xiong et al., 2025) targets streaming video understanding and interactive question answering through a memory-enhanced inference framework. It compresses redundant visual observations into historical memory and retrieves relevant information from the maintained memory when answering incoming queries.

LiveVLM. LiveVLM (Ning et al., 2025) introduces a streaming-oriented memory mechanism for online video understanding. It continuously maintains compressed historical visual information and retrieves complementary short- and long-term context when a query arrives, supporting efficient inference over continuously growing video streams.

StreamMem. StreamMem (Yang et al., 2025b) maintains a bounded, query-agnostic KV memory for streaming video. It estimates the importance of accumulated visual KV entries using generic query representations and continuously compresses historical information under a fixed memory budget.

SURGE. SURGE (Tang et al., 2026) performs training-free visual token reduction according to temporal surprise. It measures how well each incoming visual representation can be predicted from recent history and prioritizes less predictable tokens, thereby retaining temporally novel information while removing redundant observations.

FlashVID. FlashVID (Fan et al., 2026) jointly reduces spatial and temporal redundancy through a training-free compression pipeline. It first applies Attention and Diversity-based Token Selection to preserve representative tokens and subsequently performs tree-based spatiotemporal token merging to consolidate redundant visual representations.

Sparsity Forcing. Sparsity Forcing (Chen et al., 2026a) improves visual-token efficiency through reinforcement-learning-based post-training. It jointly optimizes task correctness and token efficiency under different token budgets, encouraging the model to preserve prediction quality while operating with increasingly sparse visual computation.

QDP. QDP (Zhang et al., 2026) performs token reduction by jointly considering query relevance and temporal novelty. It compares incoming visual tokens with historical representations while measuring their relevance to the current query, enabling both query-irrelevant and temporally redundant information to be discarded.

StreamingTOM. StreamingTOM (Chen et al., 2026b) is a training-free compression framework for causal streaming video understanding. Its causal temporal reduction module compresses incoming visual tokens before LLM processing, while an online quantized memory stores historical KV representations in compact form and retrieves relevant groups on demand.

EarlyTom. EarlyTom (Wang et al., 2026a) moves video token compression to an earlier stage of the Video-LLM pipeline to reduce computation within the vision encoder. It performs temporal merging among redundant frame representations and subsequently applies spatial token selection to preserve complementary dynamic and static visual information.

HAWK. HAWK (Zhu et al., 2026) performs training-free visual token pruning by accounting for the heterogeneous importance of attention heads. It combines head-level visual importance with text-guided token relevance and aggregates these signals to identify visual representations that are most useful for the current instruction.

ProtoKV. ProtoKV (Minh et al., 2026) addresses bounded-memory streaming video understanding by representing distant historical information with compact prototypes. It preserves detailed KV representations for recent observations while consolidating older visual information into a fixed-capacity prototype memory, keeping the active context bounded over long streams.

FluxMem. FluxMem (Xie et al., 2026) introduces a training-free hierarchical visual memory that progressively compresses streaming video according to temporal distance. Recent observations are maintained in short-term memory, Temporal Adjacency Selection reduces redundancy when representations are transferred to mid-term memory, and Spatial Domain Consolidation further aggregates spatially redundant information into long-term memory. Adaptive thresholds control the compression strength according to the statistics of each video.

A.2 Benchmark Details

A.2.1 Image Understanding Benchmarks

We evaluate image understanding on eight widely adopted benchmarks covering visual reasoning, general multimodal perception, hallucination, scientific reasoning, text recognition, and real-world visual question answering.

GQA. GQA (Hudson & Manning, 2019) evaluates compositional reasoning over real-world images. Its questions are constructed from scene graphs and require models to perform a variety of reasoning operations involving objects, attributes, relations, and spatial structures, making it particularly suitable for evaluating fine-grained visual reasoning.

MMBench and MMBench-CN. MMBench (Liu et al., 2024c) provides a comprehensive evaluation of multimodal perception and reasoning through carefully designed multiple-choice questions spanning diverse ability dimensions. We evaluate on both its English version (MMB) and Chinese version (MMB-CN), which enables assessment of multimodal understanding across different language settings.

MME. MME (Fu et al., 2026) is a comprehensive benchmark for evaluating both perceptual and cognitive capabilities of MLLMs. It consists of 14 subtasks covering perception-oriented abilities,

such as object recognition, counting, and spatial understanding, as well as cognition-oriented abilities including commonsense reasoning and numerical calculation. Unlike percentage-based accuracy benchmarks, MME reports an aggregated score over its constituent tasks.

POPE. POPE (Li et al., 2023) is designed specifically to evaluate object hallucination in vision-language models. It reformulates hallucination evaluation as binary object-presence questions and constructs negative samples under different sampling strategies, providing a controlled measure of whether a model generates object information unsupported by the image.

ScienceQA. ScienceQA (SQA) (Lu et al., 2022) is a multimodal multiple-choice benchmark covering a broad range of science topics. Questions may require jointly understanding visual content, textual context, and external scientific knowledge, making the benchmark useful for evaluating multimodal reasoning rather than visual perception alone.

TextVQA. TextVQA (Singh et al., 2019) evaluates the ability to understand and reason about textual information embedded in natural images. Answering its questions often requires recognizing scene text and combining the recognized textual content with other visual and linguistic cues.

VizWiz. VizWiz (Gurari et al., 2018) is a real-world visual question answering benchmark collected from blind and visually impaired users. Unlike conventionally curated VQA datasets, its images frequently contain challenging characteristics such as blur, poor framing, low illumination, and insufficient visual evidence, providing a complementary evaluation of robustness under unconstrained visual inputs.

A.2.2 Video Understanding Benchmarks

For video understanding, we evaluate RESET under both online streaming and offline long-video settings. OVO-Bench and StreamingBench measure the ability to process temporally evolving video streams, whereas Video-MME, MLVU, and LongVideoBench evaluate understanding and reasoning over complete long-form videos.

OVO-Bench. OVO-Bench (Niu et al., 2025) is designed specifically for online video understanding, where a model must reason according to the video content available at the timestamp when a query is issued. It contains 12 tasks organized into three temporal scenarios: backward tracing, which requires retrieving information from previously observed events; real-time understanding, which evaluates events occurring around the current timestamp; and forward active responding, which requires determining when sufficient future evidence has been observed before answering. We report the overall performance across its online video understanding tasks.

StreamingBench. StreamingBench (Lin et al., 2026) evaluates MLLMs under continuous streaming video inputs, where questions are issued at different timestamps rather than after the complete video has been observed. The benchmark contains 18 tasks covering real-time visual understanding, omni-source understanding, and contextual understanding. Following the setting in our main comparison, we report performance on the Real-Time Visual Understanding subset.

Video-MME. Video-MME (Fu et al., 2025a) is a comprehensive benchmark for video understanding across diverse domains and temporal scales. It contains 900 videos and 2,700 manually annotated question-answer pairs, with video durations ranging from seconds to approximately one hour. Videos are divided into short-, medium-, and long-duration subsets, allowing performance to be analyzed as the temporal context increases. Following prior token-compression studies, we evaluate Video-MME without subtitles and report results for each duration group together with the overall score.

MLVU. MLVU (Zhou et al., 2025) is a multi-task benchmark designed to evaluate long-video understanding across diverse video genres and reasoning requirements. It covers scenarios including movies, egocentric videos, surveillance footage, cartoons, and game videos, and contains tasks requiring both retrieval of individual details and integration of information distributed over long temporal contexts. We report the multiple-choice average score (M-Avg.) following the standard evaluation protocol.

LongVideoBench. LongVideoBench (Wu et al., 2024) focuses on long-context video-language understanding with videos extending up to approximately one hour. A central component of the benchmark is referring reasoning, where a question refers to particular events or contexts in a long video and requires the model to retrieve the relevant temporal evidence before performing reasoning. The benchmark contains thousands of human-annotated multiple-choice questions spanning 17 fine-grained categories, and we report accuracy on its validation split.

A.3 Implementation Details

A.3.1 Image Understanding

For image understanding, we follow the evaluation protocol and implementation of ApET (Ma et al., 2026). For the LLaVA-series models, visual token compression is performed at two stages: immediately after the vision encoder and at the 16-th layer of the LLM. Accordingly, the token budgets reported in our experiments denote the average number of visual tokens processed across LLM layers, rather than the number of tokens retained at a single compression stage. For LLaVA-1.5-7B, each image is represented by 576 visual tokens before compression, and we evaluate average token budgets of 192, 128, and 64. For LLaVA-NeXT-7B, each image is represented using four local views together with one downsampled global view. Following ApET, the reported token budget is measured as the total budget across all five views rather than the budget of an individual view. For Qwen2.5-VL-7B, the number of visual tokens varies across images due to its dynamic-resolution visual encoding. We therefore evaluate compression using different token retention ratios instead of fixed absolute token budgets. For each input, the retention ratio specifies the fraction of its original visual tokens retained under the corresponding compression setting, allowing the compression level to remain comparable across images with different input resolutions. For all image backbones, we retain the original evaluation pipeline and replace the corresponding token selection component with RESET.

A.3.2 Video Understanding

For video understanding, we follow the short–medium–long hierarchical memory organization of FluxMem (Xie et al., 2026). For online video evaluation, the short- and medium-term memory windows contain 8 and 64 frames, respectively, whereas for offline long-video evaluation they contain 8 and 512 frames. We use a per-stage retention ratio of 6.25% for online evaluation and 50% for offline evaluation at both the short-to-medium and medium-to-long compression stages. The historical buffer length T used by RESET is directly set to the corresponding medium-term memory length, i.e., $T = 64$ for online evaluation and $T = 512$ for offline evaluation. Grouped parallel selection uses a default group size of 8 throughout the video experiments.

B Additional Experimental Results

B.1 Cross-Backbone Image Understanding

We further evaluate RESET on LLaVA-NeXT-7B and Qwen2.5-VL-7B to examine its generalization across different MLLM architectures and visual tokenization schemes. As shown in Table 4, RESET consistently achieves the highest average performance retention on LLaVA-NeXT-7B, preserving 97.9%, 95.5%, and 92.0% of the vanilla performance under token budgets of 640, 320, and 160, respectively. The advantage becomes particularly evident under aggressive compression: at 160 tokens, RESET improves the average retention over ApET from 90.9% to 92.0%, while maintaining stronger performance on GQA, POPE, SQA, and TextVQA.

Table 5 further evaluates RESET on Qwen2.5-VL-7B, whose dynamic-resolution visual encoding results in a varying number of visual tokens across images. We therefore compare methods at retention ratios of 20%, 10%, and 5%. RESET again achieves the highest average retention at all three compression levels, reaching 97.8%, 93.0%, and 89.0%, respectively. In particular, at the 10% retention ratio, RESET outperforms ApET by 1.5 percentage points in average performance retention. These results demonstrate that the effectiveness of residual-based token selection is not specific to LLaVA-1.5, but transfers consistently across multi-view and dynamic-resolution MLLM architectures.

Table 4: **Performance comparison on LLaVA-NeXT-7B across five image understanding benchmarks.** The upper-bound input contains up to 2880 visual tokens. † indicates the reproduced results.

Method	GQA	POPE	SQA	MME	TextVQA	Avg.
Upper Bound, 2880 Tokens **(100%)**						
Vanilla	64.2	86.5	70.1	1851	61.3	100%
Retain 640 Tokens (↓ 77.8%)						
SparseVLM	60.3	84.2	–	1773	57.8	95.3%
VisionZip	61.3	85.9	–	1787	**60.2**	97.4%
ApET†	63.4	86.9	68.1	1798	56.2	97.0%
RESET (Ours)	**63.5**	**87.3**	**68.3**	**1819**	57.7	**97.9%**
Retain 320 Tokens (↓ 88.9%)						
SparseVLM	59.3	83.3	–	1690	58.8	94.0%
VisionZip	59.3	82.1	–	1702	**58.9**	93.8%
ApET†	**61.4**	84.9	66.8	**1754**	53.9	94.4%
RESET (Ours)	**61.4**	**86.4**	**67.0**	1749	56.2	**95.5%**
Retain 160 Tokens (↓ 94.4%)						
SparseVLM	51.2	72.7	–	1542	46.4	80.7%
VisionZip	55.5	74.8	–	1628	**56.2**	88.1%
ApET†	58.4	81.5	67.1	**1651**	51.7	90.9%
RESET (Ours)	**58.7**	**84.7**	**67.6**	1629	52.8	**92.0%**

Table 5: **Performance comparison on Qwen2.5-VL-7B across five image understanding benchmarks.** Results use dynamic-resolution visual inputs. † indicates the reproduced results.

Method	GQA	POPE	SQA	MME	MMB	Avg.
Upper Bound, Dynamic Resolution **(100%)**						
Vanilla	60.5	86.2	76.7	2327	83.3	100%
Retain 20% Tokens in Average (↓ 80%)						
VisionZip	56.8	82.4	**76.3**	2134	79.3	95.2%
MMTok	58.1	82.4	–	2217	79.3	95.5%
ApET†	59.8	85.8	75.0	**2250**	79.8	97.7%
RESET (Ours)	**59.9**	**86.1**	75.0	2221	**80.8**	**97.8%**
Retain 10% Tokens in Average (↓ 90%)						
VisionZip	52.4	78.9	**74.1**	2003	**75.6**	90.3%
MMTok	55.1	78.8	–	**2051**	74.7	90.1%
ApET†	56.0	82.8	72.4	2013	73.5	91.5%
RESET (Ours)	**58.9**	**83.7**	71.0	2035	**75.6**	**93.0%**
Retain 5% Tokens in Average (↓ 95%)						
VisionZip	46.3	66.4	–	1677	67.5	76.7%
MMTok	50.7	71.4	–	1796	65.9	80.7%
ApET†	54.2	**80.9**	**72.1**	1883	70.0	88.5%
RESET (Ours)	**55.4**	80.2	71.3	**1900**	**71.5**	**89.0%**

B.2 Detailed Results on Real-Time Video Understanding

Tables 6 and 7 provide task-level comparisons on OVO-Bench and StreamingBench, respectively. On OVO-Bench, RESET achieves an average accuracy of 71.7%, outperforming FluxMem by 4.5 percentage points. The improvement is particularly pronounced on OCR and ACR, where RESET improves accuracy from 81.2% to 92.6% and from 59.6% to 68.8%, respectively. RESET also improves STU and OJR while maintaining comparable performance on ATR. Although FluxMem performs slightly better on FPD, the overall results show that RESET provides consistent benefits across most real-time task categories rather than relying on improvements from a single task. A similar trend is observed on StreamingBench. RESET achieves an overall accuracy of 78.7%, compared with 76.4% for FluxMem, and obtains the best or tied-best results on most of the reported real-time tasks. In particular, RESET yields clear improvements on Clips Summarize, Text-Rich Understanding, Spatial Understanding, and Action Perception, with gains of 7.2, 3.4, 6.1, and 3.7 percentage points over FluxMem, respectively. It also maintains competitive performance on Object Perception, Causal Reasoning, Attribute Perception, Event Understanding, and Prospective Reasoning, while Counting remains a relatively weaker category. Overall, the task-level results across both benchmarks indicate that the performance gains of RESET extend to diverse real-time perception, spatiotemporal understanding, and reasoning tasks, supporting the robustness of residual-based token selection under streaming video compression.

Table 6: **Task-level results on the real-time subset of OVO-Bench.** The six real-time tasks are **OCR** (Optical Character Recognition), **ACR** (Action Recognition), **ATR** (Attribute Recognition), **STU** (Spatial Understanding), **FPD** (Future Prediction), and **OJR** (Object Recognition). Avg. denotes the average accuracy over the six real-time tasks. "–" denotes unavailable results. The best result in each column is shown in **bold**.

Method	Size	Frames	OCR	ACR	ATR	STU	FPD	OJR	Avg.
Open-source Offline MLLMs									
LongVU (Shen et al., 2024)	7B	1 fps	55.7	49.5	59.5	48.3	68.3	63.0	57.4
LLaVA-Video (Zhang et al., 2024c)	7B	1 fps	69.1	58.7	68.8	49.4	74.3	59.8	63.5
LLaVA-OneVision (Li et al., 2024)	7B	32	66.4	57.8	**73.3**	53.4	71.3	62.0	64.0
InternVL2.5 (Chen et al., 2024b)	8B	64	67.1	60.6	63.8	46.1	68.3	56.5	60.4
Video Token and Memory Efficiency Methods									
FluxMem (Xie et al., 2026)	7B	1 fps	81.2	59.6	70.7	53.4	**75.2**	63.0	67.2
RESET (Ours)	7B	1 fps	**92.6**	**68.8**	71.6	**56.7**	73.3	**67.4**	**71.7**

Table 7: **Task-level results on the real-time subset of StreamingBench.** The ten real-time tasks are **OP** (Object Perception), **CR** (Causal Reasoning), **CS** (Clips Summarize), **ATP** (Attribute Perception), **EU** (Event Understanding), **TR** (Text-Rich Understanding), **PR** (Prospective Reasoning), **SU** (Spatial Understanding), **ACP** (Action Perception), and **CT** (Counting). Overall denotes the accuracy reported under the corresponding StreamingBench evaluation protocol. "–" denotes unavailable results. The best result in each column is shown in **bold**.

Method	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Overall
Open-source Offline MLLMs											
LongVA (Zhang et al., 2024a)	70.0	63.3	61.2	70.9	62.7	59.5	61.1	53.7	54.7	34.7	60.0
LLaVA-OneVision (Li et al., 2024)	80.4	74.2	76.0	80.7	72.7	71.7	67.6	65.5	65.7	45.1	71.1
InternVL2.5 (Chen et al., 2024b)	68.1	60.9	69.4	77.1	67.7	62.9	59.3	53.3	55.0	**56.5**	63.7
Video Token and Memory Efficiency Methods											
Qwen2.5-VL (Bai et al., 2025)	78.3	80.5	79.8	82.4	75.5	80.4	74.1	62.6	67.6	51.1	73.9
ReKV (Di et al., 2025)	74.4	78.9	78.6	77.1	68.3	67.9	67.6	62.6	64.3	44.6	69.1
LiveVLM (Ning et al., 2025)	**81.5**	78.1	83.3	79.1	69.6	74.1	75.0	69.1	67.7	40.4	72.9
FluxMem (Xie et al., 2026)	80.2	81.1	81.4	85.3	78.0	83.8	**80.6**	65.9	69.6	52.1	76.4
RESET (Ours)	81.0	**81.9**	**88.6**	**86.5**	**78.6**	**87.2**	**80.6**	**72.0**	**73.3**	45.2	**78.7**