RISE: RECURRENT EVIDENCE CONSOLIDATION WITH KERNELIZED SELECTION FOR EFFICIENT LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although recent advancements in video large language models (Video LLMs) have enabled remarkable vision-language reasoning, long-video understanding suffers from the scalability issue due to substantial overhead of large-scale visual tokens. Existing approaches typically identify informative tokens based on local density or static saliency, retaining redundant semantically similar tokens while failing to sufficiently cover global video semantics. Towards this end, we propose a novel approach named Recurrent Evidence Consolidation with Kernelized Selection (**RISE**) for efficient long-video understanding. The core of **RISE** is to compress redundant visual tokens from both the views of temporal consistency and information utility for efficient reasoning. In particular, our **RISE** first calculates the attention-based consistency scores to quantify recurrent patterns among spatio-temporally neighboring tokens in a sliding-window manner. These scores would be utilized to guide the anchor-based aggregation of evidence-consistent tokens from neighboring frames. In addition, we measure the conditional evidence gain according to a determinantal point process (DPP) kernel and iteratively select the most informative candidates from the remaining tokens. In this way, we maximize the information utility of the selected tokens within the constraints of computational budgets. Extensive experiments on four video benchmarks demonstrate that **RISE** consistently outperforms state-of-the-art baselines across various video LLM backbones. The code is available at https://anonymous.4open.science/r/RISE_4_Video_Compression.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Yin et al., 2024; Zhang et al., 2024a; Li et al., 2024b) have emerged as powerful interfaces for vision-language understanding (Fu et al., 2026) across diverse applications (Xie et al., 2025; Lin et al., 2026; Song et al., 2025; Luo et al., 2024; Dong et al., 2026). Building upon these advances, video LLMs are tailored to sequential visual inputs composed of multiple static frames, enabling them to reason over dynamic scenes, consecutive actions, and evolving events (Nie et al., 2024; Wang et al., 2025d; Huang et al., 2024). Despite impressive progress, scaling video LLMs to long-video understanding incurs substantial computational overhead, especially for high-resolution and long-duration videos. This computational burden hinders the practical deployment of video inference in real-world scenarios, motivating growing efforts on video-token compression *to reduce redundant visual tokens during inference while preserving task-relevant visual information* (Li et al., 2024c; Qian et al., 2024). Although several methods incorporate learnable resamplers or projectors to aggregate abundant visual features into compact representations (Li et al., 2024d; Ye et al., 2025), they typically depend on additional fine-tuning to align the compressed tokens with the LLM's space, leading to computational overhead (Chen et al., 2024a; Shao et al., 2025a).

To circumvent the need for additional fine-tuning, various approaches have explored training-free video compression to accelerate Video-LLM inference without modifying the underlying model. Among these efforts, measuring the significance of visual tokens has become a central research focus. For instance, PruneVid (Huang et al., 2025) and LongVU (Shen et al., 2025b) leverage cross-modal attention scores to identify tokens or frames most relevant to the given question, based on the assumption that higher attention scores indicate greater contribution to reasoning. While

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.

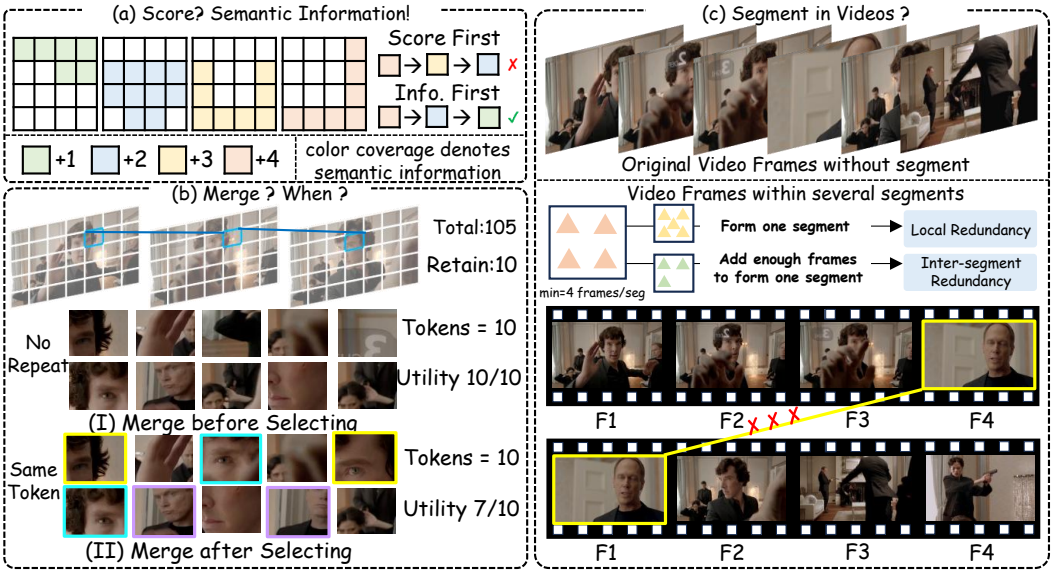


Figure 1: Motivating insights drawn from prior studies. (a) Token significance should be assessed not merely by local information density, but also by conditional evidence gain; (b) Temporal recurrence should be considered prior to token selection; (c) Event-based segmentation improves selection coverage but may fragment coherent evidence.

effective for identifying question-relevant tokens, these methods are inherently question-dependent and incur repeated computational overhead when multiple queries are posed for the same video. Beyond these question-conditioned approaches, other methods introduce heuristic scores to quantify token redundancy or local density, thereby pruning uninformative tokens or merging similar ones. While widely adopted, this pipeline raises a key question: ❶ *How can token significance be assessed beyond a single hand-crafted measure by conditioning on the existing evidence?*

In low-motion or high-resolution videos, similar visual evidence tends to recur across consecutive frames, making token merging necessary to reduce redundant representations (Tao et al., 2025; Hyun et al., 2025; Shao et al., 2025b). To address this issue, existing methods can be broadly categorized into two paradigms: merge-then-select and select-then-merge. The former paradigm, represented by FrameFusion (Fu et al., 2025b), performs similarity-based merging before importance-based pruning, whereas the latter paradigm, including VisionZip (Yang et al., 2025b) and FlashVID (Fan et al., 2026), first selects dominant or representative tokens and then merges tokens at the spatio-temporal scale. As shown in Figure 1, we illustrate these two paradigms and observe that the merge-then-select strategy is more effective at preserving temporal recurrence. In addition, several frameworks, such as METoK (Wang et al., 2025a) and FlashVID (Fan et al., 2026), divide videos into coherent event-level segments and perform token compression within each segment. Although such segmentation can improve coverage over the full video, it may fragment semantic evidence and introduce additional compression bias across different temporal segments. Motivated by this discussion, we further ask: ❷ *How can visual tokens be globally selected by jointly optimizing semantic coverage and token utility?*

To address these aforementioned challenges, we introduce a novel framework named Recurrent Evidence Consolidation with Global Kernelized Selection (**RISE**) for efficient long-video understanding. Specifically, **RISE** first measures spatio-temporal recurrence patterns via attention-based consistency in a sliding-window manner, which then serves as a weighting criterion for merging raw tokens with similar semantic evidence into compact anchors to preserve recurring evidence.. Subsequently, we formulate token compression as a fixed-budget subset selection problem, which naturally aligns with the determinantal point process (DPP) framework. Furthermore, we observe that token significance should be determined not only by local information density but also by the marginal gain conditioned on the currently selected subset. Therefore, to reduce redundancy and improve semantic coverage, **RISE** initializes the subset with the most semantically informative token and then leverages a kernel encoding both token significance and pairwise similarity to incrementally expand the subset from a global perspective until the budget is reached. Additionally, we conduct extensive experiments across multiple retention ratios and diverse video LLM backbones, demonstrating that

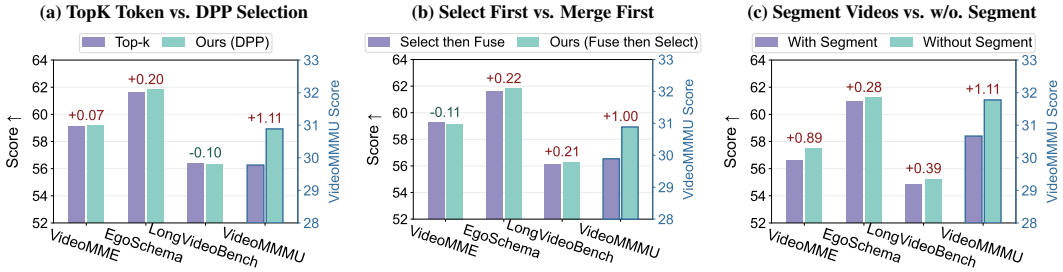


Figure 2: Empirical Observation Validation. (a) Influence of using semantic-aware selection instead of score-based selection. (b) Comparison of two processing orders. (c) Evaluation of segment-based selection against an unsegmented global strategy.

RISE consistently surpasses strong baselines while retaining robust performance under low retention ratios. In conclusion, the main contributions of RISE are summarized as follows:

❶ ***New Perspective.*** We revisit video token compression as an evidence-aware subset selection problem rather than isolated token pruning. This perspective reveals two key principles for effective long-video compression: (i) *evidence consolidation*, which treats recurrent spatio-temporal patterns as reusable semantic evidence, and (ii) *subset-aware selection*, which measures token value by its marginal semantic contribution under a limited token budget across diverse reasoning scenarios.

❷ ***Efficient Methodology.*** We propose RISE, a plug-and-play compression method for video LLMs. RISE first merges recurrent spatio-temporal tokens into compact semantic anchors, and then performs DPP-inspired global kernelized selection to preserve both important and diverse evidence.

❸ ***Empirical Validation.*** Comprehensive experiments across multiple video LLM backbones, benchmarks, and retention ratios demonstrate that RISE consistently outperforms strong baselines, validating its effectiveness and robustness for efficient long-video understanding.

2 Related Work

Video Large Language Models. Video large language models (Achiam et al., 2023; Li et al., 2024a; Bai et al., 2025; Lu et al., 2024; Shang et al., 2025) extend multimodal large language models (MLLMs) to sequential video reasoning by sampling multiple frames (Lin et al., 2024; Maaz et al., 2024; Zhang et al., 2024b; Xing et al., 2024; Zhang et al., 2025b). Video LLMs process continuous videos as discrete frame sequences, creating a trade-off between temporal coverage and computational efficiency (Li et al., 2024d). As more frames are sampled, the number of visual tokens grows rapidly, increasing the computational burden of subsequent language-model inference. This issue becomes more pronounced for long videos or dense frame sampling, which substantially increases memory consumption and prefill latency, motivating the reduction of redundant visual tokens while retaining task-relevant ones across increasingly long and densely sampled video sequences.

Token Compression in Video LLMs. Token compression (Wei et al., 2023; Liu et al., 2023; 2025; Wang et al., 2025b; Chen et al., 2025; Ju et al., 2026; Shao et al., 2025b; Rao et al., 2021; Liang et al., 2022; Cao et al., 2024) reduces computational complexity by removing redundant visual tokens (Yao et al., 2026). Existing studies introduce learnable projectors (Li et al., 2024d; Cha et al., 2024; Li et al., 2025) or memory modules (Song et al., 2025; 2024), while training-free methods (Zou et al., 2025; Tong et al., 2025; Baek et al., 2026; Yu et al., 2026; Alvar et al., 2025; Yang et al., 2025a; Wang et al., 2025c; Fayyaz et al., 2022; Bolya et al., 2023) compress tokens through frame selection, token merging, or pruning. FastVID (Shen et al., 2025a) performs spatio-temporal pruning after temporal segmentation, while AOT (Li et al., 2026) aggregates informative tokens through local and global context. Our observations suggest that consolidating recurrent evidence before kernelized subset optimization can further reduce redundancy while preserving global semantic coverage.

3 Empirical Observations on Video Compression

In this section, we revisit training-free visual token compression and identify three observations that motivate RISE. We begin by observing that token importance should be measured by its marginal semantic gain relative to the selected subset, rather than local information density alone (**Obs. ❶**). Subsequently, temporally continuous and semantically similar evidence should be aggregated before selection (**Obs. ❷**). Although conventional event-based segmentation can maintain temporal coverage, it may fragment coherent evidence and introduce bias due to intra-segment dependencies (**Obs. ❸**).

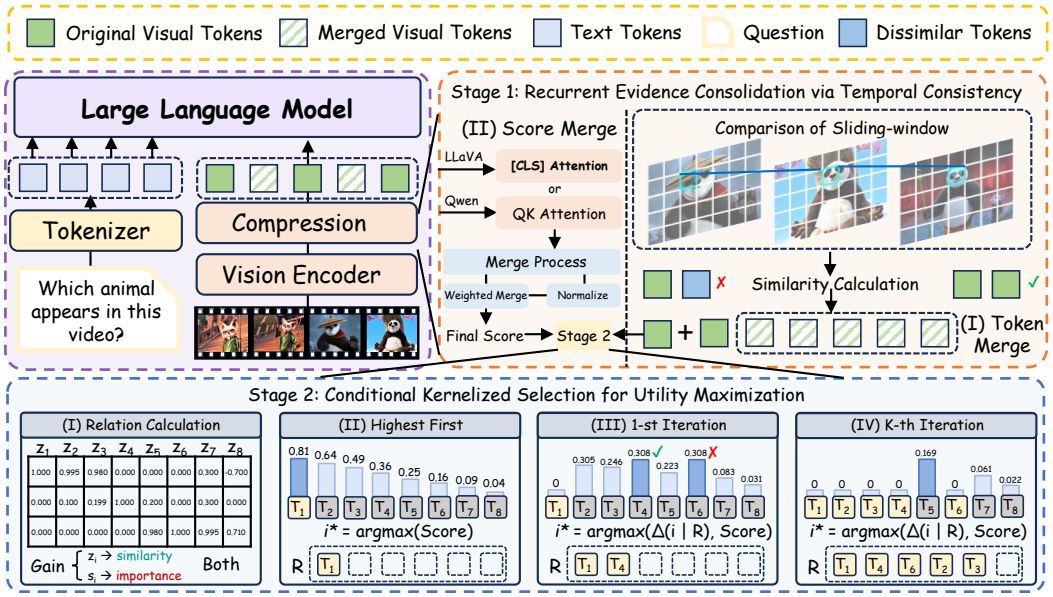


Figure 3: Overview of **RISE**. (a) **RISE** first merges tokens according to their spatio-temporal similarity across adjacent frames, thereby aggregating semantic information. (b) **RISE** selects the most important tokens as semantic anchors, and then further selects tokens based on DPP by considering both semantic information density and the marginal gain over the already selected subset.

Obs. ❶ *Token importance lies in conditional evidence gain rather than local information density.* As shown in Figure 2(a), top-K selection substantially underperforms our DPP-based strategy in relative accuracy, indicating that directly selecting the top-K high-scoring tokens leads to local information redundancy, resulting in insufficient semantic coverage of the retained tokens, which further causes performance degradation after compression under increasingly constrained token budgets.

Obs. ❷ *Temporal recurrence should be aggregated as evidence prior to token selection.* To study when token fusion should be performed, we conduct experiments by applying token fusion before and after selecting K tokens. The results in Figure 2(b) show that the merge-before-select strategy can make semantic information more compact and dense. By reducing the number of tokens while enriching their semantic representations, this strategy achieves better performance.

Obs. ❸ *Event-based segmentation improves temporal coverage but fragments coherent evidence, thereby undermining global token selection.* To assess the effect of video segmentation, we compare DySeg (Shen et al., 2025a) with a non-DySeg strategy in Figure 2(c). The non-DySeg strategy consistently performs better across all four datasets, improving relative accuracy by 1.29%. This indicates that segmenting the video may introduce semantic confusion across frames and token redundancy, thus increasing the difficulty of token selection under limited global token budgets.

Motivated by these insights, we consolidate local evidence before global token selection based on conditional marginal gain. Local consolidation reduces the token pool and produces stronger candidates, while global selection encourages complementary tokens and improves semantic coverage. Detailed information about the experiments for these observations is reported in Appendix E.1.

4 The Proposed RISE

Problem Formulation. Given a video $\mathcal{V} = \{I_1, I_2, \cdots, I_T\}$ consisting of T frames, we employ a frozen visual encoder E_v to transform $\mathcal{V}$ into a sequence of visual tokens, denoted as $\mathbf{Z} = E_v(\mathcal{V}) = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\} \in \mathbb{R}^{N \times d}$, where N and d denote the number of visual tokens and the token embedding dimension, respectively. Let $\mathcal{Q} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M\}$ represent the textual question associated with the video, where M stands for the number of question tokens. Our goal is to identify a compact, question-agnostic subset of visual tokens, denoted as $\mathbf{Z}' = \{\mathbf{z}'_1, \mathbf{z}'_2, \cdots, \mathbf{z}'_K\} \in \mathbb{R}^{K \times d}$, that maximally preserves essential spatio-temporal semantics while reducing the original token sequence to a predefined budget of K, where $K \ll N$. The subset selection is performed in a training-free manner and remains independent of the textual question $\mathcal{Q}$.

4.1 Overview of RISE

In this work, we focus on visual token compression for video LLMs, aiming to reduce redundant tokens while preserving global semantics under increasingly constrained long-video inference budget. Guided by the observations above, we adopt a fuse-then-select design that consolidates recurrent temporal evidence before globally selecting tokens based on conditional marginal gain. Within each sliding window, we identify recurrent evidence according to decision-aware token importance and local semantic similarity, and progressively merge highly similar neighboring tokens while accumulating their evidence mass. Subsequently, we construct a kernel that jointly encodes token significance and pairwise similarity, which is used to quantify the marginal gain of candidate tokens. The significance term emphasizes tokens carrying concentrated semantic evidence, while the similarity term suppresses redundant candidates, allowing the marginal gain to favor tokens that provide complementary information to the selected subset. Based on this kernel, we greedily select tokens with the largest marginal gain under the fixed budget. The framework of RISE is shown in Figure 3.

4.2 Recurrent Evidence Consolidation via Temporal Consistency

Our overall objective is to reduce redundant visual tokens without discarding informative evidence. Motivated by the observations above, we need to identify which tokens are semantically reliable and which recurrent patterns can be properly consolidated across adjacent frames in Stage 1. Therefore, before global token selection, we estimate the semantic evidence of each token and search for locally consistent correspondences across frames. Specifically, we first introduce [CLS] attention as our semantic evidence score. Following prior work (Li et al., 2026), these scores are extracted from the vision encoder. For models lacking an explicit [CLS] token, we calculate the score $q_{t,i}$ for each token i within the same frame by averaging the attention weights it receives as an attention target across all query tokens and all attention heads in the final self-attention layer. Subsequently, for each token $\mathbf{x}_{t,i}$ in frame t, we define a fixed sliding window and compute the cosine similarity between $\mathbf{x}_{t,i}$ and all tokens within the corresponding local window of the i-th token in the next frame $t+1$:

$$Sim_{i,j}^{(t)} = \cos(\mathbf{x}_{t,i}, \mathbf{x}_{t+1,j}), \tag{1}$$

where $Sim_{i,j}^{(t)}$ measures the similarity between tokens $\mathbf{x}_{t,i}$ and $\mathbf{x}_{t+1,j}$. Then, we perform a weighted merging process for tokens whose similarity exceeds a fusion threshold T_m. Specifically, we define a mass $m_{t,i} = q_{t,i}^2$ to represent the weight proportion of each token. When token i from frame t is merged into token j from frame $t+1$, the updated feature $\tilde{\mathbf{x}}_{t+1,j}$ is computed as a weighted average of the merged tokens, preserving the accumulated evidence from both temporal observations:

$$\tilde{\mathbf{x}}_{t+1,j} = \frac{m_{t+1,j}\mathbf{x}_{t+1,j} + \sum_{i\to j} m_{t,i}\mathbf{x}_{t,i}}{m_{t+1,j} + \sum_{i\to j} m_{t,i}}. \tag{2}$$

Subsequently, the updated attention score $\tilde{q}_{t+1,j}$ is derived from the accumulated mass, representing the enhanced semantic density of the fused anchor:

$$\tilde{q}_{t+1,j} = \sqrt{m_{t+1,j} + \sum_{i\to j} m_{t,i}}, \tag{3}$$

where $i \to j$ denotes the set of tokens in frame t that are merged into token j in frame $t+1$. After merging, the original tokens in frame t are removed, effectively reducing token redundancy while preserving and identifying high-value semantic information in the subsequent frame.

4.3 Conditional Kernelized Selection for Utility Maximization

After the process in Section 4.2, we denote the remaining tokens as the surviving-tokens set $\mathcal{S}$. In this stage, our goal is to identify K tokens from $\mathcal{S}$ that are both semantically representative and information-intensive across the global sequence. To achieve this, we formulate the selection as a Determinantal Point Process (DPP) based subset selection problem. Specifically, for each candidate token $i \in \mathcal{S}$, let q_i be its semantic evidence score derived from the first stage. We first project each normalized feature into a low-rank subspace, obtaining a compact representation $\mathbf{z}_i \in \mathbb{R}^r$. This representation captures the principal information directions of the token. To model the synergy between token quality and diversity, we define a quality-weighted basis vector $\mathbf{b}_i$ as:

$$\mathbf{b}_i = \sqrt{q_i}\mathbf{z}_i. \tag{4}$$

By stacking these vectors into a matrix $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{|\mathcal{S}|}]^\top$, we construct a kernel matrix $\mathbf{L} = \mathbf{B}\mathbf{B}^\top$. The entries of L are computed as:

$$L_{ij} = \mathbf{b}_i^\top b_j = \sqrt{q_i q_j}(\mathbf{z}_i^\top z_j), \tag{5}$$

where $\sqrt{q_i q_j}$ represents the joint semantic intensity of the token pair, while $\mathbf{z}_i^\top z_j$ measures their feature similarity. In this framework, the probability of selecting a subset of tokens is proportional to the determinant of the corresponding sub-matrix of L. Geometrically, this determinant corresponds to the squared volume of the parallelepiped spanned by the vectors $\{\mathbf{b}_i\}_{i \in S}$. Based on this formulation, we model the selection as maximizing a set-level objective function:

$$F(R) = \log\det\left(\mathbf{I} + \sum_{j \in R} b_j b_j^\top\right), \tag{6}$$

where $R \subseteq \{1, 2, \ldots, C\}$ $(C \leq K)$ denotes the set of selected tokens and I is the identity matrix. Intuitively, a larger information volume spanned by the selected tokens in the low-rank subspace indicates that the subset can collectively preserve more significant and complementary information directions under a limited budget. To solve this efficiently, we employ a greedy iterative strategy. We first select the token with the highest semantic evidence score q_i as the initial anchor:

$$i_0 = \arg\max_i q_i. \tag{7}$$

In each subsequent iteration, we calculate the conditional evidence gain $g_i(R)$ for each candidate token based on the currently selected set R, to guide the next token selection:

$$g_i(R) = \mathbf{b}_i^\top \mathbf{M}_R^{-1} \mathbf{b}_i, \quad \mathbf{M}_R = \mathbf{I} + \sum_{j \in R} b_j b_j^\top. \tag{8}$$

To obtain more important semantic information within a limited token budget, we then update the set by selecting the candidate with the maximum gain at each greedy selection step:

$$i^\star = \arg\max_{i \notin R} g_i(R), \quad R \leftarrow R \cup \{i^\star\}. \tag{9}$$

This process is repeated until K tokens are collected. If numerical stagnation occurs, the remaining slots are filled by candidates with the highest semantic evidence scores.

4.4 Summarization of RISE

As summarized in Algorithm 1 (shown in Appendix 1), RISE takes visual token sequence $\mathcal{X}$ and target budget K as input, and first estimates a semantic evidence score for each token. In Stage 1, the algorithm scans adjacent frames in temporal order and matches and aggregates each token to its most similar neighbor in the next frame within a local spatial window under fusion threshold T_m. In Stage 2, utilizing surviving tokens, we project each candidate token into a z-dimensional subspace and combine it with its consolidated evidence score to construct a quality-aware DPP basis. Starting from the strongest semantic anchor, RISE greedily selects tokens with the largest conditional evidence gain until the budget is reached, and returns the selected tokens as the compressed visual sequence $\hat{\mathcal{X}}$ for subsequent Video LLM inference under a fixed budget while preserving semantic diversity.

4.5 Theoretical Analysis

We provide a theoretical analysis for the two core designs of RISE: ❶ recurrent evidence consolidation before selection and ❷ conditional DPP-based budget allocation. All variables follow the definitions in Sec. 4.2 and Sec. 4.3. All proofs are given in Appendix C.1 and Appendix C.2.

Assumption 4.1. *(Coherent recurrent consolidation). For the consolidation groups induced by the local matching rule in Eq. (1) and the fusion threshold T_m, each group preserves a minimum effective coherence: $\rho_C \geq \gamma(T_m)$, where ρ_C and $\gamma(T_m)$ are formally defined in Appendix C.1.*

Theorem 4.1. *(Retention guarantee under a fixed budget). Under Assumption 4.1, recurrent evidence consolidation in Eqs. (2)–(3) preserves first-order semantic evidence before token selection. Moreover, for any selected subset R with $|R| = K$, the retained information is bounded by*

$$\gamma(T_m)\mathrm{Cov}_K(R; T_m) \leq \mathrm{Ret}_K(R; T_m) \leq \mathrm{Cov}_K(R; T_m), \tag{10}$$

where $\mathrm{Ret}_K(R; T_m)$ and $\mathrm{Cov}_K(R; T_m)$ are defined in Appendix C.1.

Table 1: Comparison of state-of-the-art methods on LLaVA-OV. Best results are marked in **bold**. Red ↑, green ↓, and gray → denote absolute score increases, decreases, and unchanged results relative to Vanilla, respectively.

Method	*Rt*	VideoMME	EgoSchema		LongVideoBench	VideoMMMU				Avg.	
		Overall	Subset	Total	Score	Adapt.	Compr.	Percep.	Total	Score	Rel. Acc. (%)
Vanilla	100%	58.5	62.4	60.4	56.6	31.0	28.0	37.3	32.1	51.9	100.0
FastV (Chen et al., 2024b)		56.5 ↓ 2.0	60.4	57.8 ↓ 2.6	55.4 ↓ 1.2	**32.0**	30.7	28.0	30.2 ↓ 1.9	50.0 ↓ 1.9	96.3
VisionZip (Yang et al., 2025b)		58.1 ↓ 0.4	63.0	60.4 → 0.0	**56.4** ↓ 0.2	**32.0**	24.3	26.3	27.6 ↓ 4.5	50.6 ↓ 1.3	97.5
FastVID (Shen et al., 2025a)	25%	57.9 ↓ 0.6	61.2	59.5 ↓ 0.9	55.9 ↓ 0.7	31.3	**32.3**	**34.3**	**32.7** ↑ 0.6	51.5 ↓ 0.4	99.2
AOT (Li et al., 2026)		57.5 ↓ 1.0	**64.4**	61.3 ↑ 0.9	56.3 ↓ 0.3	**32.0**	31.3	33.7	32.3 ↑ 0.2	51.9 → 0.0	100.0
RISE		**59.1** ↑ 0.6	64.2	**61.8** ↑ 1.4	56.3 ↓ 0.3	31.3	30.0	31.3	30.9 ↓ 1.2	**52.0** ↑ 0.1	**100.3**
FastV (Chen et al., 2024b)		56.2 ↓ 2.3	60.6	57.6 ↓ 2.8	56.6 → 0.0	30.0	28.0	34.7	30.9 ↓ 1.2	50.3 ↓ 1.6	97.0
VisionZip (Yang et al., 2025b)		58.0 ↓ 0.5	62.0	60.0 ↓ 0.4	55.4 ↓ 1.2	28.0	29.0	28.0	28.3 ↓ 3.8	50.4 ↓ 1.5	97.2
FastVID (Shen et al., 2025a)	20%	57.9 ↓ 0.6	61.2	59.5 ↓ 0.9	55.9 ↓ 0.7	31.0	31.7	33.0	31.9 ↓ 0.2	51.3 ↓ 0.6	98.9
AOT (Li et al., 2026)		57.2 ↓ 1.3	**63.4**	**61.3** ↑ 0.9	56.2 ↓ 0.4	30.0	**34.3**	33.0	32.4 ↑ 0.3	51.8 ↓ 0.1	99.8
RISE		**58.4** ↓ 0.1	60.3	61.0 ↑ 0.6	**56.3** ↓ 0.3	**33.3**	31.0	**33.7**	**32.7** ↑ 0.6	**52.1** ↑ 0.2	**100.4**
FastV (Chen et al., 2024b)		54.6 ↓ 3.9	59.8	56.6 ↓ 3.8	54.8 ↓ 1.8	**32.0**	28.3	**34.3**	31.6 ↓ 0.5	49.4 ↓ 2.5	95.2
VisionZip (Yang et al., 2025b)		55.6 ↓ 2.9	62.8	60.0 ↓ 0.4	54.1 ↓ 2.5	24.3	25.3	32.0	27.2 ↓ 4.9	49.2 ↓ 2.7	94.9
FastVID (Shen et al., 2025a)	15%	57.7 ↓ 0.8	58.8	58.9 ↓ 1.5	**56.7** ↑ 0.1	30.3	**33.0**	33.3	**32.2** ↑ 0.1	51.4 ↓ 0.5	99.0
AOT (Li et al., 2026)		56.6 ↓ 1.9	63.2	**61.3** ↑ 0.9	55.2 ↓ 1.4	31.3	32.3	33.0	32.2 ↑ 0.1	51.3 ↓ 0.6	98.9
RISE		**57.9** ↓ 0.6	**63.4**	61.0 ↑ 0.6	56.0 ↓ 0.6	30.3	30.3	33.3	31.3 ↓ 0.8	**51.6** ↓ 0.3	**99.4**
FastV (Chen et al., 2024b)		52.7 ↓ 5.8	58.8	55.9 ↓ 4.5	52.4 ↓ 4.2	32.3	28.0	32.0	30.8 ↓ 1.3	48.0 ↓ 3.9	92.5
VisionZip (Yang et al., 2025b)		53.3 ↓ 5.2	63.2	59.8 ↓ 0.6	53.7 ↓ 2.9	28.3	24.7	25.0	26.0 ↓ 6.1	48.2 ↓ 3.7	92.9
FastVID (Shen et al., 2025a)	10%	56.9 ↓ 1.6	58.6	58.7 ↓ 1.7	**55.8** ↓ 0.8	30.3	**32.3**	32.7	31.8 ↓ 0.3	50.8 ↓ 1.1	97.9
AOT (Li et al., 2026)		52.7 ↓ 5.8	62.8	59.6 ↓ 0.8	51.4 ↓ 5.2	**33.3**	31.3	33.7	**32.8** ↑ 0.7	49.1 ↓ 2.8	94.7
RISE		**57.5** ↓ 1.0	**63.4**	**61.2** ↑ 0.8	55.2 ↓ 1.4	30.3	30.3	**34.7**	31.8 ↓ 0.3	**51.4** ↓ 0.5	**99.1**

Remark 4.1. *Theorem 4.1 explains why recurrent evidence should be consolidated before selection. Under the fixed token budget K, consolidation converts repeated temporal evidence into fewer high-mass candidates, thereby improving budget coverage. Meanwhile, the factor $\gamma(T_m)$ captures the fidelity of local fusion. Thus, the effectiveness of Stage 1 is governed by the trade-off between fusion fidelity and budget coverage, selecting more informative tokens.*

Assumption 4.2. *Non-degenerate DPP bases. For each surviving token $i \in \mathcal{S}$, the quality-aware DPP basis b_i defined in Eq. (4) is non-zero and finite. Therefore, the conditional gain $g_i(R)$ in Eq. (8) is well-defined for any selected set R.*

Theorem 4.2. *Exact conditional DPP volume gain. Under Assumption 4.2, for the DPP objective $F(R)$ defined in Eq. (6), adding any candidate token $i \notin R$ yields*

$$F(R \cup \{i\}) - F(R) = \log\left(1 + g_i(R)\right), \tag{11}$$

where $g_i(R)$ is the conditional evidence gain defined in Eq. (8). Therefore, selecting the candidate with the largest $g_i(R)$ exactly maximizes the one-step DPP volume gain.

Remark 4.2. *Theorem 4.2 shows that Stage 2 is not equivalent to top-K selection by individual semantic scores. Instead, the gain in Eq. (8) is conditioned on the already selected set, so each selected token is encouraged to provide complementary information beyond the current subset.*

5 Experiments

5.1 Experimental Setup

Benchmark Datasets. We evaluate our method on four widely used video understanding benchmarks, including VideoMME (Fu et al., 2025a), EgoSchema (Mangalam et al., 2023), LongVideoBench (Wu et al., 2024) and VideoMMMU (Hu et al., 2025). These benchmarks cover a wide spectrum of video durations and scenarios, from a few seconds to hour-long videos, and from daily activities and movies to knowledge-intensive professional lectures, across diverse temporal and semantic reasoning conditions, offering an overall evaluation of RISE. Additional details can be found in Appendix D.3.

Compared Baselines. We evaluate all methods on LLaVA-OneVision (Li et al., 2024a) and Qwen3-VL (Bai et al., 2025). The compared methods include FastV (Chen et al., 2024b), VisionZip (Yang et al., 2025b), FastVID (Shen et al., 2025a), and AOT (Li et al., 2026). These baselines cover training-free token compression strategies, including attention-based pruning, redundancy reduction, and adaptive token selection. All baselines are reimplemented in LMMs-Eval (Zhang et al., 2025a), following their official implementations. More details are reported in Appendix D.

Implementation Details. We evaluate our method on LLaVA-OneVision (Li et al., 2024a) and Qwen3-VL (Bai et al., 2025), covering different video understanding architectures and tokenization mechanisms. All experiments are conducted on an NVIDIA A6000 GPU using LMMs-Eval to ensure reproducible and consistent efficiency evaluation. More details are reported in Appendix D.1.

Table 2: Comparison of state-of-the-art methods on Qwen3-VL. The best results are marked in **bold**.

Method	*Rt*	VideoMME	EgoSchema		LongVideoBench	VideoMMMU				Avg.	
		Overall	Subset	Total	Score	Adapt.	Compr.	Percep.	Total	Score	Rel. Acc. (%)
Vanilla	100%	63.7	65.0	68.2	58.4	37.7	63.0	73.3	58.0	62.1	100.0
FastV (Chen et al., 2024b)	20%	55.6 ↓ 8.1	61.0	63.4 ↓ 4.8	50.2 ↓ 8.2	30.7	45.0	52.3	42.7 ↓ 15.3	53.0 ↓ 9.1	85.3
VisionZip (Yang et al., 2025b)		59.1 ↓ 4.6	64.4	65.3 ↓ 2.9	58.1 ↓ 0.3	29.3	54.7	62.7	48.9 ↓ 9.1	57.9 ↓ 4.2	93.2
FastVID (Shen et al., 2025a)		59.1 ↓ 4.6	62.2	64.3 ↓ 3.9	55.7 ↓ 2.7	29.3	59.0	62.0	50.8 ↓ 7.2	57.5 ↓ 4.6	92.6
AOT (Li et al., 2026)		58.3 ↓ 5.4	63.2	65.0 ↓ 3.2	55.3 ↓ 3.1	17.0	20.0	40.0	25.7 ↓ 32.3	51.1 ↓ 11.0	82.3
RISE		**62.4** ↓ 1.3	**67.8**	**67.2** ↓ 1.0	**58.9** ↑ 0.5	**31.0**	**59.7**	**71.0**	**53.9** ↓ 4.1	**60.6** ↓ 1.5	**97.6**
FastV (Chen et al., 2024b)	10%	52.5 ↓ 11.2	54.4	59.7 ↓ 8.5	47.8 ↓ 10.6	28.3	44.3	48.3	40.3 ↓ 17.7	50.1 ↓ 12.0	80.7
VisionZip (Yang et al., 2025b)		56.6 ↓ 7.1	61.8	63.5 ↓ 4.7	53.9 ↓ 4.5	29.7	45.3	57.0	44.0 ↓ 14.0	54.5 ↓ 7.6	87.8
FastVID (Shen et al., 2025a)		56.8 ↓ 6.9	58.8	62.6 ↓ 5.6	53.5 ↓ 4.9	**31.0**	52.7	57.3	47.0 ↓ 11.0	55.0 ↓ 7.1	88.6
AOT (Li et al., 2026)		57.6 ↓ 6.1	62.8	64.4 ↓ 3.8	53.9 ↓ 4.5	16.3	19.0	41.7	25.7 ↓ 32.3	50.4 ↓ 11.7	81.2
RISE		**60.5** ↓ 3.2	**65.4**	**65.8** ↓ 2.4	**55.9** ↓ 2.5	30.7	**54.7**	**67.7**	**51.0** ↓ 7.0	**58.3** ↓ 3.8	**93.9**

Table 3: Comparison of state-of-the-art methods on LLaVA-OneVision under a fixed token budget. More frames can be processed with token compression. The best results are marked in **bold**.

Method	#Frames	*Rt*	VideoMME	EgoSchema		LongVideoBench	VideoMMMU					Avg.
			Overall	Subset	Total	Score	Adapt.	Compr.	Percep.	Total	Score	Rel. Acc. (%)
Vanilla	32 (1x)	100%	58.5	62.4	60.4	56.6	31.0	28.0	37.3	32.1	51.9	100.0
FastVID (Shen et al., 2025a)	160 (5x)	20%	60.4 ↑ 1.9	62.8	61.5 ↑ 1.1	56.6 → 0.0	30.7	34.0	**37.7**	**34.1** ↑ 2.0	53.1 ↑ 1.2	102.4
AOT (Li et al., 2026)			56.9 ↓ 1.6	**65.4**	61.6 ↑ 1.2	53.3 ↓ 3.3	32.3	**34.7**	33.3	33.4 ↑ 1.3	51.3 ↓ 0.6	98.9
RISE			**61.3** ↑ 2.8	63.8	**61.9** ↑ 1.5	**58.0** ↑ 1.4	**32.3**	28.3	34.7	31.8 ↓ 0.3	**53.2** ↑ 1.3	**102.6**
FastVID (Shen et al., 2025a)	320 (10x)	10%	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
AOT (Li et al., 2026)			57.5 ↓ 1.0	63.2	61.4 ↑ 1.0	54.5 ↓ 2.1	**31.3**	**31.0**	30.3	**30.9** ↓ 1.2	51.1 ↓ 0.8	98.5
RISE			**60.9** ↑ 2.4	**63.2**	**61.4** ↑ 1.0	**57.9** ↑ 1.3	29.7	27.3	**35.0**	30.7 ↓ 1.4	**52.7** ↑ 0.8	**101.6**

5.2 Performance Comparison

Performance under Different retention ratios. We compare our proposed RISE with different baselines on four datasets using LLaVA-OneVision and Qwen3-VL in Table 1 and Table 2. Based on the results, we draw two main insights as follows: **Obs.❶** *Under relatively high retention ratios on LLaVA-OneVision (e.g., 0.25 and 0.20), RISE even outperforms the vanilla model.* On Qwen3-VL, RISE also shows a substantial advantage over other baselines, outperforming FastV by 12.3% at a 0.20 retention ratio. This demonstrates that our method effectively aggregates and preserves semantic information, thereby improving model performance. **Obs.❷** RISE *remains highly competitive under lower Rt.* At 10% *Rt*, it still retains 99.1% relative performance on LLaVA-OneVision. On Qwen3-VL, RISE surpasses FastV by 13.2%, further confirming the effectiveness of our method.

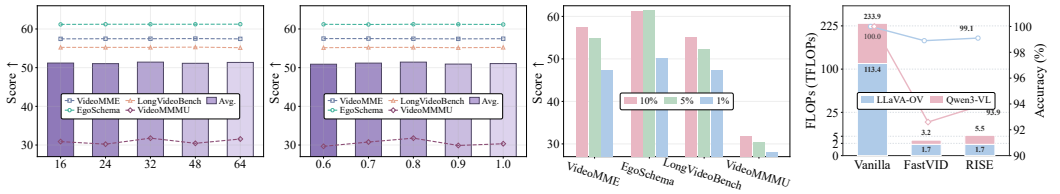
Results under Fixed Token Budget. Due to computational and memory constraints, existing Video LLMs can only process limited sampled frames, which may miss important visual cues in long videos. To examine whether token compression can extend the processable temporal context under a fixed token budget, we compare compression frameworks on LLaVA-OneVision in Table 3. We derive two remarks: **Obs.❸** *Simply increasing the number of frames does not necessarily improve performance.* Although more frames provide longer temporal context, aggressive compression may discard useful visual evidence and offset this benefit. When processing 320 frames with 10% retained tokens, AOT drops below the vanilla 32-frame setting, indicating that longer inputs require sufficiently effective token selection. **Obs.❹** RISE *more effectively exploits longer temporal context under the same token budget.* At 160 frames, RISE achieves the best average score and relative accuracy among all methods. When further extending to 320 frames, FastVID runs out of memory and AOT degrades to 98.5% relative accuracy, while RISE maintains 101.6% relative accuracy. These results demonstrate that RISE better preserves useful semantic evidence under constrained budgets, enabling Video LLMs to process longer videos more efficiently while maintaining robust task-relevant evidence coverage.

5.3 Ablation Study

Table 4: Ablation study of RISE on DPP.

Selection	VideoMME	EgoSchema	LongVideoBench	VideoMMMU	Avg.	
	Overall	Total	Score	Total	Score	Rel. Acc.
K-means	56.3	59.9	54.4	29.4	50.0	96.4
K-medoids	56.2	60.2	54.9	30.3	50.4	97.2
Spectral	56.2	59.9	54.5	28.6	49.8	96.0
RISE	**57.5**	**61.2**	**55.2**	**31.8**	**51.4**	**99.1**

As shown in Table 4, we ablate the DPP-based conditional subset selection by replacing only the DPP-based selector with K-means, K-medoids, and spectral clustering, while keeping all other settings fixed. The results lead to two findings: **Obs.❺** *Clustering-based selection is insufficient for token compression.* Although clustering methods can reduce redundancy by grouping similar tokens, they mainly rely on local feature similarity or distance-based partitioning, without optimizing the semantic gain of the selected subset. Thus, all clustering variants underperform RISE in average score and relative accuracy. **Obs.❻** *DPP-based selection better preserves complementary visual evidence.* Compared with clustering-based alternatives, RISE achieves the best overall performance across the four benchmarks, demonstrating that modeling token dependencies is crucial for selecting informative and diverse token combinations.



(a) Rank z (b) Fusion Threshold T_m (c) Retention Ratio R_t (d) Efficiency

Figure 4: Sensitivity and efficiency analyses of **RISE**. (a) Sensitivity to rank z. (b) Sensitivity to fusion threshold T_m. (c) Sensitivity to retention ratio R_t. (d) Efficiency of **RISE**, where **RISE** and FastVID use retention ratios of 10% and 20%, respectively.

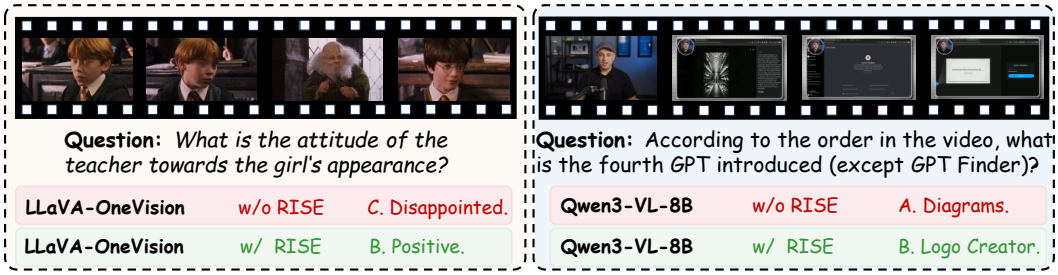


Figure 5: Case Analysis. Qualitative comparison between the Vanilla model and **RISE**.

5.4 Further Analysis

Sensitivity Study. We conduct sensitivity studies on Recurrent Evidence Consolidation (Section 4.2), the DPP rank z, and several retention ratios using LLaVA-OneVision. As shown in Figure 4, we make the following findings: **Obs.❼** ***RISE*** *is robust to hyperparameters.* For the fusion threshold T_m, the performance remains stable across several values, with the best result achieved at $T_m = 0.8$. For the rank z, the performance peaks at $z = 32$ and remains stable across different settings, indicating that **RISE** is not sensitive to its hyperparameters. **Obs. ❽** ***RISE*** *maintains competitive performance on most benchmarks at different retention ratios.* The results indicate that the selected tokens preserve compact yet informative visual evidence. Although performance decreases under the extremely aggressive 1% setting, **RISE** still retains meaningful video understanding ability.

Efficiency Analysis. As shown in Figure 4(d), we conduct efficiency experiments on VideoMME using LLaVA-OneVision and Qwen3-VL. We compare **RISE** with FastVID and the Vanilla model, making the following finding: **Obs. ❾** ***RISE*** *substantially reduces computational cost while preserving high relative accuracy across different Video LLMs.* On LLaVA-OneVision, **RISE** uses 10% visual tokens and reduces the computation from 113.4 TFLOPs to 1.7 TFLOPs, while still achieving 99.1% relative accuracy. On Qwen3-VL, **RISE** also substantially reduces the computation from 233.9 TFLOPs to 5.5 TFLOPs and preserves 93.8% relative accuracy. In contrast, although FastVID uses fewer TFLOPs, its relative accuracy drops to 79.3% on Qwen3-VL. These results demonstrate that **RISE** achieves a better trade-off between efficiency and performance across different Video LLMs.

Case Analysis. Taking LLaVA-OneVision as an example, we compare the Vanilla model with **RISE** with a notice. **Obs. ❿** ***RISE*** *preserves task-relevant semantic evidence after visual token compression.* As shown in Figure 5, the Vanilla model gives an incorrect answer, while the model equipped with **RISE** correctly identifies the teacher's attitude as positive. This comparison shows **RISE** can retain task-relevant semantic evidence even after visual token compression. Therefore, the compressed tokens remain sufficient for supporting accurate video understanding, rather than merely reducing computational cost. More visualization results are provided in Appendix E.7.1 and Appendix E.7.2.

6 Conclusion

In this paper, we introduce **RISE**, a training-free token compression method for efficient long-video understanding. Our findings show that effective compression should prioritize tokens with high marginal semantic gain rather than local scores. We further show that consolidating recurrent spatio-temporal evidence before selection preserves fragmented but related semantics. Guided by these insights, **RISE** reduces redundancy and promotes semantic diversity under fixed budgets. Experiments across multiple video LLMs and long-video benchmarks show that **RISE** consistently outperforms baselines and remains robust under aggressive compression across diverse architectures and budgets.

AI USE STATEMENT

Generative AI tools were used only for language polishing and grammar correction during manuscript preparation. They were not used to generate research ideas, methods, theoretical results, experimental data, annotations, or experimental results. All scientific content and final decisions were produced and verified by the authors.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. All experiments use publicly available datasets and pretrained models under their respective terms of use. We do not collect new data or involve human subjects, and no private, sensitive, or personally identifiable information is collected. The proposed method is an inference-time token compression technique and does not introduce new data or generative capabilities.

REPRODUCIBILITY STATEMENT

We provide an anonymized code implementation and detailed experimental settings for reproducing the main results. The evaluated backbones, datasets, frame sampling strategies, retention ratios, hyperparameters, and compute environment are specified in Sec. 5 and Appendix D. All datasets and pretrained models used in our experiments are publicly available, and the released code provides instructions for reproducing the main experiments and baselines.

REFERENCES

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.

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*, pp. 9392–9401, 2025.

Changwoo Baek, Jouwon Song, Sohyeon Kim, and Kyeongbo Kong. Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models. *arXiv preprint arXiv:2603.01236*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Alexei Borodin. Determinantal point processes. *arXiv preprint arXiv:0911.1153*, 2009.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. MADTP: multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 15710–15719. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01487. URL https://doi.org/10.1109/CVPR52733.2024.01487.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal LLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13817–13827. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01311. URL https://doi.org/10.1109/CVPR52733.2024.01311.

Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. Efficient large multi-modal models via visual context compression. *Advances in neural information processing systems*, 37:73986–74007, 2024a.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. *arXiv preprint arXiv:2509.01552*, 2025.

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, 2024b.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimomal generation for document question answering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 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.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XI*, volume 13671 of *Lecture Notes in Computer Science*, pp. 396–414. Springer, 2022. doi: 10.1007/978-3-031-20083-0_24. URL https://doi.org/10.1007/978-3-031-20083-0_24.

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*, pp. 24108–24118, 2025a.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025b.

Yiwei Fu, Hui Wan, Xiao Luo, and Minghua Deng. Star: Test-time adaptation can enhance universal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31482–31492, 2026.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

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*, pp. 19959–19973, 2025.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-granular spatio-temporal token merging for training-free acceleration of video llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23990–24000, 2025.

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, HuaiXi Wang, Yiyi Zhou, and Rongrong Ji. Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. *arXiv preprint arXiv:2603.22911*, 2026.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1193–1200, 2011.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

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*, 2024a.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024b.

Jinlong Li, Liyuan Jiang, Haonan Zhang, and Nicu Sebe. Token reduction via local and global contexts optimization for efficient video large language models. *arXiv preprint arXiv:2603.01400*, 2026.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal LLM. *Int. J. Comput. Vis.*, 133(10):6794–6812, 2025. doi: 10.1007/S11263-025-02491-7. URL https://doi.org/10.1007/s11263-025-02491-7.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024d.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *CoRR*, abs/2202.07800, 2022. URL https://arxiv.org/abs/2202.07800.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 9796–9810, 2023.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15630–15640, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, Hang Xu, and Li Zhang. Slowfocus: Enhancing fine-grained temporal understanding in video llm. *Advances in Neural Information Processing Systems*, 37:81808–81835, 2024.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 13937–13949, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22857–22867. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02122. URL https://doi.org/10.1109/ICCV51701.2025.02122.

Kele Shao, TAO Keda, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*, 2025a.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025b.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. Fastvid: Dynamic density pruning for fast video large language models. *arXiv preprint arXiv:2503.11187*, 2025a.

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. In *42st International Conference on Machine Learning, ICML 2025*. ML Research Press, 2025b.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18221–18232. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01725. URL https://doi.org/10.1109/CVPR52733.2024.01725.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18992–19001, 2025.

Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. *arXiv preprint arXiv:2505.19536*, 2025.

Mengyue Wang, Shuo Chen, Kristian Kersting, Volker Tresp, and Yunpu Ma. Metok: Multi-stage event-based token compression for efficient long video understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18881–18895, 2025a.

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. *arXiv preprint arXiv:2512.00891*, 2025b.

Zhenhailong Wang, Senthil Purushwalkam, Caiming Xiong, Silvio Savarese, Heng Ji, and Ran Xu. Dymu: Dynamic merging and virtual unmerging for efficient vlms. *arXiv preprint arXiv:2504.17040*, 2025c.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3272–3283, 2025d.

Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2092–2101, 2023.

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.

Peiyu Xie, Jinxing Li, Guangming Lu, Yong Xu, and David Zhang. Caption assisted multimodal large language model for video moment retrieval. *IEEE Transactions on Image Processing*, 2025.

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.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19803–19813, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025b.

Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, et al. Towards efficient multimodal large language models: A survey on token compression. *TechRxiv*, 2026.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29836–29846, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

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.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, 2024a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025b.

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.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase" highlighted tokens" in mllms: Revisiting visual holistic context retention. *arXiv preprint arXiv:2510.02912*, 2025.

Supplementary Material

A Implementation Preliminaries **17**

B Implementation Algorithm Pipeline **17**

C Proof of Theoretical Results **18**

C.1 Proof of Theorem 4.1 18

C.2 Proof of Theorem 4.2 19

D Implementation Details for Experiments **20**

D.1 Implementation Details 20

D.2 Implementation of Compared Methods 20

D.3 Evaluation Benchmarks 21

E More Experiments **21**

E.1 Implementation Details for Observations 21

E.2 Further efficiency analysis 22

E.3 Additional Ablation Studies 22

E.4 Repeated Experiments 22

E.5 Token Distribution 23

E.6 Complexity Analysis 23

E.7 More Visualizations 24

E.7.1 Qualitative analysis on LLaVA-OneVision 24

E.7.2 Qualitative analysis on Qwen3-VL 24

F Limitation **24**

Algorithm 1 The execution pipeline of `RISE`

Require: Visual tokens $\mathcal{X} = \{\mathbf{x}_{t,i}\}_{t=1,i=1}^{T,P}$, token budget K, fusion threshold T_m, low-rank dimension z
Ensure: Compressed visual tokens $\hat{\mathcal{X}}$
Stage 1: Recurrent Evidence Consolidation via Temporal Consistency
Estimate semantic evidence scores for all visual tokens as described in Sec. 4.2
for $t = 1$ to $T - 1$ **do**
  Match each token in frame t to its most similar neighbor in frame $t + 1$ according to Eq. (1)
  Merge valid matched pairs whose similarity exceeds T_m and update their evidence by Eqs. (2)–(3)
end for `// Aggregate temporally recurrent evidence`
Collect surviving tokens and their consolidated scores as $\mathcal{S}$
Stage 2: Conditional Kernelized Selection for Utility Maximization
Project each token in $\mathcal{S}$ into a z-dimensional subspace and build quality-aware DPP bases following Eq. (4)
Initialize the selected set $\mathcal{R}$ with the strongest semantic anchor as Eq. (7)
while $|\mathcal{R}| < \min(K, |\mathcal{S}|)$ **do**
  Evaluate each unselected token by conditional evidence gain in Eq. (8)
  Add the token with the largest gain to $\mathcal{R}$ and update the DPP statistics
end while `// Select complementary tokens under the fixed budget`
return The final selected visual tokens $\hat{\mathcal{X}} = \{\mathbf{x}_i \mid i \in \mathcal{R}\}$

A Implementation Preliminaries

Determinantal Point Processes. A point process $\mathcal{P}$ defines a probability distribution over all subsets of a discrete ground set $\mathcal{Y} = \{1, 2, \cdots, N\}$. $\mathcal{P}$ is called a determinantal point process (DPP) if there exists a marginal kernel $\mathbf{K}$ satisfying $0 \preceq \mathbf{K} \preceq \mathbf{I}$ such that, for any random subset $\mathbf{Y} \sim \mathcal{P}$ and any subset $A \subseteq \mathcal{Y}$, the inclusion probability is given by $\mathcal{P}(A \subseteq \mathbf{Y}) = \det(\mathbf{K}_A)$. Given a positive semi-definite kernel $\mathbf{L}$, an $\mathbf{L}$-ensemble DPP (Borodin, 2009) defines the probability of sampling a subset $Y \subseteq \mathcal{Y}$ as $\mathcal{P}_{\mathbf{L}}(\mathbf{Y} = Y) = \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{L}+\mathbf{I})}$, where I is the identity matrix. The determinant favors subsets whose elements span a large volume in the feature space, thereby enhancing diversity while discouraging redundancy. Each entry of the kernel matrix is defined as $\mathbf{L}_{i,j} = q_i \phi_{i,j} q_j$, where q_i measures the importance of element i, and $\phi_{i,j}$ represents the similarity between elements i and j. Since video token compression typically requires selecting a fixed-size subset under a given budget, k-DPP (Kulesza & Taskar, 2011; 2012) provides a natural formulation:

$$\mathcal{P}_{\mathbf{L}}(\mathbf{Y} = Y \mid |Y| = k) = \frac{\det(\mathbf{L}_Y)}{\sum_{|Y'|=k} \det(\mathbf{L}_{Y'})}, \tag{12}$$

where $\mathbf{L}$ is a positive semi-definite kernel. This formulation enables the selection of an informative yet diverse subset of visual tokens under a fixed-size budget, making it well suited for token compression.

B Implementation Algorithm Pipeline

We first consolidate temporally recurrent visual evidence to reduce redundant tokens while preserving information that remains consistent across adjacent frames. Specifically, we estimate the semantic evidence score of each visual token and recurrently match tokens between neighboring frames according to their local similarity. Matched tokens whose similarity exceeds the fusion threshold T_m are merged, and their evidence scores are updated throughout the temporal consolidation process. Based on the resulting token set, we then perform conditional kernelized selection under the predefined budget K. We project the consolidated tokens into a z-dimensional subspace to construct quality-aware DPP bases, initialize the selected set with the strongest semantic anchor, and iteratively add the token that provides the largest conditional evidence gain. This procedure retains complementary visual evidence while controlling the number of tokens passed to the language model. A summary of the execution pipeline of our `RISE` is presented in Algorithm 1.

C Proof of Theoretical Results

C.1 Proof of Theorem 4.1

We first introduce several auxiliary quantities for analysis. These quantities are consistent with the implementation in Sec. 4.2 and formalize the effect of recurrent consolidation. Let $\bar{x}_i = x_i/\|x_i\|_2$ denote the normalized semantic direction of token i, and let $m_i = q_i^2$ denote its semantic evidence mass, consistent with Sec. 4.2. Given the fusion threshold T_m, recurrent consolidation induces a partition $\mathcal{G}_{T_m}$ over the original tokens. Each group $C \in \mathcal{G}_{T_m}$ corresponds to one surviving token after consolidation. We define

$$M_C = \sum_{i \in C} m_i, \qquad \mu_C = \frac{1}{M_C} \sum_{i \in C} m_i \bar{x}_i, \qquad \rho_C = \|\mu_C\|_2^2. \tag{13}$$

Here, M_C is the accumulated evidence mass, μ_C is the mass-weighted semantic direction, and ρ_C measures the effective coherence of the consolidated group. For any selected subset $R \subseteq \mathcal{G}_{T_m}$ with $|R| = K$, we define the retained information and evidence coverage as

$$\mathrm{Ret}_K(R; T_m) = \frac{\sum_{C \in R} M_C \rho_C}{\sum_{C \in \mathcal{G}_{T_m}} M_C}, \qquad \mathrm{Cov}_K(R; T_m) = \frac{\sum_{C \in R} M_C}{\sum_{C \in \mathcal{G}_{T_m}} M_C}. \tag{14}$$

Proof. For each consolidation group $C \in \mathcal{G}_{T_m}$, Eq. (13) gives

$$M_C \mu_C = \sum_{i \in C} m_i \bar{x}_i. \tag{15}$$

Summing Eq. (15) over all groups, we obtain

$$\sum_{C \in \mathcal{G}_{T_m}} M_C \mu_C = \sum_{C \in \mathcal{G}_{T_m}} \sum_{i \in C} m_i \bar{x}_i = \sum_i m_i \bar{x}_i. \tag{16}$$

Therefore, recurrent consolidation preserves the first-order semantic evidence before budget selection.

Let

$$M = \sum_{C \in \mathcal{G}_{T_m}} M_C = \sum_i m_i \tag{17}$$

be the total evidence mass. By Assumption 4.1, $\rho_C \geq \gamma(T_m)$. Since μ_C is a weighted average of unit directions, we have $\|\mu_C\|_2 \leq 1$, and thus $\rho_C \leq 1$. Therefore, for each selected group $C \in R$,

$$\gamma(T_m) M_C \leq M_C \rho_C \leq M_C. \tag{18}$$

Summing Eq. (18) over $C \in R$ and dividing by M, we have

$$\gamma(T_m) \frac{\sum_{C \in R} M_C}{M} \leq \frac{\sum_{C \in R} M_C \rho_C}{M} \leq \frac{\sum_{C \in R} M_C}{M}. \tag{19}$$

Using Eq. (14), Eq. (19) becomes

$$\gamma(T_m) \mathrm{Cov}_K(R; T_m) \leq \mathrm{Ret}_K(R; T_m) \leq \mathrm{Cov}_K(R; T_m), \tag{20}$$

which proves Eq. (10).

We further decompose the information loss as

$$\begin{aligned} 1 - \mathrm{Ret}_K(R; T_m) &= 1 - \frac{\sum_{C \in R} M_C \rho_C}{M} \\ &= 1 - \frac{\sum_{C \in R} M_C}{M} + \frac{\sum_{C \in R} M_C (1 - \rho_C)}{M} \\ &= 1 - \mathrm{Cov}_K(R; T_m) + \frac{\sum_{C \in R} M_C (1 - \rho_C)}{M}. \end{aligned} \tag{21}$$

Thus, the total loss consists of a budget selection loss and a fusion distortion term.

Let

$$\mathrm{Cov}_K^\star(T_m) = \max_{|R|=K} \mathrm{Cov}_K(R; T_m), \qquad \mathrm{Ret}_K^\star(T_m) = \max_{|R|=K} \mathrm{Ret}_K(R; T_m). \tag{22}$$

From Eq. (20), we have

$$\gamma(T_m)\mathrm{Cov}_K^\star(T_m) \le \mathrm{Ret}_K^\star(T_m) \le \mathrm{Cov}_K^\star(T_m). \tag{23}$$

Therefore, a theoretically preferred threshold maximizes the guaranteed lower bound:

$$T_m^\dagger \in \arg\max_{T_m} \gamma(T_m)\mathrm{Cov}_K^\star(T_m). \tag{24}$$

In the no-fusion case, each group is a singleton and $\rho_C = 1$. The best mass-based retention under budget K is

$$\mathrm{Ret}_K^{\mathrm{raw}} = \frac{\sum_{i\in \mathrm{TopK}(m)} m_i}{\sum_i m_i}. \tag{25}$$

Thus, recurrent consolidation provides a stronger guaranteed retention than direct top-K selection whenever

$$\gamma(T_m)\mathrm{Cov}_K^\star(T_m) \ge \mathrm{Ret}_K^{\mathrm{raw}}. \tag{26}$$

This completes the proof. □

C.2 Proof of Theorem 4.2

We provide the proof for the conditional DPP selection rule in Sec. 4.3. The quantities b_i, $F(R)$, M_R, and $g_i(R)$ follow Eqs. (4), (6), and (8). For the low-rank representation used in Eq. (4), Φ_r denotes the orthonormal basis of the low-rank subspace, and each normalized surviving token feature is projected onto this subspace. Specifically, if $\bar{x}_i = x_i/\|x_i\|_2$, then

$$z_i = \frac{\Phi_r^\top \bar{x}_i}{\|\Phi_r^\top \bar{x}_i\|_2}. \tag{27}$$

This auxiliary definition only specifies the normalized low-rank direction used for analysis.

Proof. From Eq. (6), the set-level objective is

$$F(R) = \log\det\left(I + \sum_{j\in R} b_j b_j^\top\right). \tag{28}$$

For simplicity, denote

$$M_R = I + \sum_{j\in R} b_j b_j^\top. \tag{29}$$

After adding a candidate token $i \notin R$, we have

$$M_{R\cup\{i\}} = M_R + b_i b_i^\top. \tag{30}$$

By the matrix determinant lemma,

$$\det(M_R + b_i b_i^\top) = \det(M_R)\left(1 + b_i^\top M_R^{-1} b_i\right). \tag{31}$$

Taking logarithms gives

$$\begin{aligned} F(R\cup\{i\}) - F(R) &= \log\det(M_R + b_i b_i^\top) - \log\det(M_R) \\ &= \log\left(1 + b_i^\top M_R^{-1} b_i\right). \end{aligned} \tag{32}$$

Since Eq. (8) defines $g_i(R) = b_i^\top M_R^{-1} b_i$, Eq. (32) becomes

$$F(R\cup\{i\}) - F(R) = \log\left(1 + g_i(R)\right). \tag{33}$$

Because $\log(1+x)$ is strictly increasing for $x \ge 0$, maximizing the marginal DPP volume gain is exactly equivalent to maximizing $g_i(R)$. Therefore,

$$i^\star = \arg\max_{i\notin R} g_i(R) \tag{34}$$

is the exact one-step optimizer of the conditional DPP volume gain.

We further show how this gain balances semantic quality and diversity. Let P_R be the orthogonal projector onto $\mathrm{span}\{b_j : j \in R\}$, and define

$$d_i^2(R) = \|(I - P_R)z_i\|_2^2. \tag{35}$$

Since $M_R = I + \sum_{j\in R} b_j b_j^\top$, directions aligned with the selected bases are down-weighted by M_R^{-1}, while directions orthogonal to their span remain unchanged. Therefore,

$$g_i(R) = b_i^\top M_R^{-1} b_i \geq q_i d_i^2(R). \tag{36}$$

Moreover, because $M_R \succeq I$, we have $M_R^{-1} \preceq I$. Thus,

$$g_i(R) = b_i^\top M_R^{-1} b_i \leq \|b_i\|_2^2 = q_i \|z_i\|_2^2 = q_i. \tag{37}$$

Combining Eqs. (36) and (37), we obtain

$$q_i d_i^2(R) \leq g_i(R) \leq q_i. \tag{38}$$

This inequality shows that a candidate token receives a large conditional gain only when it has high semantic evidence and contributes a direction not already covered by the selected set.

Finally, the inverse statistic can be updated efficiently. After selecting $i^\star$, we have

$$M_{R\cup\{i^\star\}} = M_R + b_{i^\star} b_{i^\star}^\top. \tag{39}$$

By the Sherman–Morrison formula,

$$M_{R\cup\{i^\star\}}^{-1} = M_R^{-1} - \frac{M_R^{-1} b_{i^\star} b_{i^\star}^\top M_R^{-1}}{1 + b_{i^\star}^\top M_R^{-1} b_{i^\star}}. \tag{40}$$

Since each $b_i \in \mathbb{R}^r$, the algorithm only maintains an $r \times r$ inverse matrix and avoids constructing the full $|\mathcal{S}| \times |\mathcal{S}|$ DPP kernel. Storing all low-rank bases costs $\mathcal{O}(|\mathcal{S}|r)$, storing the inverse costs $\mathcal{O}(r^2)$, computing all conditional gains in one round costs $\mathcal{O}(|\mathcal{S}|r^2)$, and each inverse update costs $\mathcal{O}(r^2)$. This completes the proof. □

D Implementation Details for Experiments

D.1 Implementation Details

We conduct experiments on two representative Video LLMs, namely LLaVA-OneVision and Qwen3-VL, to validate the effectiveness of our method across different model architectures. Following the official setting, LLaVA-OneVision uniformly samples 32 frames, producing 32×196 visual tokens for each video. For Qwen3-VL, we adopt its default video processing pipeline and apply `RISE` to the extracted visual tokens without modifying the original model architecture. We evaluate all compression methods under matched token budgets, with retention ratios of $\{25\%, 20\%, 15\%, 10\%\}$ for LLaVA-OneVision and $\{20\%, 10\%\}$ for Qwen3-VL. To ensure a fair comparison, we keep the average token budget per transformer layer aligned across different methods. Unless otherwise stated, the same hyperparameter setting is used throughout all experiments. All experiments are conducted on an NVIDIA A6000 GPU using LMMs-Eval. For the main experiments on LLaVA-OneVision, we follow the results reported by related work on VideoMME, EgoSchema, and LongVideoBench. We conduct the remaining evaluations with LMMs-Eval following its official implementation.

D.2 Implementation of Compared Methods

FastV (ECCV 2024). FastV (Chen et al., 2024b) accelerates Vision LLM inference by exploiting the sparse attention assigned to visual tokens in deeper LLM layers. It estimates visual token importance from the attention patterns of early layers and prunes tokens with lower importance before subsequent computation. Following the official implementation, we perform token pruning at the 2nd LLM layer and adjust the pruning ratio to satisfy each target token budget.

VisionZip (CVPR 2025). VisionZip (Yang et al., 2025b) compresses visual representations by identifying a small set of dominant tokens that concentrate most visual information and retaining additional contextual tokens to preserve complementary details. The resulting compact representation is constructed before being passed to the LLM. For Video LLMs, we apply VisionZip after the original visual pooling operation to maintain compatibility with their video processing pipelines. Following the official configuration, we keep the ratio between dominant and contextual tokens fixed at $54 : 10$ within each frame.

Table 5: Statistics of the video understanding benchmarks used in our experiments.

Benchmark	# Videos	# QA Pairs	Avg. Duration	Video Domain	Evaluation Focus
VideoMME (Fu et al., 2025a)	900	2,700	1,017.9 s	6 domains / 30 subcategories	Multi-duration video perception and reasoning
EgoSchema (Mangalam et al., 2023)	5,063	5,063	180.0 s	Egocentric daily activities	Long-term temporal understanding
LongVideoBench (Wu et al., 2024)	3,763	6,678	473.0 s	Diverse web videos	Long-context referred reasoning
VideoMMMU (Hu et al., 2025)	300	900	506.2 s	6 disciplines / 30 subjects	Perception, comprehension, and adaptation

FastVID (NeurIPS 2025). FastVID (Shen et al., 2025a) exploits both temporal and visual redundancy in video inputs through dynamic density pruning. It first partitions the video into temporally ordered segments to preserve temporal structure and then performs density-based token pruning within the resulting video representation to retain informative visual content. Following the official implementation, we set the minimum number of segments to $c = 8$, the segmentation threshold to $\tau = 0.9$, the salient token ratio to $d = 0.4$, the anchor-frame step to $p = 4$, and the merging factor to $\alpha = 0.6$.

AOT (CVPR 2026). AOT (Li et al., 2026) performs video token reduction by constructing local- and global-aware token anchors and aggregating information from redundant tokens through optimal transport. At the intra-frame level, informative visual content is transported to representative anchors within each frame. At the inter-frame level, keyframe anchors aggregate shared information across consecutive frames while distinct tokens are retained to preserve temporal variations. Following the official implementation, we set the entropic regularization parameter to 0.1 and use 100 Sinkhorn-Knopp iterations for both intra-frame and inter-frame transport.

D.3 Evaluation Benchmarks

We evaluate `RISE` on four video understanding benchmarks with diverse video durations, domains, and reasoning requirements. The detailed statistics and evaluation focuses are summarized in Table 5.

VideoMME. VideoMME (Fu et al., 2025a) is a benchmark for evaluating video understanding in Video LLMs. It contains 900 videos spanning 6 domains and 30 subcategories, with video durations ranging from 11 seconds to 1 hour and a total duration of 254 hours. The benchmark includes 2,700 human-annotated multiple-choice question-answer pairs.

EgoSchema. EgoSchema (Mangalam et al., 2023) is a benchmark for very long-form egocentric video understanding. It contains over 250 hours of real egocentric video data and more than 5,000 human-curated five-choice multiple-choice questions. Each question is based on a three-minute video clip.

LongVideoBench. LongVideoBench (Wu et al., 2024) is a benchmark for long-context interleaved video-language understanding. It contains 3,763 web-collected videos with subtitles across diverse themes, with varying lengths up to 1 hour. The benchmark includes 6,678 human-annotated multiple-choice questions organized into 17 fine-grained categories.

VideoMMMU. VideoMMMU (Hu et al., 2025) is a benchmark for evaluating knowledge acquisition from educational videos. It contains 300 college-level, lecture-style videos spanning 30 subjects in 6 disciplines, together with 900 human-annotated questions. The videos have an average duration of 506.2 seconds, and the questions are designed to evaluate three stages of knowledge acquisition: perception, comprehension, and adaptation.

E More Experiments

E.1 Implementation Details for Observations

We conduct three sets of comparative experiments in Sec. 3, each designed to examine one of the three observations. Additional details of their experimental settings are provided below. In **Obs. ❶**, token fusion has already consolidated similar visual information before selection. Under a low retention ratio, only a small number of tokens can be preserved, making Top-K selection based on individual

Table 6: Prefilling efficiency under comparable final performance.

Method	Retenion ratio	Prefilling Time (s)	Score	Rel. Acc. (%)
Vanilla	100%	1220.8	51.9	100.0
FastVID	20%	1103.5	51.3	98.86
VisionZip	20%	1181.2	50.4	97.20
RISE	10%	**796.8**	**51.4**	**99.12**

Table 7: Component-wise ablation study of RISE on LLaVA-OneVision at a 10% visual token retention ratio.

Variant	VideoMME	VideoMMMU	Avg.	Rel. Acc.
V1	56.0	29.2	42.6	92.5%
V2	57.1	31.7	44.4	96.4%
RISE	**57.5**	**31.8**	**44.7**	**97.0%**

scores potentially similar to our DPP-based selection. We therefore use a retention ratio of 20% to retain a larger selected set and better reveal the difference between the two selection strategies. The experiment for **Obs. ❷** compares the ordering of fusion and selection. To maintain the same final budget of K tokens, the select-then-fuse variant first retains $2K$ tokens and subsequently merges them into K tokens. A lower retention ratio of 10% is adopted to examine whether performing fusion before selection can concentrate semantic evidence more effectively under a constrained token budget. Finally, **Obs. ❸** studies the effect of explicit temporal segmentation. Following FastVID, we employ DySeg to partition the video and compare it with the unsegmented counterpart at a 10% retention ratio, allowing us to examine whether segment-wise processing preserves additional local redundancy after compression.

E.2 Further efficiency analysis

We further compare the inference efficiency of different methods under comparable final performance. As shown in Table 6, RISE achieves a shorter overall inference time while maintaining a comparable average score. These results demonstrate that RISE enables more effective visual token compression and substantially reduces prefilling time without sacrificing overall performance.

E.3 Additional Ablation Studies

We conduct a component-wise ablation study on LLaVA-OneVision at a 10% visual token retention ratio. We construct two variants to examine the contribution of the two stages. **V1** removes recurrent evidence consolidation and directly performs token selection on the original visual tokens. **V2** retains recurrent evidence consolidation but replaces the conditional kernelized selector with Top-K selection based on individual semantic scores. As shown in Table 7, V1 achieves an average score of 42.6 with 92.5% relative accuracy, while V2 improves the results to 44.4 and 96.4%, respectively. The complete RISE further reaches the highest average score of 44.7 and preserves 97.0% relative accuracy. These results show that recurrent evidence consolidation provides the major performance gain, while conditional kernelized selection further improves the retained visual evidence under the same token budget.

E.4 Repeated Experiments

To evaluate the stability of RISE, we conduct three independent runs on LLaVA-OneVision with a 10% retention ratio. As shown in Table 8, the results exhibit limited variation across repeated runs. VideoMME, EgoSchema, and LongVideoBench remain nearly unchanged, with standard deviations of only 0.02, 0.02, and 0.05, respectively. Although VideoMMMU shows relatively larger variation, the overall average score remains stable at 51.2 ± 0.20. Moreover, even the lowest average score across the three runs reaches 51.0, remaining competitive with the reported results under the same compression setting. These results demonstrate that the performance gains of RISE are consistent across repeated experiments and are insensitive to individual runs.

Table 8: Repeated experiments of RISE on LLaVA-OneVision at a 10% retention ratio. We report the results of three independent runs and their mean and standard deviation.

Retention	Run	VideoMME	EgoSchema	LongVideoBench	VideoMMMU	Avg.
10%	1	57.5	61.2	55.2	30.8	51.2
10%	2	57.5	61.2	55.2	29.9	51.0
10%	3	57.5	61.2	55.2	31.5	51.4
10%	**Mean ± Std.**	**57.5 ± 0.02**	**61.2 ± 0.02**	**55.2 ± 0.05**	**30.7 ± 0.78**	**51.2 ± 0.20**



Figure 6: **Visualization of token distributions after video token compression.** We compare the visual tokens retained by RISE, AOT, and FastVID with the uncompressed Vanilla model on two representative video examples. Colored regions indicate the visual tokens preserved by each compression method across different frames.

E.5 Token Distribution

Figure 6 visualizes the retained visual tokens of different compression methods on two representative examples. RISE preserves only 628 tokens, yet consistently produces the same correct prediction as the uncompressed model. More importantly, the retained tokens are concentrated around informative visual regions while remaining distributed across different frames, allowing RISE to preserve both salient local evidence and temporal coverage under a compact token budget. In contrast, AOT retains substantially more tokens with a relatively dense and dispersed distribution, but still produces an incorrect prediction in the first example. FastVID achieves a similarly sparse representation, while its retained tokens fail to support the correct prediction in the second example. These examples indicate that effective compression depends on preserving complementary visual evidence across the video rather than retaining a larger number of tokens. RISE achieves such an evidence-aware distribution with substantially fewer visual tokens.

E.6 Complexity Analysis

In the first stage, each token only searches for its most similar counterpart within a local 3×3 spatial window in the adjacent frame. Let d denote the feature dimension of each visual token, which is fixed once the visual backbone is determined. Therefore, the complexity of local cross-frame fusion is $O(Nd)$, which reduces to $O(N)$ in our implementation. In the second stage, we perform global greedy selection with a low-rank DPP formulation, where each token is represented as $\mathbf{b}_i = \sqrt{q_i}\mathbf{z}_i$,

with $\mathbf{z}_i \in \mathbb{R}^r$. We set $r = 32$ in all experiments. Since the marginal gain only involves the inverse of an $r \times r$ matrix $\mathbf{M}_S^{-1}$, each greedy iteration costs $O(Nr^2)$, which reduces to $O(N)$ because r is fixed. Selecting K tokens thus requires $O(KNr^2)$, which reduces to $O(NK)$. Therefore, the overall complexity of our method is $O(N + NK)$, dominated by the greedy selection stage, and can be written as $O(NK)$.

E.7 More Visualizations

E.7.1 Qualitative analysis on LLaVA-OneVision

As illustrated in Table 1, we evaluate our **RISE** on LLaVA-OneVision at $Rt \in \{25\%, 20\%, 15\%, 10\%\}$. At higher retention ratios, **RISE** outperforms the vanilla LLaVA-OneVision. Additionally, **RISE** retains 99.1% of the uncompressed model's average score under Rt $= 10\%$. Figure 7 presents six qualitative examples comparing LLaVA-OneVision with and without **RISE**, demonstrating that **RISE** aggregates redundant tokens to improve the video LLM's comprehension.

E.7.2 Qualitative analysis on Qwen3-VL

Figure 8 demonstrates 6 representative examples comparing Qwen3-VL with and without **RISE**, highlighting its ability to retain and aggregate more semantic information as well as seize more detailed information.

F Limitation

RISE is designed as a training-free and question-agnostic compression framework, which enables efficient plug-and-play deployment across different video LLMs. However, the current design performs general-purpose video token compression and does not explicitly tailor the retained visual evidence to each specific question or task requirement. As a result, the same compressed video representation is shared across different queries for the same video. Future work may extend **RISE** with lightweight query-aware or task-aware token reweighting, enabling more targeted video processing while preserving its training-free efficiency.



Question: *What number does the person type in at the beginning of the video?*

LLaVA-OneVision w/o RISE ***B. 342432.***

LLaVA-OneVision w/ RISE ***C. 322434.***



Question: *What object is the man in the gray T-shirt pulling?*

LLaVA-OneVision w/o RISE ***C. A ladder.***

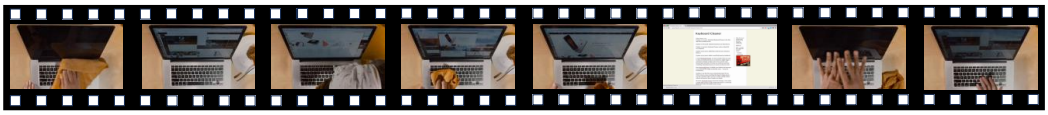
LLaVA-OneVision w/ RISE ***D. A helicopter.***



Question: *What is the attitude of the teacher towards the girl's appearance?*

LLaVA-OneVision w/o RISE ***C. Disappointed.***

LLaVA-OneVision w/ RISE ***B. Positive.***



Question: *As depicted in the video, which smart phone is advertised on the screen of the laptop?*

LLaVA-OneVision w/o RISE ***B. iPhone 14 Pro.***

LLaVA-OneVision w/ RISE ***D. iPhone 6s.***



Question: *Who was sentenced to have guilt in this video?*

LLaVA-OneVision w/o RISE ***A. The actor who triggered the gun.***

LLaVA-OneVision w/ RISE ***B. The prop master who was in charge of this gun.***



Question: *What organ did the woman in the video remove from the medical model?*

LLaVA-OneVision w/o RISE ***C. She removed the kidney.***

LLaVA-OneVision w/ RISE ***A. She removed the liver.***

Figure 7: Visualization of LLaVA-OneVision



Question: *How many times do the two people in the video give different answers to the question?*

Qwen3-VL-8B **w/o RISE** ***D. 5.***

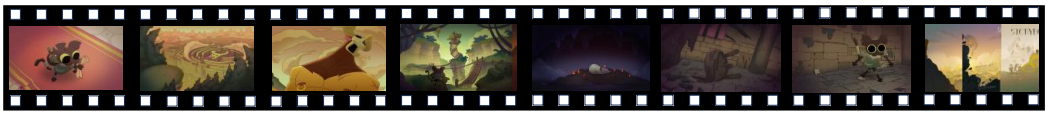
Qwen3-VL-8B **w/ RISE** ***B. 3.***



Question: *After cleaning her floor in the video, what did the woman do next?*

Qwen3-VL-8B **w/o RISE** ***D. Cutting a watermelon.***

Qwen3-VL-8B **w/ RISE** ***B. Using a cup to catch water.***



Question: *Why does the hero wait until the last moment to save the princess?*

Qwen3-VL-8B **w/o RISE** ***C. He mistakenly follows the wrong path.***

Qwen3-VL-8B **w/ RISE** ***A. She is not the princess he desires.***



Question: *According to the order in the video, what is the fourth GPT introduced (except GPT Finder)?*

Qwen3-VL-8B **w/o RISE** ***A. Diagrams.***

Qwen3-VL-8B **w/ RISE** ***B. Logo Creator.***



Question: *What is the interviewed man's opinion on cryptocurrencies in the video?*

Qwen3-VL-8B **w/o RISE** ***C. He thinks cryptocurrencies are very valuable.***

Qwen3-VL-8B **w/ RISE** ***A. He thinks it's an unregulated market and very dangerous.***



Question: *Extrapolating from the video, what team is the commentator on?*

Qwen3-VL-8B **w/o RISE** ***A. Team 1.***

Qwen3-VL-8B **w/ RISE** ***B. Team 2.***

Figure 8: Visualization of Qwen3-VL