DOVE: D-OPTIMAL VISUAL EVIDENCE FOR EXTREME VIDEO TOKEN COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

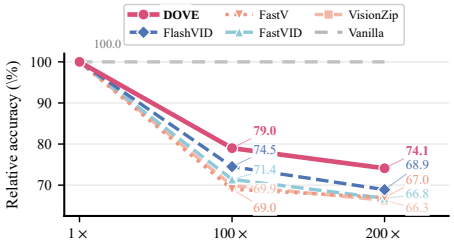
Video large language models process long videos as dense sequences of visual tokens, making multimodal prefilling increasingly expensive. At extreme compression ratios, preserving task-relevant information becomes particularly challenging because decisive visual evidence can be sparse, transient, and distributed across different spatial and temporal contexts. We introduce **DOVE** (**D**-**O**ptimal **V**isual **E**vidence), a training-free framework for extreme video-token compression. DOVE constructs a quality-weighted evidence geometry that combines visual semantics with spatial, temporal, dynamic, and optional query-dependent cues, and retains a fixed-size subset by maximizing the log-determinant of a ridge-regularized information matrix. Under this criterion, candidates aligned with already well-covered evidence directions provide limited additional volume, whereas reliable candidates spanning weakly covered directions yield larger marginal gains. We optimize the objective efficiently over a budget-proportional candidate pool using exact leverage-based greedy updates and positive-gain local refinement. A fixed-length residual readout recovers compatible discarded information without changing the selected set or output length. Across VideoMME, EgoSchema, LongVideoBench, and MVBench with three Video LLM backbones, DOVE outperforms the strongest baseline's average score by 3.5 points at 1% layer-averaged retention and by 2.9 points at 0.5% percentage. These results demonstrate the effectiveness of quality-weighted D-optimal evidence compression under extremely constrained visual-token budgets.

1 INTRODUCTION

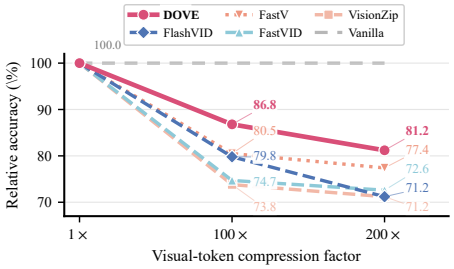
Video large language models (Video LLMs) have achieved remarkable progress in video question answering, temporal reasoning, and long-form video understanding (Lin et al., 2024; Li et al., 2024a; Zhang et al., 2024b; Bai et al., 2025). These capabilities, however, require long visual-token sequences whose length grows with both the number of sampled frames and the spatial resolution of each frame, increasing multimodal prefilling time, attention computation, and memory consumption. Visual-token compression is therefore essential for extending Video LLMs to longer videos and tighter inference budgets. Existing approaches reduce this cost through token pruning, token merging, or spatiotemporal redundancy reduction (Chen et al., 2024; Zhang et al., 2024a; Yang et al., 2025; Fan et al., 2026). Extreme compression, however, imposes a particularly stringent requirement on the retained representation. When only a tiny fraction of visual tokens can be preserved, decisive evidence may be brief, spatially localized, or distributed across temporally distant events. A compressed representation must therefore preserve reliable observations while maintaining sufficient coverage of the semantic, spatial, temporal, and dynamic evidence required for downstream reasoning.

To address this regime, we introduce **DOVE** (**D**-**O**ptimal **V**isual **E**vidence), a training-free framework for extreme video-token compression. DOVE maps each visual token to a quality-weighted evidence vector whose direction encodes visual semantics together with compact temporal, spatial, dynamic, and optional query-related context, while its magnitude reflects the reliability of that evidence. This construction defines a video-specific evidence geometry in which information volume can be measured over heterogeneous forms of video evidence.

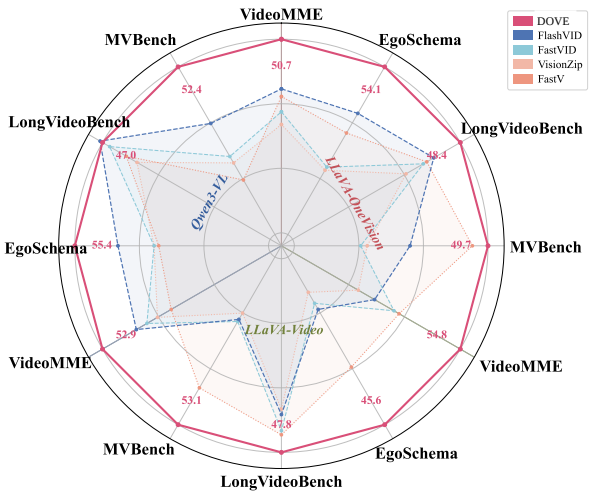
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



(a) Qwen3-VL.



(b) LLaVA-OneVision.



(c) Consistent gains across three Video LLMs and four benchmarks.

Figure 1: **Performance of DOVE under extreme video-token compression.** (a–b) On Qwen3-VL and LLaVA-OneVision, DOVE retains substantially higher relative-accuracy than competing training-free methods at layer-averaged visual-token retention budgets of 1% and 0.5%. (c) At budget of 1%, DOVE achieves the best result on 11 of 12 backbone–benchmark combinations; the only exception is Qwen3-VL on LongVideoBench, where it trails FlashVID by 0.2 points. Radial values are normalized by the corresponding DOVE score for visual comparability; the rose contour denotes DOVE, and its absolute scores are annotated around the perimeter.

Given this geometry, DOVE applies the classical D-optimal criterion from optimal experimental design (Fedorov, 1972; Pukelsheim, 2006). For a selected set $\mathcal{S}$ containing exactly B tokens, we define the ridge-regularized information matrix $\mathbf{M}_{\mathcal{S}}$ and maximize $\log\det(\mathbf{M}_{\mathcal{S}})$. Geometrically, the log-determinant measures the regularized information volume spanned by the retained evidence vectors. Candidates aligned with directions already well represented by the current information matrix provide limited additional volume, whereas reliable candidates spanning weakly covered semantic, spatial, temporal, dynamic, or query-related directions yield larger marginal gains. The effectiveness of this criterion therefore depends critically on the evidence geometry on which it operates.

Directly optimizing the D-optimal objective over all visual tokens can be costly for long sequences. DOVE makes the optimization practical using a budget-proportional candidate pool, exact leverage-based greedy updates, and bounded positive-gain exchange refinement. The selection objective is monotone and submodular on the fixed candidate pool, yielding the standard greedy approximation guarantee. After anchor selection, a fixed-length residual readout transfers compatible information from discarded tokens to the selected anchors without changing the selected set or output length. The complete framework is training-free, and both the vision encoder and language model remain frozen.

We evaluate DOVE at layer-averaged visual-token retention budgets of 1% and 0.5% on VideoMME (Fu et al., 2025), EgoSchema (Mangalam et al., 2023), LongVideoBench (Wu et al., 2024), and MVBench (Li et al., 2024b), using LLaVA-OneVision (Li et al., 2024a), LLaVA-Video (Zhang et al., 2024b), and Qwen3-VL (Bai et al., 2025). DOVE outperforms the strongest baseline by on the average score 3.5 points at 1% retention and by 2.9 points at 0.5%. As summarized in Figure 1 (c), DOVE achieves the best result on 11 of 12 backbone–benchmark combinations at budget of 1%; the only exception is Qwen3-VL on LongVideoBench, where it trails FlashVID by 0.2 points. These results demonstrate consistent effectiveness across model architectures and video-understanding tasks under extremely constrained visual-token budgets.

Our contributions are summarized as follows:

- We introduce a quality-weighted video evidence geometry for extreme token compression. It combines compact visual semantics with temporal, spatial, dynamic, and optional query-dependent coordinates, while separating evidence direction from evidence reliability.
- We develop an efficient D-optimal compression procedure over this geometry, combining a ridge-regularized log-determinant objective with a budget-proportional candidate pool, exact leverage-based greedy updates, and positive-gain exchange refinement. A fixed-length residual readout further recovers compatible discarded information without increasing the output-token budget.
- Across four video benchmarks and three Video LLM backbones, DOVE consistently outperforms representative training-free compression methods at both 1% and 0.5% layer-averaged retention.

2 METHOD

DOVE is designed to preserve heterogeneous video evidence under extremely small token budgets. As shown in Fig. 2, it first embeds each visual token into a quality-weighted evidence geometry, then applies ridge-regularized D-optimal selection within a budget-proportional candidate pool, and finally constructs fixed-length output features through a residual readout. Visual, temporal, dynamic, and query-related cues define the evidence geometry; the log-determinant criterion measures the information volume represented by a candidate subset; and the residual readout recovers compatible discarded information without changing the output length.

2.1 EVIDENCE-VOLUME FORMULATION

Let a video be encoded into T frames with P visual tokens per frame. We denote the flattened token sequence by $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $N = TP$ and $\mathbf{x}_i \in \mathbb{R}^D$ denotes the D-dimensional visual feature of the i-th token. Given a retention ratio R, our goal is to produce exactly $B = \max(1, \mathrm{round}(\rho N))$ output tokens. DOVE performs this compression without updating the parameters of either the vision encoder or the language model.

Our starting point is D-optimal experimental design (Fedorov, 1972; Pukelsheim, 2006). Suppose each visual token is represented by a design vector $\mathbf{z}_i \in \mathbb{R}^{d_z}$. Its direction describes what evidence the token provides, whereas its magnitude reflects the reliability of that evidence. For a selected set $\mathcal{S}$, the regularized information matrix is

$$\mathbf{M}_{\mathcal{S}} = \lambda \mathbf{I} + \sum_{i \in \mathcal{S}} \mathbf{z}_i \mathbf{z}_i^{\top}, \qquad \lambda > 0. \tag{1}$$

DOVE maximizes $\log\det(\mathbf{M}_{\mathcal{S}})$ under the cardinality constraint $|\mathcal{S}| = B$. Geometrically, this is the log-volume spanned by the selected evidence directions. A token nearly collinear with the current set contributes little additional volume, whereas a token along an uncovered direction produces a large gain. Token value is therefore set-dependent: the same token can be useful for one partial set and redundant for another. This property follows directly from the D-optimal objective and is exploited by the optimization procedure described in Sec. 2.3.

2.2 QUALITY-WEIGHTED EVIDENCE GEOMETRY

The determinant can reward complementarity only along directions represented in $\mathbf{z}_i$. DOVE therefore constructs a shared evidence geometry with two roles: normalized coordinates describe the type of evidence carried by a token, while a scalar reliability weight, which we refer to as evidence mass, controls its individual contribution. The visual backbone provides the dominant semantic coordinates, while compact video-context and optional query coordinates distinguish otherwise similar observations.

Evidence direction. We first map each visual feature to a compact metric embedding $\mathbf{u}_i = \mathrm{Norm}(\phi(\mathbf{x}_i))$, where ϕ reduces the channel dimension when necessary. To prevent a few high-variance visual directions from dominating the determinant, we apply shrinkage whitening. Let

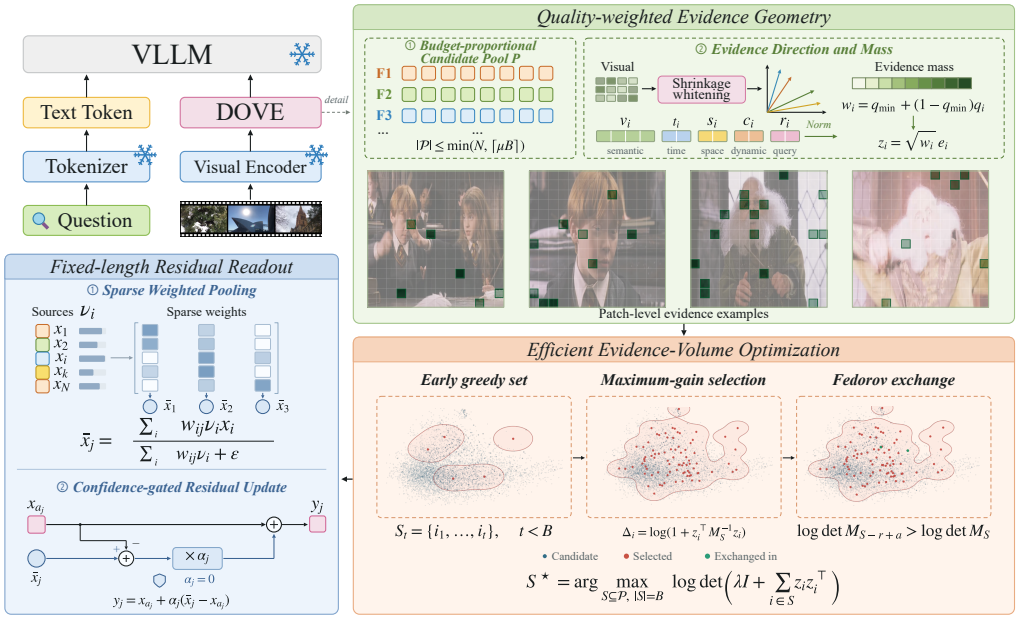


Figure 2: Overview of DOVE. DOVE constructs a compact candidate pool and embeds its tokens into a quality-weighted evidence geometry combining visual, query, dynamic, spatial, and temporal cues. A ridge-regularized D-optimal objective selects complementary anchors that maximize the covered evidence volume. Demand-weighted, protected residual fusion then recovers compatible information from discarded tokens without changing the selected set or output length. All backbone components remain frozen.

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the empirical mean and covariance of the metric features respectively, and let $\boldsymbol{\Sigma} = \mathbf{U}\,\mathrm{diag}(\boldsymbol{\sigma})\mathbf{U}^\top$. The whitened feature is

$$\mathbf{v}_i = \mathrm{Norm}\Big((\mathbf{u}_i - \boldsymbol{\mu})\mathbf{U}\,\mathrm{diag}\Big[(\boldsymbol{\sigma}/\bar{\sigma})^{-\eta/2}\Big]\Big), \tag{2}$$

where $\bar{\sigma}$ is the mean eigenvalue and $\eta \in [0,1]$ controls the whitening strength. Setting $\eta = 0$ preserves the original spectrum, while $\eta = 1$ gives full whitening.

Visual similarity alone cannot distinguish repeated content observed at different positions, times, or stages of an event. We therefore append compact video-context coordinates. The video is partitioned into K_t temporal intervals and every frame into a $K_s \times K_s$ spatial grid. For token i, $\mathbf{t}_i \in \{0,1\}^{K_t}$ is the one-hot encoding of its temporal-bin index, and $\mathbf{s}_i \in \{0,1\}^{K_s^2}$ is the one-hot encoding of its spatial-cell index. We further define

$$\mathbf{c}_i = \mathrm{Norm}[a_i, n_i, \chi_i, e_i^{\mathrm{evt}}, d_i, p_i], \tag{3}$$

where a_i is rank-normalized attention, n_i is bidirectional temporal novelty, χ_i measures trajectory curvature, e_i^{evt} is the frame-level event score, d_i measures local spatial detail, and p_i measures support from the token's temporal component. The details are provided in App. 9. These quantities are not separate selection objectives; together they form a low-dimensional context coordinate that helps the determinant distinguish different forms of video evidence. Temporal components are obtained by linking mutually matched tokens in adjacent frames while suppressing links across strong event boundaries, so component support favors trajectories that persist without being completely static.

When a query is available, using every query token directly would make the design dimension depend on question length. We instead summarize the query into at most K_q diverse atoms $\{\mathbf{g}_k\}$for $k = 1, \ldots, K_q$ through farthest-first selection. Their relevance to visual token i is

$$R_{ki} = \mathrm{Norm}_i\big(\mathbf{g}_k^\top \mathbf{u}_i\big).$$

We estimate an atom confidence from the separation between its highest-scoring visual responses and its median response, normalize the resulting atom weights ω_k, and define $\mathbf{r}_i = [\sqrt{\omega_k} R_{ki}]_{k=1}^{K_q}$.

The mean confidence $\gamma \in [0,1]$ gates the query-dependent coordinates, preventing weak query matches from distorting the visual geometry.

Evidence mass. We form a scalar reliability score by combining attention, novelty, curvature, event strength, local detail, and component support. A query score $\bar{r}_i = \sum_k \omega_k R_{ki}$ is mixed into this score in proportion to γ. After min–max normalization, the resulting $q_i \in [0,1]$ defines the evidence mass

$$w_i = q_{\min} + (1 - q_{\min}) q_i. \tag{4}$$

where $q_{\min} = 0.15$ is a fixed evidence-mass floor that prevents low-scoring tokens from receiving zero mass. The unnormalized evidence coordinates and normalized evidence direction are

$$\widetilde{\mathbf{z}}_i = [\sqrt{\alpha_v}\mathbf{v}_i; \sqrt{\alpha_t}\mathbf{t}_i; \sqrt{\alpha_s}\mathbf{s}_i; \sqrt{\alpha_c}\mathbf{c}_i; \sqrt{\alpha_q}\mathbf{r}_i], \tag{5}$$

$$\mathbf{e}_i = \mathrm{Norm}(\widetilde{\mathbf{z}}_i). \tag{6}$$

The final design vector is

$$\mathbf{z}_i = \sqrt{w_i}\, \mathbf{e}_i. \tag{7}$$

Thus, $\mathbf{e}_i$ describes what evidence the token represents, whereas w_i controls how strongly that token contributes individually. The coordinate weights $\alpha_v, \alpha_t, \alpha_s, \alpha_c$, and α_q only set relative scales within the shared geometry.

2.3 Efficient Evidence-Volume Optimization and Readout

Directly optimizing over all N visual tokens is unnecessary when $B \ll N$. DOVE therefore constructs a budget-proportional candidate pool $\mathcal{P}$ and optimizes the same evidence-volume objective within this pool. The pool reduces the search domain; it does not introduce a second selection criterion. We form $\mathcal{P}$ with at most $L = \min(N, \lceil \mu B \rceil)$ tokens by interleaving query-relevant alternatives across temporal intervals, representatives of semantic trajectory components, high-quality alternatives from temporal-spatial cells, and global quality peaks. Their exact construction is provided in App. 9. Duplicate candidates are removed, and if fewer than B candidates remain, the pool is completed with the strongest unselected tokens. This stratified construction exposes persistent objects, transient events, and query-related alternatives while keeping the determinant stage proportional to the output budget.

Given $\mathcal{P}$, the final anchor set is obtained from

$$\mathcal{S}^\star = \arg\max_{\mathcal{S} \subseteq \mathcal{P}} \log\det\left(\lambda \mathbf{I} + \sum_{i \in \mathcal{S}} \mathbf{z}_i \mathbf{z}_i^\top\right) \quad \text{s.t.} \quad |\mathcal{S}| = B. \tag{8}$$

Starting from $\mathcal{S} = \varnothing$ and $\mathbf{M}_{\mathcal{S}} = \lambda \mathbf{I}$, the matrix determinant lemma gives the exact marginal gain of candidate i:

$$\Delta(i \mid \mathcal{S}) = \log\big(1 + \mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i\big). \tag{9}$$

The quadratic term is the leverage of candidate i relative to the current set. It is large when i contributes along a weakly covered evidence direction. We greedily add the candidate with the largest marginal gain and update the inverse exactly using Sherman–Morrison:

$$\mathbf{M}_{\mathcal{S}\cup\{i\}}^{-1} = \mathbf{M}_{\mathcal{S}}^{-1} - \frac{\mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i \mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1}}{1 + \mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i}. \tag{10}$$

This avoids recomputing an inverse or determinant for every candidate. For the shifted objective $\widehat{f}(\mathcal{S}) = \log\det(\mathbf{M}_{\mathcal{S}}) - d_z \log\lambda$, positive semidefinite rank-one additions imply monotonicity and submodularity. In particular, if $\mathcal{A} \subseteq \mathcal{B}$, then $\mathbf{M}_{\mathcal{A}}^{-1} \succeq \mathbf{M}_{\mathcal{B}}^{-1}$, so Eq. 9 satisfies diminishing returns. Greedy selection therefore retains the standard $(1-1/e)$ approximation guarantee for the best size-B subset of the fixed candidate pool (Nemhauser et al., 1978).

Because greedy decisions are sequential, an early anchor can become redundant after later selections. In the default configuration, DOVE applies a small number of positive-gain Fedorov-style exchanges under the same log-determinant objective. For a selected token $r \in \mathcal{S}$, let $h_r = \mathbf{z}_r^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_r$. Removing it changes the inverse to

$$\mathbf{M}_{\mathcal{S}\setminus\{r\}}^{-1} = \mathbf{M}_{\mathcal{S}}^{-1} + \frac{\mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_r \mathbf{z}_r^\top \mathbf{M}_{\mathcal{S}}^{-1}}{1 - h_r}. \tag{11}$$

For an outside candidate a, the exact exchange gain is

$$\delta(r \to a) = \log(1 - h_r) + \log\Big(1 + \mathbf{z}_a^\top \mathbf{M}_{\mathcal{S}\setminus\{r\}}^{-1} \mathbf{z}_a\Big). \tag{12}$$

We evaluate exchanges among low-removal-loss selected tokens and high-leverage outside candidates, accepting only strictly positive gains. Each accepted exchange preserves cardinality and increases the same log-determinant objective, so the procedure is a bounded local refinement rather than a separate selection module.

Fixed-length residual readout. The determinant determines which anchors survive. After selection, DOVE constructs their output features through a conservative residual readout that does not change the selected set or output length. Each source token is sparsely assigned to at most K_a compatible anchors. For source i and anchor j, the assignment score is

$$\ell_{ij} = \mathbf{u}_i^\top \mathbf{u}_{a_j} + \beta_c \mathbf{1}[\pi_i = \pi_{a_j}], \tag{13}$$

where π_i is the temporal-component identity. Assignments are restricted to anchors in the same or an adjacent temporal interval. We retain the top K_a compatible anchors and normalize their scores with a temperature-softmax to obtain w_{ij}. Every selected anchor also retains an identity assignment. Let ν_i be a bounded source weight derived from token reliability, event strength, and temporal-component support. The details are provided in App. 9. The residual representation pooled at anchor j is

$$\overline{\mathbf{x}}_j = \frac{\sum_i w_{ij} \nu_i \mathbf{x}_i}{\sum_i w_{ij} \nu_i + \epsilon}. \tag{14}$$

The final output feature interpolates between the raw anchor and its residual representation:

$$\mathbf{y}_j = \mathbf{x}_{a_j} + \alpha_j \left(\overline{\mathbf{x}}_j - \mathbf{x}_{a_j}\right). \tag{15}$$

The coefficient α_j decreases with anchor confidence and is set to zero for strongly protected anchors. Each selected token therefore retains a direct identity path, while compatible discarded evidence can contribute to the same fixed-length output. The resulting B tokens are restored to their original video order before being passed to the language model. Overall, the evidence geometry defines what can be complementary, the log-determinant selects the final anchors, and the residual readout only constructs their fixed-length representations.

3 Experiments

We conduct extensive experiments across multiple video understanding benchmarks and VLLM architectures. We first introduce the experimental settings in Sec. 3.1, then report the main comparison in Sec. 3.2, and finally analyze the key design choices in Sec. 3.3.

3.1 Experimental Settings

Benchmarks. We evaluate our method on four widely used video understanding benchmarks: VideoMME (Fu et al., 2025), EgoSchema (Mangalam et al., 2023), LongVideoBench (Wu et al., 2024), and MVBench (Li et al., 2024b). Together, they cover videos of different durations and diverse reasoning demands, including fine-grained perception, egocentric reasoning, long-context understanding, action recognition, temporal ordering, and state change. We report both the public subset and full-test results on EgoSchema. The average score is computed over VideoMME Overall, EgoSchema Total, LongVideoBench, and MVBench, while relative accuracy normalizes this average by that of the corresponding vanilla model without compression.

Compared baselines. We compare against four representative training-free VLLM acceleration methods: 1) FastV (Chen et al., 2024), which prunes visual tokens using text-to-visual attention inside the LLM; 2) VisionZip (Yang et al., 2025), which combines attention-based dominant-token selection with spatial token merging before the LLM; 3) FastVID (Shen et al., 2025), which performs density-aware spatiotemporal pruning; and 4) FlashVID (Fan et al., 2026), which combines frame-wise representative-token selection with tree-based spatiotemporal merging. All baselines follow their official algorithmic settings. We use the same frame sampling, prompts, answer parser, and benchmark splits for all methods. Moreover, we enforce the reported token budget strictly and disable any per-frame upward rounding that would result in an actual retention ratio above the nominal value.

Table 1: Comparison on LLaVA-OneVision and LLaVA-Video. The best compressed-model result within each retention-ratio group is shown in bold.

Method	Retention Ratio R	VideoMME				EgoSchema		LongVideo Bench	MVBench	Avg.	
		Short	Medium	Long	Overall	Subset	Total			Score	Rel. Acc. (%)
LLaVA-OneVision											
Vanilla	100%	69.9	56.7	48.9	58.5	62.2	60.3	56.6	58.3	58.4	100.0
FastV	1%	49.6	47.1	41.9	46.2	46.4	47.7	45.5	48.5	47.0	80.5
VisionZip	1%	45.1	44.0	43.0	44.0	45.2	44.1	43.7	40.4	43.1	73.8
FastVID	1%	46.3	45.4	43.2	45.0	44.2	44.4	45.2	39.9	43.6	74.7
FlashVID	1%	49.1	47.1	44.3	46.8	52.8	49.6	46.1	43.7	46.6	79.8
DOVE (Ours)	1%	**55.1**	**50.3**	**46.7**	**50.7**	**56.8**	**54.1**	**48.4**	**49.7**	**50.7**	**86.8**
FastV	0.5%	45.8	46.2	41.9	44.6	45.6	45.3	45.1	45.9	45.2	77.4
VisionZip	0.5%	43.1	43.9	40.9	42.6	41.0	41.4	42.7	39.7	41.6	71.2
FastVID	0.5%	45.4	43.9	41.9	43.7	40.8	42.5	43.7	39.5	42.4	72.6
FlashVID	0.5%	43.2	42.6	41.6	42.5	40.0	41.4	43.2	39.3	41.6	71.2
DOVE (Ours)	0.5%	**53.4**	**47.0**	**43.7**	**48.0**	**51.2**	**49.9**	**45.7**	**46.0**	**47.4**	**81.2**
LLaVA-Video											
Vanilla	100%	77.0	62.1	53.3	64.2	59.4	57.3	59.5	61.9	60.7	100.0
FastV	1%	52.7	48.9	44.7	48.8	38.2	40.9	46.5	49.6	46.5	76.6
VisionZip	1%	45.9	46.6	41.8	44.8	30.4	34.8	44.6	42.5	41.7	68.7
FastVID	1%	51.4	48.8	44.7	48.3	30.0	35.7	46.2	43.3	43.4	71.5
FlashVID	1%	48.3	48.0	42.8	46.4	31.2	36.2	45.0	43.1	42.7	70.3
DOVE (Ours)	1%	**61.8**	**54.2**	**48.4**	**54.8**	**44.8**	**45.6**	**47.8**	**53.1**	**50.3**	**82.9**
FastV	0.5%	50.7	48.4	44.0	47.7	35.2	38.0	45.1	47.4	44.5	73.3
VisionZip	0.5%	46.4	44.8	42.1	44.4	28.4	33.6	44.0	42.3	41.1	67.7
FastVID	0.5%	49.9	48.6	43.6	47.4	27.6	34.2	44.9	42.2	42.2	69.5
FlashVID	0.5%	46.0	45.3	42.0	44.4	29.2	33.9	43.9	42.0	41.1	67.7
DOVE (Ours)	0.5%	**57.6**	**51.0**	**46.2**	**51.6**	**39.6**	**41.5**	**48.2**	**49.6**	**47.7**	**78.6**

Implementation details. We evaluate on three representative VLLMs with distinct architectures: LLaVA-OneVision (Li et al., 2024a), LLaVA-Video (Zhang et al., 2024b), and Qwen3-VL (Bai et al., 2025). Following their official video settings, LLaVA-OneVision and Qwen3-VL uniformly sample 32 frames, whereas LLaVA-Video samples 64 frames. For LLaVA-OneVision and LLaVA-Video, we adopt a $20{+}8$ hybrid schedule with expansion ratio 1.30 and inner retention ratio 0.1923. For Qwen3-VL, we use a $28{+}8$ schedule with expansion ratio 1.25 and inner retention ratio 0.10. The inner ratios are chosen such that the average number of visual tokens processed per transformer layer exactly matches the nominal retention budget.

DOVE applies the same quality-weighted evidence geometry and regularized evidence-volume objective, followed by the fixed-length residual readout. The method is training-free and leaves all backbone parameters frozen. All experiments are conducted on NVIDIA GeForce RTX 5090 32GB GPUs using LMMs-Eval (Zhang et al., 2025b) under the same evaluation protocol. The complete DOVE hyperparameter configuration is provided in App. 9.

3.2 Main Results

We compare our method with representative training-free baselines on three VLLMs under two extreme retention ratios, $R \in \{1\%, 0.5\%\}$. Results on the LLaVA-based models are reported in Table 1, while results on Qwen3-VL are shown in Table 2. Our method achieves the best average score under every backbone and retention-ratio setting.

Results on LLaVA-OneVision. As shown in Table 1, conventional pruning and merging methods degrade sharply under extreme compression. At $R = 1\%$, the strongest baseline obtains an average score of 47.0, whereas our method reaches 50.7, yielding a substantial gain of 3.7 points. When only 0.5% of the layer-averaged visual-token budget is retained, our method still achieves 47.4, outperforming the best baseline by 2.2 points. In particular, our method consistently outperforms on all four benchmarks, suggesting that DOVE better preserves both fine-grained and temporally distributed evidence under extreme compression.

Table 2: Comparison on Qwen3-VL. The best compressed-model result within each retention-ratio group is shown in bold.

Method	Retention Ratio R	VideoMME				EgoSchema		LongVideo Bench	MVBench	Avg.	
		Short	Medium	Long	Overall	Subset	Total			Score	Rel. Acc. (%)
Qwen3-VL											
Vanilla	100%	77.9	63.3	55.0	65.4	68.6	68.6	60.4	68.6	65.7	100.0
FastV	1%	45.2	47.0	47.0	46.4	39.6	48.2	45.0	41.8	45.4	69.0
VisionZip	1%	49.1	47.1	46.9	47.7	43.2	48.6	44.1	43.4	46.0	69.9
FastVID	1%	51.1	47.2	47.9	48.7	43.2	48.6	46.4	44.0	46.9	71.4
FlashVID	1%	53.7	47.8	47.7	49.7	49.0	51.7	**47.2**	47.1	48.9	74.5
DOVE (Ours)	1%	**59.2**	**51.0**	**48.6**	**52.9**	**53.6**	**55.4**	47.0	**52.4**	**51.9**	**79.0**
FastV	0.5%	43.7	45.7	45.6	45.0	39.2	47.5	43.8	39.9	44.1	67.0
VisionZip	0.5%	45.1	46.0	45.3	45.5	39.0	45.3	42.3	41.2	43.6	66.3
FastVID	0.5%	45.2	46.3	44.4	45.3	39.0	45.3	43.4	41.5	43.9	66.8
FlashVID	0.5%	49.1	47.3	**46.7**	47.7	42.0	47.1	43.4	43.0	45.3	68.9
DOVE (Ours)	0.5%	**54.4**	**47.9**	46.2	**49.5**	**48.4**	**52.5**	**45.0**	**47.7**	**48.7**	**74.1**

Table 3: Evidence-volume analysis on LLaVA-OneVision at 1% layer-averaged retention. Mass Top-K replaces the D-optimal objective using the same scalar evidence mass.

Method	VideoMME	EgoSchema	LongVideo Bench	MVBench	Rel. Acc.
Mass Top-K	48.8	51.9	46.0	46.9	82.9
w/o position coords	50.0	54.2	48.1	49.7	86.5
w/o residual readout	50.1	54.0	47.4	49.4	86.0
DOVE (full)	**50.7**	**54.1**	**48.4**	**49.7**	**86.8**

Table 4: Comparison of alternative anchor-selection objectives under the same DOVE evidence geometry on LLaVA-OneVision at 1% retention.

Selection Rule	VideoMME	EgoSchema	LongVideo Bench	MVBench	Rel. Acc.
k-DPP MAP	49.3	51.4	47.1	49.4	84.4
Farthest-first	49.8	53.0	46.4	49.6	85.1
FPS	49.1	53.9	47.2	49.5	85.4
Regularized D-optimal (Ours)	**50.7**	**54.1**	**48.4**	**49.7**	**86.8**

Results on LLaVA-Video. LLaVA-Video uses 64 sampled frames and therefore presents a longer visual sequence than the other evaluated backbones. DOVE outperforms the strongest baseline from 46.5 to 50.3 at $R = 1\%$ and from 44.5 to 47.7 at $R = 0.5\%$. These gains show that the method remains effective under the evaluated higher-frame-count setting.

Results on Qwen3-VL. To further validate generalization, we evaluate our method on Qwen3-VL, which differs substantially from the LLaVA family in its visual token layout and M-RoPE positional encoding. As shown in Table 2, our method outperforms the best baseline average score by 3.0 points at $R = 1\%$ and by 3.4 points at $R = 0.5\%$. It achieves the best result on all but two individual metrics, with only marginal gaps of 0.2 on LongVideoBench at 1% and 0.5 on VideoMME Long at 0.5%. Across the three evaluated architectures and six backbone–budget settings, DOVE outperforms the best baseline by 3.2 points on average and ranks first on 46 of 48 reported metrics, showing consistent cross-architecture gains.

3.3 Ablation Analysis

Contribution of the D-optimal objective. Table 3 isolates the effect of the regularized log-determinant objective while keeping the evidence-mass construction fixed. Mass Top-K selects the B tokens with the largest scalar evidence mass, whereas DOVE uses the same reliability information within its D-optimal evidence representation. Replacing the D-optimal objective with Mass Top-K reduces the average score from 50.7 to 48.4 and relative accuracy from 86.8% to 82.9%, with consistent degradation across the four benchmark summaries. This result shows that, within the proposed evidence geometry, explicitly accounting for coverage and redundancy provides a substantial benefit beyond the scalar reliability signal alone.

Geometry support and fixed-length readout. The remaining variants preserve the evidence-volume objective and test two supporting choices within the same formulation. Removing the temporal and spatial position coordinates reduces the average score only slightly, from 50.7 to 50.5, with task-dependent changes across benchmarks. This shows that the visual evidence geometry carries most of the signal, while explicit position coordinates provide a modest bias toward broader

Table 5: Quality–efficiency comparison with existing visual token compression methods. We conduct the efficiency analysis on LLaVA-OneVision, and report the prefilling time and Time-To-First-Token (TTFT) in milliseconds (ms).

Method	Retention Ratio R	TFLOPs	Vision Encoding	Prefilling Time			TTFT	Avg.	
				Compression	LLM Forward	Total		Score	Rel. Acc. (%)
Vanilla	100%	48.8	260.3	–	496.3	496.3 (1.0×)	756.6 (1.0×)	58.4	100.0
FastV	2%	3.9	259.5	0.3	66.8	67.0 (7.4×)	326.5 (2.3×)	48.3	82.7
FlashVID	2%	0.8	259.5	40.7	38.9	79.6 (6.2×)	339.2 (2.2×)	49.9	85.4
DOVE (Ours)	1%	0.4	259.9	22.4	30.9	53.3 (**9.3**×)	313.3 (**2.4**×)	**50.7**	**86.8**

temporal-spatial coverage. Removing the residual readout leaves the selected anchor set unchanged but decreases the average score from 50.7 to 50.2, with the largest drop on LongVideoBench (48.4 to 47.4). The readout therefore acts as a post-selection recovery mechanism: it transfers compatible discarded evidence at fixed output length, whereas the residual readout provides an additional post-selection recovery gain on top of the D-optimal anchor selection.

Alternative subset objectives. We replace only the anchor-selection rule while keeping the evidence geometry, candidate pool, token budget, and residual readout fixed. We made comparisons with representative methods such as k-DPP MAP (Chen et al., 2018), Farthest-first (Gonzalez, 1985), and FPS (Qi et al., 2017). Regularized D-optimal selection achieves the best result on all four benchmarks and the highest relative accuracy of 86.8%, exceeding the strongest alternative by 1.4 percentage points. The largest benchmark-level improvements occur on VideoMME and LongVideoBench, where DOVE exceeds the best alternative by 0.9 and 1.2 points, respectively. Distance-based selectors reward pairwise separation, whereas k-DPP MAP uses a determinant-based diversity criterion. DOVE instead evaluates each candidate by its marginal leverage under the ridge-regularized information matrix, directly measuring which evidence directions remain weakly covered by the current set. These results suggest that the gains are not explained by generic subset diversity alone: under the same DOVE evidence geometry, ridge-regularized D-optimal selection provides a better empirical criterion for the evaluated extreme-compression regime.

3.4 Efficiency Analysis

Token compression reduces LLM prefilling cost but can introduce additional selection overhead. We therefore measure end-to-end efficiency on LLaVA-OneVision over VideoMME under a shared hardware and evaluation protocol. Total prefilling time includes both compression and the subsequent LLM forward pass. As shown in Table 5, DOVE reduces total prefilling time from 496.3 ms for the vanilla model to 53.3 ms, corresponding to a $9.3\times$ speedup, and reduces Time-To-First-Token (TTFT) from 756.6 ms to 313.3 ms ($2.4\times$). Its 22.4 ms compression overhead is offset by reducing the LLM-forward time to 30.9 ms and is lower than FlashVID's 40.7 ms overhead. At its 1% operating point, DOVE achieves an average score of 50.7 and a relative score of 86.8%, compared with 48.3/82.7% for FastV and 49.9/85.4% for FlashVID at 2% retention. Because the retention ratios differ, Table 5 should be interpreted as an operating-point comparison rather than an equal-budget latency comparison. Overall, DOVE provides the strongest quality–efficiency operating point among the reported settings.

4 Conclusion

We presented DOVE, a training-free framework for extreme video-token compression based on quality-weighted D-optimal evidence preservation. DOVE constructs a shared video evidence geometry from visual, spatial, temporal, dynamic, and optional query-dependent cues, and couples this representation with a ridge-regularized log-determinant objective. Efficient leverage-based greedy updates and positive-gain exchanges make the optimization practical under small output budgets, while a fixed-length residual readout recovers compatible discarded information without increasing the number of output tokens. Across four benchmarks and three Video LLM backbones, DOVE consistently improves over representative training-free compression methods at both 1% and 0.5% layer-averaged retention while substantially reducing multimodal prefilling cost. These results demonstrate the potential of D-optimal evidence modeling as a practical mechanism for preserving video information under extreme visual-token budgets.

REFERENCES

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

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. *arXiv preprint arXiv:2210.09461*, 2022.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in neural information processing systems*, 31, 2018.

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, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22826–22835, 2025.

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.

Valerii V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

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, 2025.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

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.

Alex Kulesza and Ben Taskar. k-DPPs: Fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning*, 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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal LLM. *International Journal of Computer Vision*, 133(10):6794–6812, 2025.

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, 2024c.

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.

Xinying Lin, Xuyang Liu, Yiyu Wang, Teng Ma, and Wenqi Ren. V-CAST: Video curvature-aware spatio-temporal pruning for efficient video large language models. *arXiv preprint arXiv:2603.27650*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024.

Xinxin Liu, Shiwei Gan, Xiao Liu, Yafeng Yin, Lei Xie, and Sanglu Lu. InfoMerge: Information-aware token compression for efficient video large language models. *arXiv preprint arXiv:2606.02161*, 2026.

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*, 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.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

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.

Michael S Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, et al. xGen-MM-Vid (BLIP-3-Video): You Only Need 32 Tokens to Represent a Video Even in VLMs. *arXiv preprint arXiv:2410.16267*, 2024.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. FastVID: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, 2025.

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.

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.

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, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023.

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, 2025b.

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, 2025c.

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*, 2024a.

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*, 2024b.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. AIM: Adaptive inference of multi-modal LLMs via token merging and pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20180–20192, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Supplementary Material

CONTENTS

5 Ethics statement **14**

6 AI use statement **14**

7 Reproducibility statement **14**

8 Related Work **14**

9 Implementation Details and Reproducibility **15**

9.1 Budget and Layer-wise Allocation 15

9.2 Evidence Geometry and Mass 15

9.3 Volume Optimization and Residual Readout 17

10 Complexity Analysis **17**

11 Theoretical Analysis **17**

12 Additional Qualitative Visualizations **19**

12.1 Qualitative Comparisons with Existing Methods 21

5 ETHICS STATEMENT

This work studies training-free, inference-time token compression for existing Video LLMs and evaluates the method on established benchmarks; it does not introduce a new dataset or modify the backbone parameters. By reducing the number of visual tokens processed by the language model, DOVE may reduce inference cost and associated resource consumption. However, the method inherits potential biases, privacy concerns, and licensing constraints from the underlying models and datasets. Moreover, extreme compression may discard subtle or infrequent visual evidence, which could be consequential in high-stakes applications. We therefore recommend task-specific auditing, conservative retention settings for safety-critical use, and compliance with the applicable dataset and model licenses.

6 AI USE STATEMENT

Large Language Models (LLMs) were used solely to improve the clarity, grammar, and presentation of the manuscript. They were not involved in research ideation, methodological or experimental design, data analysis, interpretation of results, or the formulation of conclusions. All substantive scientific and intellectual contributions were made entirely by the authors. The authors have carefully reviewed and verified all content following language polishing and take full responsibility for the accuracy, integrity, and final form of the manuscript.

7 REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of this work. Sec. 2 specifies the DOVE formulation and its complete execution path, including the fixed-cardinality evidence-volume objective, exact marginal-gain computation, rank-one inverse updates, local exchange refinement, and fixed-length residual readout. Sec. 3.1 reports the evaluated benchmarks and backbones, frame-sampling settings, layer-averaged retention budgets, hybrid compression schedules, hardware, and evaluation framework. All methods are evaluated using the same frame sampling, prompts, answer parser, benchmark splits, and strictly enforced token budgets. App. 9 provides the fixed implementation and budget configurations used across all reported backbones and retention ratios. DOVE is training-free, keeps all backbone parameters frozen, and involves no training- or validation-time tuning. The supplementary material further provides the time and memory complexity analysis and states the scope and complete proofs of the theoretical results, including exact marginal gains, monotonicity, submodularity, the greedy approximation guarantee, and monotonic exchange refinement. Together, these materials document the algorithmic, experimental, and theoretical details required to independently reproduce the reported results.

8 RELATED WORK

Vision-Language Models and Video-LLMs. Vision-language models (VLMs) couple pretrained visual encoders with large language models to support open-ended visual understanding and language generation. Representative systems include Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2024), and LLaVA (Liu et al., 2023). More recent general-purpose models, such as LLaVA-OneVision (Li et al., 2024a) and Qwen3-VL (Bai et al., 2025), further broaden this paradigm across image and video tasks. Video-focused models, including Video-ChatGPT (Maaz et al., 2024), Video-LLaMA (Zhang et al., 2023), Video-LLaVA (Lin et al., 2024), and LLaVA-Video (Zhang et al., 2024b), extend multimodal language modeling to temporally ordered visual observations. Subsequent work improves long-form or streaming video understanding through longer temporal contexts, online processing, and stronger video-oriented representations (Qian et al., 2024; Zhang et al., 2025a). Despite these advances, the number of visual tokens still grows with both the number of sampled frames and the spatial resolution of each frame, making multimodal prefilling, attention computation, and memory increasingly expensive for long-video inference.

Vision Token Compression for VLMs. Visual-token compression reduces this cost through compact representation learning, token merging, and token pruning. Architecture-level approaches con-

struct compact visual representations or memories, as in TokenPacker (Li et al., 2025), LLaMA-VID (Li et al., 2024c), MovieChat (Song et al., 2024), and xGen-MM-Vid (Ryoo et al., 2024). Plug-and-play approaches instead operate on the token sequences of existing models. Token Merging (Bolya et al., 2022) and LLaVA-PruMerge (Shang et al., 2025) aggregate redundant tokens; FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024a), PyramidDrop (Xing et al., 2024), and MustDrop (Liu et al., 2024) prune low-value tokens during inference; and VisionZip (Yang et al., 2025) and AIM (Zhong et al., 2025) combine token selection with merging. Recent studies further examine visual cues, contextual relevance, adaptive inference, and the design of token-pruning criteria (Zhang et al., 2025c; Endo et al., 2025; Wen et al., 2025).

Video inputs introduce additional redundancy both within individual frames and across time. Video-specific approaches such as PruneVID (Huang et al., 2025), FastVID (Shen et al., 2025), and FlashVID (Fan et al., 2026) exploit spatial and temporal redundancy through video-aware pruning or merging. More recent methods incorporate trajectory dynamics (Lin et al., 2026) or temporal fingerprints (Liu et al., 2026) to preserve motion-sensitive evidence. Although effective, many existing approaches primarily rely on token-wise importance, pairwise similarity, or local redundancy. Such criteria do not explicitly optimize whether the retained subset jointly covers heterogeneous semantic, spatial, temporal, dynamic, and query-relevant evidence. This limitation becomes particularly consequential under extreme compression, where omitting a single transient event or localized object may change the downstream answer.

Determinantal point processes. Determinant-based diversity underlies DPPs (Kulesza & Taskar, 2012), k-DPPs (Kulesza & Taskar, 2011), and fast greedy MAP inference (Chen et al., 2018), with applications to video summarization (Gong et al., 2014). For quality-weighted features, the DOVE objective is closely related to fixed-cardinality DPP MAP inference through determinant identities. Subset-level diversity and determinant-based selection are therefore not new in themselves. DOVE instead focuses on adapting the D-optimal information-volume criterion to extreme Video LLM compression through a video-specific quality-weighted evidence geometry, budget-proportional optimization, and fixed-length residual information recovery.

9 Implementation Details and Reproducibility

This section reports the fixed implementation configuration used in all experiments. Parameters are grouped by their role in the shared quality-weighted evidence geometry, volume optimization, and fixed-length residual readout. Unless a backbone-specific setting is stated explicitly, the same configuration is used for all models, benchmarks, and retention ratios. At each selection point, DOVE receives an effective retention ratio R and returns exactly $B = \max(1, \mathrm{round}(RN))$ tokens, clipped to $[1, N]$. The hybrid wrapper only redistributes the reported layer-averaged token budget across transformer depth. All backbone parameters remain frozen, and no training or validation-time tuning is performed.

9.1 Budget and Layer-wise Allocation

The reported ratio R measures the average number of visual tokens processed per transformer layer relative to the uncompressed model. The expansion and inner-retention values determine how this fixed budget is allocated before and after the inner pruning point; they do not change the evidence geometry or the volume objective.

9.2 Evidence Geometry and Mass

Coordinate construction. When the original channel dimension exceeds 96, visual features are average-pooled to a 96-dimensional metric embedding and then ℓ_2-normalized. Each token is augmented with one temporal coordinate and one spatial coordinate from a 3×3 grid. Adjacent-frame temporal associations retain mutual matches with cosine similarity at least 0.82 and apply a spatial-distance penalty of 0.08. Event boundaries are detected from the 90th percentile of the frame-event score, subject to a minimum threshold of 0.65. Bidirectional novelty weights forward and backward matches equally, and local detail is computed with a 3×3 average-pooling window.

Table 6: Backbone-specific frame sampling and layer-wise token-budget allocation.

Setting	Value	Description
Frame sampling	32 / 64 / 32	LLaVA-OneVision / LLaVA-Video / Qwen3-VL
Reported budget	$R \in \{1\%, 0.5\%\}$	Layer-averaged visual-token retention
Depth allocation	$20 + 8$	LLaVA-OneVision and LLaVA-Video: initial tokens are processed for 20 layers before the inner pruning point, followed by 8 layers
Depth allocation	$28 + 8$	Qwen3-VL: initial tokens are processed for 28 layers before the inner pruning point, followed by 8 layers
Expansion ratio	1.30 / 1.25	LLaVA-OneVision and LLaVA-Video / Qwen3-VL
Inner retention	0.1923 / 0.10	LLaVA-OneVision and LLaVA-Video / Qwen3-VL
Budget rounding	nearest integer	$B = \max(1, \mathrm{round}(\rho N))$, clipped to $[1, N]$

Table 7: Default configuration grouped by its role in the evidence-volume formulation.

Group	Parameters	Values
Evidence geometry	Metric dimension; temporal intervals; spatial grid	96; $K_t = \min(12, T)$; 3×3
Temporal association	Mutual-match threshold; spatial penalty	0.82; 0.08
Evidence scaling	Structural/query coordinate mass; whitening strength; evidence-mass floor	0.32; $\eta = 0.50$; $q_{\min} = 0.15$
Query summarization	Maximum query atoms; base query weight; maximum query contribution	$K_q = 8$; 0.18; 0.30
Volume search	Candidate multiplier; ridge; exchange sweeps	$\mu = 2.5$; $\lambda = 0.50$; 6
Residual readout	Anchors per source; temperature; component bonus; base readout strength	$K_a = 2$; 0.07; 0.08; 0.12

Fixed cue aggregation. All scalar cues are normalized before aggregation. The visual evidence-mass score uses fixed normalized weights $(0.28, 0.20, 0.14, 0.12, 0.12, 0.14)$ for attention, novelty, curvature, frame-event strength, local detail, and temporal-component support, respectively. The event-oriented score uses $(0.34, 0.28, 0.18, 0.10, 0.10)$ for novelty, curvature, frame-event strength, local detail, and query relevance. These weights are shared by all backbones, benchmarks, and retention budgets.

Query gating. The query is summarized by farthest-first selection with at most eight atoms. For atom k, confidence is estimated from the separation between its upper 5% visual responses and its median response:

$$\gamma_k = \mathrm{clip}_{[0,1]}\left(\frac{\big(\mathrm{top5}(R_k) - \mathrm{median}(R_k)\big)/(\mathrm{std}(R_k) + \epsilon) - 0.75}{2}\right). \tag{16}$$

The query coordinates use base weight 0.18 and are capped at a total contribution of 0.30. Farthest-first selection stops when the next atom has cosine distance below 0.01.

For the residual readout, the unnormalized source-demand score is

$$\widetilde{d}_i = 0.20 + 0.42 q_i + 0.20 e_i^{\mathrm{evt}} + 0.18 p_i, \qquad d_i = \frac{\widetilde{d}_i}{\sum_j \widetilde{d}_j}, \tag{17}$$

where q_i, e_i^{evt}, and p_i denote evidence mass, frame-event strength, and temporal-component support, respectively. All min–max and ℓ_2 normalization operations use $\epsilon = 10^{-6}$. Covariance eigenvalues are floored at 10^{-4} times the mean eigenvalue, and numerical denominators and temperatures are floored at 10^{-4}.

9.3 Volume Optimization and Residual Readout

The candidate pool contains at most $L = \min(N, \lceil 2.5B \rceil)$ tokens after duplicate removal. A single stratified proposal pass interleaves high-quality representatives across query–time, temporal-component, and temporal–spatial strata, with global quality peaks used as fallback candidates. DOVE initializes the regularized information matrix with ridge $\lambda = 0.50$, performs exact greedy log-determinant maximization, and applies at most six positive-gain Fedorov-style exchange sweeps. For exact reproducibility, each sweep examines at most 24 selected removal candidates and 24 unselected insertion candidates, and accepts an exchange only when its log-determinant gain exceeds 10^{-4}.

After anchor selection, each source token is assigned to its two most similar compatible anchors in the same or an adjacent temporal interval. A same-component assignment receives a similarity bonus of 0.08, and the assignment softmax uses temperature 0.07. The bounded source mass is

$$\nu_i = \mathrm{clip}_{[0.25,2.0]}(0.5 + 0.5 N d_i)\,. \tag{18}$$

The readout coefficient is initialized to 0.12 and clipped to $[0, 0.75]$. The top 15% of anchors, ranked by the maximum of attention and query relevance, preserve an unchanged identity path by setting their coefficient to zero. For the remaining anchors, the coefficient is attenuated by $1 - 0.65 p_j^{\mathrm{prot}}$, where p_j^{prot} is the normalized protection score. The readout changes only the features associated with the selected anchors and never changes the selected set or its output length.

All settings in this section are fixed across the reported backbones, benchmarks, and retention ratios.

10 Complexity Analysis

Let $L = |\mathcal{P}|$ and let d_z be the design dimension. A straightforward implementation that recomputes every candidate quadratic form $\mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i$ at every greedy step costs $\mathcal{O}(BLd_z^2 + Bd_z^2)$ time and stores $\mathcal{O}(Ld_z + d_z^2)$ values. We instead cache all candidate leverages and update them after each selection using the rank-one Sherman–Morrison update; the candidate-scoring part then costs $\mathcal{O}(BLd_z)$, while maintaining the inverse costs $\mathcal{O}(Bd_z^2)$. Thus the implemented greedy stage has cost $\mathcal{O}(B(Ld_z + d_z^2))$ and memory $\mathcal{O}(Ld_z + d_z^2)$. Since L is proportional to B rather than N, the determinant stage scales with the output budget. If a bounded exchange pool of size E is examined for J refinement sweeps, a straightforward exact implementation costs at most $\mathcal{O}(JE^2 d_z^2)$ time and $\mathcal{O}(Ed_z + d_z^2)$ additional memory. Finally, the residual readout computes similarities between N source tokens and B anchors at $O(NBd_u)$ cost for d_u-dimensional metric features, followed by $O(NK_aD)$ pooling cost; only K_a assignments are retained per source token.

11 Theoretical Analysis

This section formalizes the guarantee for the anchor-selection stage. The analysis is relative to a fixed candidate pool, which is the set on which the greedy and exchange procedures operate. It does not make a claim about the quality of the candidate-pool construction or about the subsequent residual readout.

Let $\mathcal{P}$ be a candidate pool, let d_z be the design dimension, and let $\mathbf{z}_i \in \mathbb{R}^{d_z}$ denote the design vector of candidate i. For any $\mathcal{S} \subseteq \mathcal{P}$, define

$$\mathbf{M}_{\mathcal{S}} = \lambda \mathbf{I} + \sum_{i \in \mathcal{S}} \mathbf{z}_i \mathbf{z}_i^\top, \qquad \lambda > 0, \tag{19}$$

and the normalized objective

$$F(\mathcal{S}) = \log\det(\mathbf{M}_{\mathcal{S}}) - d_z \log\lambda = \log\det\left(\mathbf{I} + \lambda^{-1} \sum_{i \in \mathcal{S}} \mathbf{z}_i \mathbf{z}_i^\top\right). \tag{20}$$

The constant shift does not change the maximizing subset and gives $F(\varnothing) = 0$.

Lemma 1 (Exact marginal gain)**.** *For any $\mathcal{S} \subseteq \mathcal{P}$ and $i \notin \mathcal{S}$,*

$$F(\mathcal{S} \cup \{i\}) - F(\mathcal{S}) = \log\big(1 + \mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i\big)\,. \tag{21}$$

Proof. The matrix $\mathbf{M}_{\mathcal{S}}$ is positive definite because $\lambda > 0$. Applying the matrix determinant lemma to $\mathbf{M}_{\mathcal{S}} + \mathbf{z}_i \mathbf{z}_i^\top$ gives

$$\det(\mathbf{M}_{\mathcal{S}} + \mathbf{z}_i \mathbf{z}_i^\top) = \det(\mathbf{M}_{\mathcal{S}}) \left(1 + \mathbf{z}_i^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_i\right). \tag{22}$$

Taking logarithms and cancelling the constant shift in Eq. 20 proves the result. □

Theorem 1 (Monotonicity and submodularity)**.** *The objective F in Eq. 20 is normalized, nonnegative, monotone, and submodular on $2^{\mathcal{P}}$. In particular, for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{P}$ and $i \notin \mathcal{B}$,*

$$F(\mathcal{A} \cup \{i\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{i\}) - F(\mathcal{B}). \tag{23}$$

Proof. Normalization follows directly from Eq. 20. Every summand $\mathbf{z}_i \mathbf{z}_i^\top$ is positive semidefinite, so $\mathbf{M}_{\mathcal{A}} \preceq \mathbf{M}_{\mathcal{B}}$ whenever $\mathcal{A} \subseteq \mathcal{B}$. Since both matrices are positive definite, their inverses satisfy $\mathbf{M}_{\mathcal{A}}^{-1} \succeq \mathbf{M}_{\mathcal{B}}^{-1}$. Therefore,

$$\mathbf{z}_i^\top \mathbf{M}_{\mathcal{A}}^{-1} \mathbf{z}_i \geq \mathbf{z}_i^\top \mathbf{M}_{\mathcal{B}}^{-1} \mathbf{z}_i. \tag{24}$$

Lemma 1 and monotonicity of $\log(1 + x)$ then give Eq. 23. The same lemma shows that every marginal gain is nonnegative, proving monotonicity and nonnegativity. □

Theorem 2 (Greedy approximation on the candidate pool)**.** *Let $\mathcal{S}_{\mathcal{P}}^\star$ be an optimal subset of $\mathcal{P}$ with $|\mathcal{S}_{\mathcal{P}}^\star| = B$. Let $\mathcal{G}_B$ be the size-B subset produced by exact greedy selection, starting from $\mathcal{G}_0 = \varnothing$ and adding the candidate with the largest marginal gain at each step. Then*

$$F(\mathcal{G}_B) \geq \left[1 - \left(1 - \frac{1}{B}\right)^B\right] F(\mathcal{S}_{\mathcal{P}}^\star) \geq (1 - 1/e) F(\mathcal{S}_{\mathcal{P}}^\star). \tag{25}$$

Proof. Fix an iteration $k \in \{1, \ldots, B\}$. By submodularity and monotonicity, the remaining improvement from $\mathcal{G}_{k-1}$ to the optimum is bounded by the sum of the marginal gains of the at most B elements in $\mathcal{S}_{\mathcal{P}}^\star \setminus \mathcal{G}_{k-1}$. Hence,

$$F(\mathcal{S}_{\mathcal{P}}^\star) - F(\mathcal{G}_{k-1}) \leq B \max_{i \in \mathcal{P} \setminus \mathcal{G}_{k-1}} \left[F(\mathcal{G}_{k-1} \cup \{i\}) - F(\mathcal{G}_{k-1})\right]. \tag{26}$$

Greedy chooses the maximizer, so

$$F(\mathcal{G}_k) - F(\mathcal{G}_{k-1}) \geq \frac{1}{B} \left[F(\mathcal{S}_{\mathcal{P}}^\star) - F(\mathcal{G}_{k-1})\right]. \tag{27}$$

Equivalently, the optimality gap contracts as

$$F(\mathcal{S}_{\mathcal{P}}^\star) - F(\mathcal{G}_k) \leq \left(1 - \frac{1}{B}\right) \left[F(\mathcal{S}_{\mathcal{P}}^\star) - F(\mathcal{G}_{k-1})\right]. \tag{28}$$

Applying this recurrence for B iterations and using $(1 - 1/B)^B \leq e^{-1}$ proves Eq. 25. □

Lemma 2 (Exact exchange gain)**.** *Let $r \in \mathcal{S}$ and $a \in \mathcal{P} \setminus \mathcal{S}$. Define $h_r = \mathbf{z}_r^\top \mathbf{M}_{\mathcal{S}}^{-1} \mathbf{z}_r$. Then $1 - h_r > 0$, and the objective change from exchanging r for a is*

$$\begin{aligned} F(\mathcal{S} \setminus \{r\} \cup \{a\}) - F(\mathcal{S}) = \; & \log(1 - h_r) \\ & + \log\left(1 + \mathbf{z}_a^\top \mathbf{M}_{\mathcal{S} \setminus \{r\}}^{-1} \mathbf{z}_a\right). \end{aligned} \tag{29}$$

Proof. Since $\mathbf{M}_{\mathcal{S} \setminus \{r\}}$ is positive definite, the matrix determinant lemma gives

$$\det(\mathbf{M}_{\mathcal{S}}) = \det(\mathbf{M}_{\mathcal{S} \setminus \{r\}}) \left(1 + \mathbf{z}_r^\top \mathbf{M}_{\mathcal{S} \setminus \{r\}}^{-1} \mathbf{z}_r\right). \tag{30}$$

Writing $q_r = \mathbf{z}_r^\top \mathbf{M}_{\mathcal{S} \setminus \{r\}}^{-1} \mathbf{z}_r$, the Sherman–Morrison identity gives $h_r = q_r/(1+q_r)$. Hence $1 - h_r = (1 + q_r)^{-1}$ is the positive ratio $\det(\mathbf{M}_{\mathcal{S} \setminus \{r\}}) / \det(\mathbf{M}_{\mathcal{S}})$. Applying Lemma 1 to the addition of a after removing r, and subtracting the objective of $\mathcal{S}$, yields Eq. 29. □

Corollary 1 (Monotonicity of accepted exchanges)**.** *Suppose each Fedorov exchange is evaluated with Eq. 29 and accepted only when its gain is strictly positive. Then every exchange preserves cardinality, and after any finite number of accepted exchanges the resulting set $\mathcal{S}_{\mathrm{ex}}$ satisfies*

$$|\mathcal{S}_{\mathrm{ex}}| = B, \qquad F(\mathcal{S}_{\mathrm{ex}}) \geq F(\mathcal{G}_B). \tag{31}$$

Consequently, the final anchor set retains the greedy guarantee in Eq. 25 relative to the best size-B subset of the fixed candidate pool.

Proof. Each exchange removes one selected element and adds one outside element, so cardinality is unchanged. By Lemma 2, accepting only a positive gain strictly increases F. Induction over the accepted exchanges gives the result. □

Remark 1 (Scope of the guarantee)**.** *The theorem also concerns the selected anchor indices before the residual readout. The readout changes the output features but does not alter the selected indices, cardinality, or selection objective, so its effect on downstream language-model performance is empirical rather than covered by the D-optimal guarantee.*

12 Additional Qualitative Visualizations

We provide three sports examples from VideoMME to illustrate how DOVE differs from scalar quality-based selection under the same extreme token budget. Each video is represented by 32 sampled frames and contains 6,272 raw visual tokens. Both Quality Top-K and DOVE retain exactly 82 tokens before inner-LLM pruning. The filmstrip summarizes the video and highlights the four frames enlarged below. Within each retention map, outlined patches denote retained tokens, whereas dimmed patches are discarded. Quality Top-K serves as a diagnostic baseline that uses only the scalar evidence-mass component of DOVE, whereas the full method additionally applies the D-optimal objective over the evidence geometry.

Question: **How many athletes are doing high jumps in the video?**

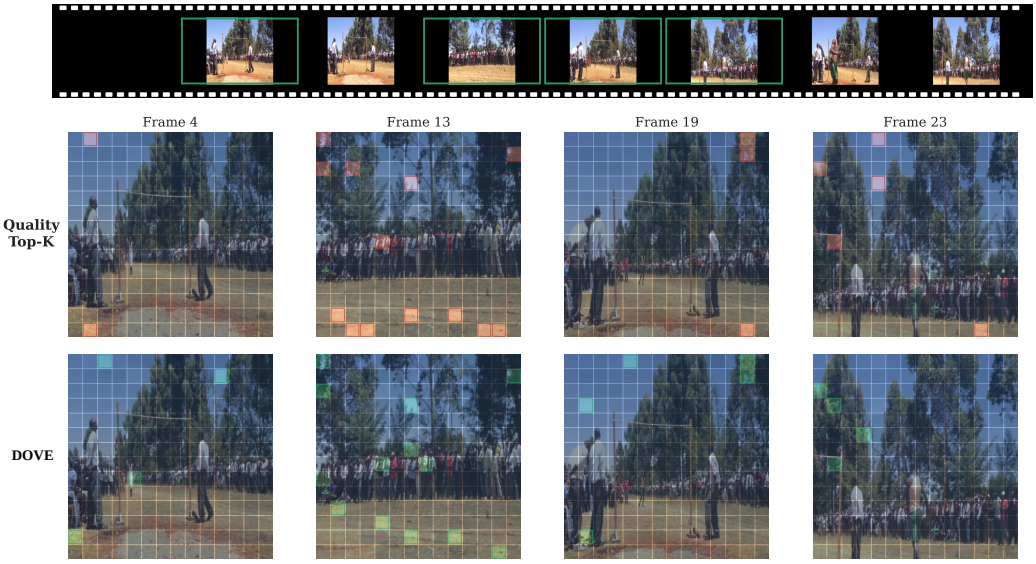


Figure 3: **Equal-budget token selection for multi-event counting.** Answering this question requires evidence from multiple athletes and moments. DOVE distributes its limited budget across complementary states, while Quality Top-K is more concentrated on individually strong regions.

Question: **How many athletes in the video were eliminated while jumping a height of 2 metres 27?**

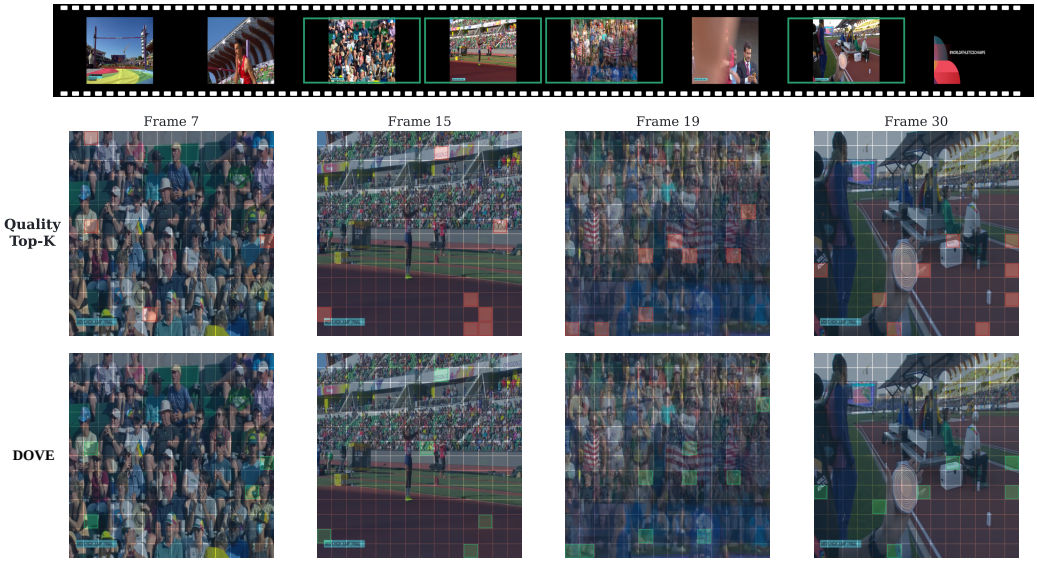


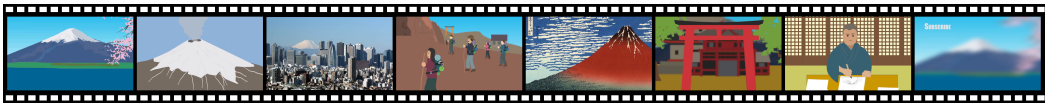
Figure 4: **Equal-budget token selection for temporally distributed evidence.** The elimination count depends on observations from distinct attempts. DOVE retains complementary evidence across these events under the same global token budget as Quality Top-K.

Question: **What sport are the two athletes playing?**



Figure 5: **Equal-budget token selection for sport recognition.** DOVE preserves athlete, equipment, and scene evidence across complementary directions in its evidence geometry, whereas the Mass/Quality Top-K diagnostic baseline concentrates more of its budget on a smaller number of high-quality regions.

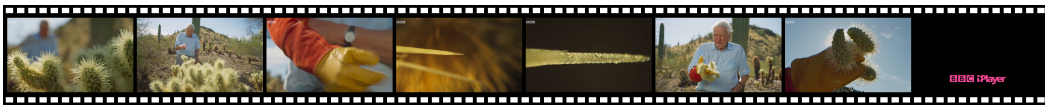
12.1 Qualitative Comparisons with Existing Methods



Question: What color of clothing is the painter wearing?

"A. Blue."
"B. Yellow."
"C. Black."
"D. White."

FastV	VisionZip	FastVID
"C.." ✗	*"C.."* ✗	*"C.."* ✗
FlashVID	**DOVE**	
"C.." ✗	*"A.."* ✓	



Question: Which hand did the person in the video wear a glove on?

"A. His left hand."
"B. His right hand."
"C. Neither hand."
"D. Both hands."

FastV	VisionZip	FastVID
"A.." ✗	*"A.."* ✗	*"A.."* ✗
FlashVID	**DOVE**	
"A.." ✗	*"B.."* ✓	



Question: Which player is running the pick-and-roll for the other players?

"A. Player number 5."
"B. Player number 4."
"C. Player number 2."
"D. Player number 3."

FastV	VisionZip	FastVID
"D.." ✗	*"D.."* ✗	*"D.."* ✗
FlashVID	**DOVE**	
"D.." ✗	*"B.."* ✓	

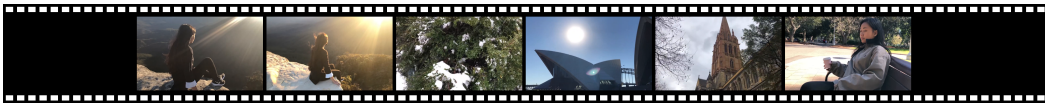
Figure 6: **Qualitative comparison under extreme token compression (Examples 1–3).** All methods use LLaVA-OneVision with a 1% visual-token retention ratio. DOVE correctly answers each question, while all four competing compression methods produce incorrect predictions.



Question: Which option correctly indicates the number of microphones on the stage?

"A. 4."
"B. 3."
"C. 2."
"D. 1."

FastV	VisionZip	FastVID
"B.." ✗	*"B.."* ✗	*"B.."* ✗
FlashVID	**DOVE**	
"B.." ✗	*"C.."* ✓	



Question: Which animal is shown in the video?

"A. Kangaroo."
"B. Giraffe."
"C. Donkey."
"D. Horse."

FastV	VisionZip	FastVID
"B.." ✗	*"B.."* ✗	*"B.."* ✗
FlashVID	**DOVE**	
"B.." ✗	*"A.."* ✓	



Question: What is the performance of Brazil's player number 9 in the match?

"A. He saved a penalty as the team's goalkeeper."
"B. He provided an assist for the winning goal."
"C. He received a red card and was sent off."
"D. He scored two goals and led the team to victory."

FastV	VisionZip	FastVID
"B.." ✗	*"B.."* ✗	*"B.."* ✗
FlashVID	**DOVE**	
"B.." ✗	*"D.."* ✓	

Figure 7: **Qualitative comparison under extreme token compression (Examples 4–6).** The examples involve object counting, animal recognition, and temporally dependent action understanding. DOVE preserves the decisive evidence required by the questions despite the extremely limited token budget.