VOLT: SUBSPACE VOLUME MAXIMIZATION FOR TRAINING-FREE VIDEO TOKEN COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free video token compression has emerged as a widely used approach for efficient video MLLM inference. Existing compressors either distribute the token budget uniformly across frames or use an auxiliary scorer to allocate per-frame budgets before applying a separate token selector. However, uniform allocation can waste budget on redundant frames, while two-stage allocation can decouple frame scoring from token selection, leaving sparse event-bearing frames under-represented. We propose VOLT, which formulates video token compression as constrained subspace volume maximization for joint token selection and cross-frame budget allocation. A single log-determinant criterion scores complementary tokens across frames; a per-frame retention floor preserves temporal support, and the remaining budget is allocated by the same selection rule. Pruned tokens are then merged into their nearest retained neighbors within the same frame, recovering local visual detail without increasing the token count. Across five video understanding benchmarks and three video MLLM backbones, VOLT retains 96.0% of LLaVA-Video-7B's uncompressed accuracy with 10% of its visual tokens and 8.4% of its LLM prefill FLOPs, without retraining or backbone modification. Code is available at `https://anonymous.4open.science/r/ICLR_VOLT-A12E`.

1 INTRODUCTION

Recent video MLLMs Li et al. (2024a); Zhang et al. (2024b); Bai et al. (2025); Zhu et al. (2025); Lin et al. (2024) encode each sampled frame as hundreds of visual tokens and pass the resulting sequence to a language model for spatiotemporal reasoning. Despite their strong performance, the number of visual tokens grows linearly with the number of sampled frames, making long-video inference computationally expensive. Because retraining or modifying strong video backbones is often impractical, training-free video token compression provides a practical route to efficient video MLLM deployment.

Existing training-free compressors reduce visual tokens by leveraging attention signals Chen et al. (2024); Zhang et al. (2024a); Xing et al. (2024); Shang et al. (2024); Zhang et al. (2025c), feature-space diversity Alvar et al. (2025); Yang et al. (2025b); Bolya et al. (2023), or video-specific temporal statistics Tao et al. (2025); Fu et al. (2024); Shen et al. (2025); Shao et al. (2025); Huang et al. (2024); Jin et al. (2024). These methods improve efficiency, but aggressive compression exposes a more global allocation problem: locally salient or diverse tokens do not necessarily form a complementary set across the full video. Under a tight token budget, a compressor must suppress repeated visual evidence across frames while preserving sufficient support for sparse but semantically important events.

Motivation. As illustrated in Figure 1, representative attention-based, text-aware, and diversity-based selectors often spend budget on visually similar regions. Although such tokens may appear salient or diverse within individual frames, they can occupy already-saturated directions in the global video feature subspace and contribute little additional evidence. We refer to this failure mode as *Redundant Subspace Sampling*. A complementary failure occurs in temporal allocation. When per-token saliency is aggregated into frame-level scores, frames containing only a few sparse but important tokens can receive small total budgets, even though those tokens may provide large marginal

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.

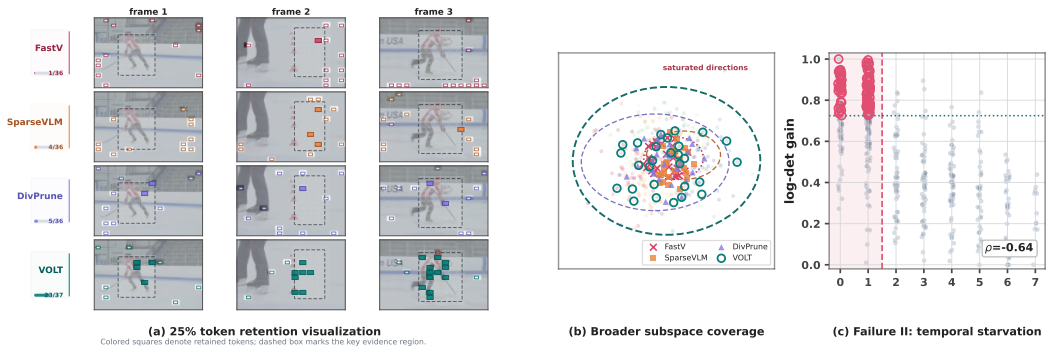


Figure 1: Failure modes in training-free video token compression. (a) Attention-based, text-aware, and diversity-based selectors repeatedly preserve visually similar regions, while VOLT favors tokens with larger marginal volume gains. (b) Existing selectors can concentrate on saturated subspace directions, indicating redundant subspace sampling. (c) Token-level marginal gains vary across frames: gray points show individual tokens, pink markers and shading highlight high-gain early frames, and ρ is the correlation annotated in the plot.

volume gains. We refer to this under-representation of event-bearing frames as *Temporal Starvation*. These two failure modes suggest that video token compression should jointly decide which tokens to retain and how to allocate the budget across frames. This raises the central question: how can we select globally complementary video tokens while maintaining minimum visual support for every frame?

Our Solution. We propose VOLT, a training-free video token compression framework based on constrained subspace volume maximization. Instead of first allocating per-frame budgets and then applying a separate selector, VOLT uses one volume-guided rule to select tokens and allocate the budget remaining after per-frame reservation. The objective measures the geometric volume spanned by the retained tokens in feature space. Tokens that introduce new complementary directions receive large marginal gains, whereas tokens nearly collinear with previously selected ones are naturally suppressed. Because the objective is evaluated over the global cross-frame token pool, visually similar tokens from different frames directly compete for the same budget.

The resulting framework has three steps. First, volume-guided orthogonal selection favors tokens that add complementary feature directions without explicitly constructing large Gram matrices. Second, temporal reservation encodes a per-frame retention floor in the feasible set, ensuring that each frame keeps at least m tokens before the remaining budget is globally contested. Third, within-frame cluster merging routes each pruned token to its nearest retained neighbor in the same frame, recovering local visual detail without increasing the token count. In this formulation, volume maximization provides the global selection criterion, the retention floor maintains frame-level coverage, and within-frame merging recovers local context after pruning.

Experiments across five video understanding benchmarks and three video MLLM backbones show that VOLT improves the accuracy-efficiency trade-off under aggressive compression. Our contributions are summarized as follows:

- We identify two coupled failure modes of aggressive training-free video token compression: *Redundant Subspace Sampling*, where retained tokens occupy already-saturated feature directions, and *Temporal Starvation*, where sparse event-bearing frames receive insufficient retained support.
- We formulate video token compression as constrained subspace volume maximization, where a single log-determinant objective jointly performs token selection and cross-frame budget allocation under a per-frame retention floor.
- We realize the volume objective with efficient volume-guided orthogonal selection, a per-frame retention floor, and within-frame merging, yielding a training-free compressor validated across video MLLM backbones, benchmarks, and retention ratios.

2 Related Work

Video multimodal LLMs. Video MLLMs extend image MLLMs from single images to frame sequences by encoding each frame as visual tokens and projecting them into a language model. LLaVA-OneVision Li et al. (2024a), LLaVA-Video Zhang et al. (2024b), Qwen2.5-VL Bai et al. (2025), and InternVL3 Zhu et al. (2025) follow this paradigm and use spatial pooling, token merging, or related visual-token reduction strategies to keep the per-frame token count manageable. Another line of work integrates compression into the model architecture through joint training, such as LongVU Shen et al. (2024) and matryoshka-style designs Cai et al. (2024). While effective, these methods modify the backbone or require retraining. VOLT instead operates as a training-free, plug-and-play compression module between the vision encoder and the LLM, leaving both components intact.

Training-free visual token compression. Training-free visual token compressors rank, allocate, or merge tokens using different signals. Attention-based methods use LLM-side or vision-encoder attention to prune visual tokens (FastV Chen et al. (2024), SparseVLM Zhang et al. (2024a), PyramidDrop Xing et al. (2024), FitPrune Ye et al. (2024), VisPruner Zhang et al. (2025c), LLaVA-PruMerge Shang et al. (2024), VScan Zhang et al. (2025a)), while diversity-based methods select representative tokens directly in feature space (DivPrune Alvar et al. (2025), VisionZip Yang et al. (2025b), TopV Yang et al. (2025a)). Video-specific compressors further introduce temporal segmentation, chunk-level re-weighting, token merging, or query-driven clustering (DyCoke Tao et al. (2025), FrameFusion Fu et al. (2024), FastVID Shen et al. (2025), HoliTom Shao et al. (2025), PruneVid Huang et al. (2024)). DivPrune optimizes max-min feature diversity Alvar et al. (2025); MMG-Vid allocates budgets by segment before token selection Ma et al. (2025); and HoliTom combines temporal segmentation with token merging Shao et al. (2025). VOLT scores tokens in one cross-frame pool, using the same volume-guided rule for token selection and allocation of the budget remaining after per-frame reservation.

3 Method

Figure 2 gives an overview of VOLT. Given per-frame visual tokens, VOLT first forms a global cross-frame token pool and selects a fixed-size subset by constrained subspace volume maximization. A volume-guided selection rule chooses tokens from the cross-frame pool and distributes the remaining budget, while a per-frame retention floor maintains minimum support for each frame. After selection, pruned tokens are merged into nearby retained tokens within the same frame, recovering local context without increasing the token count.

3.1 Problem Formulation

Let the input video contain F sampled frames. The frozen vision encoder and projector produce a visual-token matrix for each frame,

$$X_f = [x_{f,1}; \ldots; x_{f,T}] \in \mathbb{R}^{T \times d}, \qquad f = 1, \ldots, F, \tag{1}$$

where each row $x_{f,i} \in \mathbb{R}^d$ is a token feature, T is the number of tokens per frame, and d is the feature dimension. When no ambiguity arises, we also use X_f to denote the indexed rows of this matrix. The global token pool is

$$\mathcal{X} = \bigcup_{f=1}^{F} X_f = \{x_i\}_{i=1}^{M} \subset \mathbb{R}^d, \qquad M = FT. \tag{2}$$

Token identities retain their frame and position indices even when feature vectors coincide.

Given a retention ratio $r \in (0, 1]$, we define the reference per-frame budget as $k = \mathrm{round}(rT)$ and the total budget as $B = Fk$. Let $S \subseteq \mathcal{X}$ denote the retained token set, and let $k_f = |S \cap X_f|$ be the number of retained tokens from frame f. Uniform compression fixes $k_f = k$ for all frames. VOLT keeps the same total budget but allows the budget to move across frames subject to a per-frame floor:

$$\sum_{f=1}^{F} k_f = B, \qquad k_f \geq m, \quad \forall f \in \{1, \ldots, F\}, \tag{3}$$

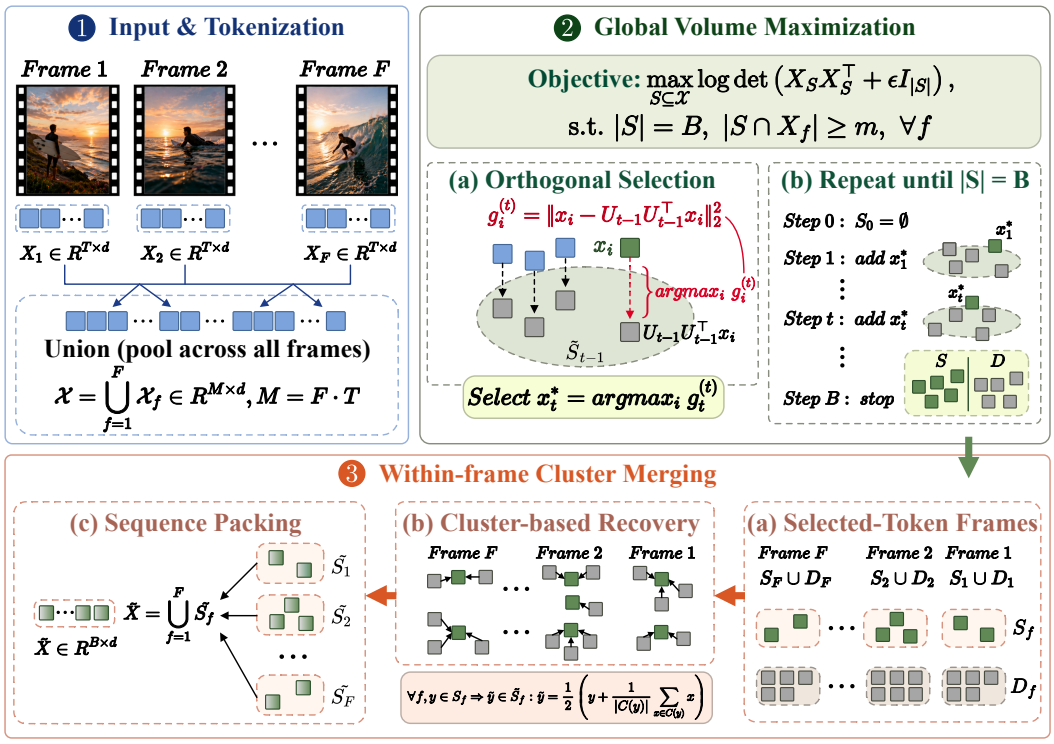


Figure 2: Overview of VOLT. **(1) Input and tokenization:** per-frame token matrices $\{X_f\}_{f=1}^{F}$ form a global token pool $\mathcal{X} = \{x_i\}_{i=1}^{M}$ with $M{=}F \cdot T$. **(2) Volume-guided orthogonal selection:** tokens form a retained set S under the total budget $|S|{=}B$ and the per-frame retention floor $|S \cap X_f| \geq m$. **(3) Within-frame merging:** pruned tokens are routed to their nearest retained neighbors within the same frame to recover local context before sequence packing.

where $m \in \mathbb{Z}_{\geq 0}$ is the per-frame retention floor and $0 \leq m \leq k \leq T$ is required for feasibility. The goal is to select S with $|S| = B$ such that the retained tokens form a compact and informative representation for the downstream LLM.

3.2 Global Volume Maximization

Objective. Within the feasible set in Eq. (3), VOLT scores a candidate retained set by the volume spanned by its token features Taskar (2013):

$$\max_{S\subseteq\mathcal{X}} \log\det\left(I_d + \epsilon^{-1} X_S^\top X_S\right), \quad \text{s.t. } |S| = B, \quad |S \cap X_f| \geq m\ ,\ \forall f, \tag{4}$$

where $X_S \in \mathbb{R}^{|S|\times d}$ stacks the retained token features, I_d is the d-by-d identity matrix, and $\epsilon > 0$ is used for numerical stability. At the fixed budget $|S| = B$, this objective differs from $\log\det(X_S X_S^\top + \epsilon I_B)$ by the constant $B\log\epsilon$. It favors subsets that span complementary feature directions Boyd & Vandenberghe (2004). Nearly collinear tokens add little volume, while tokens that introduce new directions produce larger marginal gains. Since the objective is computed over the full cross-frame pool, visually similar tokens from different frames compete for the same token budget.

Volume-guided orthogonal selection. Directly enumerating feasible subsets is intractable. We therefore score each candidate by the feature component orthogonal to the selected span. Token features are ℓ_2-normalized before selection. Let S_{t-1} be the tokens selected after step $t-1$, and let $U_{t-1} \in \mathbb{R}^{d\times q_{t-1}}$ be an orthonormal basis for their span, where $q_{t-1} \leq \min(t-1, d)$ is the current numerical rank and U_0 is empty. For each candidate $x_i \in \mathcal{X} \setminus S_{t-1}$, its residual energy is

$$g_i^{(t)} = \|x_i - U_{t-1}U_{t-1}^\top x_i\|_2^2. \tag{5}$$

Writing $F(S) = \log\det(I_d + \epsilon^{-1} X_S^\top X_S)$ and $\Sigma_S = I_d + \epsilon^{-1} X_S^\top X_S$, the exact marginal gain is

$$\Delta F(x_i \mid S) = \log\big(1 + \epsilon^{-1} x_i^\top \Sigma_S^{-1} x_i\big)\,. \tag{6}$$

This gain is nonnegative and decreases as the retained set grows, so F is monotone submodular. In the small-ϵ regime, its quadratic term approaches the orthogonal residual in Eq. (5); the derivation is given in Appendix A.2.

At step t, the selector chooses the feasible candidate with the largest orthogonal component, which tracks the log-determinant marginal in the small-ϵ regime. If the selected residual norm is above a threshold η, the normalized residual is appended to the basis; otherwise, the token is retained but the basis is left unchanged. This avoids normalizing a zero vector when the selected span has reached full numerical rank.

The solver performs B greedy steps and costs $\mathcal{O}(MBd)$ outside the backbone. Because the greedy steps are sequential, this becomes $\mathcal{O}(M^2 d)$ when $B = \Theta(M)$. We include this pre-LLM compression cost in the wall-clock efficiency measurements in Sec. 4.

Per-frame retention floor. We enforce Eq. (3) with a two-stage selection procedure. First, the volume-guided selection rule is applied independently within each frame to select m base tokens, forming S_f^{base}. If $m = 0$, this stage is skipped. The base tokens consume Fm slots and provide minimum per-frame support. Second, the global basis is initialized from the union of the base tokens, and the greedy solver continues over the remaining cross-frame pool until $|S| = B$. The remaining $B - Fm$ slots are therefore allocated by the same volume criterion across all frames. The floor m controls the balance between frame-level coverage and global competition: $m = k$ recovers uniform per-frame retention, while $m = 0$ recovers unconstrained global selection.

3.3 Within-frame Merging

Hard pruning can remove local visual context around the retained support. To recover part of this information without increasing the token count Ryoo et al. (2021), VOLT applies within-frame merging after the retained set is fixed. Let $S_f = S \cap X_f$ be the retained tokens in frame f, and let $D_f = X_f \setminus S_f$ be the pruned tokens. For frames with $S_f = \emptyset$, which can occur only when $m = 0$, no within-frame merge is performed. Otherwise, each pruned token is assigned to its cosine-nearest retained token in the same frame:

$$\mathrm{nn}(x) = \arg\max_{y \in S_f} \frac{\langle x, y\rangle}{(\|x\|_2 + \delta)(\|y\|_2 + \delta)}, \qquad x \in D_f, \tag{7}$$

where $\delta > 0$ is a small constant for numerical stability.

For each retained token $y \in S_f$, define its assigned cluster as

$$\mathcal{C}(y) = \{x \in D_f : \mathrm{nn}(x) = y\}. \tag{8}$$

We aggregate the routed tokens by

$$h(y) = \begin{cases} \frac{1}{|\mathcal{C}(y)|} \sum_{x \in \mathcal{C}(y)} x, & |\mathcal{C}(y)| > 0, \\ y, & |\mathcal{C}(y)| = 0, \end{cases} \tag{9}$$

and update the retained token as

$$\tilde{y} = \tfrac{1}{2}\big(y + h(y)\big). \tag{10}$$

Thus, retained tokens with empty clusters remain unchanged.

Both assignment and merging are restricted to tokens from the same frame. This keeps the cross-frame allocation produced by Eq. (3) unchanged and avoids mixing features from different time steps. The final visual sequence is obtained by replacing each retained token y with $\tilde{y}$ and concatenating the retained tokens in their original temporal order. The entire pipeline introduces no trainable parameters, does not modify the vision encoder or the LLM, and is independent of the text query.

Table 1: Comparison on LLaVA-OneVision-7B Li et al. (2024a). Best per retention is shown in **bold**, and second best is underlined. Avg. is computed over all five benchmarks.

Method	FLOPs (T) ↓	FLOPs Ratio ↓	Retain Ratio	MVBench	EgoSchema	LongVideo Bench	VideoMME	MLVU	Avg. ↑	
									Score	%
LLaVA-OV-7B	40.8	100%	100%	58.3	60.4	56.4	58.6	64.0	59.5	100.0
FastV Chen et al. (2024)	10.7	26.3%	25%	54.7	55.5	56.2	57.0	60.8	56.8	95.5
PDrop Xing et al. (2024)	10.5	25.7%	25%	56.1	58.0	54.1	56.4	61.5	57.2	96.1
DyCoke Tao et al. (2025)	8.7	21.3%	25%	53.1	59.5	49.5	54.3	58.7	55.0	92.4
VisionZip Yang et al. (2025b)	8.7	21.3%	25%	57.9	60.3	56.5	58.2	63.5	59.3	99.6
PruneVid Huang et al. (2024)	8.7	21.3%	25%	57.4	59.9	55.7	57.4	63.0	58.7	98.6
FastVID Shen et al. (2025)	8.7	21.3%	25%	56.5	60.6	56.3	58.0	63.4	59.0	99.0
AOT Li et al. (2026)	8.7	21.3%	25%	58.4	60.4	56.7	58.7	63.9	59.6	100.1
FlashVID Fan et al. (2026)	8.7	21.3%	25%	58.3	60.5	56.5	58.5	63.8	59.5	100.0
VOLT (Ours)	8.7	21.3%	25%	**58.6**	**61.1**	**57.0**	**59.1**	**64.5**	**60.1**	**100.9**
VisionZip Yang et al. (2025b)	6.9	16.9%	20%	57.2	60.0	55.6	57.4	62.7	58.6	98.4
FastVID Shen et al. (2025)	6.9	16.9%	20%	56.3	60.3	56.3	57.9	63.0	58.8	98.7
AOT Li et al. (2026)	6.9	16.9%	20%	58.2	60.7	56.5	58.5	63.7	59.5	100.0
FlashVID Fan et al. (2026)	6.9	16.9%	20%	58.3	60.8	56.6	58.6	63.8	59.6	100.1
VOLT (Ours)	6.9	16.9%	20%	**58.5**	**61.1**	**57.2**	**58.8**	**64.4**	**60.0**	**100.8**
VisionZip Yang et al. (2025b)	5.2	12.7%	15%	56.5	59.8	54.4	56.1	61.5	57.7	96.8
PruneVid Huang et al. (2024)	5.2	12.7%	15%	56.8	59.7	55.4	56.6	62.0	58.1	97.6
FastVID Shen et al. (2025)	5.2	12.7%	15%	56.0	60.0	56.2	57.7	62.8	58.5	98.3
AOT Li et al. (2026)	5.2	12.7%	15%	58.0	60.8	56.3	58.3	63.5	59.4	99.7
FlashVID Fan et al. (2026)	5.2	12.7%	15%	58.1	60.9	56.5	58.4	63.6	59.5	99.9
VOLT (Ours)	5.2	12.7%	15%	**58.3**	**61.2**	**57.4**	**58.6**	**64.4**	**60.0**	**100.7**
VisionZip Yang et al. (2025b)	3.4	8.3%	10%	53.5	58.0	49.3	53.4	58.0	54.4	91.4
PruneVid Huang et al. (2024)	3.4	8.3%	10%	56.2	59.8	54.5	56.0	61.4	57.6	96.7
FastVID Shen et al. (2025)	3.4	8.3%	10%	55.9	60.0	56.3	57.3	62.6	58.4	98.1
AOT Li et al. (2026)	3.4	8.3%	10%	57.4	60.7	56.6	58.0	63.4	59.2	99.5
FlashVID Fan et al. (2026)	3.4	8.3%	10%	57.6	60.9	56.8	58.1	63.5	59.4	99.7
VOLT (Ours)	3.4	8.3%	10%	**57.9**	**61.2**	**57.1**	**58.3**	**64.4**	**59.8**	**100.4**

4 EXPERIMENTS

Benchmarks. We evaluate VOLT on five video understanding benchmarks: MVBench Li et al. (2024b), EgoSchema Mangalam et al. (2023) (full set), LongVideoBench Wu et al. (2024), MLVU Zhou et al. (2025), and VideoMME Fu et al. (2025).

Implementation Details. VOLT is evaluated as a drop-in module on three video MLLM backbones: LLaVA-OneVision-7B Li et al. (2024a), LLaVA-Video-7B Zhang et al. (2024b), and Qwen2.5-VL-7B Bai et al. (2025). All measurements are conducted on a single NVIDIA H200 GPU under matched or comparable prefilling FLOPs within each operating point. LLaVA-OneVision samples 32 frames (N_v=196), and LLaVA-Video samples 64 frames (N_v=169). Unless otherwise stated, we set m=1 across backbones and retention ratios. All runs use LMMs-Eval Zhang et al. (2025b).

Inference Cost Analysis. Following Chen et al. (2024); Xing et al. (2024); Tao et al. (2025), processing n_i visual tokens at layer i incurs $4n_id^2 + 2n_i^2d + 2n_idd_{\text{ffn}}$ FLOPs, where d_{ffn} denotes the feed-forward hidden width. With output length R=100, decoding accounts for only $\sim$2% of total FLOPs in our setting, so reducing visual tokens before the LLM is the primary source of inference savings.

Results on LLaVA-OneVision. Table 1 compares VOLT with representative baselines under four retention budgets. At 10% retention, VOLT remains above the vanilla baseline, reaching 100.4% relative accuracy and exceeding FlashVID and AOT by 0.4 and 0.6 Avg. points, respectively. Across all retention budgets, VOLT achieves the best average score over the five benchmarks.

Results on LLaVA-Video. Table 2 reports results on LLaVA-Video-7B, which uses a different visual-token configuration from LLaVA-OneVision (64 frames with 169 tokens per frame). VOLT achieves the best average score at every retention budget and retains 96.0% of vanilla performance at 10% retention, where the prefilling FLOPs are reduced to 8.4% of the vanilla cost. These results indicate that VOLT transfers consistently across LLaVA-family backbones.

Table 2: Comparison on LLaVA-Video-7B Zhang et al. (2024b). Avg. is computed over all five benchmarks.

Method	FLOPs (T) ↓	FLOPs Ratio ↓	Retain Ratio	MVBench	EgoSchema	LongVideo Bench	VideoMME	MLVU	Avg. ↑	
									Score	%
LLaVA-Video-7B	80.2	100%	100%	60.4	57.2	58.9	64.3	70.8	62.3	100.0
FastV Chen et al. (2024)	17.1	21.3%	25%	54.3	54.1	55.0	58.8	65.5	57.5	92.3
PDrop Xing et al. (2024)	19.5	24.3%	25%	55.9	54.3	54.7	61.9	66.4	58.6	94.1
VisionZip Yang et al. (2025b)	15.2	18.9%	25%	56.7	54.7	54.7	60.7	67.0	58.8	94.3
FastVID Shen et al. (2025)	15.2	18.9%	25%	57.5	55.5	56.4	61.7	68.0	59.8	96.0
AOT Li et al. (2026)	15.2	18.9%	25%	58.5	56.0	57.4	62.6	69.0	60.7	97.4
FlashVID Fan et al. (2026)	15.2	18.9%	25%	58.3	56.1	57.2	62.4	68.9	60.6	97.2
VOLT (Ours)	15.2	18.9%	25%	**58.7**	**56.4**	**57.8**	**62.9**	**69.5**	**61.1**	**98.0**
VisionZip Yang et al. (2025b)	12.1	15.1%	20%	57.4	55.0	55.6	61.0	67.4	59.3	95.1
FastVID Shen et al. (2025)	12.1	15.1%	20%	57.9	55.3	56.7	61.9	68.2	60.0	96.3
AOT Li et al. (2026)	12.1	15.1%	20%	58.2	55.7	57.0	62.3	68.7	60.4	96.9
FlashVID Fan et al. (2026)	12.1	15.1%	20%	58.4	55.9	57.1	62.4	68.8	60.5	97.1
VOLT (Ours)	12.1	15.1%	20%	**58.6**	**56.2**	**57.5**	**62.7**	**69.3**	**60.9**	**97.7**
VisionZip Yang et al. (2025b)	9.3	11.6%	15%	56.7	54.7	54.7	60.7	67.0	58.8	94.3
FastVID Shen et al. (2025)	9.3	11.6%	15%	56.9	55.0	55.8	61.0	67.5	59.2	95.1
AOT Li et al. (2026)	9.3	11.6%	15%	57.8	55.2	55.0	62.0	68.2	59.6	95.7
FlashVID Fan et al. (2026)	9.3	11.6%	15%	58.0	55.3	55.5	62.2	68.4	59.9	96.1
VOLT (Ours)	9.3	11.6%	15%	**58.4**	**55.6**	**56.0**	**62.5**	**69.0**	**60.3**	**96.8**
VisionZip Yang et al. (2025b)	6.7	8.4%	10%	55.4	53.8	53.5	59.5	65.6	57.6	92.4
FastVID Shen et al. (2025)	6.7	8.4%	10%	56.4	54.5	55.0	60.4	66.7	58.6	94.0
AOT Li et al. (2026)	6.7	8.4%	10%	57.3	54.9	54.6	61.4	67.5	59.1	94.9
FlashVID Fan et al. (2026)	6.7	8.4%	10%	57.5	55.0	54.9	61.6	67.7	59.3	95.2
VOLT (Ours)	6.7	8.4%	10%	**58.0**	**55.3**	**55.5**	**61.9**	**68.4**	**59.8**	**96.0**

Table 3: Comparison on Qwen2.5-VL-7B Bai et al. (2025) under standard retention budgets. Avg. is computed over all five benchmarks.

Method	FLOPs (T) ↓	FLOPs Ratio ↓	Retain Ratio	MVBench	EgoSchema	LongVideo Bench	VideoMME	MLVU	Avg. ↑	
									Score	%
Qwen2.5-VL-7B	21.5	100%	100%	67.0	58.0	56.5	60.5	60.0	60.4	100.0
FastV Chen et al. (2024)	5.6	26.0%	25%	64.5	55.8	54.0	57.8	56.5	57.7	95.6
VisionZip Yang et al. (2025b)	4.6	21.4%	25%	65.8	56.5	55.0	58.7	58.2	58.8	97.4
FastVID Shen et al. (2025)	4.6	21.4%	25%	66.0	57.0	55.5	59.0	58.7	59.2	98.1
AOT Li et al. (2026)	4.6	21.4%	25%	66.4	57.4	55.8	59.5	59.1	59.6	98.7
FlashVID Fan et al. (2026)	4.6	21.4%	25%	66.5	57.5	55.9	59.6	59.2	59.7	98.9
VOLT (Ours)	4.6	21.4%	25%	**66.8**	**57.8**	**56.1**	**60.0**	**59.7**	**60.1**	**99.5**
VisionZip Yang et al. (2025b)	3.7	17.2%	20%	65.0	56.0	54.4	58.0	57.4	58.2	96.3
FastVID Shen et al. (2025)	3.7	17.2%	20%	65.5	56.5	54.9	58.5	58.0	58.7	97.2
AOT Li et al. (2026)	3.7	17.2%	20%	66.0	57.0	55.3	59.0	58.5	59.2	97.9
FlashVID Fan et al. (2026)	3.7	17.2%	20%	66.1	57.1	55.4	59.1	58.6	59.3	98.1
VOLT (Ours)	3.7	17.2%	20%	**66.5**	**57.5**	**55.8**	**59.6**	**59.2**	**59.7**	**98.9**
VisionZip Yang et al. (2025b)	2.8	13.0%	15%	63.8	55.0	53.5	57.0	56.3	57.1	94.6
FastVID Shen et al. (2025)	2.8	13.0%	15%	64.5	55.8	54.2	57.8	57.2	57.9	95.9
AOT Li et al. (2026)	2.8	13.0%	15%	65.2	56.5	54.7	58.5	57.8	58.5	96.9
FlashVID Fan et al. (2026)	2.8	13.0%	15%	65.4	56.7	54.9	58.6	58.0	58.7	97.2
VOLT (Ours)	2.8	13.0%	15%	**65.9**	**57.0**	**55.3**	**59.0**	**58.6**	**59.2**	**97.9**
VisionZip Yang et al. (2025b)	1.9	8.8%	10%	61.5	53.5	52.0	55.5	54.6	55.4	91.8
FastVID Shen et al. (2025)	1.9	8.8%	10%	63.0	54.5	52.8	56.5	55.7	56.5	93.5
AOT Li et al. (2026)	1.9	8.8%	10%	64.0	55.5	53.5	57.4	56.7	57.4	95.1
FlashVID Fan et al. (2026)	1.9	8.8%	10%	64.2	55.7	53.7	57.6	56.9	57.6	95.4
VOLT (Ours)	1.9	8.8%	10%	**64.8**	**56.2**	**54.2**	**58.0**	**57.5**	**58.1**	**96.3**

Results on Qwen2.5-VL. Table 3 evaluates VOLT on Qwen2.5-VL-7B, whose MRoPE-based visual representation differs from the LLaVA family. VOLT achieves the best average score across all retention budgets, retaining 96.3% relative accuracy at 10% retention. The consistent gains over FlashVID, AOT, and FastVID suggest that the proposed selection rule generalizes beyond a single video backbone.

Frame scaling on Qwen2.5-VL. Token compression can redirect a fixed LLM visual-token input budget from per-frame tokens to more input frames. Table 4 and Appendix Table 7 evaluate this trade-off on Qwen2.5-VL-7B. As the frame count increases from $3\times$ to $10\times$, VOLT improves relative accuracy from 104.0% to 106.3%, with the largest gains appearing on long-video benchmarks such as VideoMME-Long and MLVU. Compared with VisionZip, FastVID, AOT, and

Table 4: Frame scaling on Qwen2.5-VL-7B at 5× and 10×. Avg. is computed over VideoMME Overall, EgoSchema, LongVideoBench, MVBench, and MLVU.

Method	#Frames	Retain	VideoMME				EgoSchema	LongVideo Bench	MVBench	MLVU	Avg.	
			Short	Medium	Long	Overall					Score	Rel. %
Vanilla	16 (1×)	100%	71.0	60.0	50.5	60.5	58.0	56.5	67.0	60.0	60.4	100.0
VisionZip Yang et al. (2025b)	80 (5×)	20%	70.5	58.5	48.0	59.0	57.5	56.0	66.5	60.2	59.8	99.1
FastVID Shen et al. (2025)	80 (5×)	20%	71.0	59.0	50.0	60.0	58.5	57.0	67.5	61.0	60.8	100.7
AOT Li et al. (2026)	80 (5×)	20%	72.0	60.5	52.0	61.5	59.0	57.5	68.0	62.0	61.6	102.0
FlashVID Fan et al. (2026)	80 (5×)	20%	72.5	61.0	52.5	62.0	59.5	58.0	68.5	62.5	62.1	102.8
VOLT (Ours)	80 (5×)	20%	**76.0**	**64.0**	**53.5**	**64.5**	**60.5**	**59.0**	**70.0**	**64.4**	**63.7**	**105.4**
VisionZip Yang et al. (2025b)	160 (10×)	10%	70.0	57.0	47.0	58.0	56.5	56.0	64.5	60.5	59.1	97.8
FastVID Shen et al. (2025)	160 (10×)	10%	71.5	60.5	50.0	60.7	58.0	57.0	67.0	61.6	60.9	100.8
AOT Li et al. (2026)	160 (10×)	10%	71.5	61.0	52.0	61.5	59.0	57.5	68.0	62.5	61.7	102.2
FlashVID Fan et al. (2026)	160 (10×)	10%	72.0	62.5	53.0	62.5	60.0	58.5	69.0	63.5	62.7	103.8
VOLT (Ours)	160 (10×)	10%	**74.0**	**65.5**	**55.0**	**64.8**	**61.0**	**59.5**	**70.5**	**65.1**	**64.2**	**106.3**

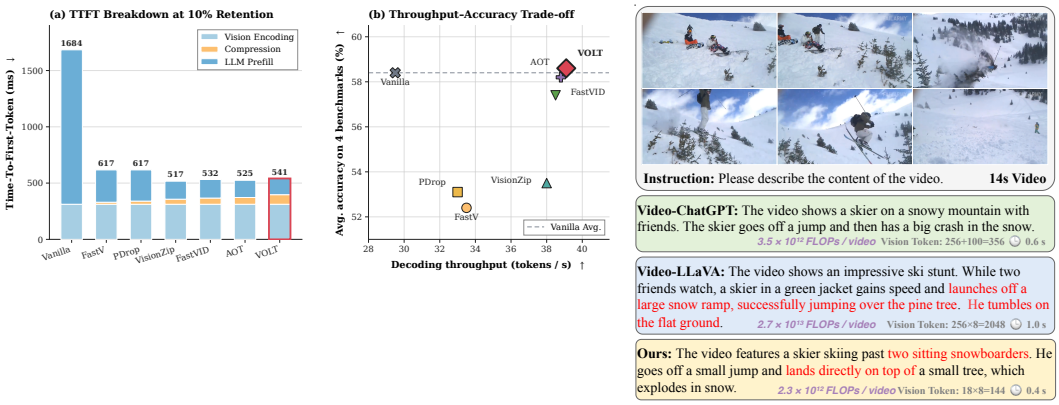


Figure 3: Efficiency and qualitative comparison. (a) End-to-end TTFT on LLaVA-OneVision-7B at 10% retention. (b) Decoding throughput versus average accuracy across four video-understanding benchmarks. (c) A token-compression example showing that VOLT preserves event-relevant evidence under a smaller visual-token budget.

FlashVID, VOLT more effectively converts the additional temporal context into reasoning accuracy at a matched LLM input budget.

Superior performance after token reduction. Across Tables 1–3, VOLT can surpass the vanilla baseline at moderate retention. This suggests that the full visual-token sequence may contain redundant evidence that dilutes task-relevant information at the LLM input. By allocating the budget to tokens that introduce complementary visual directions and by recovering local context through within-frame fusion, VOLT produces a compact token set that remains effective for video reasoning.

Efficiency analysis. Figure 3 reports end-to-end TTFT on LLaVA-OneVision-7B, including vision encoding, compression, and LLM prefill. Vanilla inference spends 1684 ms on TTFT, dominated by LLM prefill. Inner-LLM methods such as FastV and PDrop reduce part of the forward cost but still process full visual tokens in early layers, whereas pre-LLM compressors reduce the token sequence before the LLM. VOLT reaches 541 ms TTFT while maintaining accuracy above the vanilla reference, placing it on the Pareto frontier of speed and accuracy.

Component-wise ablation. Table 5 ablates the three components of VOLT: volume maximization (V), per-frame floor (R), and within-frame merging (M). The combined design delivers the strongest result. Adding R to V raises Avg. from 57.5 to 58.2; the per-frame floor ensures that every frame retains visual support while the remaining budget is allocated globally. Adding M on top of V+R further improves Avg. to 58.6, indicating that local fusion complements the selected support. Table 6 further shows that within-frame neighborhood fusion outperforms cross-frame fusion variants, which can disrupt temporal position alignment.

Table 5: Component-wise ablation at 10% retention. **V**: volume maximization; **R**: per-frame floor; **M**: within-frame merging. Avg. is computed over the four displayed benchmarks.

V	R	M	MVBench	EgoSchema	LongVideo Bench	VideoMME	Avg. Score	Avg. Δ
Vanilla (full)			58.3	60.4	56.4	58.6	58.4	–
Random			53.8	58.2	53.0	54.6	54.9	−3.5
✓			56.7	60.0	56.0	57.5	57.5	−0.9
	✓		55.4	59.6	55.1	56.8	56.7	−1.7
		✓	55.9	60.1	55.6	57.1	57.2	−1.2
✓	✓		57.4	60.7	56.6	57.9	58.2	−0.2
✓		✓	57.0	60.5	56.3	57.6	57.9	−0.5
	✓	✓	56.3	60.0	55.7	57.2	57.3	−1.1
✓	✓	✓	**57.9**	**61.2**	**57.1**	**58.3**	**58.6**	**+0.2**

Table 6: Within-frame fusion strategy at 10% retention.

Fusion strategy	VideoMME	MLVU
Drop only	57.90	64.00
Cross-frame nearest	57.30	63.50
Cross-frame top-k	57.60	63.70
Within-frame avg.	58.10	64.20
Within-frame nbhd.	**58.33**	**64.40**

Hyperparameter sensitivity. The per-frame floor has its highest average score at m=1, preserving temporal support while leaving the remaining budget for global selection (Appendix Table 8).

Token-retention example. Figure 4 shows the spatial support retained at progressively tighter token budgets. At 10% retention, VOLT keeps the food region visually coherent under the same budget.

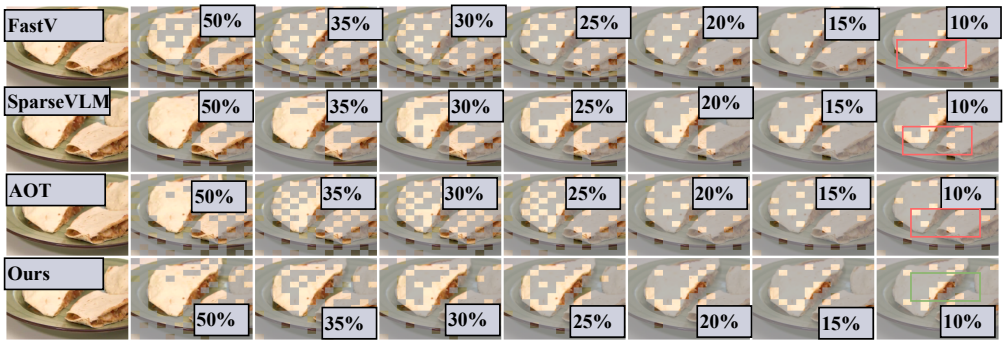


Figure 4: VOLT token-retention example at 50%, 35%, 30%, 25%, 20%, 15%, and 10% of the original tokens. Rows compare FastV, SparseVLM, AOT, and VOLT on the same frame.

5 Conclusion

We have presented VOLT, a training-free video-token compression framework that unifies cross-frame budget allocation and token selection under a single subspace volume maximization objective, subject to a hard per-frame retention floor. By replacing soft temporal weighting with a hard feasibility constraint and by recovering local detail through within-frame cluster merging, VOLT jointly addresses the *Redundant Subspace Sampling* and *Temporal Starvation* failure modes that limit existing compressors under aggressive token budgets. At 10% token retention on LLaVA-Video-7B, VOLT preserves 96.0% of uncompressed accuracy with 8.4% of the LLM prefill FLOPs, illustrating the value of globally complementary tokens for efficient video reasoning.

AI use statement

During manuscript preparation, large language models were used only for language polishing and grammar refinement. The authors reviewed the final text and take responsibility for its content.

REFERENCES

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 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*, 2023.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *ArXiv*, abs/2405.17430, 2024. URL https://api.semanticscholar.org/CorpusID:270063538.

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. *arXiv preprint arXiv:2403.06764*, 2024.

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=H6rDX4w6Al.

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 Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

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 visual language models. *arXiv preprint arXiv:2501.01986*, 2024.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, pp. 13700–13710, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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.

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 *CVPR*, pp. 22195–22206, 2024b.

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.

Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang. MMG-Vid: Maximizing marginal gains at segment-level and token-level for efficient video llms, 2025. URL https://arxiv.org/abs/2508.21044.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, volume 36, pp. 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(1):265–294, 1978.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

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*, 2025.

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*, 2025.

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

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

Alex Kulesza Ben Taskar. Determinantal point processes for machine learning. *stat*, 1050:10, 2013.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, volume 37, pp. 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.

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 *CVPR*, 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 *CVPR*, pp. 19792–19802, 2025b.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Yaqi Xie, Katia Sycara, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *arXiv preprint arXiv:2505.22654*, 2025a.

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 *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. *arXiv preprint arXiv:2412.01818*, 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. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *CVPR*, pp. 13691–13701, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Additional Frame Scaling and Ablations

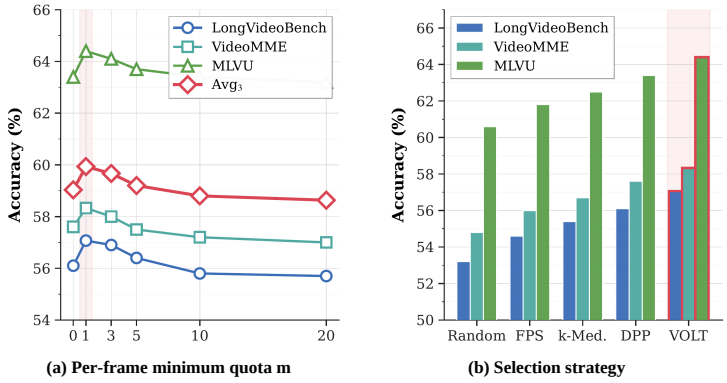
Table 7: Frame scaling on Qwen2.5-VL-7B at 3× and 4×. Avg. is computed over VideoMME Overall, EgoSchema, LongVideoBench, MVBench, and MLVU.

Method	#Frames	Retain	VideoMME				EgoSchema	LongVideo Bench	MVBench	MLVU	Avg.	
			Short	Medium	Long	Overall					Score	Rel. %
Vanilla	16 (1×)	100%	71.0	60.0	50.5	60.5	58.0	56.5	67.0	60.0	60.4	100.0
VisionZip Yang et al. (2025b)	48 (3×)	33%	71.0	60.5	51.5	61.0	58.5	56.5	67.0	60.5	60.7	100.5
FastVID Shen et al. (2025)	48 (3×)	33%	72.0	61.0	51.5	61.5	58.5	57.0	67.0	61.0	61.0	101.0
AOT Li et al. (2026)	48 (3×)	33%	72.5	61.5	52.0	62.0	59.0	57.5	68.0	61.5	61.6	102.0
FlashVID Fan et al. (2026)	48 (3×)	33%	73.0	62.0	52.5	62.5	59.5	58.0	68.5	62.0	62.1	102.8
VOLT (Ours)	48 (3×)	33%	**74.0**	**62.0**	**53.0**	**63.0**	**60.0**	**58.5**	**69.5**	**63.0**	**62.8**	**104.0**
VisionZip Yang et al. (2025b)	64 (4×)	25%	72.0	60.0	51.0	61.0	58.0	56.5	67.0	61.0	60.7	100.5
FastVID Shen et al. (2025)	64 (4×)	25%	72.0	60.5	52.0	61.5	58.5	57.0	67.5	61.5	61.2	101.3
AOT Li et al. (2026)	64 (4×)	25%	73.0	62.0	52.5	62.5	59.5	58.0	68.5	62.0	62.1	102.8
FlashVID Fan et al. (2026)	64 (4×)	25%	73.5	62.5	53.0	63.0	60.0	58.5	69.0	62.5	62.6	103.6
VOLT (Ours)	64 (4×)	25%	**74.5**	**63.0**	**54.5**	**64.0**	**60.5**	**58.5**	**70.0**	**63.5**	**63.3**	**104.8**

Table 8: Sensitivity to per-frame floor m at 10% retention. m=0 disables the floor; m=20 recovers uniform retention. Avg. is computed over the four displayed benchmarks.

m	MVBench	EgoSchema	LongVideo Bench	VideoMME	Avg.
0	57.0	60.5	56.3	57.6	57.9
1	**57.9**	**61.2**	**57.1**	**58.3**	**58.6**
3	57.7	61.0	56.9	58.1	58.4
5	57.4	60.8	56.7	57.9	58.2
10	57.1	60.6	56.6	57.7	58.0
20	56.5	60.2	56.1	57.4	57.6

Selection and reservation ablations. Figure 5 summarizes two ablations on LLaVA-OneVision-7B. **(a)** The per-frame floor is most effective at m=1, balancing temporal coverage and global competition. **(b)** Volume-guided orthogonal selection achieves the highest score among the displayed selection strategies: Random, FPS, k-Medoids, and DPP-MAP.



(a) Per-frame minimum quota m (b) Selection strategy

Figure 5: Ablations on LLaVA-OneVision-7B at 10% retention. (a) Per-frame quota m (m=1 default); Avg_3 averages LongVideoBench, VideoMME, and MLVU. (b) Selection strategies: VOLT compared with Random, FPS, k-Medoids, and DPP-MAP.

A.2 Theoretical Analysis of Subspace Volume Maximization

We characterize the log-determinant criterion in Eq. (4) and its connection to volume-guided orthogonal selection. The normalized criterion makes the fixed-budget objective explicit, while the marginal analysis explains the small-ϵ orthogonal score used by VOLT.

A.2.1 Normalized Objective and Exact Marginal Gain

Let $\mathcal{X} = \{x_1, \ldots, x_M\} \subset \mathbb{R}^d$ be the cross-frame token pool. For $S \subseteq \mathcal{X}$, define

$$F(S) = \log\det\left(I_d + \epsilon^{-1} X_S^\top X_S\right), \qquad \epsilon > 0. \tag{11}$$

Sylvester's identity gives

$$\log\det\left(X_S X_S^\top + \epsilon I_{|S|}\right) = |S|\log\epsilon + F(S). \tag{12}$$

Thus, the two objectives have the same maximizer at the fixed budget $|S| = B$. The matrix determinant lemma gives the marginal in Eq. (6). For $A \subseteq C$, $\Sigma_C \succeq \Sigma_A$ implies $\Sigma_A^{-1} \succeq \Sigma_C^{-1}$ and therefore $\Delta F(x_i \mid A) \geq \Delta F(x_i \mid C)$. Hence F is normalized, monotone, and submodular.

For the cardinality-only problem, exact greedy maximization of $\Delta F(x_i \mid S)$ satisfies the standard bound $F(S_{\text{greedy}}) \geq (1 - 1/e)F(S^\star)$ Nemhauser et al. (1978). This bound applies to exact greedy selection under $|S| = B$. VOLT uses the orthogonal score characterized next and initializes with per-frame base tokens to satisfy Eq. (3).

A.2.2 Orthogonal-Selection Limit

Let $G_S = X_S X_S^\top + \epsilon I_{|S|}$ and $P_S(\epsilon) = X_S^\top G_S^{-1} X_S$. The Schur complement of G_S and Eq. (12) give the exact normalized-objective marginal

$$\Delta F(x_i \mid S) = \log\left(1 + \epsilon^{-1} x_i^\top (I_d - P_S(\epsilon)) x_i\right). \tag{13}$$

For a compact singular value decomposition $X_S = W\Lambda V^\top$,

$$P_S(\epsilon) = V\,\mathrm{diag}\left(\frac{\sigma_j^2}{\sigma_j^2 + \epsilon}\right) V^\top. \tag{14}$$

As $\epsilon \to 0$, this matrix approaches the orthogonal projector $UU^\top$ onto the span of the selected tokens. Consequently,

$$\lim_{\epsilon \to 0} x_i^\top (I_d - P_S(\epsilon)) x_i = \|x_i - UU^\top x_i\|_2^2 = g_i. \tag{15}$$

For positive singular values, the finite-ϵ difference in the quadratic term is at most $\epsilon\|x_i\|_2^2/(\sigma_{\min}^2 + \epsilon)$, where $\sigma_{\min}$ is the smallest positive singular value. This relationship motivates the orthogonal-selection rule in the high-compression regime.

A.2.3 Discussion: Subspace Saturation and Complexity Reduction

We conclude by addressing two critical implications of our theoretical framework.

1. Subspace saturation ($B > d$): When the selected span reaches rank d, further orthogonal residuals are zero. The regularized objective in Eq. (4) remains defined, and the rank threshold η in the solver prevents zero-vector normalization. The residual criterion is most discriminative in the aggressive-compression settings where the retained span has not saturated the feature space.

2. Complexity Reduction: Naively recomputing a determinant for every candidate at every step is costly. VOLT instead maintains an orthonormal basis and updates each candidate's residual by $x_i^{\perp(t)} = x_i^{\perp(t-1)} - (u_t^\top x_i^{\perp(t-1)})u_t$. This costs $\mathcal{O}(d)$ per candidate and $\mathcal{O}(MBd)$ over B selection steps, enabling global cross-frame selection before LLM inference.

Backbone	Frames	Tokens/frame	Vanilla visual tokens	Default setting
LLaVA-OneVision-7B	32	196	6,272	$m = 1$
LLaVA-Video-7B	64	169	10,816	$m = 1$
Qwen2.5-VL-7B	16	model-native	dynamic	$m = 1$

Table 9: Backbone and default configuration summary. Fixed-token LLaVA-family models expose a constant number of visual tokens per sampled frame, while Qwen2.5-VL uses its model-native visual-tokenization pipeline.

Backbone	25% retain	20% retain	15% retain	10% retain
LLaVA-OneVision-7B	49/frame, 1,568 total	39/frame, 1,248 total	29/frame, 928 total	20/frame, 640 total
LLaVA-Video-7B	42/frame, 2,688 total	34/frame, 2,176 total	25/frame, 1,600 total	17/frame, 1,088 total

Table 10: Token-budget configurations for fixed-token video backbones. The final allocation is not forced to be uniform: the listed per-frame value defines the global total budget, while VOLT reallocates the non-reserved slots across frames.

B Additional Implementation Details

This section records the implementation choices used throughout the main paper and the supplementary visualizations. VOLT is training-free: it introduces no trainable parameters, does not alter the frozen vision encoder or LLM, and operates after visual features have been produced by the backbone-specific video encoder/projector stack. Unless stated otherwise, we use a per-frame reservation floor $m = 1$.

Token Budget. For a fixed-token model with F frames and T visual tokens per frame, the vanilla visual-token count is FT. At retention ratio r, VOLT sets the total budget to $B = F\,\mathrm{round}(rT)$. The per-frame floor consumes Fm slots first; all remaining $B - Fm$ slots then compete globally under the volume-guided selection rule. For Qwen2.5-VL, retention is computed from its model-native visual-token count. For the two fixed-token backbones, the standard budgets are:

Computational Cost. Following the main manuscript, processing n_i visual tokens in one decoder layer with hidden size d and FFN intermediate size d_{ffn} costs

$$C_i(n_i) = 4n_i d^2 + 2n_i^2 d + 2n_i d d_{\mathrm{ffn}}. \tag{16}$$

Reducing visual tokens before the LLM lowers every subsequent prefill layer. In contrast, inner-LLM pruning leaves the early layers exposed to the full visual sequence. This is why the main-paper efficiency figure separates selector overhead from LLM prefill time: VOLT spends a small amount of CPU/GPU time on orthogonal selection, but it substantially reduces the larger downstream prefill cost.

C Reproducible Procedure

This section gives a compact implementation-level description of VOLT. It is included to make the supplementary material self-contained and to clarify the exact order in which volume maximization, temporal reservation, and within-frame merging are applied.

Input: video-token features $\{X_f\}_{f=1}^{F}$, where $X_f = \{x_{f,i}\}_{i=1}^{T}$; retention ratio r; per-frame floor m.
Output: compressed video-token sequence $\tilde{X}$.

1. Set the per-frame nominal budget $k = \mathrm{round}(rT)$ and total budget $B = Fk$.
2. Normalize token features and initialize the retained set $S = \emptyset$.
3. For each frame f, apply volume-guided orthogonal selection within X_f until m base tokens are selected:
$$g(x \mid U) = \left\| x - U(U^\top x) \right\|_2^2,$$
where U is the current orthonormal basis. Add the selected base tokens to S.
4. Build an orthonormal basis from all base tokens. Continue the same selection rule over the full cross-frame pool $\mathcal{X} = \cup_f X_f$ until $|S| = B$.
5. For every pruned token $x_{f,i} \notin S$, find its nearest retained token in the same frame by cosine similarity and merge it into that retained representative by a weighted average.
6. Concatenate retained-and-merged tokens in temporal order and feed the compressed sequence to the frozen LLM.

The first selection pass enforces $|S \cap X_f| \geq m$ for every frame. The second pass lets all remaining slots compete globally, so two similar tokens from different frames no longer receive separate slots unless both provide meaningful orthogonal contribution. The final merge is restricted to the same frame to preserve positional consistency and avoid mixing unrelated temporal evidence.

D MASKED TOKEN-RETENTION EXAMPLES

We additionally provide a masked-token view of the retained support.

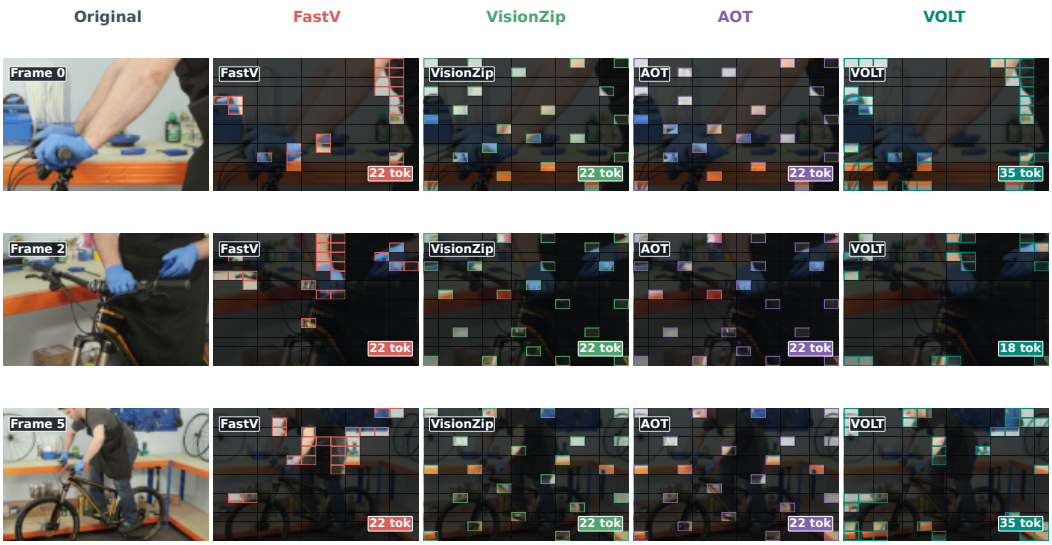


Figure 6: Masked token-retention examples on a second input video. Black patches denote discarded visual tokens; colored patch borders denote retained tokens. The total budget is comparable across methods, while VOLT redistributes tokens under temporal reservation instead of forcing a uniform per-frame allocation.

We also visualize the retained support under progressively tighter budgets.

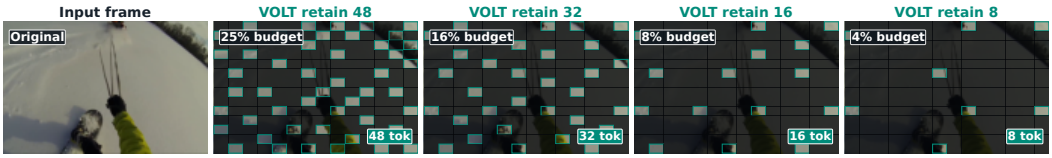


Figure 7: Token-retention progression under progressively tighter budgets. The same frame is compressed to 48, 32, 16, and 8 retained tokens. VOLT maintains a sparse but distributed support, illustrating how the volume objective discourages repeatedly selecting near-duplicate local evidence.

E Temporal Reservation and Subspace Diagnostics

We provide complementary diagnostics for temporal reservation and subspace saturation.

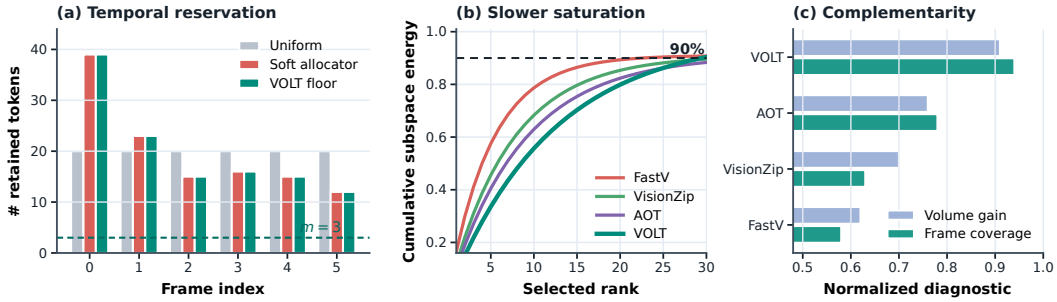


Figure 8: Diagnostics for temporal reservation and subspace complementarity. (a) The temporal floor prevents starvation under uneven frame scores. (b) Slower saturation indicates that selected tokens continue to add complementary subspace directions. (c) VOLT improves the balance between volume gain and frame coverage.

F Within-frame Cluster Merging

We conclude with a visualization of within-frame cluster merging.

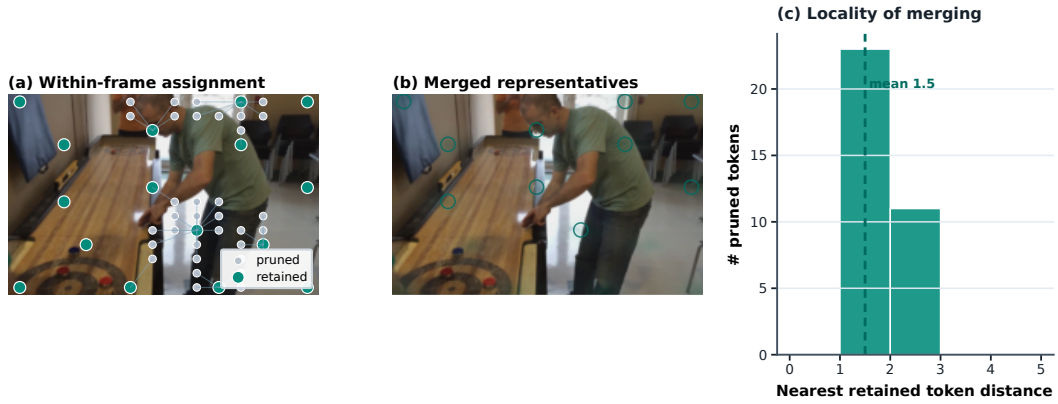


Figure 9: Visualization of within-frame cluster merging. Pruned tokens are assigned to nearby retained representatives in the same frame; most assignments remain local, supporting detail-preserving refinement rather than cross-frame mixing.

G Broader Impacts

VOLT reduces the inference cost of video MLLMs without retraining, which can make long-video understanding more accessible under limited compute budgets. This is beneficial for deployment scenarios where latency, memory, and energy are practical bottlenecks. For safety-critical deployments, task-specific stress validation remains appropriate.

H Assets, Licenses, and Consent

Our experiments use the public LLaVA-OneVision, LLaVA-Video, and Qwen2.5-VL backbones with LMMs-Eval and the five video-understanding benchmarks cited in Sec. 4. The supplementary visualizations use local ActivityNet clips for qualitative mechanism illustration. The respective project pages provide the applicable license and usage terms.