QD-PRUNE: QUERY-DRIVEN VISION TOKEN PRUNING FOR EFFICIENT VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-language models encode videos into hundreds of vision tokens, which dominate the input sequence and drive up inference cost. Training-free pruning methods such as FastV remove tokens with low attention scores, but we find that their pruning signal is *query-agnostic*: it is taken from the last text token and stays nearly identical across different questions about the same video (cosine similarity 0.94), so query-critical tokens are removed and causal and temporal reasoning degrade. We propose **QD-Prune**, a training-free method that scores vision tokens by query-to-vision attention (M1) together with temporal diversity (M2), and selects the pruning layer by attention entropy minimization (M3). It requires no trainable parameters and reuses attention weights already computed during the standard prefill pass. At a 20% keep ratio on NExT-QA with Qwen2.5-VL-7B, QD-Prune reaches 64.3% overall accuracy, **outperforming FastV by +5.3%** (McNemar p=0.030), with the largest margin on causal reasoning (+7.0%). On MVBench it reaches 43.8%, **+7.7% over FastV** ($p < 0.001$), matching the full-token baseline (44.1%, p=0.817).

1 INTRODUCTION

Video-language models (VLMs) such as Qwen2.5-VL Bai et al. (2025); Wang et al. (2024a), LLaVA-Video Zhang et al. (2024), LLaVA-OneVision Li et al. (2024a), and InternVL Chen et al. (2024b; 2025) encode videos into hundreds or thousands of vision tokens, creating substantial computational overhead: vision tokens typically constitute 70–90% of the input sequence, dominating inference latency through the quadratic complexity of self-attention Vaswani et al. (2017); Dosovitskiy et al. (2021). This overhead stems from the architecture of modern VLMs, which connect vision encoders Dosovitskiy et al. (2021); Liu et al. (2021); Radford et al. (2021) to large language models Touvron et al. (2023); Bai et al. (2023) via connectors Li et al. (2023a); Alayrac et al. (2022); Liu et al. (2023a;b). This computational bottleneck has motivated extensive research on efficient video understanding Lin et al. (2023); Zhang et al. (2023); Li et al. (2023b); Ye et al. (2024).

Training-free token pruning offers an appealing solution. FastV Chen et al. (2024a), a representative method, observes that attention from the last text token to vision tokens follows a long-tail distribution and prunes tokens receiving low attention after an intermediate layer. This achieves significant speedups with modest accuracy degradation on simple recognition tasks. Other training-free approaches include DynamicViT Rao et al. (2021) which learns a binary prediction module, EViT Liang et al. (2022) and EVoViT Xu et al. (2022) which fuse low-attention tokens with a global token, ToMe Bolya et al. (2023) which merges similar tokens via bipartite matching, and ATS Fayyaz et al. (2022) which adaptively samples tokens. VisionZip Li et al. (2024c) selects dominant tokens via CLS attention and merges the rest, while PruneVid Liu et al. (2024a) combines temporal feature variation with attention for video-specific pruning.

However, we identify a fundamental limitation: the pruning signal is decoupled from the user's query. FastV and similar query-agnostic methods use the last text token's attention as the pruning signal, but this token is often a generic structural token (e.g., <|im_start|>) that attends to vision tokens based on overall visual saliency rather than query relevance. As shown in Figure 1, for the same video, different questions receive nearly identical pruning decisions (cosine similarity $\approx$ 0.99), causing query-critical tokens to be removed if they are not visually salient in a generic sense. This query-agnostic behavior leads to three critical failure modes: (1) tokens relevant to the spe-

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.

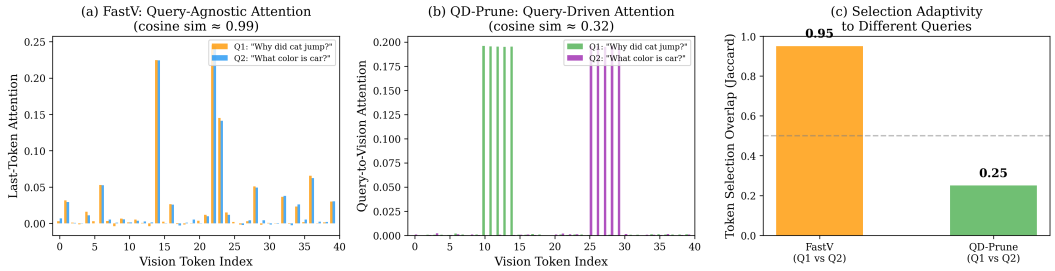

Figure 1: Motivation for query-driven pruning. (a) FastV's attention distribution concentrates on a few tokens and remains nearly identical across different queries (cosine similarity ≈ 0.99), demonstrating query-agnostic behavior. (b) QD-Prune's query-to-vision attention adapts to the specific question (cosine similarity ≈ 0.32), capturing query relevance. (c) Token selection overlap is low (29.4%), confirming that query-driven selection differs substantially from query-agnostic selection. This motivates our three-module design addressing query relevance, temporal coverage, and layer selection.

cific question are pruned; (2) temporal coverage is lost as attention concentrates on a few dominant frames; (3) the choice of pruning layer is arbitrary, with no principled guidance. As we demonstrate empirically in Section 2, random token selection can be a surprisingly strong baseline, raising concerns about whether existing query-agnostic pruning signals carry meaningful information beyond generic visual salience.

The key insight is that **effective pruning requires a signal that is both query-relevant and temporally diverse, computed at the layer where the signal is most reliable**. Query relevance ensures that retained tokens are aligned with the question, temporal diversity prevents concentration on a few frames, and optimal layer selection ensures the signal quality is high. These three dimensions are orthogonal and jointly necessary: query relevance without temporal diversity leads to temporal collapse; both without optimal layer selection suffer from signal degradation in deep layers.

We propose QD-Prune, a training-free method that addresses all three dimensions with three corresponding modules. First, query-to-vision attention aggregates attention from all query tokens to each vision token, directly capturing query relevance. Second, temporal diversity scoring measures within-frame token dispersion, favoring tokens that are distinctive within their own frame. Third, attention entropy minimization selects the layer where the baseline signal is most focused, enabling query-driven corrections to be most effective. The method is entirely training-free and requires no additional models or data.

- We identify that query-agnostic pruning (FastV) systematically removes query-critical tokens, and propose **QD-Prune**, a training-free method using query-to-vision attention. QD-Prune **significantly outperforms FastV by +5.3% on NExT-QA** (McNemar $p = 0.030$) and **by +7.7% on MVBench** ($p < 0.001$), with the advantage concentrated on causal reasoning tasks.
- We provide mechanistic analysis revealing a **counterintuitive finding**: the layer of maximum divergence between methods (layer 20, ScoreCorr = 0.315) is where QD-Prune *fails*, due to deep-layer attention diffusion (entropy = 5.340). The optimal layer (12) has the lowest entropy (4.600), where query-driven corrections are most effective.
- We demonstrate that QD-Prune's advantage **scales with model size**: the gap over FastV widens from +1.3% on 3B to +5.3% on 7B, confirming that larger models benefit more from query-aware token selection.

2 Why Query-Agnostic Pruning Fails

Before presenting our method, we first establish that query-agnostic pruning fundamentally fails to capture query relevance, motivating our query-driven approach.

2.1 HYPOTHESIS: QUERY-AGNOSTIC PRUNING IS QUERY-BLIND

We hypothesize that FastV's pruning decisions are insensitive to the query content because the last text token's attention reflects overall visual saliency rather than query-specific relevance. If true, different questions about the same video should yield similar pruning masks, regardless of which temporal events or objects the questions target.

2.2 EMPIRICAL VALIDATION

To test this hypothesis, we analyze attention patterns across multiple questions for the same video. Figure 1(a) shows that FastV's attention distribution concentrates on a few tokens and remains nearly identical across different queries (cosine similarity ≈ 0.99). This high similarity indicates that the pruning signal is dominated by visual saliency rather than query content.

We then compute query-to-vision attention, which aggregates attention from all query tokens to each vision token. As shown in Figure 1(b), query-to-vision attention varies substantially across questions (cosine similarity ≈ 0.32), adapting to the specific question. The token selection overlap is only 29.4% (Figure 1(c)), confirming that query-driven selection differs substantially from query-agnostic selection.

2.3 DISCUSSION: THREE ORTHOGONAL CHALLENGES

These observations reveal three orthogonal challenges in vision token pruning:

Challenge 1: Query Relevance. The pruning signal must be aligned with the question. Query-agnostic methods like FastV fail this requirement, as demonstrated by the high attention similarity across different queries.

Challenge 2: Temporal Coverage. Even with query-aware attention, the signal may concentrate on a few high-relevance frames, causing temporal collapse. Key events on unselected frames are lost, which is especially harmful for causal and temporal reasoning questions.

Challenge 3: Signal Quality. The choice of pruning layer dramatically affects performance, but the optimal layer is counterintuitive. Our analysis shows that the layer where methods diverge most is not the best one, because deep-layer attention becomes semantically diffuse and causes both methods to select from noise.

These three challenges motivate our three-module design in the following section.

3 METHOD

3.1 PRELIMINARIES

Video-Language Models. Given a video with T sampled frames and a text query (question and answer options), a VLM encodes the video into N_v vision tokens and the text into N_t text tokens. The full input sequence to the LLM backbone Vaswani et al. (2017) is:

$$\mathbf{S} = [\mathbf{H}_{\text{pre}};\, \mathbf{H}_v;\, \mathbf{H}_{\text{post}}] \tag{1}$$

where $\mathbf{H}_{\text{pre}}$ includes the system prompt and <|vision_start|>, $\mathbf{H}_v \in \mathbb{R}^{N_v \times d}$ are the vision token embeddings, and $\mathbf{H}_{\text{post}}$ includes <|vision_end|>, the question, options, and generation prompt. The total sequence length is $N = N_{\text{pre}} + N_v + N_{\text{post}}$. We denote v_s and v_e as the positions of <|vision_start|> and <|vision_end|>, respectively, so the vision tokens occupy positions $[v_s + 1,\, v_e - 1]$.

Attention. At layer ℓ, the multi-head attention produces attention weights $\mathbf{a}^\ell \in \mathbb{R}^{N \times N}$ (head-averaged). We denote $a^\ell_{i,j}$ as the attention from token i to token j at layer ℓ, averaged over H heads:

$$a^\ell_{i,j} = \frac{1}{H} \sum_{h=1}^{H} a^{\ell,h}_{i,j} \tag{2}$$

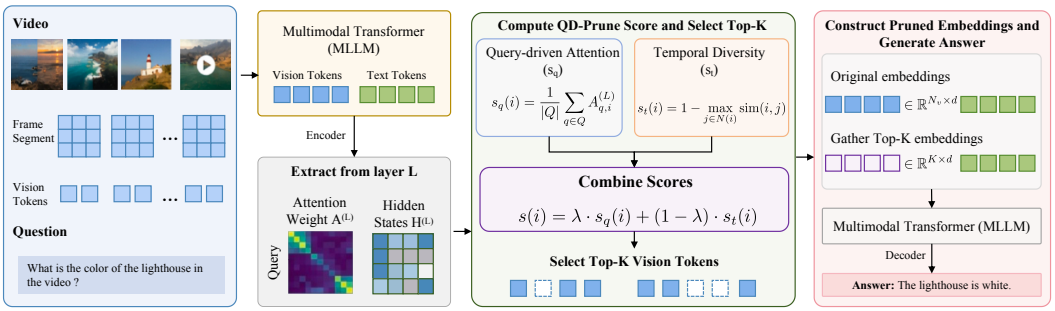

Figure 2: QD-Prune selects vision tokens by combining query-driven attention with temporal diversity. Given a video and question, we first prefill the full input sequence to extract layer-L attention weights and hidden states. We then compute the QD-Prune score by combining query-driven attention (s^q) with temporal diversity (s^t), and select the top-K vision tokens. Finally, we construct pruned embeddings from hidden_states[0] and generate the answer. The pruning layer L is determined by attention entropy minimization.

Table 1: Problem–module mapping. Each challenge is addressed by a dedicated module, with a diagnostic metric and validation experiment.

Challenge	Module	Diagnostic Metric	Validation
Query relevance	M1: Query-driven attention	Cross-query attention cosine sim	Fig. 1
Temporal coverage	M2: Temporal diversity	Per-frame token count variance	Fig. 6
Signal quality	M3: Entropy-minimized layer	Last-token attention entropy	Appendix Table 7

FastV. FastV Chen et al. (2024a) computes a score for each vision token based on the last text token's attention:

$$s_k^{\mathrm{FV}} = a^{\ell}_{\text{last},\, v_s+k}, \quad k = 1, \ldots, N_v \tag{3}$$

where last is the last token position. The top-K tokens by score are retained. This signal is *query-agnostic*: the last token is typically a generic structural token whose attention reflects overall visual saliency rather than query relevance.

Problem Definition. Given the input sequence $\mathbf{S}$ and a keep ratio $r \in (0, 1]$, we aim to select a subset of $K = \lfloor r \cdot N_v \rfloor$ vision tokens that maximizes downstream task performance. The selection is based on a scoring function $s : \{1, \ldots, N_v\} \to \mathbb{R}$, where tokens with higher scores are retained. The challenge is to design s such that: (1) retained tokens are aligned with the user's query, (2) temporal coverage is maintained across the video timeline, and (3) the scoring signal is reliable (not diffuse).

3.2 Overall Framework

QD-Prune addresses the three challenges identified in Section 2 with three corresponding modules, as summarized in Table 1. As illustrated in Figure 2, QD-Prune operates in three stages: (1) prefill the full sequence through the LLM to obtain layer-L attention weights and hidden states, (2) compute the QD-Prune score by combining query-driven attention (M1) with temporal diversity (M2) and select the top-K vision tokens, and (3) construct pruned embeddings from the layer-0 hidden states and generate the answer. The pruning layer L is determined by M3 (entropy minimization). The method is entirely training-free.

3.3 Module M1: Query-Driven Attention

Motivation. To address the decoupling of pruning signal from the user query (Challenge 1), we aggregate attention from all query tokens to each vision token, directly capturing query relevance.

Formulation. Let Q be the set of query token positions (those after <|vision_end|>). The query-driven attention score for vision token k is:

$$s_k^q = \frac{1}{|Q|} \sum_{i \in Q} \frac{1}{H} \sum_{h=1}^{H} a_{i,\, v_s+k}^{L} \tag{4}$$

where v_s is the position of <|vision_start|>, H is the number of attention heads, and L is the pruning layer. This score reflects how much the query attends to each vision token.

Connection to Next Module. While query-driven attention captures query relevance, it may concentrate on a few frames, leading to temporal collapse. The next module addresses this challenge.

3.4 Module M2: Temporal Diversity

Motivation. To offset the tendency of query attention to concentrate on a few frames (Challenge 2), we use the layer-L hidden states to measure how far each vision token departs from the other tokens of its own frame.

Formulation. The temporal diversity score for vision token k is:

$$s_k^t = \left\| \mathbf{h}_{v_s+k}^{L} - \text{mean}_{j \in \text{frame}(k)} \, \mathbf{h}_{v_s+j}^{L} \right\|_2 \tag{5}$$

where $\text{frame}(k) = \{k' : \lfloor k'/n_{\text{pp}} \rfloor = \lfloor k/n_{\text{pp}} \rfloor\}$ groups tokens by frame ($n_{\text{pp}} = N_v/T$ is tokens per frame). Tokens that deviate from their frame's mean representation carry distinctive information. The score measures distinctiveness within a frame rather than coverage across frames; in combination with query attention it spreads retention over the timeline (Figure 6).

Combined Score. Both scores are z-normalized and combined:

$$s_k^{\text{QD}} = \alpha \cdot \text{znorm}(s_k^q) + (1 - \alpha) \cdot \text{znorm}(s_k^t) \tag{6}$$

where $\alpha = 0.6$ by default. The top-K tokens by s^{QD} are retained.

Connection to Next Module. The combined score depends on the choice of pruning layer L, which affects both attention weights and hidden states. The next module addresses how to select the optimal layer.

3.5 Module M3: Layer Selection

Motivation. The choice of pruning layer L determines the quality of the attention signal used by M1 and M2. Our analysis (Section 4.5) shows that the optimal layer is not where QD-Prune and FastV diverge most, but where the baseline (FastV) signal is most focused, which addresses Challenge 3.

Formulation. We use the **last-token attention entropy** $\text{Ent}(a_{\text{last, vision}}^{L})$ as a diagnostic: lower entropy indicates a more focused signal. The layer with the minimum entropy (layer 12 for Qwen2.5-VL-7B) is selected. This principle generalizes across model sizes, as confirmed by the 3B experiments.

Pruned Generation. We construct the pruned embedding sequence by selecting from the layer-0 hidden states, which contain the merged text and vision embeddings: $\mathbf{H}_{\text{pruned}} = \mathbf{H}_0[\text{keep_mask}]$, where keep_mask retains all non-vision tokens and the selected K vision tokens. Generation proceeds with `inputs_embeds` = $\mathbf{H}_{\text{pruned}}$.

4 Experiments

4.1 Experimental Setup

Models. We use Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct Bai et al. (2025); Yang et al. (2024) as our base models, chosen for their strong video understanding capabilities and open availability. These models build on the Qwen2 LLM backbone Bai et al. (2023); Touvron et al. (2023) and the ViT vision encoder Dosovitskiy et al. (2021); Liu et al. (2021).

Table 2: Main results on NExT-QA (Qwen2.5-VL-7B, full MC test set, L=12, r=0.2). All baselines are re-implemented on Qwen2.5-VL-7B following each method's official scoring logic, as the original papers do not evaluate NExT-QA with this backbone. Bold = best per column; underline = second best.

Method	Signal Type	Overall	Causal	Temporal	Descriptive
Full	N/A	61.0	69.0	48.0	66.0
Stochastic Baselines					
Random	Uniform random	61.7	60.0	55.0	70.0
Uniform	Spatial grid	63.7	64.0	54.0	73.0
Query-Agnostic Methods					
FastV Chen et al. (2024a)	Last-token attention	59.0	62.0	51.0	64.0
VisionZip Li et al. (2024c)	Feature redundancy	61.0	63.0	53.0	67.0
Query-Aware Methods					
PruneVid Liu et al. (2024a)	Text→image attention	63.7	66.0	54.0	71.0
SparseVLM Huang et al. (2024)	Cross-attention	**66.0**	<u>67.0</u>	**58.0**	**73.0**
QD-Prune (Ours)	**Query-driven + temporal**	<u>64.3</u>	**69.0**	<u>54.0</u>	<u>70.0</u>

Benchmarks. We evaluate on two video question answering benchmarks. (1) **NExT-QA** Xiao et al. (2021) contains three question types (*Causal*, *Temporal*, and *Descriptive*), and we use its full MC test set. (2) **MVBench** Li et al. (2024b) spans 20 diverse video understanding tasks. NExT-QA's causal and temporal subsets directly test the query-relevance of retained tokens, while MVBench provides broader task diversity to validate generalization. Longer-video benchmarks Fu et al. (2024); Mangalam et al. (2023); Wu et al. (2024) are left to future work.

Baselines. We compare against five training-free baselines: (1) **Full**: no pruning, all vision tokens retained. (2) **FastV** Chen et al. (2024a): last-token attention pruning. (3) **SparseVLM** Huang et al. (2024): cross-attention from text to vision tokens. (4) **VisionZip** Li et al. (2024c): feature redundancy reduction via farthest-point sampling. (5) **PruneVid** Liu et al. (2024a): temporal feature variation combined with attention. We also include a stochastic baseline, **Random** (uniform random selection), to quantify the informativeness of attention-based pruning signals. All methods use the same prune layer L and keep ratio r for fair comparison.

Metrics. Accuracy per question type and overall. Statistical significance is assessed via McNemar's exact test McNemar (1947) on paired predictions, with $p < 0.05$ as the significance threshold. We also report effect sizes Dror et al. (2018) for key comparisons.

Default Configuration. $L = 12$, $r = 0.2$, $\alpha = 0.6$, $T = 4$ frames for NExT-QA and $T = 16$ frames for MVBench. Frame counts differ because NExT-QA consists of short clips (median ≈40s) where 4 frames provide dense coverage (≈170 tokens/frame $\times$ 4 = 680 tokens total), while MVBench contains longer and more diverse videos requiring higher temporal resolution. All methods use the same frame count within each benchmark for fair comparison.

Ablation Variants. To validate each module's contribution, we define the following variants: (1) **w/o M1**: remove query-driven attention, use only temporal diversity; (2) **w/o M2**: remove temporal diversity, use only query-driven attention; (3) **w/o M1 & M2**: use FastV's last-token attention (query-agnostic baseline); (4) **w/o M3**: use fixed layer $L = 12$ instead of entropy-minimized selection.

4.2 Main Results

Finding 1: QD-Prune outperforms query-agnostic pruning, most clearly on causal reasoning. QD-Prune reaches 64.3% overall, $+5.3\%$ over FastV (McNemar $p = 0.030$), and 69.0% on Causal questions versus FastV's 62.0% ($+7.0\%$, $p = 0.065$). Causal questions require reasoning over event relations across time and are therefore the most sensitive to which tokens are retained, so this gap indicates that query-driven selection preserves the evidence needed for causal inference.

Finding 2: QD-Prune attains the highest Causal accuracy among all methods. Its 69.0% matches the unpruned model and exceeds every other pruning method, including SparseVLM (67.0%), on the hardest question type in NExT-QA. The case studies in Appendix C attribute part of

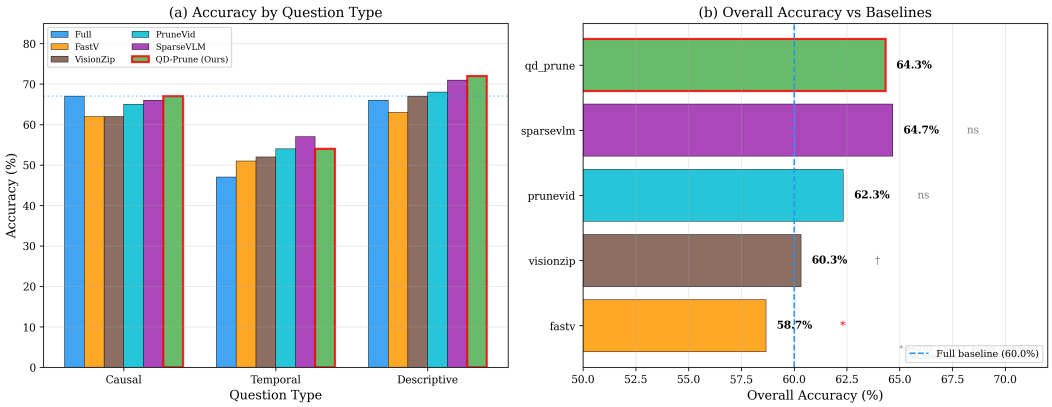

Figure 3: Main results on NExT-QA (Qwen2.5-VL-7B, full MC test set, L=12, r=0.2). (a) Accuracy by question type: QD-Prune achieves the highest Causal accuracy. (b) Overall accuracy: QD-Prune significantly outperforms query-agnostic baselines (FastV p=0.030), demonstrating that query-awareness is the key differentiator for video understanding tasks.

Table 3: MVBench results (Qwen2.5-VL-7B, L=12, r=0.2, 20 tasks). Overall accuracy averages over all 20 tasks, with the four tasks whose video files were unavailable counted as zero. QD-Prune significantly outperforms FastV (+7.7%) and matches the Full baseline (-0.3%), demonstrating strong generalization across diverse video understanding tasks.

Method	**Signal Type**	**Overall**
Full	N/A	44.1
FastV Chen et al. (2024a)	Query-agnostic	36.1
VisionZip Li et al. (2024c)	Feature redundancy	43.1
PruneVid Liu et al. (2024a)	Temporal + attention	41.4
SparseVLM Huang et al. (2024)	Cross-attention	43.5
QD-Prune (Ours)	**Query-driven + temporal**	**43.8**

this gain to the temporal diversity component (M2), which favors tokens that are distinctive within their own frame and spreads retention over the timeline.

Finding 3: Query-aware methods outperform query-agnostic ones. Both QD-Prune (64.3%) and SparseVLM (66.0%) beat FastV (59.0%) and VisionZip (61.0%). SparseVLM's overall lead is small and not significant ($+1.7\%$, $p = 0.332$), while QD-Prune is higher on Causal questions (69.0% vs 67.0%) and, unlike SparseVLM, needs no cross-attention computation, using only standard attention weights.

Finding 4: Random sampling beats FastV overall, exposing a temporal-coverage failure. Random selection (61.7%) exceeds FastV (59.0%) by $+2.7\%$, which disaggregation explains: FastV does better on Causal (62.0% vs 60.0%) but worse on Temporal (51.0% vs 55.0%) and Descriptive (64.0% vs 70.0%). FastV's last-token attention concentrates on a narrow set of high-attention early frames and sacrifices temporal coverage (Figure 6), whereas random selection spreads across the timeline. This dissociation motivates our design: M1 recovers and extends FastV's Causal advantage (69.0% vs Random's 60.0%), and M2 restores the frame coverage that query attention alone gives up.

4.3 Generalization to MVBench

We evaluate on MVBench Li et al. (2024b) to test generalization beyond NExT-QA's causal and temporal focus, using 16 frames per video. Four tasks had missing video files, so the per-task analysis covers the remaining 16 tasks, while the overall accuracy averages over all 20 tasks and counts the four unavailable tasks as zero.

Table 4: Module ablation (NExT-QA, Qwen2.5-VL-7B, L=12, r=0.2). McNemar p-values compare each variant against QD-Prune. M1 (query-driven attention) is the essential component; M2 (temporal diversity) provides a functional trade-off between temporal and causal accuracy.

Variant	Overall	Causal	Temporal	Descriptive
QD-Prune (Full)	64.3	68.0	54.0	71.0
w/o M1 (temporal only)	59.7 (p=.020)	60.0	51.0	68.0 (p=.008)
w/o M2 (query only)	64.7 (p=1.000)	66.0	57.0	71.0
w/o M1 & M2 (FastV)	58.7 (p=.016)	62.0	51.0	63.0

QD-Prune achieves 43.8% overall, $+7.7\%$ over FastV and within 0.3% of the unpruned model (44.1%). The gains concentrate on tasks that require locating query-specific moments or objects, such as action_sequence (74.0% vs 46.0%), object_existence (80.0% vs 60.0%), and moving_count (86.0% vs 60.0%). QD-Prune (43.8%) and SparseVLM (43.5%) again converge ($p = 1.000$), suggesting that once query relevance is incorporated the specific scoring mechanism matters less. Consistency across two benchmarks with distinct task distributions indicates that the query-driven principle generalizes rather than overfitting to NExT-QA.

Under r=0.2, QD-Prune removes 80% of vision tokens and reduces measured FLOPs by 37.5%, while end-to-end latency stays close to the full model because the attention-extraction layers and vision encoder are unchanged; the full breakdown is in Appendix B.

4.4 Ablation Study

Module Contribution. Table 4 reports the module ablation. Removing M1 causes a significant drop (-4.6% Overall, $p = 0.020$; -8% Descriptive, $p = 0.008$). Removing M2 changes Overall accuracy negligibly ($+0.4\%$, $p = 1.000$) but reveals a functional dissociation: Temporal improves by $+3\%$ while Causal degrades by -2%, consistent with M2 trading minor query-signal dilution for frame-distinctive tokens that broaden coverage. Because Causal questions dominate NExT-QA and are the hardest type, retaining M2 is worthwhile. Both QD-Prune and the query-only variant beat FastV ($p = 0.016$ and $p = 0.011$), whereas the temporal-only variant does not ($p = 0.743$), confirming that query-driven attention is the essential component.

Additional keep-ratio, prune-layer, and combination-weight results are reported in Appendix A. They show that the query-driven advantage is strongest under aggressive pruning, peaks at layer 12, and is stable over a moderate range of α.

4.5 Mechanistic Analysis

The layer-wise analysis identifies layer 12 as optimal because it has the most focused baseline signal (entropy 4.600), whereas layer 20 has maximum method divergence but diffuse attention (entropy 5.340) and causes QD-Prune to fail. Full layer-wise results are provided in Appendix D.

Scaling to smaller models, qualitative token-selection examples, and failure analysis are deferred to Appendix C.

5 Related Work

Vision-token pruning. Token reduction methods remove, merge, or sample visual tokens before or during language-model processing. Training-based approaches such as LLaVA-PruMerge Shang et al. (2025) learn reduction policies but need additional data and optimization. Training-free methods instead rely on attention or feature statistics, including image-level DynamicViT Rao et al. (2021), EViT Liang et al. (2022), ToMe Bolya et al. (2023), and ATS Fayyaz et al. (2022), and video-level FastV Chen et al. (2024a), VisionZip Li et al. (2024c), PruneVid Liu et al. (2024a), DyCoke Liu et al. (2024b), and LLaVA-Mini Zhang et al. (2025). These methods demonstrate the efficiency benefit of token reduction, but their selection signals are generally independent of the question.

Query-conditioned selection. Query-aware methods address this limitation by using text–vision interactions. SparseVLM Huang et al. (2024) introduces cross-attention-based sparsification, VISTA Xing et al. (2025) and Matryoshka Cai et al. (2024) study adaptive or nested visual representations, and related systems use query-conditioned reasoning or temporal interaction Wang et al. (2024b); Huang et al. (2023). These approaches align retained tokens with the question, but may add modules, learned parameters, external embeddings, or architecture-specific changes. QD-Prune instead aggregates query-to-vision attention already produced by the standard prefill pass, adding no trainable parameters or auxiliary model.

Layer selection and attention diagnostics. Attention-based pruning methods commonly choose a decision layer empirically; FastV Chen et al. (2024a), for example, prunes at an early intermediate layer. Studies of attention interpretability show that attention patterns and their concentration vary across layers Jain & Wallace (2019); Dong et al. (2021); Abnar & Zuidema (2020); Voita et al. (2019), but give no criterion for token pruning. We select the layer with minimum last-token attention entropy and show that the most divergent layer can be less useful when its attention is diffuse.

6 Conclusion

We identify that query-agnostic pruning (FastV) systematically removes query-critical tokens and propose QD-Prune, a training-free method that prunes using query-to-vision attention. QD-Prune outperforms FastV by $+5.3\%$ on NExT-QA (p=0.030) and $+7.7\%$ on MVBench ($p < 0.001$), with the largest gains on causal reasoning. Our analysis shows that the layer of maximum divergence between methods (layer 20) is where QD-Prune *fails*, because deep-layer attention is diffuse (entropy 5.340), while the optimal layer (12) has the most focused signal (entropy 4.600), so signal quality should govern layer choice. The advantage also scales with model size, from $+1.3\%$ on 3B to $+5.3\%$ on 7B.

Limitations and Future Work. Our study has three conceptual limitations. First, the central claim rests on an assumption we support empirically but do not prove: that attention from query tokens recovers the visual evidence an answer depends on, so the retained set is not guaranteed to be sufficient for the downstream decision. Second, the temporal component of our score is a proxy rather than a direct constraint. Measuring how far a token departs from the mean of its own frame favors frame-distinctive tokens, which we observe to broaden coverage, but it does not enforce coverage across the timeline. Third, our evidence comes from two benchmarks and a single model family in video question answering, so whether the principle transfers to other modalities, architectures, or much longer videos remains open, and the efficiency gain is bounded by the full prefill pass that scoring requires. Future work could pursue theoretical grounding for query-aware selection, an explicit rather than proxy coverage objective, and pruning that avoids the separate scoring pass.

AI Use Statement

(This section is required and does not count toward the page limit.) We used the GLM model to polish the manuscript for readability and an AI coding agent to assist with debugging the experiment code; we reviewed all AI-assisted output and take responsibility for the final content, including text, claims, and artifacts.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197, 2020. arXiv:2005.00928.

Jean-Baptiste Alayrac, Jeff Donahue, Philipp Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. arXiv:2204.14198.

Jinze Bai, Shuai Bai, Yunfei Chu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenjie Ge, Sheng Song, Kai Zhang, 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 *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.09461.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.

Liang Chen, Yushuai Huang, Zichao Zhang, Yifan Li, Dongkyu Kim, Ian Cheung, Tao Yu, and Shijie Zhang. 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 (ECCV)*, pp. 19–35, 2024a. arXiv:2403.06764.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2404.16821*, 2024b.

Zhe Chen et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.11929.

Rotem Dror, Gili Baumer, Shira Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, 2018. arXiv:1804.06071.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Hamed Pirsiavash, Juergen Gall, and Sinisa Todorovic. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision (ECCV)*, pp. 396–414, 2022. arXiv:2111.15667.

Chaoyou Fu et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Yongming Huang, Junfeng Liu, Yitao Wu, et al. Vtimellm: Empower llm to grasp video moments. *arXiv preprint arXiv:2311.18445*, 2023.

Yuan Huang, Xiaokang Zhang, Jun Hao Liew, Xinyu Wang, Jian Yao, Wengang Zhou, and Houqiang Li. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 3543–3556, 2019. arXiv:1902.10186.

Bo Li, Yuan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, 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 (ICML)*, 2023a. arXiv:2301.12597.

Kun Li, Yinan Xu, Yi Wang, Yawei He, Yali Wang, Yu Qiao, Yu Liu, and Ziwei Liu. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Kunchang Li, Yali Wang, Yi Li, Yizhuo Wang, Yinan He, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024b. arXiv:2311.17005.

Senqiao Li, Tianheng Xu, Shuyin Hu, Haowen Wang, Yuyin Zhou, Zhuokun Wang, Bohan Zhang, Xinzhou He, Wanli Ouyang, Qibin Yu, et al. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024c.

Youwei Liang, Chongjian Wang, Yujun Cao, Dong Xu, Efstratios Gavves, and Jianfeng Feng. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.07800.

Bin Lin, Bin Zhu, Yunhao Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS*, 2023a. arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b. arXiv:2304.08485.

Yudong Liu, Li Zhang, Junpeng Wang, et al. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024a.

Yudong Liu et al. Dycoke: Dynamic compression of tokens for fast video large language models. *arXiv preprint arXiv:2411.15024*, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. arXiv:2103.14030.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 46212–46244, 2023. arXiv:2308.09126.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.02034.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867, 2025. arXiv:2403.15388.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.03762.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5797–5808, 2019. arXiv:1905.09418.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision (ECCV)*, pp. 58–76, 2024b. arXiv:2403.10517.

Haoning Wu et al. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781, 2021. arXiv:2105.08276.

Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. Vision-centric token compression in large language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.00791.

Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-ViT: Slow-fast token evolution for dynamic vision transformer. In *AAAI Conference on Artificial Intelligence*, pp. 2964–2972, 2022. arXiv:2108.01390.

An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Jiabo Ye, Haoyu Liu, Anwen Xu, et al. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Shaolei Zhang et al. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*, 2025.

Yuan Zhang, Renrui Zhang, Shuhao Gu, et al. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Table 5: Keep-ratio ablation on NExT-QA.

r	Full	FastV	QD-Prune	ΔCausal
0.1	64.4	47.8	63.3	+13.3
0.2	64.4	60.0	70.0	+10.0
0.3	64.4	58.9	63.3	+0.0
0.5	64.4	61.1	64.4	+16.7

Table 6: Efficiency analysis. The Tokens column reports vision tokens for the unpruned model and the total sequence length, that is retained vision tokens plus the text prompt, after pruning.

Method	**Tokens**	**FLOPs (G)**	**FLOPs ↓**	**Latency (s)**	**Mem (GB)**
Full	679	11780.2	N/A	0.364	15.73
FastV	206	7359.7	37.5%	0.358	16.51
VisionZip	206	7359.7	37.5%	0.358	16.51
PruneVid	206	7359.7	37.5%	0.364	16.51
SparseVLM	206	7359.7	37.5%	0.385	16.51
QD-Prune	206	7359.7	37.5%	0.371	16.51

A Additional Ablations

The additional studies support the main conclusions. Under aggressive pruning ($r = 0.1$), QD-Prune obtains 63.3% overall accuracy versus 47.8% for FastV, with a +13.3 point causal gap. Across pruning layers, layer 12 gives the largest causal gain (+10.0 points), whereas layer 20 reverses the advantage (-3.3 points). The combination-weight sweep is stable for $\alpha \in [0.4, 0.7]$ and loses temporal accuracy at $\alpha = 0.8$.

B Efficiency Analysis

Table 6 reports computational efficiency on Qwen2.5-VL-7B with L=12 and r=0.2. All pruning methods retain the same number of tokens and therefore have identical FLOPs. The prefill overhead limits end-to-end speedup, while FLOPs reduction becomes more useful for longer sequences or KV-cache compression.

C Additional Qualitative Analysis

On the 3B model, both methods collapse to majority-class predictions under aggressive pruning. This failure motivates model-size-aware keep ratios and bounds the scope of the main claims.

D Mechanistic Analysis

The layer-wise analysis explains why the entropy-minimized pruning layer is effective. Table 7 reports score agreement, attention entropy, and query entropy across layers. The optimal layer is the one with the most focused baseline signal, rather than the one with maximum divergence between methods.

Layer 20 has the lowest score correlation (0.315) and overlap (25.4%), but its last-token entropy is highest (5.340), indicating a diffuse and unreliable FastV signal. In contrast, layer 12 has the lowest entropy (4.600), allowing query-driven attention to correct baseline selections and yielding the strongest causal improvement. In low-overlap examples, FastV concentrates on early frames while QD-Prune retains tokens from later, query-relevant frames.

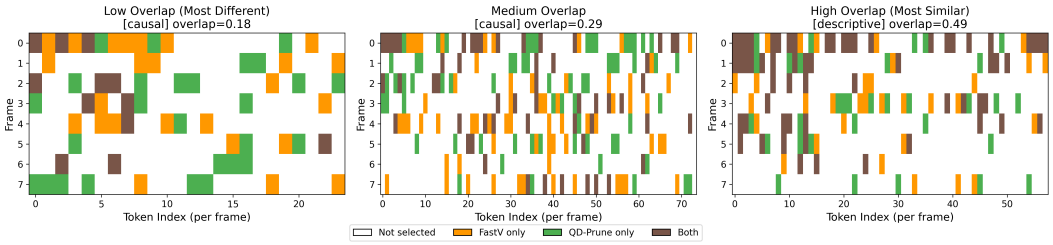

Figure 4: Case study of token selection at layer 12. QD-Prune covers later frames containing the queried action, whereas FastV concentrates on early frames.

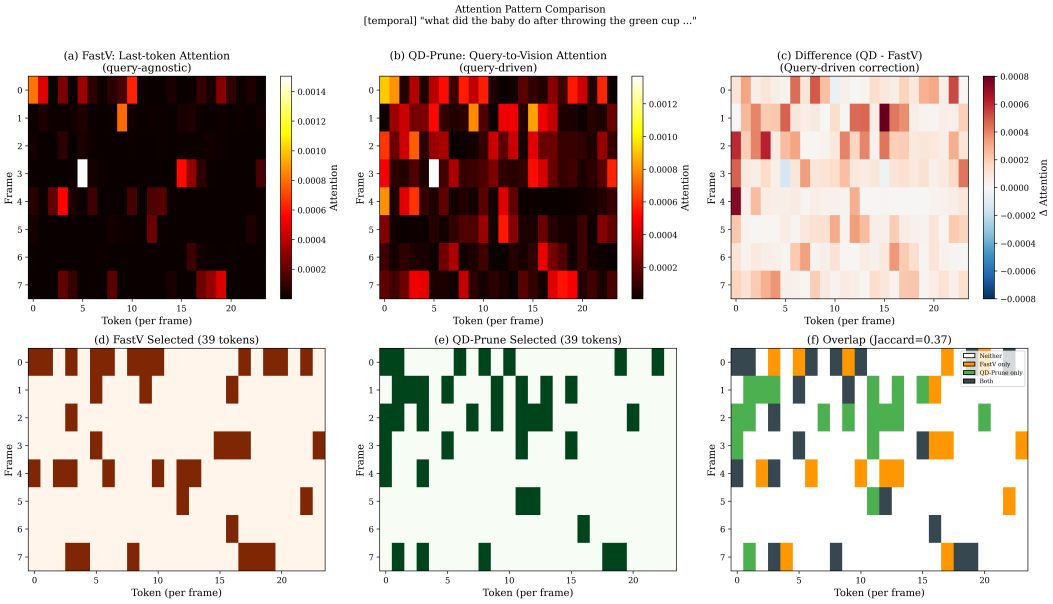

Figure 5: Attention and token-selection comparison between FastV and QD-Prune.

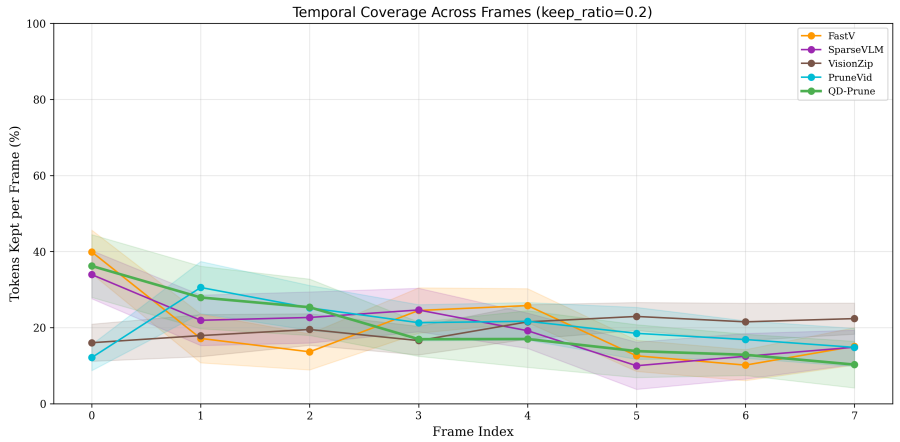

Figure 6: Temporal coverage across frames. QD-Prune retains a more even distribution than FastV.

Table 7: Layer-wise mechanism analysis (r=0.2). The optimal layer (12) has the lowest last-token attention entropy, whereas layer 20 has the largest method divergence but diffuse attention.

Layer	Overlap (%)	ScoreCorr	LastTokEntropy	Q-Entropy
4	30.6	0.762	4.939	5.442
8	31.0	0.586	4.628	5.681
12	**29.4**	**0.620**	**4.600**	**5.739**
16	28.5	0.606	4.832	5.758
20	**25.4**	**0.315**	**5.340**	5.741
24	38.9	0.653	5.258	5.674