AnchorPrune: Relevance-Anchored Contextual Expansion for Visual Token Pruning

Kyuan Oh and Bumsoo Kim†

Chung-Ang University, Seoul, Korea
{oka04108,bumsoo}@cau.ac.kr

Abstract. Large vision-language models incur substantial inference costs because high-resolution inputs introduce thousands of visual tokens, many of which are redundant for a given query. Existing pruning methods often combine query relevance and token diversity, yet these objectives can conflict under aggressive compression: relevance-driven selection may overconcentrate the budget on correlated local evidence, while diversity-driven selection may suppress indispensable tokens or retain distinct but uninformative regions. We introduce **AnchorPrune**, a training-free framework that first constructs a protected relevance anchor and then expands it with complementary visual context. AnchorPrune adaptively determines the anchor size from the novelty profile of relevance-ranked tokens, preserving a compact set of query-critical evidence, and allocates the remaining budget through importance-weighted novelty to recover informative, non-redundant context relative to the anchor. This ordered design prevents contextual expansion from displacing indispensable query cues while improving overall visual coverage. AnchorPrune is lightweight, architecture-aware, and requires neither retraining nor model modification. Across image and video vision-language models and benchmarks, it consistently improves the accuracy–efficiency trade-off over training-free baselines, particularly under severe compression. On LLaVA-NeXT-7B, AnchorPrune preserves 97.6% of full-token performance using only 160 of 2,880 visual tokens. These results establish relevance-anchored contextual expansion as an effective principle for efficient multimodal inference. Code is available at https://github.com/MULTI-cau/AnchorPrune.

Keywords: Visual Token Pruning · Vision-Language Models · Efficient Multimodal Inference · Relevance-Anchored Contextual Expansion

1 Introduction

Vision-language models (VLMs) achieve strong performance across image and video understanding tasks, but their long visual-token sequences incur substantial inference costs [2, 11, 12, 29]. High-resolution, multi-crop, and multi-frame inputs can introduce thousands of visual tokens, many of which are redundant for a given query yet must still be processed by the language model. This overhead is especially pronounced during prefilling, where visual tokens dominate

† : corresponding author

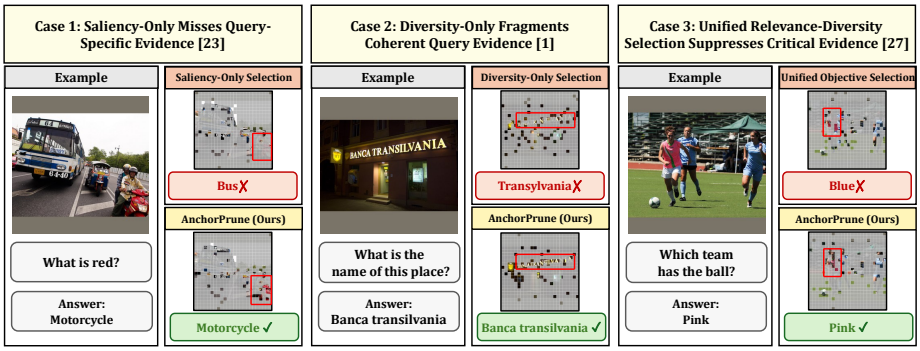

Fig. 1: Qualitative comparison on LLaVA-1.5-7B with 64 retained visual tokens. VisionZip [23] uses query-agnostic saliency, DivPrune [1] emphasizes query-agnostic diversity, and CDPruner [27] combines relevance and diversity within a unified objective. Under severe compression, these strategies may miss, fragment, or suppress query-critical evidence. AnchorPrune instead protects a relevance anchor before expanding it with informative, non-redundant context, yielding correct predictions across diverse tasks.

sequence length, memory consumption, and attention computation. Visual token pruning therefore provides a direct route to efficient multimodal inference by retaining only the evidence needed for accurate prediction [1, 3, 20, 22, 23, 27, 28].

The central challenge is not merely estimating token importance, but determining which evidence must be protected before the remaining budget is allocated. As illustrated in Fig. 1, saliency-based methods may retain redundant high-importance regions while overlooking query-specific cues [3, 20, 22, 23, 28]. Diversity-based methods distribute selections across the visual feature space, but can fragment locally coherent evidence or preserve visually distinct yet query-irrelevant regions [1]. Query-aware methods improve instruction alignment, but relevance-driven selection may still overconcentrate the budget on correlated local evidence. Recent approaches address this limitation by combining relevance and diversity within a unified objective [27]. Under aggressive compression, however, such coupling provides no explicit protection for indispensable evidence: diversity can displace query-critical tokens, while weak or diffuse relevance guidance cannot be compensated by a separate contextual expansion stage.

We argue that query-critical evidence and complementary context play asymmetric roles and should therefore be selected in a fixed order. Evidence directly required by the query, such as a small object, scene text, fine-grained attribute, or spatial relation, is largely non-substitutable: once removed, it cannot be recovered by subsequent selection. Contextual evidence, by contrast, is more substitutable, as multiple informative subsets may provide comparable support for grounding, disambiguation, or holistic reasoning. Effective pruning should therefore first establish a protected relevance anchor and only then expand it with complementary context. This expansion should prioritize tokens that are both

informative and non-redundant with respect to the retained anchor, rather than rewarding visual distinctiveness alone.

Based on this principle, we introduce **AnchorPrune**, a training-free framework for relevance-anchored contextual expansion. AnchorPrune first ranks visual tokens using architecture-appropriate query-conditioned priority scores and constructs a protected relevance anchor. Rather than assigning a fixed fraction of the token budget to relevance, it determines the anchor size adaptively from the novelty profile of relevance-ranked tokens, accommodating both concentrated and distributed query evidence. The remaining budget is then allocated through importance-weighted novelty, expanding the retained set with tokens that are globally informative and complementary to the current selection. By separating anchor construction from contextual expansion, AnchorPrune protects indispensable query evidence while allowing broader context to compensate when relevance guidance is incomplete or spatially diffuse.

AnchorPrune requires neither retraining nor architectural modification and applies to both CLIP-aligned and non-CLIP VLMs through architecture-appropriate representation spaces. Across image and video benchmarks, it consistently improves the accuracy–efficiency trade-off over representative training-free pruning methods, with the largest gains under aggressive compression. On LLaVA-NeXT-7B, AnchorPrune preserves 97.6% of full-token performance while retaining only 160 of 2,880 visual tokens. These results establish relevance-anchored contextual expansion as an effective principle for efficient multimodal inference.

Our contributions are threefold:

- We identify an ordering limitation in existing visual-token pruning methods: jointly optimizing relevance and diversity can both displace non-substitutable query evidence and leave weak or diffuse relevance guidance without a separate mechanism for contextual compensation.
- We introduce **AnchorPrune**, a training-free framework that constructs an adaptive protected relevance anchor and expands it through importance-weighted novelty to retain informative, non-redundant context.
- We validate AnchorPrune across CLIP-aligned and non-CLIP image and video VLMs, demonstrating consistent performance preservation and strong accuracy–efficiency trade-offs under severe visual-token compression.

2 Preliminaries

2.1 Visual Token Pruning in VLMs

Given a visual input I and a text query q, a vision encoder produces a sequence of visual tokens

$$X = x_i * i = 1^N, \qquad x_i \in \mathbb{R}^d. \tag{1}$$

Visual token pruning selects an index set $S \subseteq 1, \ldots, N$ with $|S| = K \ll N$ and passes the reduced sequence $X_S = x_i : i \in S$ to the language model. The objective is to reduce inference cost while approximately preserving the conditional

output distribution:

$$p * \theta(y \mid X, q) \approx p_\theta(y \mid X_S, q). \tag{2}$$

Because direct subset optimization is combinatorial, existing methods rely on surrogate criteria such as token importance, query relevance, and feature diversity.

2.2 Relevance, Diversity, and Contextual Expansion

Let ρ_i denote a query-conditioned priority score for token x_i. Relevance-driven selection prioritizes instruction-aligned evidence, but may concentrate the retained budget on correlated local tokens. Diversity-driven selection instead favors candidates that contribute information not already represented by the retained set. Given visual features h_i, the novelty of candidate token i relative to a selected set S can be measured as

$$\Delta(i; S) = \min_{j \in S} \left(1 - \frac{h_i^\top h_j}{|h_i|_2 |h_j|_2}\right). \tag{3}$$

Recent query-aware methods combine relevance and diversity within a unified subset-selection objective [27]. For example, let $\widetilde{r} \in \mathbb{R}^N$ denote normalized token-relevance scores. A visual-similarity kernel L can be conditioned on relevance as

$$\widetilde{L} = \mathrm{diag}(\widetilde{r}) L \, \mathrm{diag}(\widetilde{r}), \qquad S^\star = \arg\max_{|S|=K} \log\det(\widetilde{L}_S). \tag{4}$$

This formulation jointly promotes instruction relevance and subset diversity, but it provides no explicit mechanism for protecting indispensable evidence before diversity influences selection. Moreover, when relevance guidance is weak or spatially diffuse, the unified objective offers no separate stage for compensating through broader contextual selection.

We use *contextual expansion* to denote the addition of visual tokens that are both informative and complementary to a protected set of query-critical evidence. This notion is stricter than diversity alone: a visually distinct but uninformative region may increase feature dispersion without providing useful contextual support. In AnchorPrune, contextual expansion is therefore performed only after a relevance anchor has been established.

3 Method

3.1 Overview

AnchorPrune is a training-free visual token pruning framework that establishes a protected relevance anchor before expanding it with complementary visual context. Given visual tokens $X = x_i * i = 1^N$, a text query q, and a target token budget K, AnchorPrune constructs the retained set as

$$S = S * \mathrm{rel} \cup S_{\mathrm{ctx}}, \qquad |S_{\mathrm{rel}}| = K_{\mathrm{rel}}, \qquad |S_{\mathrm{ctx}}| = K_{\mathrm{ctx}}, \tag{5}$$

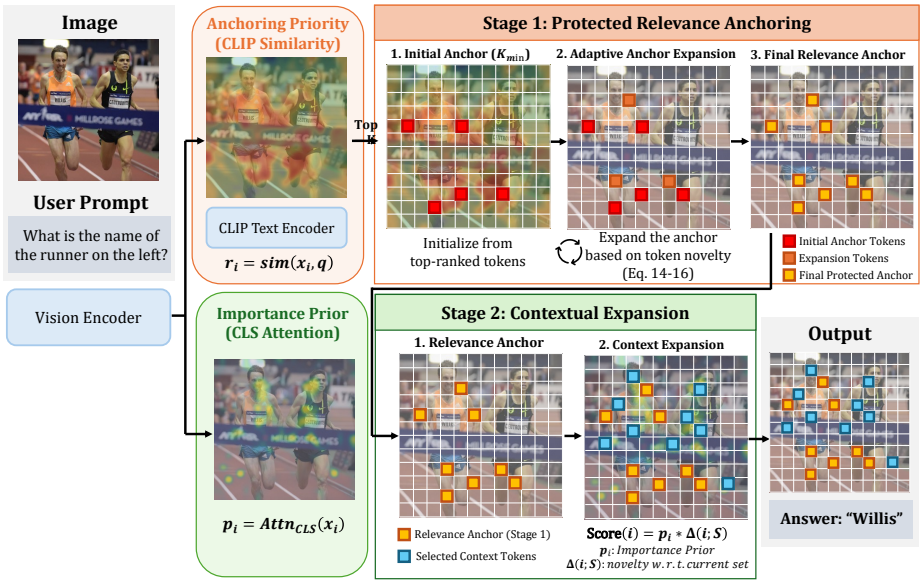

Fig. 2: Overview of AnchorPrune. Given an image and an instruction, Stage 1 ranks visual tokens using an architecture-specific query-conditioned priority signal and constructs a protected relevance anchor whose size is determined adaptively from the novelty profile of the ranked tokens. Stage 2 initializes the retained set with this anchor and allocates the remaining budget through importance-weighted novelty, selecting tokens that are both globally informative and complementary to the evidence already retained. The selected tokens are restored to their native visual order and passed to the unchanged language model. The illustrated example uses LLaVA-1.5-7B with 64 of 576 visual tokens retained.

where $K_{\text{rel}} + K_{\text{ctx}} = K$.

The two subsets are selected sequentially. Stage 1 constructs a protected relevance anchor S_{rel} from tokens prioritized by the input instruction. Stage 2 uses the remaining budget to expand this anchor with tokens that are globally informative and non-redundant with respect to the current retained set. Because S_{rel} is fixed before contextual expansion, subsequent selections cannot displace query-critical evidence.

3.2 Stage 1: Protected Relevance Anchoring

Stage 1 assigns each visual token an anchoring-priority score s_i, where a larger value indicates higher priority for inclusion in the protected relevance anchor. The score is computed in an architecture-appropriate representation space in which visual and textual features are comparable.

CLIP-aligned VLMs. For models equipped with a paired CLIP vision–language encoder [19], we compute the priority score before the multimodal projector. Let

$$v_i = \frac{\phi(x_i)}{|\phi(x_i)| * 2} \tag{6}$$

denote the normalized visual embedding of token i, and let $t_m * m = 1^M$ denote normalized text embeddings produced by the corresponding text encoder. If the instruction exceeds the encoder context length, it is divided into M segments. We define

$$s_i = -\frac{1}{M} \sum_{m=1}^{M} v_i^\top t_m. \tag{7}$$

Equation (7) defines the *negated patch–text similarity*. Prior analysis found that raw CLIP local similarities can respond more strongly to background than to discriminative foreground regions, and that reversing these maps can improve localization [9]. We therefore use negated similarity as a training-free anchoring-priority signal, without interpreting it as calibrated semantic relevance. The two similarity directions are compared in the supplementary material.

Non-CLIP VLMs. For architectures without a comparable pre-projector vision–language space, we compute the priority score after the multimodal projector. Let

$$z_i = g_{\text{mm}}(x_i) \tag{8}$$

denote the projected visual token, and let $e_m * m = 1^M$ denote the language-model embeddings of the instruction tokens. We define

$$s_i = \max * 1 \le m \le M \left\langle \frac{z_i}{|z_i|_2}, \frac{e_m}{|e_m|_2} \right\rangle. \tag{9}$$

Token-wise maximum matching preserves strong alignment with individual semantic components of the instruction, which can be obscured when all instruction embeddings are averaged into a single vector.

Relevance ranking. Using the architecture-specific score s_i, we sort the visual tokens in descending priority:

$$\pi = (\pi_1, \ldots, \pi_N), \qquad s_{\pi_1} \ge s_{\pi_2} \ge \cdots \ge s_{\pi_N}. \tag{10}$$

For an anchor budget K_{rel}, the protected relevance anchor is

$$S_{\text{rel}} = \pi_1, \ldots, \pi_{K_{\text{rel}}}. \tag{11}$$

No diversity penalty is applied in this stage. Consequently, correlated tokens may remain in the anchor when they jointly contain useful local evidence for answering the query.

3.3 Adaptive Anchor Budget

A fixed relevance ratio is unnecessarily rigid because the structure of query-critical evidence varies across image–query pairs. Some queries depend on a compact local region, while others require evidence distributed across multiple objects, attributes, or textual regions. AnchorPrune therefore determines K_{rel} from the novelty profile of the relevance-ranked sequence.

We begin with a minimum anchor size $K_{\min}$ and impose an upper bound

$$K_{\max} = \left\lfloor \frac{K}{2} \right\rfloor, \tag{12}$$

ensuring that at least half of the total budget remains available for contextual expansion.

Let

$$A_{\min} = \pi_1, \ldots, \pi_{K_{\min}} \tag{13}$$

denote the initial relevance anchor. Using visual features h_i, we measure the novelty of each subsequent relevance-ranked candidate π_t relative to $A_{\min}$:

$$\Delta(\pi_t; A_{\min}) = \min_{j \in A_{\min}} \left(1 - \frac{h_{\pi_t}^{\top} h_j}{|h_{\pi_t}| * 2 |h_j| * 2} \right). \tag{14}$$

For $K * \min < t \leq K * \max$, we count the number of sufficiently novel candidates encountered after the initial anchor:

$$c_t = \sum_{u=K_{\min}+1}^{t} \mathbb{1}\left[\Delta(\pi_u; A_{\min}) > \tau\right], \tag{15}$$

where τ is a novelty threshold. Given a patience parameter P, the anchor budget is selected as

$$K_{\text{rel}} = \begin{cases} t, & \text{if } t \text{ is the first index satisfying } c_t \geq P, \\ K_{\max}, & \text{otherwise.} \end{cases} \tag{16}$$

The remaining context budget is

$$K_{\text{ctx}} = K - K_{\text{rel}}. \tag{17}$$

When sufficiently novel candidates appear early in the relevance-ranked sequence, AnchorPrune expands the anchor through the P-th novelty event. If fewer than P such events are observed, it conservatively retains the maximum anchor budget $K_{\max}$.

3.4 Stage 2: Contextual Expansion

After constructing the protected relevance anchor, AnchorPrune allocates the remaining budget to contextual expansion. Unlike diversity-only sampling, Stage 2 does not favor novelty in isolation. Each candidate must be both globally informative and complementary to the evidence already retained.

Global importance prior. We assign each visual token a prior p_i estimating its contribution to the global visual representation. The prior is instantiated according to the vision-encoder architecture.

For encoders with a dedicated `[CLS]` token, let $A_{h,\mathrm{cls},i}$ denote the attention from the `[CLS]` token to visual token i at attention head h. We compute

$$p_i = \frac{1}{H} \sum_{h=1}^{H} A_{h,\mathrm{cls},i}. \tag{18}$$

For encoders without a dedicated global token, we use the average attention mass received by token i:

$$\widetilde{p} * i = \frac{1}{HN} \sum * h = 1^H \sum_{j=1}^{N} A_{h,j,i}. \tag{19}$$

If the architecture merges multiple encoder tokens into a single candidate token, let $G(i)$ denote the corresponding pre-merge token group. The aligned prior is

$$p_i = \frac{1}{|G(i)|} \sum_{j \in G(i)} \widetilde{p}_j. \tag{20}$$

When no merging is applied, $G(i) = i$.

Anchor-conditioned novelty. We initialize the retained set with the protected relevance anchor:

$$S \leftarrow S_{\mathrm{rel}}. \tag{21}$$

For candidate token $i \notin S$, its novelty relative to the current retained set is

$$\Delta(i; S) = \min_{j \in S} \left(1 - \frac{h_i^\top h_j}{|h_i|_2 |h_j| * 2} \right). \tag{22}$$

Because S is initialized with $S * \mathrm{rel}$, novelty is measured relative to both the protected anchor and the contextual tokens selected in preceding iterations.

Importance-weighted expansion. At each iteration, AnchorPrune selects

$$i^\star = \arg\max_{i \notin S} p_i, \Delta(i; S), \qquad S \leftarrow S \cup i^\star, \tag{23}$$

until $|S| = K$. The resulting context subset is

$$S_{\mathrm{ctx}} = S \setminus S_{\mathrm{rel}}. \tag{24}$$

The multiplicative objective in Eq. (23) suppresses both visually distinct but uninformative tokens and globally important but redundant ones. Each selected token must therefore be simultaneously informative and complementary to the current retained set, producing contextual expansion rather than feature dispersion alone. The selected tokens are restored to their native visual order and passed to the unchanged language model without retraining or architectural modification.

4 Experiments

We evaluate AnchorPrune in terms of performance preservation under aggressive visual-token compression, generalization across heterogeneous image and video VLM architectures, and the effectiveness of relevance-anchored contextual expansion. Our evaluation spans two CLIP-aligned image VLMs with substantially different visual-sequence lengths, a non-CLIP image VLM, and a video VLM. Detailed implementation settings, video breakdowns, and additional ablations are provided in the supplementary material.

4.1 Experimental Setup

Models and evaluation. We evaluate AnchorPrune on LLaVA-1.5-7B [11], LLaVA-NeXT-7B [12], Qwen2.5-VL-7B [2], and LLaVA-Video-7B [29].

For LLaVA-1.5-7B, we retain 128, 64, or 32 tokens from 576 visual tokens. For LLaVA-NeXT-7B, we retain 640, 320, or 160 tokens from 2,880 visual tokens. Both models are evaluated on VQAv2 [6], TextVQA [21], GQA [7], ScienceQA-IMG [15], MME [4], POPE [10], MMBench-EN/CN [13], and MM-Vet [24].

For Qwen2.5-VL-7B, we retain 256, 128, or 64 tokens from 1,296 visual tokens and evaluate on MME, TextVQA, DocVQA [18], AI2D [8], MMMU [25], and MMBench-EN/CN.

For LLaVA-Video-7B, we evaluate 16-frame inputs and retain 1,024 or 512 tokens from 2,704 visual tokens on Video-MME [5], EgoSchema [16], and TempCompass [14].

Results for all evaluated model–budget settings are reported in Tables 1, 2, 3, and 4. We additionally conduct a controlled Stage-2 ablation on LLaVA-1.5-7B.

Baselines and protocol. We compare against representative training-free methods based on attention or saliency, progressive token reduction, token merging, diversity, and unified relevance–diversity selection [1,3,20,22,23,27,28]. All methods use matched model checkpoints, input processing, decoding settings, benchmark protocols, and retained-token budgets. We use `lmms-eval` [26] whenever supported. Architecture-specific implementation details, hyperparameters, and hardware configurations are provided in the supplementary material.

Relative retained performance. To aggregate benchmarks with different metric scales, we report the average performance retained relative to the unpruned backbone:

$$\text{Rel.}(m) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\text{Score}_b(m)}{\text{Score}_b(\text{Full})} \times 100, \tag{25}$$

where $\mathcal{B}$ denotes the set of metrics reported in the corresponding table.

Table 1: Main results on LLaVA-1.5-7B across visual-token budgets. We evaluate AnchorPrune on a compact CLIP-aligned VLM containing 576 visual tokens. Best and second-best *Rel.* values within each budget are highlighted in **bold** and underlined, respectively.

Method	VQAv2	TextVQA	GQA	SQA^{IMG}	MME	POPE	MMB^{EN}	MMB^{CN}	MM-Vet	Rel.
Upper Bound: All 576 Tokens (100%)										
LLaVA-1.5-7B	76.7	58.3	61.9	69.6	1509.1	85.8	64.0	55.6	31.3	100.0
Retain 128 Tokens (↓ 77.8%)										
FastV (ECCV'24)	72.7	56.4	56.9	68.9	1456.1	76.2	62.5	53.6	28.2	94.7
PyramidDrop (CVPR'25)	73.4	55.9	56.5	69.2	1382.5	78.8	62.7	51.8	26.3	93.4
SparseVLM (ICML'25)	74.4	56.7	58.6	68.7	1435.6	85.0	63.5	54.4	30.4	97.3
PruMerge+ (ICCV'25)	72.3	54.7	57.7	69.5	1406.3	81.0	60.1	51.4	26.6	93.3
VisionZip (CVPR'25)	73.8	56.8	57.5	68.7	1435.1	83.1	62.4	53.5	32.7	97.3
DivPrune (CVPR'25)	74.2	56.4	59.2	69.0	1405.3	86.8	62.6	52.5	29.2	96.5
CDPruner (NeurIPS'25)	75.0	56.3	59.6	68.6	1413.3	87.1	62.4	51.9	28.1	96.1
AnchorPrune	75.1	57.0	59.8	68.9	1454.3	86.8	63.3	53.2	32.8	**98.7**
Retain 64 Tokens (↓ 88.9%)										
FastV (ECCV'24)	68.2	54.1	53.6	69.3	1357.8	68.1	60.1	50.3	26.4	89.5
PyramidDrop (CVPR'25)	69.0	53.0	52.8	68.1	1204.7	69.3	58.2	45.5	20.2	84.7
SparseVLM (ICML'25)	68.7	53.4	53.7	69.4	1318.7	77.5	59.9	48.9	23.6	89.1
PruMerge+ (ICCV'25)	69.7	54.1	55.3	69.5	1318.7	74.9	58.8	49.6	24.1	89.5
VisionZip (CVPR'25)	70.6	55.4	55.0	69.0	1373.1	77.1	60.0	51.1	30.0	93.0
DivPrune (CVPR'25)	72.5	54.8	57.8	68.3	1343.8	85.5	59.2	49.4	24.2	91.9
CDPruner (NeurIPS'25)	73.9	55.5	58.9	68.4	1400.6	87.3	61.2	49.8	26.1	94.2
AnchorPrune	73.8	56.1	58.6	68.8	1420.2	85.8	61.3	52.0	30.1	**96.2**
Retain 32 Tokens (↓ 94.4%)										
PruMerge+ (ICCV'25)	65.5	52.0	53.3	69.2	1227.0	69.6	56.4	45.2	22.1	84.7
VisionZip (CVPR'25)	65.7	53.2	51.8	68.8	1255.8	68.8	57.2	48.0	24.0	86.1
DivPrune (CVPR'25)	69.4	53.0	54.7	68.6	1267.9	81.9	57.9	45.8	24.3	88.7
CDPruner (NeurIPS'25)	72.1	52.9	57.2	69.1	1362.6	87.5	60.1	46.0	27.7	92.6
AnchorPrune	71.7	54.2	56.9	69.5	1394.7	85.3	60.1	49.9	28.7	**93.9**

4.2 Main Results

CLIP-aligned image VLM. Table 1 reports results on LLaVA-1.5-7B, whose visual representation contains 576 tokens. This compact setting provides a demanding test because the original sequence contains substantially less redundancy than high-resolution multi-crop representations.

AnchorPrune achieves the highest aggregate retention at all three budgets. It retains 98.7% and 96.2% of the full-token baseline with 128 and 64 tokens, respectively. Under the most aggressive setting, it retains 93.9% using only 32 visual tokens, exceeding the second-best method by 1.3 percentage points.

The gains are distributed across benchmarks rather than being driven by a single metric. At 32 tokens, AnchorPrune is particularly strong on MME, TextVQA, MMBench-CN, and MM-Vet, showing that relevance-anchored contextual expansion remains effective even when the original visual sequence is comparatively compact.

Scaling to high-resolution visual sequences. Table 2 evaluates AnchorPrune on LLaVA-NeXT-7B, whose high-resolution multi-crop representation

Table 2: Main results on LLaVA-NeXT-7B across visual-token budgets. We evaluate scalability to a high-resolution multi-crop representation containing 2,880 visual tokens. Best and second-best *Rel.* values within each budget are highlighted in **bold** and underlined, respectively.

Method	VQAv2	TextVQA	GQA	SQA^{IMG}	MME	POPE	MMB^{EN}	MMB^{CN}	MM-Vet	Rel.
Upper Bound: All 2,880 Tokens (100%)										
LLaVA-NeXT-7B	80.2	61.3	64.2	70.2	1528.8	86.4	67.2	56.5	39.3	100.0
Retain 640 Tokens (↓ 77.8%)										
FastV (ECCV'24)	77.2	57.6	60.5	67.9	1513.6	82.6	62.5	54.0	25.5	92.1
PyramidDrop (CVPR'25)	78.7	59.8	61.0	68.4	1470.3	84.6	65.3	57.3	26.6	94.3
SparseVLM (ICML'25)	77.1	57.8	59.4	67.8	1481.7	84.7	64.9	56.9	32.6	95.0
PruMerge+ (ICCV'25)	70.5	51.0	55.4	70.0	1295.0	67.1	62.8	55.2	31.9	88.0
VisionZip (CVPR'25)	77.3	60.1	61.3	68.2	1461.5	86.0	64.6	56.3	35.8	96.6
DivPrune (CVPR'25)	78.9	52.6	58.1	67.4	1314.6	79.6	62.8	53.8	28.8	90.1
CDPruner (NeurIPS'25)	78.2	58.8	62.5	68.2	1491.3	87.4	65.9	56.6	33.5	<u>96.7</u>
AnchorPrune	80.2	61.3	64.3	70.3	1528.8	86.4	67.1	56.5	37.0	**99.4**
Retain 320 Tokens (↓ 88.9%)										
FastV (ECCV'24)	71.0	54.1	55.7	67.0	1327.2	73.9	60.2	50.8	21.7	85.1
PyramidDrop (CVPR'25)	76.2	56.8	57.9	67.7	1451.6	77.8	64.2	54.3	29.3	91.7
SparseVLM (ICML'25)	73.4	55.5	57.4	67.6	1419.1	81.2	63.2	54.0	29.8	91.1
PruMerge+ (ICCV'25)	67.7	49.6	53.7	69.6	1221.2	61.2	61.6	52.9	25.4	83.2
VisionZip (CVPR'25)	74.4	58.9	58.9	67.6	1410.1	82.2	63.1	55.0	31.5	92.9
DivPrune (CVPR'25)	76.6	50.8	56.1	67.3	1277.6	74.4	60.4	50.4	26.7	86.5
CDPruner (NeurIPS'25)	76.8	57.3	61.3	67.2	1482.9	87.6	64.6	55.2	35.7	<u>95.9</u>
AnchorPrune	79.4	59.6	62.9	69.0	1507.5	87.2	65.3	56.9	37.2	**98.3**
Retain 160 Tokens (↓ 94.4%)										
PruMerge+ (ICCV'25)	63.7	47.7	51.8	70.2	1153.1	56.8	58.0	47.3	27.3	79.8
VisionZip (CVPR'25)	70.0	56.0	55.3	67.8	1326.3	74.8	58.8	51.8	26.7	86.9
DivPrune (CVPR'25)	73.9	48.1	53.2	67.3	1211.7	67.9	57.4	47.1	19.7	80.7
CDPruner (NeurIPS'25)	75.2	55.7	60.8	67.4	1430.8	86.7	64.2	54.0	29.9	<u>92.9</u>
AnchorPrune	78.6	60.1	62.6	68.7	1481.1	87.5	65.5	56.9	35.8	**97.6**

contains 2,880 visual tokens. This setting tests whether the same selection principle scales to substantially longer and more redundant visual sequences.

AnchorPrune achieves the highest retained performance at every evaluated budget. With 640 tokens, it retains 99.4% of the full-token baseline, outperforming the second-best method by 2.7 percentage points in *Rel.* At 320 tokens, AnchorPrune achieves 98.3%, compared with 95.9% for the strongest competing method.

The advantage becomes larger under the most aggressive setting. With only 160 of 2,880 tokens retained, AnchorPrune preserves 97.6% of the full-token baseline, exceeding the second-best result by 4.7 percentage points. It also remains close to the unpruned model on TextVQA, MME, POPE, and both MMBench variants. Together with the LLaVA-1.5 results, this demonstrates that relevance-anchored contextual expansion is effective across both compact and high-resolution CLIP-aligned visual sequences.

Generalization to a non-CLIP architecture. Table 3 evaluates AnchorPrune on Qwen2.5-VL-7B. Unlike the LLaVA models, Qwen2.5-VL does not

Table 3: Main results on Qwen2.5-VL-7B across visual-token budgets. We evaluate cross-architecture generalization on a non-CLIP VLM under matched inference settings. Best and second-best *Rel.* values within each budget are highlighted in **bold** and underlined, respectively.

Method	MME	TextVQA	DocVQA	AI2D	MMMU	MMB^{EN}	MMB^{CN}	Rel.
Upper Bound: All 1,296 Tokens (100%)								
Qwen2.5-VL-7B	1687.9	77.3	94.3	83.2	50.6	83.1	80.5	100.0
Retain 256 Tokens (↓ 80.2%)								
FastV (ECCV'24)	1638.9	74.9	60.0	78.3	49.3	79.8	77.3	91.6
DivPrune (CVPR'25)	1671.9	70.0	67.9	77.8	48.6	80.0	77.2	<u>91.9</u>
CDPruner (NeurIPS'25)	1621.9	63.2	52.7	78.0	47.7	78.0	77.0	87.3
AnchorPrune	1657.5	72.6	72.9	78.8	48.6	80.8	78.2	**93.5**
Retain 128 Tokens (↓ 90.1%)								
FastV (ECCV'24)	1554.8	70.7	40.0	72.9	47.1	76.3	72.1	84.0
DivPrune (CVPR'25)	1642.4	66.3	49.8	75.2	47.3	76.9	75.4	<u>86.6</u>
CDPruner (NeurIPS'25)	1572.1	57.7	36.3	76.9	45.9	77.9	74.7	82.3
AnchorPrune	1663.9	68.4	51.0	77.7	46.8	79.6	75.7	**88.1**
Retain 64 Tokens (↓ 95.1%)								
FastV (ECCV'24)	1381.4	63.7	27.7	68.4	45.4	67.5	64.7	75.3
DivPrune (CVPR'25)	1544.1	60.7	31.6	70.7	46.9	74.6	71.7	<u>80.0</u>
CDPruner (NeurIPS'25)	1431.7	51.6	22.8	73.6	44.7	73.9	72.9	76.0
AnchorPrune	1563.2	61.5	31.3	74.2	45.2	75.7	73.2	**80.8**

expose the same paired pre-projector CLIP representation. AnchorPrune therefore computes its Stage-1 priority in the post-projector embedding space while preserving the same protected-anchor design.

AnchorPrune achieves the highest *Rel.* at all three budgets. At 256 tokens, it retains 93.5% of the full-token baseline, improving over the second-best result by 1.6 percentage points. At 128 tokens, AnchorPrune obtains 88.1%, compared with 86.6% for the strongest baseline. Even when only 64 of 1,296 tokens remain, AnchorPrune achieves the best overall result at 80.8%.

The improvements are consistent across general, text-oriented, and document-oriented benchmarks, indicating that the anchoring principle transfers beyond CLIP-aligned architectures. These results demonstrate that AnchorPrune is not tied to a specific visual encoder or relevance representation.

Video Generalization. Table 4 evaluates AnchorPrune on LLaVA-Video-7B under 16-frame inputs. AnchorPrune achieves the highest aggregate retained performance at both budgets, preserving 98.3% and 94.1% of the full-token baseline with 1,024 and 512 tokens, respectively. This result indicates that relevance-anchored contextual expansion extends to temporally structured visual inputs.

Table 4: Video results on LLaVA-Video-7B. We report aggregate performance under 16-frame evaluation. Detailed short-, medium-, and long-video results are provided in the supplementary material. Best and second-best *Rel.* values within each budget are highlighted in **bold** and underlined, respectively.

Method	Video-MME w/o Subtitles	Video-MME w/ Subtitles	EgoSchema 500-subset	TempCompass Avg.	Rel.
Upper Bound: All 2,704 Visual Tokens					
LLaVA-Video-7B	59.96	61.33	55.4	67.1	100.0
Retain 1,024 Tokens (↓ 62.1%)					
FastV (ECCV'24)	59.26	60.59	52.8	65.6	<u>97.7</u>
DivPrune (CVPR'25)	57.00	59.15	53.2	64.9	96.1
CDPruner (NeurIPS'25)	57.85	59.81	52.6	65.6	96.8
AnchorPrune	58.19	60.89	55.0	65.4	**98.3**
Retain 512 Tokens (↓ 81.1%)					
FastV (ECCV'24)	56.37	57.63	50.2	61.1	92.4
DivPrune (CVPR'25)	55.89	58.19	52.0	61.3	<u>93.4</u>
CDPruner (NeurIPS'25)	54.22	57.00	51.4	60.6	91.7
AnchorPrune	56.26	58.07	53.0	61.8	**94.1**

Overall trend. Across compact and high-resolution CLIP-aligned image VLMs, a non-CLIP image VLM, and a video VLM, AnchorPrune consistently achieves the strongest retained performance. Its advantage generally becomes clearer as the token budget decreases, supporting the principle that query-critical evidence should be protected before complementary context is expanded.

4.3 Efficiency Analysis

Finally, we evaluate the practical efficiency of AnchorPrune, including token-selection overhead. To summarize the joint trade-off between prefill latency and resident memory in a single quantity, we define an *Efficiency Score* as

$$\text{Efficiency Score} = \text{Prefill Speedup} \times \frac{\text{Memory}_{\text{Full}}}{\text{Memory}_{\text{Method}}}, \tag{26}$$

where Prefill Speedup denotes the speedup relative to the unpruned model, $\text{Memory}_{\text{Full}}$ is the resident GPU memory usage of the full-token baseline, and $\text{Memory}_{\text{Method}}$ is the resident GPU memory usage of the pruned method.

Figure 3 plots this score against MME performance under a matched budget of 32 retained visual tokens. AnchorPrune achieves the highest MME score among the evaluated pruning methods while maintaining a competitive efficiency score of 5.24. VisionZip and DivPrune obtain slightly higher efficiency scores of 5.28 and 5.41, respectively, due to marginally lower resident memory, but their MME scores are substantially lower. CDPruner reaches a similar prefill speedup but has a lower efficiency score of 4.68 because of its larger resident

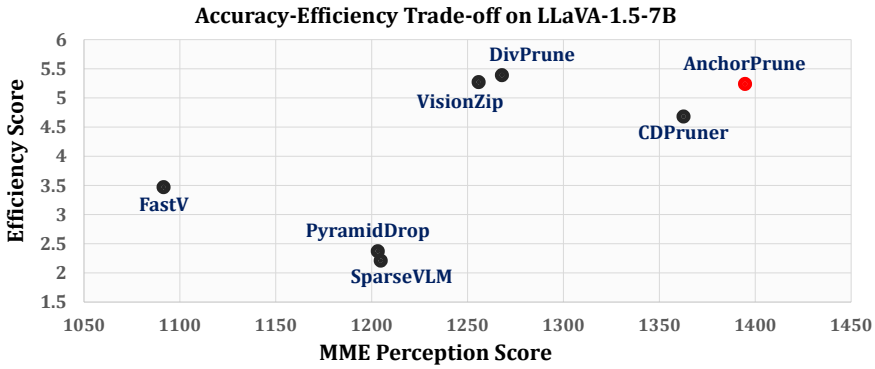

Fig. 3: Accuracy–efficiency trade-off on LLaVA-1.5-7B. The efficiency score jointly reflects prefill speedup and resident GPU memory relative to the full-token baseline. AnchorPrune occupies the favorable high-accuracy, high-efficiency region, combining the strongest task performance with competitive practical efficiency.

memory footprint. These results show that AnchorPrune provides the strongest accuracy–efficiency operating point rather than improving accuracy through a costly selection mechanism.

This result is consistent with its lightweight design: Stage 1 uses simple anchoring-priority scoring, while Stage 2 performs greedy importance-weighted expansion without requiring kernel-based subset optimization. Detailed FLOPs, latency, memory, and performance measurements are provided in the supplementary material.

4.4 Ablation Study

Stage-2 expansion strategy. We isolate the contribution of the Stage-2 selection criterion in Table 5. All variants share the same protected Stage-1 relevance anchor and differ only in the criterion used to allocate the remaining token budget. We compare diversity-only selection, additive relevance–diversity coupling, multiplicative relevance–diversity coupling, and the proposed importance-weighted contextual expansion rule. For this ablation, *Rel.* is computed over MME [4], ChartQA [17], DocVQA [18], TextVQA [21], and MMBench-CN [13], as reported in the table.

At 64 retained tokens, diversity-only selection obtains 85.4% *Rel.* The additive and multiplicative relevance–diversity formulations improve *Rel.* to 87.3% and 87.2%, respectively. In contrast, importance-weighted contextual expansion achieves 91.0%, outperforming the strongest alternative by 3.7 percentage points.

The same conclusion holds under stronger compression. At 32 retained tokens, the proposed rule achieves 85.1% *Rel.*, compared with 81.1% for the strongest alternative and 79.4% for diversity-only selection. The gains are especially pronounced on MME, TextVQA, and MMBench-CN.

Table 5: Controlled Stage-2 ablation on LLaVA-1.5-7B. All variants use the same protected relevance anchor and differ only in how the remaining budget is allocated. Best and second-best *Rel.* values within each budget are highlighted in **bold** and underlined, respectively.

Method	MME	ChartQA	DocVQA	TextVQA	MMB^{CN}	Rel.
Upper Bound: All 576 Tokens (100%)						
LLaVA-1.5-7B	1509.1	18.2	21.5	58.3	55.6	100.0
Retain 64 Tokens						
Diversity Only	1391.0	15.1	15.5	54.5	48.0	85.4
Additive Relevance–Diversity	1380.9	15.9	16.4	54.8	48.5	<u>87.3</u>
Multiplicative Relevance–Diversity	1379.9	15.9	16.6	54.5	48.2	87.2
Importance-Weighted Contextual Expansion **(Ours)**	1420.2	16.7	17.1	56.1	52.0	**91.0**
Retain 32 Tokens						
Diversity Only	1317.9	14.6	12.8	52.6	44.2	79.4
Additive Relevance–Diversity	1344.5	14.2	14.2	52.7	45.2	81.0
Multiplicative Relevance–Diversity	1329.4	14.4	14.4	52.7	45.0	<u>81.1</u>
Importance-Weighted Contextual Expansion **(Ours)**	1394.7	15.1	14.5	54.2	49.9	**85.1**

These results show that contextual expansion is not equivalent to diversity maximization. Diversity-only selection may allocate capacity to visually distinct but weakly informative regions, while direct relevance–diversity coupling permits strength in one criterion to compensate for weakness in the other. Starting from the same protected relevance anchor, importance-weighted contextual expansion selects tokens that are simultaneously informative and complementary to the evidence already retained.

5 Conclusion

We presented **AnchorPrune**, a training-free visual token pruning framework that constructs an adaptive protected relevance anchor before expanding it through importance-weighted novelty. This sequential design preserves query-critical evidence while adding informative, non-redundant context. Across CLIP-aligned and non-CLIP image and video VLMs, AnchorPrune consistently improves performance preservation under matched token budgets, with the largest gains under severe compression. On LLaVA-NeXT-7B, it preserves 97.6% of full-token performance using only 160 of 2,880 visual tokens. These results establish relevance-anchored contextual expansion as an effective principle for efficient multimodal inference.

Acknowledgements. This research was supported by the Electronics and Telecommunications Research Institute (ETRI) Grant funded by the Korean Government (Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems) under Grant 25ZB1200.

References

1. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9392–9401 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
3. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024)
4. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: Advances in Neural Information Processing Systems. vol. 38 (2025)
5. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
6. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
7. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
8. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
9. Li, Y., Wang, H., Duan, Y., Xu, H., Li, X.: Exploring visual interpretability for contrastive language-image pre-training. arXiv preprint arXiv:2209.07046 (2022)
10. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
11. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
12. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024)
13. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
14. Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: Tempcompass: Do video llms really understand videos? In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 8731–8772 (2024)
15. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in neural information processing systems **35**, 2507–2521 (2022)

16. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
17. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
18. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22857–22867 (2025)
21. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
22. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. arXiv preprint arXiv:2410.17247 (2024)
23. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19792–19802 (2025)
24. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
25. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
26. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)
27. Zhang, Q., Liu, M., Li, L., Lu, M., Zhang, Y., Pan, J., She, Q., Zhang, S.: Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. Advances in Neural Information Processing Systems **38**, 25438–25468 (2025)
28. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. arXiv preprint arXiv:2410.04417 (2024)
29. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)