RECASTPRUNE: RETHINKING TOKEN SELECTION AND INFORMATION RETENTION IN VISUAL TOKEN PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free visual token pruning reduces the number of visual tokens processed by multimodal large language models without updating model parameters. However, existing methods often rely on image-only importance signals that may be misaligned with the user query, greedily select tokens without revisiting earlier decisions, and permanently discard the information carried by unselected tokens. We propose RecastPrune, which aligns token importance with the user query, revises the initial token selection, and selectively fuses information from unselected tokens. RecastPrune first applies query-aware relevance calibration by combining CLS-to-patch attention with patch–query similarity from the frozen CLIP encoders. It then performs coverage-aware initialization followed by restricted local exchange. Finally, saliency-aware selective absorption fuses relevant, compatible information from unselected tokens into retained representatives. Across LLaVA-1.5 and LLaVA-NeXT, RecastPrune achieves the highest average accuracy relative to the full-token upper bound at every evaluated token budget.

Keywords. multimodal large language model, visual token pruning, training-free acceleration, token aggregation

1 INTRODUCTION

Multimodal large language models (MLLMs) encode images as patch-level tokens and process them jointly with textual instructions, allowing a single model to handle visual question answering, document understanding, and multimodal reasoning (Liu et al., 2023; 2024a; Gurari et al., 2018; Yu et al., 2023). This versatility comes at a substantial inference cost: LLaVA-1.5 contributes 576 visual tokens per image, while higher-resolution models may contribute thousands. Because every retained token participates in LLM prefill and occupies the key–value cache, redundant visual tokens directly increase computation, memory use, and response latency.

Visual token compression addresses this cost through pruning, merging, or hybrids of the two (Shao et al., 2025; Jin et al., 2025). Training-free pruning is particularly attractive because it can accelerate an existing MLLM without updating its parameters. Its central difficulty, however, is not simply to identify individually important patches. Under a tight budget, the retained set must be relevant to the current query, cover complementary visual evidence, and preserve useful information from discarded regions. Importance-only selection can concentrate on one salient object; diversity-oriented selection can retain distinctive but irrelevant background; and indiscriminate recovery can reintroduce noise or smooth fine details (Alvar et al., 2025; Wen et al., 2025b; Wu et al., 2026). Treating relevance, coverage, and recovery independently therefore leaves a coupled source of compression error. SCOPE (Deng et al., 2026) partially addresses this problem by combining CLS-derived saliency with the marginal gain of a set coverage objective. However, three limitations remain: its saliency is image-generic rather than query-aware, its greedy selections cannot be revised, and its unselected features are discarded without information recovery. These limitations compound under aggressive compression. An irrelevant but globally salient patch can consume a slot early and distort subsequent coverage gains; the resulting omission is then irreversible because the excluded feature is not preserved.

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.

We propose RecastPrune, a training-free framework that coordinates representative selection, set refinement, and information recovery with a shared query-aware relevance signal. Query-Aware Relevance Calibration combines CLS attention with patch–query similarity from the frozen CLIP encoders (Radford et al., 2021). Coverage-Aware Initialization constructs a representative set, Restricted Local Exchange revises weak representatives, and Saliency-Aware Selective Absorption (SASA) fuses only discarded tokens that are both query-relevant and compatible with a refined representative. All pretrained parameters remain frozen, and the output contains exactly the configured number of visual tokens.

Our main contributions are summarized as follows:

- We introduce Query-Aware Relevance Calibration, which fuses CLS attention with frozen CLIP patch–query similarity to derive the query-aware relevance vector $\mathbf{r}$ that is reused throughout the framework.
- We design a two-stage selection strategy: Coverage-Aware Initialization constructs a relevant and representative token set, while Restricted Local Exchange revisits weak greedy selections without changing the visual token budget.
- We propose Saliency-Aware Selective Absorption (SASA), which enriches retained representatives by selectively absorbing information from discarded tokens that are both query-relevant and compatible with those representatives.
- Experiments across LLaVA-1.5, LLaVA-NeXT, and Qwen2-VL demonstrate the effectiveness of RecastPrune under matched token budgets. With only 160 of the 2,880 LLaVA-NeXT tokens, it reaches 95.3% of the full-token upper-bound accuracy. Under the 150 token Qwen2-VL setting, RecastPrune again achieves the highest average accuracy. On text-rich benchmarks, its improvement over SCOPE reaches up to 2.4%.

2 Related Work

2.1 Visual Token Pruning

Training-free pruning retains a compact subset of visual tokens using signals already available at inference. FastV ranks tokens with early-layer language-model attention (Chen et al., 2024); SparseVLM uses relevant text tokens and layer-wise sparsity (Zhang et al., 2024); and VisionZip selects dominant tokens from vision-encoder attention (Yang et al., 2025). PyramidDrop and MustDrop distribute reduction across model stages, while SAINT and diagnostic studies emphasize similarity, duplication, and evaluation design (Xing et al., 2024; Liu et al., 2024b; Jeddi et al., 2025; Wen et al., 2025a). Instruction- and content-aware approaches include IVTP, LVPruning, FlashVLM, and ICCTP, while VisPruner relies on visual cues (Huang et al., 2024; Sun et al., 2025; Cai et al., 2025; Yu et al., 2026b; Zhang et al., 2025). AgilePruner, D^2Pruner, IDPruner, and ZOO-Prune adaptively combine importance with diversity or sensitivity (Baek et al., 2026; Zhang et al., 2026; Tan et al., 2026; Kim et al., 2026), while V^2Drop exploits token variation within the LLM (Chen et al., 2026). Beyond per-token importance, DivPrune formulates selection as max–min diversity (Alvar et al., 2025), DART retains tokens with low duplication to a set of pivots (Wen et al., 2025b), and SCOPE greedily combines CLS saliency with marginal set coverage (Deng et al., 2026). Our work starts from SCOPE's set-level objective but replaces image-generic saliency with query-aware relevance and revises the greedy set through local exchange.

2.2 Merging and Hybrid Compression

Token merging preserves discarded information through aggregation. ToMe introduced similarity-based merging for Vision Transformers (Bolya et al., 2022); PruMerge+ selects tokens using sparse CLS attention and supplements them by key-similarity clustering (Shang et al., 2025). Hybrid MLLM compressors couple selection with recovery: TokenCarve uses an information-preservation criterion for both pruning and merging (Tan et al., 2025), VLM-Pruner combines spatially constrained near-to-far selection with selective fusion of discarded features (Wu et al., 2026), and VisionTrim pairs global–local selection with text-guided complementary tokens (Yu et al., 2026a). Unlike indiscriminate or similarity-only aggregation, our method first constructs a query-aware representative set and then

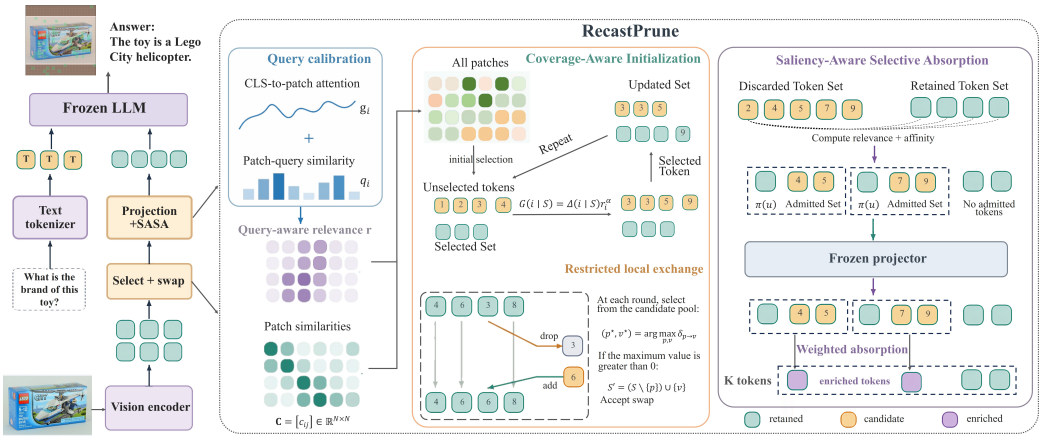

Figure 1: Overview of RecastPrune. Query-Aware Relevance Calibration guides Coverage-Aware Initialization, Restricted Local Exchange, and Saliency-Aware Selective Absorption while preserving a fixed K-token output.

selectively absorbs relevant discarded features into affinity-compatible representatives, increasing the information density of retained tokens.

3 METHOD

3.1 PROBLEM SETTING AND OVERVIEW

We consider a frozen MLLM comprising a vision encoder, a vision-to-language projector, and an LLM. Given an image I, the encoder produces

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d_v}, \tag{1}$$

where N is the number of patches and d_v is the feature dimension. For query tokens Q, frozen projector $P : \mathbb{R}^{d_v} \to \mathbb{R}^{d_\ell}$, and budget $K \ll N$, let $s_1 < \cdots < s_K$ denote the final representative indices. The compressor outputs

$$\mathcal{C}(\mathbf{X}, Q, K) = [\hat{\mathbf{z}}_{s_1}, \ldots, \hat{\mathbf{z}}_{s_K}] \in \mathbb{R}^{K \times d_\ell} \tag{2}$$

and thus replaces the original visual sequence with exactly K embeddings. Within this fixed budget, RecastPrune uses a shared query-aware relevance signal throughout the pruning process to preserve visual evidence relevant to the current query. Query-Aware Relevance Calibration first derives a shared relevance signal from visual saliency and patch–query alignment; Coverage-Aware Initialization and Restricted Local Exchange then construct and refine the representative set; finally, SASA absorbs relevant, compatible discarded features into the refined representatives. As summarized in Figure 1, all operations run at inference time with frozen parameters and no learned auxiliary scorer; the final embeddings retain the representatives' original spatial order.

3.2 QUERY-AWARE RELEVANCE CALIBRATION

CLS attention highlights visually prominent regions, but these regions are not always the ones needed to answer a query; small text, local attributes, or relational evidence may receive little attention. Query similarity provides a task-specific signal, but it can be unreliable for vague prompts and may overlook useful context. We therefore treat patch–instruction similarity as a calibration signal for the CLS-derived visual prior, rather than as a replacement for it.

We encode the complete text prompt as a pooled CLIP embedding and project each patch $\mathbf{x}_i$ into the same space.

Following the visual-saliency term in SCOPE's selection rule (Deng et al., 2026), we normalize the CLS-to-patch attention as $g_i = a_i/(\sum_j a_j + \epsilon_{\text{num}})$, where $\epsilon_{\text{num}} = 10^{-6}$ is a numerical stabilizer. The

resulting g_i measures the image-intrinsic saliency of patch i. Let q_i denote the softmax-normalized cosine similarity between the pooled prompt and patch i, providing its text-guided relevance. Query calibration combines the two signals into the query-aware relevance vector $\mathbf{r}$, whose i-th element is

$$r_i = (1 - \lambda_q)g_i + \lambda_q q_i, \qquad \lambda_q \in [0, 1]. \tag{3}$$

With $\lambda_q = 0$, RecastPrune retains the image-intrinsic visual saliency used by SCOPE; with $\lambda_q = 1$, relevance is determined entirely by text-guided patch relevance. Intermediate values preserve globally informative regions while reallocating part of the budget toward task-specific evidence.

3.3 Coverage-Aware Initialization

A relevance-only top-K rule can spend several slots on similar patches from the same region. We therefore adapt SCOPE's central idea: a patch should be selected only when it is relevant to the query and adds coverage beyond the current representatives.

We normalize the patch embeddings and compute pairwise cosine similarity:

$$c_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2 + \epsilon_{\text{num}}}. \tag{4}$$

For a selected index set S, we maintain the current coverage of each patch u as

$$m_u(S) = \max\left(0, \max_{s \in S} c_{us}\right), \qquad m_u(\varnothing) = 0. \tag{5}$$

Thus, $m_u(S)$ measures how well patch u is represented by its most similar member of S, while the outer clamp prevents negative similarity from counting as useful coverage. The marginal gain of a candidate $i \notin S$ is

$$\Delta(i \mid S) = \sum_{u=1}^{N} \left[\max(m_u(S), c_{ui}) - m_u(S)\right]. \tag{6}$$

This quantity measures the additional visual content covered by adding patch i. The greedy initializer combines it with query-aware relevance:

$$i^\star = \arg\max_{i \notin S} \Delta(i \mid S)\, r_i^\alpha. \tag{7}$$

Starting from $S = \varnothing$, Eq. (7) is repeated until $|S| = K$. The two factors favor candidates that both expand the represented visual content and carry high query-aware relevance, suppressing irrelevant but distinctive patches as well as relevant patches already covered by the selected set. The exponent $\alpha \geq 0$ controls the effect of relevance: $\alpha = 0$ yields coverage-only initialization, while larger values place more weight on r_i.

This retains the main benefit of SCOPE—jointly optimizing importance and diversity—while replacing image-only saliency with the calibrated score. However, the procedure remains forward greedy: as the set grows, a patch selected early may become redundant and cannot be reconsidered.

3.4 Restricted Local Exchange

Solving the full cardinality-constrained subset problem is expensive. Instead, we revisit the greedy set through one-swap updates, which can correct its membership while keeping exactly K representatives. Low-relevance selected tokens form the drop candidates, high-relevance unselected tokens form the add candidates, and coverage determines whether a replacement improves the set as a whole. The refinement evaluates sets with a relevance-regularized surrogate:

$$J(S) = F_{\text{raw}}(S) + \lambda_s \sum_{s \in S} r_s^\alpha, \qquad F_{\text{raw}}(S) = \sum_{u=1}^{N} \max_{s \in S} c_{us}. \tag{8}$$

The first term measures how well the selected set represents all visual patches, whereas the second rewards representatives with high query-aware relevance. The coefficient $\lambda_s \geq 0$ controls their balance. Consequently, a replacement is favored only when its coverage improvement is sufficient to

compensate for any loss of relevant visual evidence. We use J only to compare local replacements; because it differs from the greedy criterion in Eq. (7), an accepted swap is guaranteed to improve J, not the original initialization objective.

Rather than evaluating all $K(N-K)$ possible membership pairs, each round considers the $n_d = \lfloor \zeta K \rfloor$ lowest-relevance selected tokens as drop candidates and the $n_a = \mu K$ highest-relevance unselected tokens as add candidates. Here, $\zeta \in (0, 1]$ is the drop ratio and $\mu \in \mathbb{N}_+$ is the add-candidate multiplier. For a drop candidate $p \in S$ and an add candidate $v \notin S$, the surrogate gain

$$\delta_{p \to v} = \underbrace{F_{\text{raw}}((S \setminus \{p\}) \cup \{v\}) - F_{\text{raw}}(S)}_{\text{coverage change}} + \lambda_s \underbrace{\left(r_v^{\alpha} - r_p^{\alpha}\right)}_{\text{relevance change}}. \tag{9}$$

The two terms respectively measure how the replacement changes set-level coverage and total relevance. We evaluate all one-for-one replacements between the two candidate pools and apply only the swap with the largest positive gain. Thus, every accepted swap improves J without changing the number of selected tokens. The search stops after $R \in \mathbb{N}_0$ rounds or when none of the candidate swaps improves the score.

3.5 Saliency-Aware Selective Absorption

After Restricted Local Exchange, S contains the indices of the K refined representatives, and $D = \{1, \ldots, N\} \setminus S$ contains the discarded token indices. SASA recovers information from a limited number of these discarded tokens without increasing the sequence length. We first define an assignment map $\pi : D \to S$ that associates every discarded token $u \in D$ with its most similar representative. The corresponding maximum similarity $\bar{c}_u$ measures their feature affinity:

$$\pi(u) = \arg\max_{s \in S} c_{us}, \qquad \bar{c}_u = c_{u,\pi(u)} = \max_{s \in S} c_{us}. \tag{10}$$

Not every discarded token should be fused back: a useful candidate should be both compatible with its assigned representative and relevant to the query. We therefore min–max normalize the query-aware relevance as $\tilde{r}_u$ and define the absorption score

$$h_u = [\bar{c}_u]_+^{\eta} (\tilde{r}_u + \epsilon_{\text{num}})^{\gamma}. \tag{11}$$

Here, $[x]_+ = \max(x, 0)$ denotes the positive-part operator, and $\eta, \gamma \geq 0$ control the effects of feature affinity and query relevance, respectively. Among the discarded tokens, we admit at most $\lfloor \rho K \rfloor$ tokens with the highest h_u scores, where $\rho \geq 0$ controls the admission budget. Let $A \subseteq D$ denote this admitted set and $D_s = \{u \in A : \pi(u) = s\}$ the admitted tokens assigned to representative s. For each $u \in D_s$, its aggregation weight is

$$w_{u \to s} = \frac{\exp(c_{us}/\tau) h_u}{\sum_{v \in D_s} \exp(c_{vs}/\tau) h_v + \epsilon_{\text{num}}}. \tag{12}$$

Here, $\tau > 0$ controls the concentration of the affinity-based weights. After projecting the representative and admitted discarded features into the LLM embedding space through the frozen multimodal projector P, the final embedding of representative token s is

$$\hat{\mathbf{z}}_s = \begin{cases} \beta P(\mathbf{x}_s) + (1-\beta) \sum_{u \in D_s} w_{u \to s} P(\mathbf{x}_u), & D_s \neq \varnothing, \\ P(\mathbf{x}_s), & D_s = \varnothing. \end{cases} \tag{13}$$

The parameters control distinct parts of the procedure. The ratio ρ sets the number of discarded tokens that may be absorbed, while $\beta \in [0, 1]$ balances the original representative token against the absorbed information. The temperature τ controls how concentrated the aggregation weights are. The exponents η and γ determine the influence of token similarity and query relevance on admission, respectively. Since discarded tokens are merged into existing representatives, the output remains exactly K visual tokens. The representatives are finally restored to their original spatial order before being passed to the LLM.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models, token budgets, and benchmarks. We evaluate RecastPrune on LLaVA-1.5-7B (Liu et al., 2023) with 32, 64, or 128 of 576 visual tokens; LLaVA-NeXT-7B (Liu et al., 2024a) with 160, 320, or 640 of 2880 tokens; and Qwen2-VL-7B-Instruct with 150, 300, or 450 of 1350 tokens. The paired LLaVA budgets correspond to matched reductions of 94.4%, 88.9%, and 77.8%. The benchmarks are GQA (Hudson & Manning, 2019), MMBench (Liu et al., 2024c), MME (Fu et al., 2026), POPE (Li et al., 2023), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and OCRBench (Liu et al., 2024d).

Baselines and protocol. We compare with the training-free compression methods listed in Tables 1 and 2, including SCOPE (Deng et al., 2026), the direct basis of our method. Within each model, all locally reproduced methods use the same checkpoint, preprocessing, prompts, decoding settings, and evaluator. Comparisons are matched by the number of visual positions entering the LLM; published results obtained under different environments are treated as reference values rather than controlled measurements. We report each benchmark's official metric and average retention, computed by normalizing every score by its full-token counterpart and then averaging across the available benchmarks. Deterministic decoding uses temperature zero, while randomized baselines use at least three seeds.

Implementation details. Unless otherwise stated, we set $\lambda_q = 0.5$, $\alpha = 1$, $R = 2$, $\lambda_s = 0.2$, $\zeta = 0.2$, $\mu = 2$, $\rho = 1$, $\beta = 0.5$, $\tau = 0.1$, and $\eta = \gamma = 1$. The efficiency evaluation uses one RTX 4090 and 100 samples; software versions, numerical precision, and batch size remain to be documented.

4.2 MAIN RESULTS

4.2.1 RESULTS ON LLAVA-1.5-7B

As shown in Table 1, RecastPrune achieves the highest average accuracy relative to the full-token upper bound at all three token budgets, reaching 93.6%, 96.1%, and 98.2% with 32, 64, and 128 tokens, respectively. Under the extreme 32-token setting, it retains only 5.6% of the visual sequence yet leads on GQA, MMB, MME, and TextVQA, exceeding the strongest competing average by 0.4 percentage points. At 64 tokens, RecastPrune recovers the full-model POPE score and gives the best TextVQA result. At 128 tokens, the adapted variant improves the strongest non-RecastPrune results by 1.1 points on MMB and 21 points on MME. These results show that the method preserves task-relevant and fine-grained evidence even under severe compression.

4.2.2 RESULTS ON LLAVA-NEXT-7B

We next test whether the improvement transfers to the denser 2880-token representation of LLaVA-NeXT-7B.

Table 2 shows the same aggregate trend on LLaVA-NeXT: RecastPrune improves average retention over SCOPE by 0.2–0.4 percentage points at every budget. With only 160 of 2880 tokens, it retains 95.3% of full-token performance and leads on GQA, MME, and TextVQA. The gains remain benchmark dependent—VisionZip or SCOPE is stronger on several individual metrics, but the consistent average improvement indicates that relevance-guided refinement and selective recovery also transfer to longer visual sequences.

4.2.3 RESULTS ON QWEN2-VL-7B-INSTRUCT

We further evaluate whether the method transfers beyond the LLaVA family to Qwen2-VL-7B-Instruct, whose evaluated input configuration contains 1350 visual tokens.

Table 3 reveals a budget-dependent transfer pattern. At 450 and 300 tokens, RecastPrune remains competitive but trails VLM-Pruner in average retention by 0.3 and 0.1 percentage points, respectively; its strongest result in these regimes is POPE, where it exceeds the full-token baseline at both budgets. Under the tightest 150-token budget, however, RecastPrune retains 95.8% of baseline performance,

Table 1: Performance on LLaVA-1.5-7B under different visual token budgets. Avg. is the mean benchmark-wise retention relative to the 576-token baseline.

Method	GQA	MMB	MME	SQA	POPE	TextVQA	Avg.
Upper bound: 576 tokens (100%)							
Baseline	61.9	64.7	1862	69.5	85.9	58.2	100.0%
Retain 32 tokens (↓94.4%)							
PruMerge+ (ICCV'25)	52.9	55.1	1550	67.9	66.7	49.2	85.6%
VisionZip (CVPR'25)	51.8	57.0	1579	69.1	69.4	53.1	88.0%
VisPruner (ICCV'25)	52.2	58.4	1538	69.2	72.7	53.9	89.0%
DivPrune (CVPR'25)	54.9	57.6	1594	68.6	**81.5**	52.9	91.3%
ICCTP (AAAI'26)	52.9	58.0	–	**69.4**	80.4	53.3	92.0%
SCOPE (NeurIPS'25)	56.1	60.7	1645	69.1	80.1	54.4	93.2%
AgilePruner (ICLR'26)	54.1	60.4	1603	69.0	80.1	54.5	92.2%
RecastPrune	**56.3**	**61.0**	**1651**	69.3	80.5	**54.7**	**93.6%**
Retain 64 tokens (↓88.9%)							
PruMerge+ (ICCV'25)	55.4	59.6	1640	**69.5**	75.7	52.0	91.2%
VisionZip (CVPR'25)	55.1	60.1	1690	68.7	77.0	55.5	92.8%
VisPruner (ICCV'25)	55.4	61.3	1650	69.1	80.4	55.8	93.6%
DivPrune (CVPR'25)	57.5	60.1	1615	68.0	85.5	54.5	93.9%
SCOPE (NeurIPS'25)	58.3	**61.7**	1698	68.6	83.9	56.6	95.7%
VLM-Pruner (CVPR'26)	**59.1**	61.3	**1752**	68.37	82.2	56.05	95.8%
ICCTP (AAAI'26)	55.6	60.3	–	68.9	82.8	55.5	94.8%
AgilePruner (ICLR'26)	57.4	60.7	1703	68.6	84.1	56.0	95.1%
ZOO-Prune (CVPR'26)	58.5	60.2	1675	68.3	85.8	55.4	95.1%
V^2Drop (CVPR'26)	50.5	55.2	1470	68.9	75.1	51.8	86.9%
RecastPrune	58.3	61.2	1707	68.7	**85.9**	**56.7**	**96.1%**
Retain 128 tokens (↓77.8%)							
PruMerge+ (ICCV'25)	58.2	61.8	1554	69.1	83.1	54.0	93.7%
VisionZip (CVPR'25)	57.6	62.1	1763	68.7	83.3	56.9	96.2%
VisPruner (ICCV'25)	58.2	62.7	1768	69.0	84.6	57.0	96.9%
DivPrune (CVPR'25)	59.4	60.1	1707	68.6	85.5	54.5	95.4%
SCOPE (NeurIPS'25)	59.7	62.5	1776	68.4	86.1	57.2	97.6%
VLM-Pruner (CVPR'26)	61.0	62.5	1768	69.31	85.4	57.32	98.0%
ICCTP (AAAI'26)	57.8	61.6	–	69.0	85.2	56.8	96.9%
AgilePruner (ICLR'26)	59.4	61.8	1748	68.6	**87.4**	57.0	97.3%
ZOO-Prune (CVPR'26)	59.5	61.9	1752	68.9	87.1	**57.8**	97.6%
V^2Drop (CVPR'26)	56.3	61.8	1712	68.8	80.9	53.8	94.0%
RecastPrune	60.0	63.5	1780	**69.4**	86.7	57.0	98.2%
RecastPrune†	**61.2**	**63.8**	**1797**	**69.4**	86.0	**57.8**	**98.9%**

ICCTP does not report MME; its Avg. is computed over the remaining five benchmarks. † denotes the adapted-parameter variant.

outperforming VLM-Pruner by 0.8 percentage points in average retention and achieving the best result on five of the six benchmarks. The exception is TextVQA, where VLM-Pruner remains 0.85 points higher. This shift suggests that Query-Aware Relevance Calibration, Coverage-Aware Initialization, Restricted Local Exchange, and SASA become most useful when the representative set is severely capacity constrained, while milder compression leaves less room for these mechanisms to improve over a strong spatially regularized baseline.

4.2.4 Text-Rich Visual Understanding

We further evaluate ChartQA, DocVQA, and OCRBench on LLaVA-1.5-7B, where small text regions and document layout make query-aware evidence preservation especially important.

Table 4 shows that RecastPrune gives the highest average retention at every budget, with the largest margin over SCOPE (2.4 percentage points) at 64 tokens. The gain is most pronounced on DocVQA, consistent with query calibration preserving small task-relevant text that image-generic saliency may overlook. The smaller margin at 32 tokens suggests that representation capacity becomes the dominant constraint under extreme compression.

Table 2: Performance on LLaVA-NeXT-7B under different visual token budgets. Avg. is the mean benchmark-wise retention relative to the 2880-token baseline. Bold indicates the best compressed result within each budget.

Method	GQA	MMB	MME	SQA	MMMU	TextVQA	Avg.
Upper bound: 2880 tokens (100%)							
Baseline	64.2	67.9	1842	70.2	35.1	61.3	100.0%
Retain 160 tokens (↓94.4%)							
SparseVLM (ICML'25)	51.2	63.1	1542	67.5	32.8	46.4	86.9%
VisionZip (CVPR'25)	55.5	60.1	1630	**68.3**	**36.1**	56.2	92.5%
SCOPE (NeurIPS'25)	60.0	**64.3**	1700	67.4	35.6	56.8	95.1%
RecastPrune	**60.3**	63.8	**1701**	**68.3**	35.6	**57.0**	**95.3%**
Retain 320 tokens (↓88.9%)							
SparseVLM (ICML'25)	57.7	64.3	1694	67.3	34.4	55.9	93.6%
VisionZip (CVPR'25)	59.3	63.1	1702	67.3	35.3	**58.9**	95.0%
SCOPE (NeurIPS'25)	61.0	**65.9**	1789	**67.7**	35.6	58.4	97.1%
RecastPrune	**61.1**	65.6	**1808**	67.6	**36.2**	58.8	**97.5%**
Retain 640 tokens (↓77.8%)							
SparseVLM (ICML'25)	60.3	65.7	1772	67.7	34.6	57.8	96.0%
VisionZip (CVPR'25)	61.3	**66.3**	1787	**68.1**	34.7	60.2	97.4%
SCOPE (NeurIPS'25)	61.9	66.2	**1842**	67.8	**36.9**	60.1	98.9%
RecastPrune	**62.0**	65.9	1835	**68.1**	**36.9**	**61.0**	**99.1%**

Table 3: Performance on Qwen2-VL-7B-Instruct under different visual token budgets. Avg. is the mean benchmark-wise retention relative to the 1350-token baseline. Bold indicates the best compressed result within each budget.

Method	GQA	MMB	MME	SQA	POPE	TextVQA	Avg.
Upper bound: 1350 tokens (100%)							
Baseline	61.64	81.6	2321	85.42	88.5	82.73	100.0%
Retain 150 tokens (↓88.9%)							
FastV (ECCV'24)	55.65	73.5	2174	78.53	81.6	72.86	91.0%
DART (EMNLP'25)	55.86	73.0	2087	78.14	81.6	65.10	88.7%
DivPrune (CVPR'25)	59.17	74.8	2059	80.22	86.5	71.65	92.4%
VLM-Pruner (CVPR'26)	60.49	76.1	2158	80.81	87.4	**76.15**	95.0%
RecastPrune	**61.00**	**76.8**	**2194**	**81.61**	**88.8**	75.30	**95.8%**
Retain 300 tokens (↓77.8%)							
FastV (ECCV'24)	58.9	77.6	**2290**	82.35	86.4	79.43	96.6%
DART (EMNLP'25)	57.64	77.0	2181	80.47	86.0	74.10	93.8%
DivPrune (CVPR'25)	60.59	77.8	2202	82.10	87.8	77.74	96.3%
VLM-Pruner (CVPR'26)	**61.49**	78.1	2266	**83.49**	88.2	**79.51**	**97.8%**
RecastPrune	61.33	**78.2**	2278	82.85	**88.6**	78.90	97.7%
Retain 450 tokens (↓66.7%)							
FastV (ECCV'24)	60.1	**79.6**	**2319**	82.94	87.2	**81.28**	98.1%
DART (EMNLP'25)	59.55	78.4	2259	82.25	86.9	78.06	96.5%
DivPrune (CVPR'25)	61.11	79.4	2199	83.39	88.0	80.07	97.5%
VLM-Pruner (CVPR'26)	**61.54**	79.5	2273	**84.28**	87.9	81.01	**98.5%**
RecastPrune	61.45	78.4	2263	83.50	**88.8**	80.70	98.2%

5 Analysis

5.1 Component Ablation

Table 5 isolates the three modules on OCRBench at 64 tokens. Query calibration increases NES from 0.1667 to 0.1800, indicating improved partial text fidelity. Exchange is accepted on 88% of samples and raises coverage from 0.89 to 0.90, although its exact-match score decreases by one point. Finally, SASA absorbs 72% of discarded saliency with a mean assignment similarity of 0.88, giving the full method the highest score of 297. These results support complementary mechanism-level effects rather than monotonic standalone accuracy gains.

Table 4: Text-rich visual understanding on LLaVA-1.5-7B under different visual token budgets.

Method	ChartQA	DocVQA	OCRBench	Avg.
Upper bound: 576 tokens (100%)				
Baseline	18.2	28.0	31.1	100.0%
Retain 128 tokens (↓77.8%)				
VisionZip	17.1	25.1	30.0	93.4%
SCOPE	17.3	**25.9**	30.7	95.4%
RecastPrune	**17.6**	25.7	**31.0**	**96.1%**
Retain 64 tokens (↓88.9%)				
VisionZip	16.0	21.1	28.2	84.6%
SCOPE	16.7	23.2	29.5	89.8%
RecastPrune	**16.8**	**24.9**	**29.7**	**92.2%**
Retain 32 tokens (↓94.4%)				
VisionZip	15.0	18.1	26.3	77.2%
SCOPE	15.1	20.0	27.1	80.5%
RecastPrune	**15.3**	**20.2**	**27.5**	**81.5%**

Table 5: Component ablation on OCRBench with 64 visual tokens. OCR is the official score out of 1000; NES denotes normalized edit similarity.

Config.	Cal.	Exch.	SASA	OCR ↑	NES ↑
SCOPE	–	–	–	295	0.1667
+Calibration	✓	–	–	296	**0.1800**
+Cal.+Exchange	✓	✓	–	295	0.1790
RecastPrune	✓	✓	✓	**297**	0.1750

Mechanism diagnostics

Module	Diagnostic	Result
Calibration	ΔNES	+0.0133
Exchange	Acceptance rate	88.0%
Exchange	ΔCoverage	+0.010
SASA	Absorbed saliency mass	0.720
SASA	Assignment similarity	0.880
SASA	W→C / C→W	7 / 4

Table 6: Efficiency on LLaVA-1.5-7B using one RTX 4090 and 100 samples. Latencies are mean ± standard deviation; TTFT is measured from visual encoding to the first token.

Method	Tok.	Comp. (ms)	Prefill (ms)	TTFT (ms)	Peak (GB)
Full model	576	0.0	66.6 ± 0.2	86.7 ± 0.3	14.22
SCOPE	64	10.5 ± 1.0	23.4 ± 1.1	54.2 ± 1.7	14.35
RecastPrune	64	33.8 ± 3.2	23.8 ± 1.5	75.7 ± 3.0	14.82

5.2 Efficiency

Table 6 reports the number of visual tokens entering the LLM (Tok.), compression overhead (Comp.), LLM prefill latency, time to first token (TTFT), and peak GPU tensor memory (Peak). Comp. covers Query-Aware Relevance Calibration, Coverage-Aware Initialization, Restricted Local Exchange, and SASA when applicable; it is 0 for the full model because no compression stage is executed. Prefill measures LLM processing after multimodal input assembly, whereas TTFT spans visual encoding through selection of the first output token. Peak is measured over the same TTFT interval.

At 64 tokens, SCOPE and RecastPrune accelerate LLM prefill by $2.85\times$ and $2.80\times$, respectively. The additional refinement increases RecastPrune's compression overhead to 33.8 ms, yielding a net TTFT speedup of $1.15\times$. Thus, RecastPrune preserves the downstream benefit of token reduction, although its preprocessing overhead is only partially amortized at this scale.

Coverage-Aware Initialization and SASA are the main overheads (13.6 and 9.3 ms). Peak memory rises from 14.22 to 14.82 GB due to the similarity workspace and CLIP text branch; we therefore claim lower prefill latency rather than lower total memory.

6 Conclusion

RecastPrune treats training-free visual token compression as a coordinated selection-and-recovery problem. Query-Aware Relevance Calibration produces query-aware relevance from CLS saliency and frozen CLIP patch–query similarity; Coverage-Aware Initialization and Restricted Local Exchange refine the representative set; and SASA selectively absorbs relevant discarded features without increasing the token budget. Across LLaVA-1.5-7B and LLaVA-NeXT-7B, RecastPrune achieves the best average retention at every evaluated budget, including 95.3% with 160 of 2880 LLaVA-NeXT tokens. On Qwen2-VL-7B-Instruct, it performs best under the tightest 150-token budget while remaining competitive at 300 and 450 tokens. Gains on text-rich benchmarks further support query-aware compression. Remaining challenges include quadratic patch similarity, prompt-dependent recomputation, and possible over-smoothing. Future work will study subquadratic similarity, adaptive budgets, and broader model families.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the proposed algorithm and translating portions of the manuscript. Additionally, we used generative AI tools to diagnose and fix software bugs and to edit the manuscript for clarity, grammar, and consistency. All AI-assisted code was reviewed and tested by the authors for correctness and consistency with the described method. All AI-assisted text was checked for technical accuracy and consistency with the implementation, reported experiments, and cited sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Kaitong Cai, Jusheng Zhang, Jing Yang, Yijia Fan, Pengtao Xie, Jian Wang, and Keze Wang. Flashvlm: Text-guided visual token selection for large multimodal models. *arXiv preprint arXiv:2512.20561*, 2025.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3489–3499, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. Scope: Saliency-coverage oriented token pruning for efficient multimodel llms. *Advances in Neural Information Processing Systems*, 38: 161527–161552, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European conference on computer vision*, pp. 214–230. Springer, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Ahmadreza Jeddi, Negin Baghbanzadeh, Elham Dolatabadi, and Babak Taati. Similarity-aware token pruning: Your vlm but faster. *arXiv preprint arXiv:2503.11549*, 2025.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. Zoo-prune: training-free token pruning via zeroth-order gradient estimation in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39572–39582, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 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.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. 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.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *arXiv preprint arXiv:2507.20198*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Theresa Batista-Navarro. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4299–4308, 2025.

Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025.

Yifan Tan, Yifu Sun, Shirui Huang, Hong Liu, Guanghua Yu, Jianchen Zhu, and Yangdong Deng. Idpruner: Harmonizing importance and diversity in visual token pruning for mllms. *arXiv preprint arXiv:2602.13315*, 2026.

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

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025b.

Zhenkai Wu, Xiaowen Ma, Zhenliang Ni, Dengming Zhang, Han Shu, Xin Jiang, and Xinghao Chen. Vlm-pruner: Buffering for spatial sparsity in an efficient vlm centrifugal token pruning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31952–31961, 2026.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026a.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Yunqian Yu, Biao Chen, Yunya Zhang, Tonglan Xie, Mengmeng Jing, and Lin Zuo. Instruction-guided cross-modal clustering for training-free visual token pruning in vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12213–12221, 2026b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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.

Evelyn Zhang, Fufu Yu, Aoqi Wu, Zichen Wen, Ke Yan, Shouhong Ding, Biqing Qi, and Linfeng Zhang. D^2pruner: Debiased importance and structural diversity for mllm token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12412–12420, 2026.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025.

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, 2024.