GROUNDPRUNE: EFFICIENT VISUAL TOKEN PRUNING WITHOUT LOSING FINE-GRAINED VISUAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token pruning reduces the inference cost of vision-language models (VLMs), and existing methods are evaluated by whether downstream answer accuracy is preserved. We argue this criterion is insufficient: a pruned model can still answer correctly while its internal representation has already diverged from that of the original model. To measure this directly, we compare pruned and original models at two levels: the KL divergence between their next-token distributions on general VQA, and the IoU between the boxes that a linear probe, trained on the original model, regresses from the last prompt token's hidden state of the pruned and of the original model. Across FastV, SparseVLM, VisionZip, and Nüwa, methods that retain most VQA accuracy still incur substantial distribution shift and degraded box decodability, an erosion masked by VQA accuracy but amplified on referring expression comprehension (REC). We trace it to two causes: saliency mode collapse at the vision stage, where retained tokens cluster around a few dominant hotspots, and the choice of routing point in the decoder, where only a few attention heads can guide pruning without disturbing the original model. We propose GroundPrune, a training-free cascaded framework that addresses each in turn: Visual Evidence Preservation (VEP) selects feature-diverse tokens from a saliency-guided candidate pool, and Behavior-Preserving Decoder Pruning (BPDP) prunes inside the decoder using the attention of a single routing head, located offline as the one whose pruning keeps the model closest to the original, with no supervision at inference. Across eight REC subsets and eight VQA benchmarks, GroundPrune attains the lowest KL divergence and probe error among compared methods; at 25% retention on Qwen2.5-VL-7B it reaches 89.7% relative REC accuracy, 5.1 points above the strongest baseline, while keeping 96.1% relative VQA accuracy, with consistent gains on LLaVA-1.5-7B and LLaVA-NeXT-7B.

1 INTRODUCTION

Vision-language models (VLMs) have become the standard interface for understanding images and text together (Liu et al., 2023; 2024a; Bai et al., 2025), and much of their recent progress comes from representing an image with more, and more detailed, visual tokens. LLaVA-NeXT (Liu et al., 2024b) spends up to 2,880 tokens on a single image, and Qwen2.5-VL (Bai et al., 2025) up to 16,384. These extra tokens succeed in carrying fine-grained perceptual information, such as smaller objects, the parts of each object, and the spatial relations between them, but they have also become the dominant cost in the latency of the language decoder. Visual token pruning (Chen et al., 2024a; Zhang et al., 2025c; Yang et al., 2025) has therefore become an active research direction: by removing visual tokens that look unnecessary at inference time, it cuts computation substantially while retaining accuracy. Here two trends collide. VLMs add tokens for accuracy, while pruning removes them for efficiency. If what higher resolution buys is precisely fine-grained information, then the tokens that pruning discards first as unnecessary may be exactly the information that was worth adding.

Existing pruning methods answer the question *which visual tokens matter?* by reading different signals at different places. Vision-stage methods (Yang et al., 2025; Zhang et al., 2024) rank tokens by image-intrinsic saliency such as `[CLS]` attention and remove them before the LLM, which yields large savings; however, they must decide before seeing the query, and are therefore prone

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.

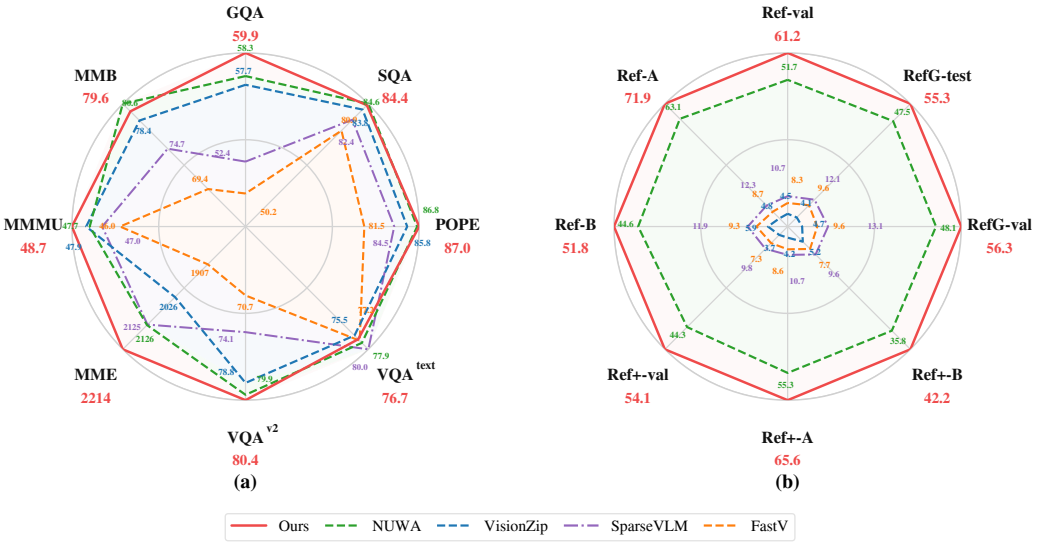
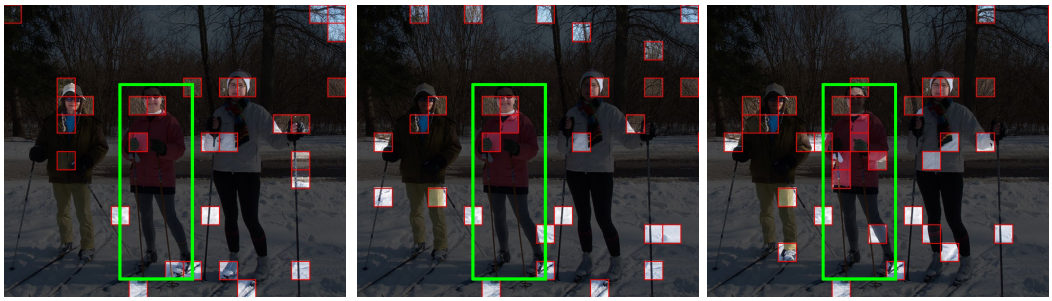


Figure 1: Benchmark comparison under constrained token budgets. (a) General VQA performance on Qwen2.5-VL-7B at a 25% visual-token retention ratio. (b) REC performance on LLaVA-1.5-7B with 128 retained visual tokens.

to discarding information the answer requires. Decoder-stage methods (Chen et al., 2024a; Zhang et al., 2025c; Xing et al., 2024) condition the selection on the query through the attention of its last token, but this signal is usually averaged over all heads of a particular layer. More recent methods additionally encourage diversity or spatial coverage to keep the selection from concentrating on a few regions (Alvar et al., 2025; Huang et al., 2026). Despite relying on different signals, all of these methods are ultimately evaluated by the same criterion: final answer accuracy. Accuracy is a reasonable metric, easy to compare across methods and directly tied to what end users experience, but until an answer actually flips, it cannot show how far the internal representation of the original model has been distorted. This hidden damage is most pronounced on tasks that demand fine-grained localization. In our experiments, FastV (Chen et al., 2024a) removes 75% of the visual tokens of Qwen2.5-VL-7B and still keeps 87.6% of the original general VQA accuracy, yet on referring expression comprehension (REC) it collapses to 26.9% of the original. High VQA accuracy is thus less evidence that pruning preserves the original model than a sign that the benchmark never asked about the difference (Endo et al., 2025).

In this paper, we shift the goal of token pruning from preserving the answer to preserving the behavior of the original model. Let f be the original VLM, which maps visual input X_V and textual query X_Q to a response, and let $f_{\mathcal{P}}$ be its pruned counterpart. Rather than only requiring $f_{\mathcal{P}}(X_V, X_Q)$ to produce the correct answer, we ask how closely $f_{\mathcal{P}}$ follows f itself, and compare the two directly at two levels. At the **output level**, we measure the KL divergence between their next-token distributions on general VQA. At the **representation level**, we train a linear probe on the original model to regress the box of the referred object from the hidden state of the last prompt token, and measure the IoU between the boxes it regresses from the hidden states of the pruned and of the original model. From this perspective, even existing methods that retain most of their VQA accuracy already suffer considerable distribution shift and lose much of their box recoverability (Table 1).

Building on this observation, we propose GROUNDPRUNE, a training-free pruning framework with two stages. (i) **Visual Evidence Preservation (VEP)** forms an informative candidate pool from the vision encoder outputs using saliency and representational energy, and selects representationally diverse tokens from it. (ii) **Behavior-Preserving Decoder Pruning (BPDP)** prunes the remaining tokens once more inside the decoder, using the attention of a single decoder head, chosen as the routing point whose pruning disturbs the original model the least. The cascade separates two decisions: before the query is known, it preserves *what is distinct*; only after the query is seen does it decide *what to discard*.



(a) Saliency Only (b) Saliency + Spatial Coverage (c) Saliency + Feature Diversity (Ours)

Figure 2: **Failure modes of vision-stage pruning.** We visualize retained visual tokens for the query "woman in the pink jacket." in different pruning methods.

The defining feature of GROUNDPRUNE is that it decides *where* to read the decoder signal by measuring representational shift rather than by tuning for accuracy. On a small subset of each benchmark, we use each head as the routing point and measure how far the resulting pruned model departs from the original at both levels, in its output distribution and in the box its hidden state encodes. Only a few heads keep both small, and filtering tokens along their attention removes first the tokens that would not have influenced the model's decision in the first place, so the behavior of the original model is disturbed the least. The selected routing point is determined once per backbone and fixed at inference. Since it is selected based on how closely the pruned model follows the output distribution of the original, regardless of the correct answer, preserving the original model's behavior becomes the guiding principle of our pruning design.

We evaluate GROUNDPRUNE on three backbones, Qwen2.5-VL (Bai et al., 2025), LLaVA-1.5 (Liu et al., 2024a), and LLaVA-NeXT (Liu et al., 2024b), across general VQA and REC benchmarks (Yu et al., 2016; Mao et al., 2016). On Qwen2.5-VL-7B with 25% of the visual tokens retained, GROUNDPRUNE reaches 89.7% relative REC accuracy, 5.1 percentage points above the strongest baseline Nüwa (Huang et al., 2026), while keeping 96.1% relative general VQA accuracy. This REC advantage holds at every token budget on LLaVA-1.5-7B and LLaVA-NeXT-7B as well, at a computational cost on par with existing methods. More notably, distribution shift tracks accuracy. On Qwen2.5-VL-7B, the ranking of the five methods by average KL shift coincides exactly with their rankings by REC and by general VQA accuracy, and on both backbones where we measure shift, GROUNDPRUNE, with the smallest KL shift, achieves the highest accuracy. This suggests that preserving the original behavior carries over to preserving accuracy, and that confirming it requires diagnosis at the level of the model's latent representations.

GROUNDPRUNE aims to minimize the damage to the model's internal inference state that VQA accuracy does not reveal, and through it we propose moving the criterion of token pruning from maximizing accuracy to preserving the behavior of the original model. In summary, the contributions of this paper are as follows:

- We introduce a two-level diagnostic, the KL divergence of output distributions and a linear box probe on hidden states, that evaluates token pruning by preservation of the original behavior rather than of the answer, and show that even methods that retain VQA accuracy are already damaged at the hidden-state level, damage that REC amplifies.
- We identify two sources of this damage: saliency mode collapse at the vision stage, and the choice of routing point at the decoder stage, where only a few heads can guide pruning while keeping the model close to the original.
- We propose GROUNDPRUNE, a training-free cascaded framework that filters tokens along the routing point with the smallest shift from the original model at both levels.
- Across three backbones and 15 evaluation settings, GROUNDPRUNE attains the lowest average KL shift in every setting where shift is measured, and the best (or tied-best) REC and VQA accuracy at every token budget.

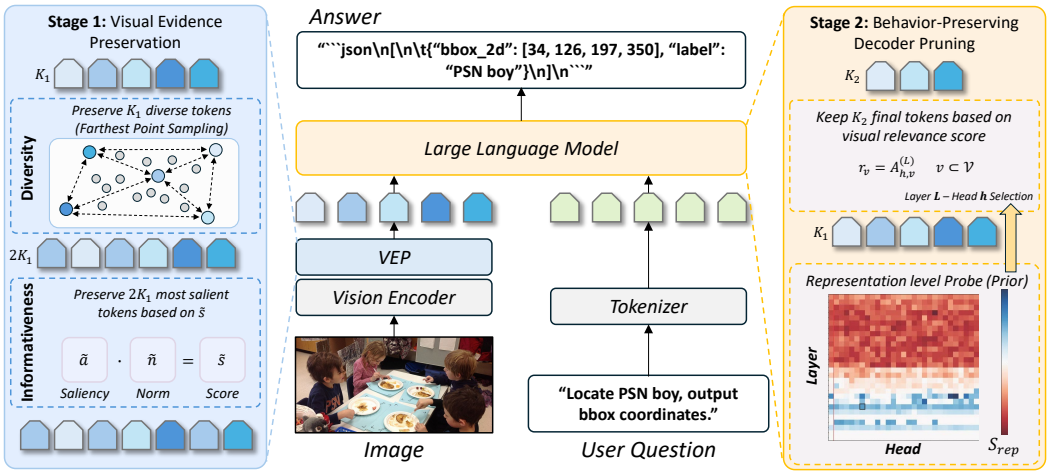


Figure 3: **Overview of GROUNDPRUNE.** We prune visual tokens in two cascaded steps. VEP keeps feature-diverse tokens from an informative candidate pool before the query is seen; BPDP then prunes inside the decoder along the attention of a routing head selected offline for the smallest departure from the original model.

2 RELATED WORKS

2.1 LARGE VISION-LANGUAGE MODELS

Modern Large Vision-Language Models (LVLMs) commonly encode images into visual tokens, project them into the language embedding space, and concatenate them with text tokens for autoregressive decoding Liu et al. (2023; 2024a). Recent models improve fine-grained perception by supporting higher or dynamic image resolutions, such as AnyRes-style image partitioning in LLaVA-NeXT Liu et al. (2024b), tiled high-resolution processing in InternVL Chen et al. (2024c;b), and native dynamic-resolution encoding in Qwen2.5-VL Bai et al. (2025). While effective for localization-sensitive tasks, these designs substantially increase the number of visual tokens, making visual-token reduction important for efficient inference.

2.2 VISUAL TOKEN PRUNING FOR LVLMS

Visual token pruning reduces inference cost by removing or merging redundant image tokens without retraining the LVLM. Decoder-side methods such as FastV Chen et al. (2024a), SparseVLM Zhang et al. (2025c), and PyramidDrop Xing et al. (2024) prune tokens using attention-based importance signals inside the LLM, while vision-side methods such as VisionZip Yang et al. (2025), Faster-VLM Zhang et al. (2024), and DivPrune Alvar et al. (2025) reduce tokens before they enter the decoder using attention, similarity, or diversity criteria. Selection by image-intrinsic saliency tends to concentrate the budget on a few dominant regions, whereas enforcing spatial coverage spends part of it on uninformative background; decoder-side methods typically read importance from attention averaged over all heads of a fixed layer. Although these approaches report strong accuracy-efficiency trade-offs on general VQA benchmarks, recent analyses suggest that such benchmarks may under-measure fine-grained localization failures Endo et al. (2025). More recent grounding-aware methods, including Nüwa Huang et al. (2026) and VScan Zhang et al. (2025a), therefore attempt to preserve spatial reference cues and refine pruning with language-conditioned signals. Our work follows this direction, but combines feature-diverse evidence preservation at the vision stage with decoder pruning along a routing head selected by how little pruning departs from the original model.

3 METHOD

As shown in Fig. 3, GROUNDPRUNE prunes visual tokens in two cascaded stages. Visual Evidence Preservation (VEP) reduces the post-encoder visual tokens before the query is seen, and Behavior-Preserving Decoder Pruning (BPDP) prunes the remaining visual tokens inside the decoder along the

Table 1: Departure of pruned models from the original model. General VQA: KL divergence between next-token distributions (lower is better). REC: IoU (%) of probe-regressed boxes with ground truth (higher is better).

	General VQA KL ↓								REC IoU ↑			
Method	GQA	MMBench	MME	MMMU	SQA	VQA^{text}	VQA^{v2}	Avg	RefCOCO	RefCOCO+	RefCOCOg	Avg
Qwen2.5-VL-7B (R=25%)												
Base	–	–	–	–	–	–	–	–	57.37	50.89	51.22	53.16
FastV	0.3482	0.1611	0.2053	0.0993	0.1295	0.1596	0.3385	0.2059	34.47	29.19	31.27	31.64
SparseVLM	0.2792	0.1078	0.1144	0.0799	0.1033	**0.0859**	0.2336	0.1434	38.01	32.94	35.12	35.36
VisionZip	0.1029	0.0604	0.1395	0.0464	0.0574	0.2042	0.1069	0.1025	47.56	41.13	43.09	43.93
Nüwa	0.0905	0.0458	0.0907	**0.0363**	0.0670	0.1262	0.0759	0.0760	50.10	43.07	45.16	46.11
Ours	**0.0682**	**0.0379**	**0.0785**	0.0405	**0.0533**	0.1504	**0.0757**	**0.0721**	**51.75**	**45.30**	**46.67**	**47.91**
LLaVA-1.5-7B (R=22.2%)												
Base	–	–	–	–	–	–	–	–	50.19	41.66	43.48	45.11
FastV	0.1227	0.0417	0.0393	0.0248	0.0372	0.2658	0.1303	0.0945	28.31	23.03	24.91	25.41
SparseVLM	0.1058	0.0342	0.0318	0.0205	0.0292	0.2118	0.1095	0.0775	30.49	24.81	26.62	27.31
VisionZip	0.0615	0.0276	0.0168	0.0316	0.0218	**0.0775**	0.0467	0.0405	25.87	21.99	21.84	23.23
Nüwa	0.0493	0.0133	0.0105	0.0182	0.0115	0.1180	0.0397	0.0372	**39.18**	32.42	33.08	34.89
Ours	**0.0410**	**0.0107**	**0.0087**	**0.0168**	**0.0090**	0.1029	**0.0263**	**0.0308**	38.35	**32.67**	**33.93**	**34.98**

attention of a single routing head. The routing head is located once per backbone by searching all layer–head pairs for the one whose pruning departs least from the original model (Section 3.4).

3.1 Preliminaries

Token budget. Let T_v be the number of visual tokens produced by the vision encoder and N_L the number of decoder layers. VEP keeps $K_1 = \max(1, \lfloor r\, T_v \rfloor)$ tokens, and BPDP, applied after decoder layer L, keeps $K_2 = \max(1, \lfloor \lambda K_1 \rfloor)$ of them for the remaining $N_L - L - 1$ layers. The effective retention ratio, the fraction of visual tokens averaged over decoder layers, is

$$R = r\left(\frac{L+1}{N_L} + \lambda\,\frac{N_L - L - 1}{N_L}\right). \tag{1}$$

Departure from the original model. Let $y = (y_1, \ldots, y_T)$ be the greedy answer of the original model f to (X_V, X_Q). At the output level, we teacher-force y into both f and its pruned counterpart $f_{\mathcal{P}}$ and average the KL divergence of their next-token distributions over the first $T' = \min(T, 64)$ answer positions,

$$D_{\text{out}} = \frac{1}{T'}\sum_{t=1}^{T'} \mathrm{KL}\big(p_f(\cdot \mid X_V, X_Q, y_{<t}) \,\|\, p_{f_{\mathcal{P}}}(\cdot \mid X_V, X_Q, y_{<t})\big). \tag{2}$$

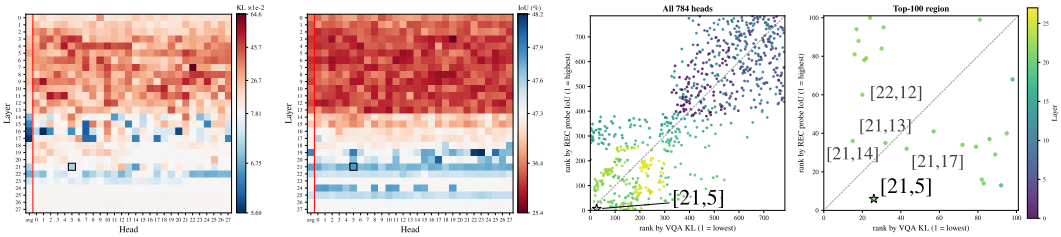
At the representation level, a linear probe W fit on the original model maps the final-layer hidden state h of the last prompt token to the box of the referred object, and we measure how well the box read from the pruned model agrees with the one read from the original,

$$S_{\text{rep}} = \mathrm{IoU}\big(W h_{f_{\mathcal{P}}},\, W h_f\big). \tag{3}$$

D_{out} is averaged within each general VQA benchmark and then across benchmarks, and S_{rep} across REC splits (Appendix A). S_{rep}, which compares only against the original model, is used to select the routing point (Section 3.4); to compare methods in Table 1, we instead report the IoU of $W h_{f_{\mathcal{P}}}$ with the ground-truth box.

3.2 Stage 1: Visual Evidence Preservation (VEP)

VEP runs before the query is seen, and tokens it discards cannot be recovered by the decoder. Since it must serve any query, it should keep *what is distinct* in the image. Selecting the top-K tokens by image-intrinsic saliency (Zhang et al., 2024; Yang et al., 2025) concentrates the budget on a few dominant regions, a behavior we call *saliency mode collapse*: in Fig. 2(a), most retained tokens fall on faces and few on the referred pink jacket. Enforcing spatial coverage (Alvar et al., 2025; Huang et al., 2026) spreads the tokens but spends many of them on background such as trees and snow (Fig. 2(b)).



(a) D_{out} (b) S_{rep} (c) All heads (d) Top-100 region

Figure 4: **Departure from the original model for every routing head** (Qwen2.5-VL-7B, $R = 25\%$). (a) D_{out} on general VQA (lower is better) and (b) S_{rep} on REC (higher is better) for each layer and head. (c, d) Distribution of the ranks of each head by D_{out} and by S_{rep}.

VEP therefore first restricts the selection to an informative candidate pool and then enforces diversity in feature space within it (Fig. 2(c)).

Let $X = [x_1, \ldots, x_N] \in \mathbb{R}^{N \times d}$ denote the post-merger visual features produced by the vision encoder, where N is the number of visual tokens and d is the feature dimension. We compute an image-intrinsic saliency score from the last full-attention block of the vision encoder. Since Qwen2.5-VL (Bai et al., 2025) does not contain an explicit CLS token, we use the column-wise mean of the self-attention matrix as a CLS-attention analogue. For visual token i, the saliency score is

$$a_i = \frac{1}{H_v} \sum_{h=1}^{H_v} \text{softmax}\left(\frac{Q^{(h)} K^{(h)\top}}{\sqrt{d_h}} \right)_i, \tag{4}$$

where H_v is the number of vision-attention heads and d_h is the head dimension. This score measures how much the final vision-encoder layer globally attends to each token. As a lightweight proxy for representational energy, we also compute the L2 norm $n_i = \|x_i\|_2$. The two scores are min-max normalized and combined multiplicatively:

$$\begin{gathered} \tilde{a}_i = \text{minmax}(a_i), \quad \tilde{n}_i = \text{minmax}(n_i), \\ s_i = \tilde{a}_i \cdot \tilde{n}_i. \end{gathered} \tag{5}$$

so that a selected token must be both attentionally salient and representationally informative.

Given a target keep count K_1, we first select the top-$2K_1$ tokens according to s_i to form an informative candidate pool $\mathcal{P}$. We then apply farthest point sampling (FPS) in the visual feature space to choose K_1 diverse tokens from this pool. Specifically, candidate features are L2-normalized as $\hat{x}_i = x_i / \|x_i\|_2$ for $i \in \mathcal{P}$. The selected set $\mathcal{S}$ is initialized with the highest-scoring token in $\mathcal{P}$, and at each subsequent step we add the token whose maximum cosine similarity to the current selected set is smallest:

$$j^\star = \arg \min_{j \in \mathcal{P} \setminus \mathcal{S}} \max_{i \in \mathcal{S}} \hat{x}_j^\top \hat{x}_i, \quad \mathcal{S} \leftarrow \mathcal{S} \cup \{j^\star\}. \tag{6}$$

The final selected indices are sorted in their original sequence order before being passed to the LLM.

3.3 Stage 2: Behavior-Preserving Decoder Pruning (BPDP)

After VEP, the retained visual tokens enter the decoder together with the query. Decoder-side methods read the query-conditioned importance of visual tokens from the attention of the last query token, usually averaged over all heads of a fixed layer (Zhang et al., 2025c; Xing et al., 2024). BPDP instead reads it from a single routing head (L, h).

Let $A^{(L)} \in \mathbb{R}^{H_q \times T}$ denote the attention probabilities from the last query position to all key positions at decoder layer L, where H_q is the number of query heads and T is the sequence length. Let $\mathcal{V} \subset [T]$ denote the set of visual-token positions. For a selected head h, BPDP defines the visual relevance score as

$$r_v = A^{(L)}_{h,v}, \qquad v \in \mathcal{V}. \tag{7}$$

Given a keep count K_2, BPDP retains the top-K_2 visual tokens according to r_v and preserves all non-visual tokens. The hidden states, attention mask, position ids, and other sequence metadata are then sliced consistently before the next decoder layer.

Table 2: Performance comparison under different retention ratios on Qwen2.5-VL-7B across referring expression comprehension and general VQA benchmarks.

Method	Retention Ratio R	Referring Expression Comprehension									General VQA								
		RefCOCO			RefCOCO+			RefCOCOg		Rel. Acc. (%)	GQA	MMBench	MMMU	MME	VQA^{v2}	VQA^{text}	POPE	SQA	Rel. Acc. (%)
		val	testA	testB	val	testA	testB	val	test										
Qwen2.5-VL-7B																			
Vanilla	100%	88.9	91.9	84.9	82.4	87.8	74.2	86.3	86.3	100.0	60.9	82.8	51.1	2325.7	82.9	83.0	87.8	88.4	100.0
FastV		86.2	88.5	82.3	78.7	83.0	72.2	83.8	83.7	96.4	60.1	82.0	**51.3**	2317.0	82.2	82.8	87.7	86.1	99.2
SparseVLM		87.0	89.7	82.7	79.5	84.9	72.8	84.7	84.6	97.5	60.0	82.2	50.7	2335.3	82.3	**82.9**	**87.9**	86.5	99.3
VisionZip	75%	**88.9**	91.9	84.4	82.6	**88.2**	74.5	84.8	85.2	99.7	60.5	82.8	49.9	2318.1	82.4	82.8	87.7	86.7	99.2
Nüwa		88.7	**92.1**	**84.6**	**82.9**	87.8	74.6	84.9	85.3	99.7	60.4	**83.0**	51.0	2332.3	82.5	82.8	87.7	86.7	99.6
Ours		88.8	91.7	84.1	82.3	87.7	**74.6**	**85.9**	**86.3**	**99.8**	**60.7**	82.9	50.4	**2352.3**	**82.6**	82.8	87.8	**88.0**	**99.8**
FastV		72.8	71.0	72.5	64.9	65.2	62.0	72.0	71.5	80.8	58.1	79.1	49.6	2207.6	80.2	82.5	87.2	84.1	96.7
SparseVLM		76.0	76.4	75.5	68.8	70.4	66.2	75.6	75.2	85.6	58.5	80.8	49.3	2264.0	80.8	**82.7**	87.4	84.4	97.5
VisionZip	50%	85.9	89.2	81.4	79.2	84.9	71.6	82.1	82.0	96.1	59.9	81.9	49.1	2243.2	81.6	81.5	**87.6**	85.7	97.9
Nüwa		87.2	90.4	**82.5**	80.7	86.0	72.2	**84.0**	83.6	97.6	60.0	**82.5**	**49.7**	2302.7	81.7	81.6	**87.6**	86.9	**98.7**
Ours		**87.7**	**90.8**	82.4	**80.8**	**86.2**	**72.4**	83.6	**84.5**	**97.9**	**60.2**	81.7	49.4	**2316.8**	**81.8**	81.7	87.2	**87.8**	**98.7**
FastV		23.9	22.1	26.0	20.8	19.9	21.8	24.9	24.1	26.9	50.2	69.4	46.0	1907.3	70.7	77.2	81.5	80.9	87.6
SparseVLM		34.0	32.7	36.9	30.5	30.0	31.3	36.2	36.3	39.3	52.4	74.7	47.0	2124.7	74.1	**80.0**	84.5	82.4	91.9
VisionZip	25%	67.5	73.1	61.7	60.2	67.3	52.5	62.4	62.1	74.2	57.7	78.4	47.9	2025.5	78.8	75.5	85.8	83.8	93.6
Nüwa		76.6	80.2	72.4	68.5	73.2	61.2	73.0	72.7	84.6	58.3	**80.6**	47.7	2126.5	79.9	77.9	86.8	**84.6**	95.3
Ours		**81.2**	**85.7**	**74.7**	**74.4**	**80.0**	**64.0**	**76.3**	**76.1**	**89.7**	**59.9**	79.6	**48.7**	**2213.9**	**80.4**	76.7	**87.0**	84.4	**96.1**

3.4 Routing-Point Selection

Which head to read the pruning signal from is the one design decision of BPDP, and we make it with the same diagnostic used to evaluate pruning. Every layer–head pair (L, h) is a candidate routing point. For each candidate, we prune along (L, h) at a fixed R and measure $D_{\text{out}}(L, h)$ and $S_{\text{rep}}(L, h)$ on data randomly sampled from each benchmark. Since the two diagnostics lie on different scales, we combine their ranks among all candidates, where $\text{rk}_D(L, h)$ ranks D_{out} in ascending and $\text{rk}_S(L, h)$ ranks S_{rep} in descending order, and select the head with the smallest rank sum:

$$(L^\star, h^\star) = \arg\min_{(L,h)} \ \text{rk}_D(L, h) + \text{rk}_S(L, h). \tag{8}$$

Fig. 4 shows both diagnostics over all 784 heads of Qwen2.5-VL-7B at $R = 25\%$ (other backbones in Appendix B). The departure varies widely with the routing point, and heads that keep both diagnostics small are rare: only 8 of the 784 heads rank in the top 10% on both, all in layers 21–22 (Fig. 4d). Reading the attention averaged over all heads of a layer matches the selected head on both diagnostics at no layer. The selected head is thus a routing point shared across tasks that preserves the behavior of the original model on both general VQA and REC, and BPDP prunes along it.

4 Experiments

Experimental Setup. We evaluate GroundPrune on referring expression comprehension (REC) and general VQA. For REC, we report Acc@0.5 on all standard splits of RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016; Mao et al., 2016). For general VQA, we evaluate GQA, MMBench, MMMU, MME, VQA^{v2}, VQA^{text}, POPE, and ScienceQA (Hudson & Manning, 2019; Liu et al., 2024c; Yue et al., 2024; Fu et al., 2026; Goyal et al., 2017; Singh et al., 2019; Li et al., 2023; Lu et al., 2022). We compare against four training-free pruning baselines: FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2025c), VisionZip (Yang et al., 2025), and Nüwa (Huang et al., 2026). Experiments are conducted on Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-1.5-7B Liu et al. (2024a), and LLaVA-NeXT-7B (Liu et al., 2024b) under matched visual-token budgets. We additionally report *relative accuracy* (*Rel. Acc.*) by normalizing each pruned model by its unpruned counterpart; for REC, scores are averaged over the eight RefCOCO-series splits, and for VQA, normalized scores are averaged across benchmarks. We use the RefCOCO training split for probe calibration, keeping it strictly disjoint from all validation and test evaluation sets.

4.1 Main Results

Tables 2 and 3 report the main comparison on Qwen2.5-VL-7B and LLaVA-1.5-7B across REC and general VQA benchmarks. We highlight three findings.

General VQA preservation does not imply grounding preservation. Existing pruning methods often maintain strong general VQA performance while suffering severe degradation on REC. For example, on LLaVA-1.5-7B with 192 retained tokens, VisionZip preserves 98.3% relative VQA

Table 3: Performance comparison under different retained-token budgets on LLaVA-1.5-7B across referring expression comprehension and general VQA benchmarks.

Method	Retained Tokens	Referring Expression Comprehension									General VQA								
		RefCOCO			RefCOCO+			RefCOCOg		Rel. Acc. (%)	GQA	MMBench	MMMU	MME	VQA^{v2}	VQA^{text}	POPE	SQA	Rel. Acc. (%)
		val	testA	testB	val	testA	testB	val	test										
LLaVA-1.5-7B																			
Vanilla	576	56.2	64.0	47.2	49.5	58.7	38.8	52.2	51.9	100.0	62.0	64.1	36.1	1870.1	76.7	46.0	87.0	69.6	100.0
FastV	192 (33.3%)	26.4	23.4	21.7	18.1	21.1	18.1	21.2	20.9	40.8	52.7	61.2	34.3	1612.0	67.1	52.5	64.8	67.3	89.5
SparseVLM		23.3	26.2	24.0	20.7	23.8	19.9	23.6	23.3	44.1	57.6	62.5	33.8	1721.0	75.6	56.1	83.6	69.1	96.1
VisionZip		6.2	7.0	8.3	5.5	6.2	6.5	6.5	5.9	12.4	59.3	63.0	**36.6**	1782.0	**76.8**	**57.3**	85.3	**68.9**	98.3
Nüwa		59.4	68.2	49.4	51.7	62.4	40.5	54.7	53.7	105.1	**61.1**	63.8	35.7	1848.8	76.1	45.6	85.9	68.7	98.9
Ours		**68.0**	**76.3**	**57.8**	**59.3**	**70.8**	**46.1**	**61.7**	**61.7**	**119.9**	**61.1**	**64.2**	36.0	**1852.2**	76.0	44.4	**86.8**	68.8	**99.0**
FastV	128 (22.2%)	8.3	8.7	9.3	7.3	8.6	7.7	9.6	9.6	16.5	49.6	56.1	34.9	1490.0	61.8	50.6	59.6	60.2	85.0
SparseVLM		10.7	12.3	11.9	9.8	10.7	9.6	13.1	12.1	21.5	56.0	60.0	33.8	1696.0	73.8	54.9	80.5	67.1	93.4
VisionZip		4.5	4.8	5.9	3.7	4.2	5.2	4.7	4.1	8.9	57.6	62.0	**37.9**	1761.0	75.6	**56.8**	83.2	**68.9**	97.6
Nüwa		51.7	63.1	44.6	44.3	55.3	35.8	48.1	47.5	93.3	60.4	63.0	36.3	**1814.0**	75.2	43.5	85.4	68.2	97.8
Ours		**61.2**	**71.9**	**51.8**	**54.1**	**65.6**	**42.2**	**56.3**	**55.3**	**109.6**	**60.7**	**63.7**	36.9	1790.9	**75.8**	44.2	**86.5**	68.3	**98.4**
FastV	64 (11.1%)	0.1	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.1	46.1	48.0	34.0	1256.0	55.0	47.8	59.6	51.1	79.4
SparseVLM		0.5	0.5	0.9	0.4	0.5	0.7	0.5	0.8	1.1	53.8	60.1	35.4	1589.0	68.2	53.4	77.5	**69.8**	89.9
VisionZip		3.9	3.6	4.0	3.7	3.3	3.9	3.7	4.0	7.2	55.1	60.1	36.2	**1690.0**	72.4	**55.5**	77.0	69.0	94.0
Nüwa		32.0	40.8	**25.1**	26.4	33.8	**20.1**	28.7	29.5	56.5	**58.5**	**62.1**	35.8	1676.7	72.6	39.4	82.5	68.0	94.1
Ours		**32.9**	**43.9**	24.4	**28.5**	**38.5**	19.8	**31.1**	**30.6**	**59.7**	58.1	62.0	**36.8**	1538.5	**73.8**	42.6	**86.6**	67.7	**95.0**

Table 4: Referring expression comprehension performance under different retained-token budgets on LLaVA-Next-7B.

Method	Retained Tokens	RefCOCO			RefCOCO+			RefCOCOg		Rel. Acc. (%)
		val	testA	testB	val	testA	testB	val	test	
LLaVA-Next-7B										
Vanilla	2880	84.9	90.5	77.7	77.1	86.1	65.9	80.5	80.1	100.0
Nüwa	640 (22.2%)	66.2	71.1	59.8	59.0	66.5	48.8	59.9	59.5	76.4
Ours		**70.5**	**76.5**	**61.8**	**63.4**	**72.0**	**50.2**	**65.0**	**65.4**	**81.7**
Nüwa	320 (11.1%)	37.2	39.8	36.0	29.1	34.4	26.8	31.0	31.2	41.3
Ours		**48.1**	**51.8**	**41.4**	**41.4**	**46.7**	**32.2**	**45.8**	**45.0**	**54.8**
Nüwa	160 (5.6%)	18.2	19.3	17.5	13.6	14.6	11.8	15.6	15.5	19.6
Ours		**19.9**	**20.6**	**19.8**	**16.7**	**18.1**	**14.9**	**21.6**	**20.9**	**23.7**

Table 5: Efficiency on LLaVA-1.5-7B at R=11.1%

Method	Tokens	TFLOPs	Prefill (ms)
Vanilla	576	9.59	99
FastV	64	2.60	33
SparseVLM	64	2.84	42
VisionZip	64	2.58	33
Nüwa	64	2.57	33
Ours	64	2.59	39

accuracy but only 12.4% relative REC accuracy; FastV similarly retains 89.5% relative VQA accuracy but only 40.8% relative REC accuracy. The same pattern appears on Qwen2.5-VL-7B at R=25%, where FastV maintains 87.6% relative VQA accuracy but drops to 26.9% relative REC accuracy. These results show that general VQA benchmarks can substantially overestimate the reliability of visual token pruning for fine-grained grounding.

Our method preserves REC performance under aggressive pruning. Across both Qwen2.5-VL-7B and LLaVA-1.5-7B, GROUNDPRUNE consistently achieves the strongest REC robustness at low token budgets. On Qwen2.5-VL-7B, most methods remain competitive at R=75%, but the gap widens as the retention ratio decreases. At R=25%, FastV collapses to 26.9% relative REC accuracy and VisionZip retains 74.2%. Nüwa substantially improves the result to 84.6%, while GROUNDPRUNE further raises relative REC accuracy to 89.7%, achieving the best performance across all RefCOCO-series splits.

On LLaVA-1.5-7B, GROUNDPRUNE also provides consistent gains under more aggressive token budgets. At 128 retained tokens, GROUNDPRUNE reaches 109.6% relative REC accuracy, outperforming Nüwa by +16.3 percentage points. Even at only 64 retained tokens, where FastV and SparseVLM nearly fail on REC, GROUNDPRUNE maintains the highest relative REC accuracy among all methods. These results indicate that preserving fine-grained grounding requires more careful token selection than what is needed for coarse visual understanding.

Grounding robustness is achieved without sacrificing general VQA. Since GROUNDPRUNE aims to preserve the behavior of the original model rather than grounding alone, it also maintains strong general VQA performance. On Qwen2.5-VL-7B, GROUNDPRUNE achieves 99.8%, 98.7%, and 96.1% relative VQA accuracy at R=75%, 50%, and 25%, respectively. On LLaVA-1.5-7B, it similarly retains 99.0%, 98.4%, and 95.0% relative VQA accuracy at 192, 128, and 64 retained tokens. Thus, the gains on REC do not come from sacrificing general visual question answering performance; instead, GROUNDPRUNE provides a more balanced pruning behavior across coarse semantic understanding and localization-sensitive grounding.

Extension to LLaVA-NeXT. We further evaluate whether the proposed pruning strategy generalizes to a higher-resolution VLM. Table 4 reports REC performance on LLaVA-NeXT-7B, which uses the

Table 6: Ablation on VEP pruning method.

Method	VQA			RefCOCO		RefCOCO+		RefCOCOg
	GQA	MMB	SQA	testA	testB	testA	testB	test
Random	60.7	61.6	68.5	47.5	34.5	41.6	26.5	36.3
Uniform	60.9	63.2	**69.1**	49.9	35.4	44.6	26.6	38.2
Saliency only	60.9	63.6	68.2	55.4	39.4	50.6	32.4	44.1
Spatial coverage	61.0	**63.9**	68.4	49.0	34.6	44.6	28.9	38.0
Saliency + Feature Diversity	**61.1**	63.7	68.2	**65.2**	**45.5**	**60.1**	**37.8**	**50.0**

Table 7: Ablation on BPDP routing point.

Config	VQA			RefCOCO		RefCOCO+		RefCOCOg
	GQA	MMB	SQA	testA	testB	testA	testB	test
Layer 5, heads Avg.	50.2	54.3	**70.2**	1.8	0.8	1.6	0.5	1.9
Layer 10, heads Avg.	54.0	59.1	67.5	10.8	7.9	9.6	7.3	8.7
Layer 15, heads Avg.	**60.9**	63.5	68.1	67.2	48.4	61.4	38.5	52.2
Layer 20, heads Avg.	60.7	63.1	68.8	61.7	40.0	56.3	33.2	45.4
Layer 25, heads Avg.	60.3	63.1	68.8	56.9	36.6	51.1	30.0	41.9
Layer 30, heads Avg.	59.6	63.0	68.4	53.0	31.7	47.2	26.4	37.5
Layer 14, head 24	60.7	**63.7**	68.3	**71.9**	**51.8**	**65.6**	**42.2**	**55.3**

AnyRes mechanism and can process images with dynamic resolutions, producing up to 2880 visual tokens per image. We evaluate retention ratios of 22.2%, 11.1%, and 5.6% relative to the original visual-token count.

Across all retention ratios, GROUNDPRUNE consistently outperforms Nüwa, improving relative REC accuracy by $+5.3$, $+13.5$, and $+4.1$ percentage points at 22.2%, 11.1%, and 5.6% retention, respectively. The proposed pruning strategy also generalizes to dynamic-resolution VLMs with larger visual-token budgets.

4.2 EFFICIENCY ANALYSIS

Table 5 reports inference efficiency on LLaVA-1.5-7B at R=11.1%. All measurements include both visual and language tokens in the input sequence. Compared with the unpruned vanilla model, pruning methods substantially reduce end-to-end computation, lowering TFLOPs from 9.59 to 2.57–2.84 and prefill latency from 99 ms to 33–42 ms.

GROUNDPRUNE achieves 2.59 TFLOPs and 39 ms prefill latency, which is comparable to existing pruning baselines at the same token budget. Importantly, these gains require no additional learnable parameters and no extra forward pass. Thus, the improved grounding robustness of GROUNDPRUNE does not come at the cost of sacrificing the efficiency benefits of visual token pruning.

4.3 ABLATION STUDY

We conduct ablation studies to analyze the contribution of each component in GROUNDPRUNE. All experiments in this section are conducted on LLaVA-1.5-7B with 128 retained visual tokens.

Ablation on VEP. Table 6 compares vision-stage selection rules. Random and uniform sampling preserve general VQA but perform poorly on REC, showing that arbitrary token retention is insufficient for grounding. Saliency-only Top-K improves REC, but still suffers from redundant local clustering, while coordinate-space coverage wastes tokens on uninformative regions. By selecting feature-diverse tokens from a saliency- and energy-guided candidate pool, VEP achieves the best REC performance with comparable VQA accuracy.

Ablation on BPDP. Table 7 compares the routing head used in BPDP with head-averaged attention from different decoder layers. Head-averaged attention is highly layer-sensitive: early layers nearly fail on REC, and even the best averaged variant at layer 15 remains below our selected single head. The selected head (14, 24) achieves the strongest REC performance across all RefCOCO-series splits, showing that reading a single routing head is more reliable than averaging over all heads of a layer.

5 CONCLUSION

In this work, we proposed GROUNDPRUNE, a training-free cascaded visual-token pruning framework that preserves the behavior of the original model for efficient VLM inference. VEP retains informative and feature-diverse visual tokens at the vision stage, while BPDP further prunes tokens along the attention of a routing head selected for the smallest departure from the original model. Across Qwen2.5-VL, LLaVA-1.5, and LLaVA-NeXT, GROUNDPRUNE improves REC robustness under constrained token budgets while maintaining strong general VQA performance and inference efficiency. Our results suggest that efficient VLM pruning should be evaluated not only by answer accuracy but also by how closely the pruned model follows the original, a departure that general VQA accuracy alone can hide.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024c.

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

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.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. N\" uwa: Mending the spatial integrity torn by vlm token pruning. *arXiv preprint arXiv:2602.02951*, 2026.

Drew A Hudson and Christopher D Manning. 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.

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. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

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

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

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

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

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

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.

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

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pp. arXiv–2412, 2024.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, pp. 74840–74857. PMLR, 2025c.

A Experimental Details

A.1 Benchmarks

We evaluate GroundPrune on two complementary groups of benchmarks.

A.1.1 Referring Expression Comprehension (REC)

First, for REC, we report Acc@0.5 on all standard splits of RefCOCO Series:

RefCOCO. RefCOCO (Yu et al., 2016) is a referring expression comprehension dataset collected via the ReferItGame, containing short and often location-based expressions on MS-COCO images. It provides three evaluation splits: **val** for validation, **testA** consisting of images with multiple people, and **testB** consisting of images with multiple non-person objects.

RefCOCO+. RefCOCO+ (Yu et al., 2016) is collected with the same protocol as RefCOCO but disallows absolute location words, forcing expressions to rely on appearance-based attributes such as color, shape, and texture. It shares the same split structure as RefCOCO, with **val**, **testA** (person-centric), and **testB** (object-centric) subsets.

RefCOCOg. RefCOCOg (Mao et al., 2016) contains longer and more descriptive referring expressions collected in a non-interactive setting, with an average expression length roughly twice that of RefCOCO. We follow the UMD partition and report results on the **val** and **test** splits.

A.1.2 General Visual Question Answering (VQA)

Second, to assess whether visual pruning preserves general visual understanding, we evaluate on eight widely used VQA benchmarks:

GQA. GQA (Hudson & Manning, 2019) is a compositional visual question answering benchmark built on Visual Genome scene graphs, designed to evaluate visual reasoning over object attributes, relations, and spatial configurations.

MMBench. MMBench (Liu et al., 2024c) is a multimodal benchmark that evaluates VLMs across diverse fine-grained ability dimensions, including perception, reasoning, and knowledge, using multiple-choice questions with a circular evaluation protocol.

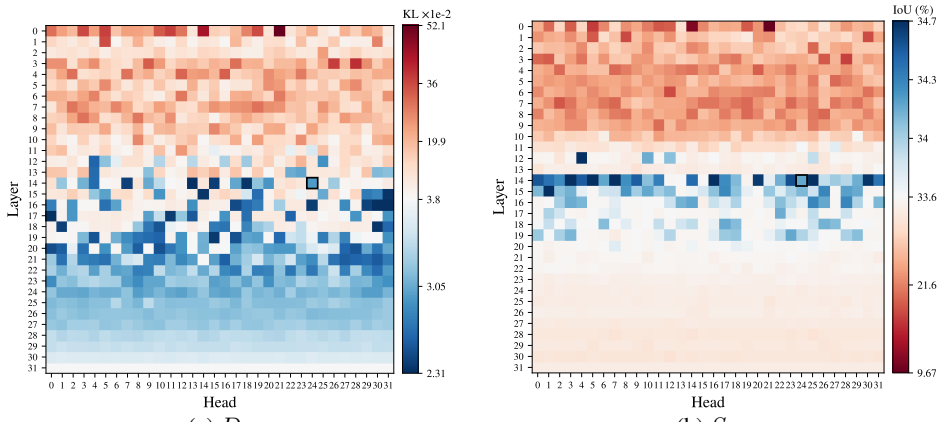
MMMU. MMMU (Yue et al., 2024) is a massive multi-discipline benchmark that tests college-level multimodal understanding and expert-level reasoning across six disciplines, including science, engineering, and humanities.

MME. MME (Fu et al., 2026) is a comprehensive evaluation benchmark for VLMs covering 14 subtasks across perception and cognition, using yes/no questions to assess capabilities ranging from coarse recognition to commonsense reasoning.

VQA^{v2}. VQA^{v2} (Goyal et al., 2017) is a large-scale open-ended visual question answering benchmark that reduces language priors by pairing each question with two visually similar images leading to different answers.

VQA^{text}. VQA^{text} (Singh et al., 2019), also known as TextVQA, evaluates a model's ability to read and reason about text appearing within natural images, requiring OCR-style understanding alongside visual perception.

POPE. POPE (Li et al., 2023) is a benchmark for evaluating object hallucination in VLMs, formulating polling-based binary questions about object presence under random, popular, and adversarial sampling strategies.



(a) D_{out} (b) S_{rep}

Figure 5: **Departure from the original model for every routing head** (LLaVA-1.5-7B, $R = 22.2\%$). (a) D_{out} on general VQA (lower is better) and (b) S_{rep} on REC (higher is better) for each layer and head.

ScienceQA. ScienceQA (Lu et al., 2022) is a multimodal question answering benchmark spanning natural science, social science, and language science topics, with image-text questions and chain-of-thought style explanations.

A.2 Implementation Details

We evaluate three representative VLM backbones: Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-1.5-7B (Liu et al., 2024a), and LLaVA-NeXT-7B (Liu et al., 2024b). For Qwen2.5-VL-7B, we vary the visual-token retention ratio $R \in \{75\%, 50\%, 25\%\}$. For LLaVA-1.5-7B, which uses a fixed budget of 576 visual tokens, we evaluate retained-token budgets of 192, 128, and 64, corresponding to $R \in \{33.3\%, 22.2\%, 11.1\%\}$. For LLaVA-NeXT-7B, which uses the AnyRes mechanism and can generate up to 2880 visual tokens, we report REC results under $R \in \{22.2\%, 11.1\%, 5.6\%\}$. All methods are evaluated by lmms-eval (Zhang et al., 2025b) and COCO evaluation protocol, under matched visual-token budgets for each backbone.

Artifact Use and Licenses. We use publicly available VLM backbones, benchmark datasets, and pruning baselines only for research and evaluation purposes, following their respective licenses and terms of use. We do not redistribute the original datasets, model weights, or baseline implementations. Our use of these artifacts is consistent with their intended research use, and all reported experiments are conducted as comparative evaluations of visual token pruning methods.

Data Privacy and Documentation. We do not collect new data or introduce new annotations in this work. All datasets used in our experiments are publicly available benchmark datasets, and we use their official evaluation splits and protocols. Some image benchmarks may contain people, but our experiments use only the provided images and task annotations for aggregate evaluation; we do not attempt to identify individuals, infer personal attributes, or redistribute the original data. Dataset coverage, task definitions, and split details are documented in the original dataset papers, which we cite in the experimental setup.

B Routing-Point Selection Details

B.1 Selected Routing Heads

The selected routing heads are $(L, h) = (21, 5)$ for Qwen2.5-VL-7B and $(14, 24)$ for LLaVA-1.5-7B. Fig. 5 shows the maps for LLaVA-1.5-7B.

Algorithm 1 Visual Evidence Preservation (VEP)

Require: Visual features $X \in \mathbb{R}^{N \times d}$, last-block attention score $a \in \mathbb{R}^N$, keep count K_1
Ensure: Selected visual-token indices $\mathcal{S}$
$n_i \leftarrow \|x_i\|_2$ for all i
$\tilde{a} \leftarrow \mathrm{minmax}(a)$, $\tilde{n} \leftarrow \mathrm{minmax}(n)$
$s_i \leftarrow \tilde{a}_i \cdot \tilde{n}_i$
$\mathcal{P} \leftarrow \mathrm{TopK}(s, 2K_1)$
$\hat{X}_{\mathcal{P}} \leftarrow \ell_2$-normalize$(X_{\mathcal{P}})$
$\mathcal{S} \leftarrow \{\arg\max_{j \in \mathcal{P}} s_j\}$
while $|\mathcal{S}| < K_1$ **do**
    $j^\star \leftarrow \arg\min_{j \in \mathcal{P} \setminus \mathcal{S}} \max_{i \in \mathcal{S}} \hat{x}_j^\top \hat{x}_i$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{j^\star\}$
end while
return sorted $\mathcal{S}$

Algorithm 2 Behavior-Preserving Decoder Pruning (BPDP)

Require: Hidden states H, visual-token mask $\mathcal{V}$, decoder layer L, head h, keep count K_2
Ensure: Pruned hidden states and corresponding sequence metadata
Run decoder layer L and obtain last-query attention $A^{(L)} \in \mathbb{R}^{H_q \times T}$
$r_v \leftarrow A^{(L)}_{h,v}$ for all $v \in \mathcal{V}$
$\mathcal{M} \leftarrow \mathrm{TopK}(r, K_2)$ over visual positions
$\mathcal{I} \leftarrow \mathcal{M} \cup ([T] \setminus \mathcal{V})$
Sort $\mathcal{I}$ in sequence order
Slice hidden states, attention mask, position ids, and position embeddings by $\mathcal{I}$
return pruned sequence states and metadata

C Algorithmic Details

We provide the detailed procedures of our Visual Evidence Preservation (VEP) Alg. 1 and Behavior-Preserving Decoder Pruning (BPDP) in Alg. 2, respectively.

D Limitations

Although GroundPrune is training-free and does not use ground-truth boxes at test time, it requires a one-time offline search for each backbone to select the routing head whose pruning departs least from the original model. This calibration uses a small set of sampled examples, but the selected routing point may not directly transfer to different VLM architectures or substantially different input processing pipelines. In addition, although GroundPrune improves grounding robustness under pruning, its localization predictions can still be imperfect and should be treated as informative references rather than standalone decisions, especially in sensitive or high-impact applications. Future work could remove this calibration step through fully automatic or sample-adaptive routing-point selection.