DIAGNOSING QUERY AWARENESS IN VISUAL TOKEN PRUNING: FROM BIASED IMPORTANCE SIGNALS TO CALIBRATED SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual-token pruning for VLMs relies on importance scores that are often treated as proxies for question-relevant evidence. Popular scores can be divided into image-conditioned visual priors and query-conditioned signals represented by early text-visual attention. Their high aggregate fidelity has been reported, yet aggregate fidelity alone cannot reveal what these scores measure and limits understanding. In this work, we conduct a structured diagnostic analysis along three axes: prompt sensitivity, selection-behavior preference, and depth-wise query formation. Across multiple VLMs, early text-visual attention exhibits little query discrimination despite being prompt-conditioned in construction. It behaves as an image-conditioned local-detail prior, while visual priors favor broader spatial coverage; the two families succeed on complementary global- and local-evidence regimes. Diversity-aware selection reduces redundancy but does not remove these inherited biases. Reliable query-discriminative attention emerges only in middle-to-late language-model layers. These findings motivate Query-Calibrated Pruning (Q-Cal), a two-stage framework for multi-round VLM inference. Q-Cal first forms a reusable visual prefix from complementary early priors, then refines it with a calibrated mixture of deeper text-visual attention once query-specific evidence has formed. Across six Qwen-family VLMs and eight benchmarks, Q-Cal improves both the accuracy and efficiency of visual pruning for multi-round questions. Code is available at `https://github.com/anonymous733882/q-cal-submission`.

1 INTRODUCTION

Multimodal large language models (MLLMs) represent an image as hundreds or thousands of visual tokens, making visual processing a major source of inference cost. Visual token pruning addresses this bottleneck by ranking visual tokens with an importance signal and retaining only a subset for subsequent language-model computation.

Existing pruning signals fall into two broad families. Query-conditioned signals are most commonly early text-visual attention or prompt-image similarity, to estimate which visual tokens are relevant to the current question (Chen et al., 2024; Zhu et al., 2026; Huang et al., 2026; Zhang et al., 2025a). The second family consists of image-conditioned visual priors, such as visual-encoder saliency or representation leverage, to preserve visually salient or structurally representative tokens without considering the prompt (Apedo et al., 2026; Alvar et al., 2025; Zhang et al., 2025b). These signals can be applied near the language model (LM) input to maximize savings or used in progressive pruning that refines token retention across layers or decoding steps (Xing et al., 2025; Zhang et al., 2025c; Khaki et al., 2025).

Although they are widely adopted, their behavior and applicability boundaries remain insufficiently understood. Query-conditioned signals assume that incorporating the prompt enables them to identify query-relevant evidence, for example, early text-visual attention. However, reported results already expose a tension. On Qwen2.5-VL-7B and 90% visual-token pruning, the image-conditioned method DivPrune scores 81.5% on POPE, whereas HAWK, which uses first-layer text-visual attention logits with head-importance weighting, scores 77.9% (Zhu et al., 2026; Alvar et al., 2025).

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.

Why can image-conditioned pruning outperform an early attention-based selector that has access to the question? A benchmark score alone cannot answer this question. High or comparable aggregate fidelity can arise from retaining broadly useful image regions, from complementary failures, or from larger token budgets compensating for inaccurate ranking. It therefore remains unclear whether early rankings respond when different questions require different evidence, whether visually derived and text-attention scores preserve the same evidence, and whether diversity-aware selection corrects a biased estimate or merely reduces redundancy.

This work asks three diagnostic questions. First, do nominally query-conditioned early attention scores respond to different questions about the same image? Second, if early scores are not fully query-discriminative, what visual preferences do they encode and where do they fail? Third, at what point in the decoder does text-visual attention become reliable enough for aggressive query-specific pruning? Our controlled diagnostics provide a consistent answer. Early text-visual attention is almost prompt-invariant and behaves as an image-conditioned local-detail prior. In contrast, visual priors favor broader spatial coverage. These two signal families are complementary and therefore have complementary failure regions under low retention, explaining how they can achieve similar aggregate fidelity while succeeding on different task subsets. Diversity suppression reduces redundancy to improve performance but cannot eliminate the inherited bias or make the resulting ranking reliably query-discriminative. In contrast, query-discriminative relevance emerges gradually in deeper decoder layers, where text-visual attention becomes a substantially stronger pruning signal.

These findings reveal a timing dilemma. Pruning around the LM input is cheap but only utilizes signals that mostly behave as image-conditioned priors; deeper layer pruning gives access to more reliable question-specific relevance but processes the full visual sequence for longer. We resolve this trade-off with Query-Calibrated Pruning (Q-Cal), a diagnosis-driven two-stage framework for multi-round VLM inference. Q-Cal first treats complementary early local- and global-coverage signals as conservative visual priors for constructing a reusable candidate prefix. Once sufficient cross-modal interaction has occurred, Q-Cal applies a calibrated mixture of deeper text-visual attention layers to refine this candidate set into a compact query-specific subset for the remaining prefill computation and decoding. Across six Qwen-family VLMs and eight benchmarks, Q-Cal improves the accuracy-efficiency trade-off under strict cost constraints while retaining strong fidelity.

This work makes three main contributions:

- We introduce a structured diagnostic framework that evaluates pruning signals along three axes: same-image question sensitivity, selection-behavior preference, and depth-wise query formation. Across multiple VLMs, early text-visual attention exhibits limited query discrimination despite being prompt-conditioned by construction.
- We characterize the encoded preferences and applicability boundaries of representative early pruning signals. Early text-visual attention favors local details, whereas visual representation signals favor broader spatial coverage. Diversity-aware selection reduces redundancy without eliminating these underlying biases, and substantially stronger query discrimination emerges in deeper LM layers.
- Guided by this diagnosis, we propose Q-Cal, which constructs a reusable candidate prefix from complementary early priors and performs query-specific calibration using deep text-visual attention during prefill. Experiments across six Qwen-family VLMs and eight benchmarks demonstrate an improved accuracy–efficiency trade-off at low retention rates.

2 Related Work

Progressive pruning. Progressive visual-token pruning refines token selection across LM layers or decoding steps, allowing later activations to guide the decision (Xing et al., 2025; Zhang et al., 2025c). Layerwise pruning recipes can also be fitted from attention statistics (Ye et al., 2025). While later context can improve reliability, it trades off some computational savings.

One-shot pruning. One-shot methods rank visual tokens once, typically near the LM input, using either query-conditioned scores or image-conditioned visual priors. Query-conditioned scores mainly rely on early text-visual attention or prompt–image similarity (Chen et al., 2024; Zhu et al., 2026; Huang et al., 2026; Zhang et al., 2025a). In contrast, visual priors use saliency, representation leverage, or structural coverage criteria (Apedo et al., 2026; Alvar et al., 2025; Zhang et al., 2025b).

Other selectors combine attention-derived importance with diversity (Zhang et al., 2026; Baek et al., 2026) or use attention-free pruning and clustering (Dhouib et al., 2025). A query-conditioned signal might be expected to capture query-relevant tokens, but this assumption has received limited direct validation.

Hybrid and staged selection. Beyond simple one-shot selection, prior work has explored instruction-guided pruning (Huang et al., 2024), positional-bias correction (Wang et al., 2026), staged token pruning (Takezoe et al., 2026; Khaki et al., 2025). These studies emphasize aggregate accuracy, redundancy, cache efficiency, or positional bias, but leave open how signal behavior changes with the question and decoder depth. Our work studies these questions directly and uses the diagnosis to design signal roles.

3 METHOD

3.1 PROBLEM SETUP

Given an image I and a sequence of questions $Q = \{q_1, \ldots, q_R\}$, a VLM constructs visual tokens $V = \{v_1, \ldots, v_n\}$ and text tokens for each question. For any question q_r, we assume there exists a minimal visual-token set $E_r^\star \subseteq V$ that contains the visual evidence needed to support the correct answer after pruning. Visual token pruning aims to retain a subset that covers this evidence while reducing the computation and inference cost. We quantify the inference cost of the retained tokens by the cumulative number of token instances processed across layers:

$$C = \sum_{\ell=1}^{L} |\mathcal{T}_\ell|, \tag{1}$$

where $\mathcal{T}_\ell$ is the activated token set processed at LM layer ℓ in both pre-filling and decoding stage.

3.2 DIAGNOSING IMPLICIT PRUNING PRIORS

The minimal evidence set $E_r^\star$ is unobserved, so pruning methods rely on proxy scores to decide which visual tokens should be retained. Existing studies usually validate these scores with a fixed pruning recipe, using aggregate accuracy or efficiency as evidence. This leaves a diagnostic gap: good average fidelity does not reveal whether a score is selecting question-specific evidence, preserving broadly useful image priors, or merely retaining enough redundant tokens to mask ranking errors. For better pruning, it is necessary to identify what each score actually measures.

Our analysis focuses on two representative signals: image-based visual priors and nominally query-conditioned signals, represented by early text-visual attention. The diagnosis identifies two properties that directly determine whether a score can approximate $E_r^\star$ under aggressive pruning. First, a query-relevance score should change when the question about a fixed image changes; otherwise it cannot distinguish different evidence sets required by different questions. Second, different score families may preserve different visual evidence even when their average accuracy is similar, so we test whether the two signals exhibit different task preferences.

Prompt invariance. We first test whether a nominally query-conditioned signal actually responds to questions. For two prompts on the same image, let $s_a, s_b \in \mathbb{R}^n$ be the corresponding visual-token score vectors over the same n visual tokens, and let $r_a(i)$ and $r_b(i)$ be the ranks of token i. We measure agreement by top-k Jaccard overlap and full-ranking Spearman correlation:

$$J_k(s_a, s_b) = \frac{|\mathrm{Top}_k(s_a) \cap \mathrm{Top}_k(s_b)|}{|\mathrm{Top}_k(s_a) \cup \mathrm{Top}_k(s_b)|}, \qquad \rho(s_a, s_b) = \frac{\sum_{i=1}^{n} \big(r_a(i) - \bar{r}_a\big)\big(r_b(i) - \bar{r}_b\big)}{\sqrt{\sum_{i=1}^{n} \big(r_a(i) - \bar{r}_a\big)^2} \sqrt{\sum_{i=1}^{n} \big(r_b(i) - \bar{r}_b\big)^2}}. \tag{2}$$

We conducted cross-model diagnostics with Qwen2.5-VL-3B/7B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and LLaVA-1.5-7B (Liu et al., 2023a) using 100 VQAv2 validation images (Goyal et al., 2017), each paired with 5 different prompts. The results are shown in Table 1. Surprisingly, the early text–visual attention score—which has been widely considered as a query-conditioned signal in previous studies—exhibited strong consistency in ranking distribution across

Table 1: Prompt-invariance and cross-signal agreement of visual-token pruning scores across four VLMs on 100 VQAv2 validation images with five questions per image. J_k is top-k token Jaccard overlap; ρ is full-ranking Spearman correlation.

Category	Signal detail	$J_{10\%}$ mean	$J_{10\%}$ var.	$J_{20\%}$ mean	$J_{20\%}$ var.	ρ mean	ρ var.
Early text–visual attention	Decoder layer 0	0.960	0.004883	0.957	0.005518	0.989	0.000370
	Decoder layer 1	0.924	0.002811	0.939	0.000946	0.990	0.000134
Visual representation prior	ViT received-attention saliency	1.000	0.000000	1.000	0.000000	1.000	0.000000
	SVD-Prune leverage saliency	1.000	0.000000	1.000	0.000000	1.000	0.000000
Deep text–visual attention	Mid-depth decoder layer	0.428	0.013579	0.503	0.011231	0.775	0.010403
	75%-depth decoder layer	0.349	0.003954	0.442	0.002396	0.714	0.001090
Cross-signal agreement	Deep attention vs early attention	0.191	0.002329	0.225	0.001170	0.265	0.003712
	Deep attention vs visual representation priors	0.110	0.001099	0.170	0.002867	0.163	0.044289

Note. This table reports the representative signal families used in the main diagnostic analysis. Appendix C provides the expanded signal set and the corresponding qualitative top-token visualization.

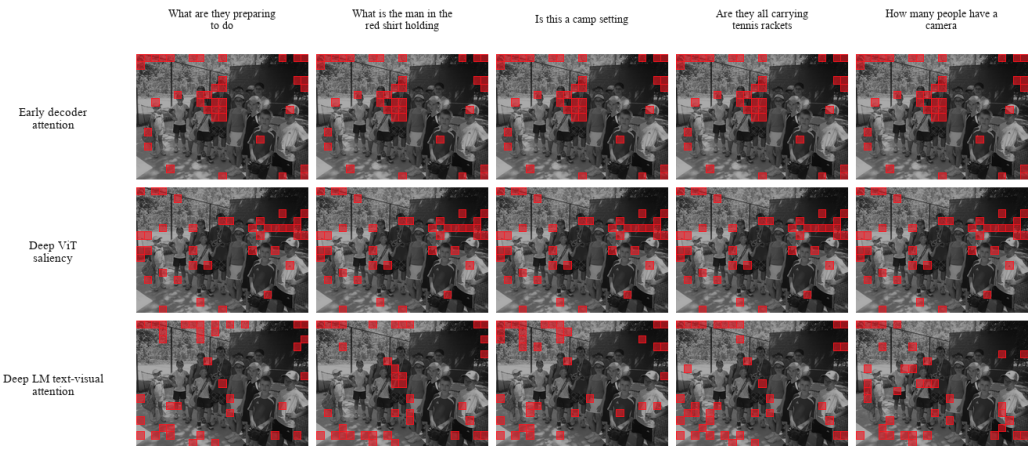


Figure 1: Top-10% visual-token selections for one VQAv2 image under five questions. Columns are questions, and rows are signals. Unselected regions are dimmed; selected token regions are marked in red. A more comprehensive signal visualization is provided in Appendix C.

prompts for the same image. The average J10%, J20%, and Spearman's correlation coefficients were close to 1.00, with variance close to 0. This indicates that the early text–visual attention score is closer to an image-conditioned prior rather than a query-conditioned signal. In contrast, the deep text–visual attention score varies closely with the prompt, exhibiting query-conditioned characteristics. Figure 1 qualitatively illustrates the differences among early attention, ViT saliency, and deep attention by showing the top 10% of visual tokens retained by each signal on the same image.

Selection-behavior preference. Prompt sensitivity reveals whether a score changes with the question, but it does not explain what visual property the score measures. To make this distinction observable, we use the following diagnostic assumption. Since early text-visual attention and visual representation priors are prompt-invariant or only weakly prompt-dependent, and since their token rankings have low cross-signal agreement, a task boundary on which the two signals systematically succeed or fail can reveal the visual evidence each signal tends to preserve. We therefore construct a 240-task diagnostic pool and group questions by the type of visual retrieval needed to answer them. Global-like tasks require broad spatial search or scene-level evidence; local-like tasks require focused recognition of fine details, text, or small regions; and no-difference tasks can be answered by either retrieval pattern, such as negative existence questions where both broad coverage and local inspection lead to the same answer. Examples are listed in Appendix F.

Figure 2 shows that the two prompt-invariant signal families preserve different kinds of evidence. At 10% retention, the visual prior remains close to the full-token reference on global-like tasks but fails on local-like tasks, indicating a preference for broad spatial coverage and scene-level evidence. Early attention shows the reverse pattern: it preserves local-like evidence much more reliably, but drops sharply on global-like tasks. Diversity suppression improves both families by reducing redundant selections, yet the arrows in Figure 2(a) show that it does not reverse their basic preference.

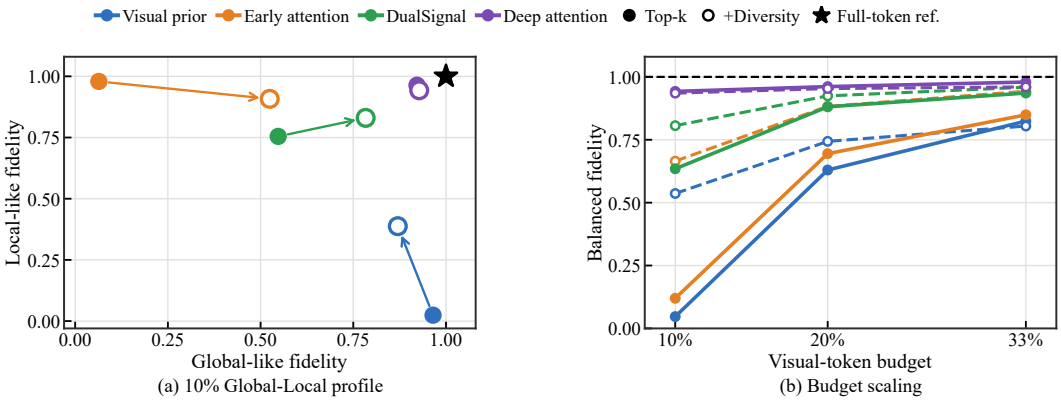


Figure 2: Selection-behavior preference. (a) Global–Local signal profile at 10% visual-token retention. Filled circles denote Top-k selection, open circles add diversity suppression, arrows show the Top-k →+Diversity shift, and the black star marks the full-token reference. (b) Budget scaling, where balanced fidelity is the harmonic mean of global-like and local-like fidelity.

Increasing the budget from 10% to 33% narrows the gap because both signals can retain more evidence, but the scaling trend still reflects the same global-versus-local bias. The complete 240-task table, including overall and no-difference categories, is given in Appendix D.

This complementarity motivates DualSignal as a practical way to mitigate early signal bias: combining visual-prior and early-attention scores gives a more balanced selection than either component alone, especially as the token budget increases. However, DualSignal is still built from two weakly prompt-dependent rankings, so at low budgets it remains a bias-mitigation strategy rather than a fully query-specific relevance signal. Deep text–visual attention serves as the stronger reference point in this diagnostic. It stays closest to the full-token reference across both global-like and local-like tasks, showing that once sufficient cross-modal computation has occurred, attention can provide a more accurate and balanced pruning signal than either early signal family or their simple mixture.

Depth-wise query formation. As a third diagnostic, we ask when text-visual attention becomes reliable enough to serve as a query-discriminative pruning signal. On the same diagnostic pool, text-visual attention from deeper decoder layers was used as the pruning score, with 10%, 20%, and 33% visual-token retention. Figure 3 summarizes the cross-model behavior: early decoder layers remain clearly below the full-token reference, especially under the 10% budget, while the mean curves and envelope rise sharply around the middle-to-late layers. Accuracy then saturates around the 60–75% depth range, and moving to the very last layers does not provide a consistent additional gain. Thus, query-discriminative visual relevance is not present immediately after the image enters the LM; it forms gradually through the prefill computation and becomes practically useful in the middle-to-late decoder layers.

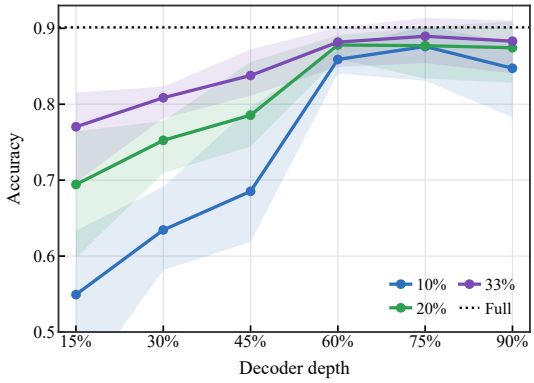


Figure 3: Depth-wise text–visual attention on the 240-task diagnostic pool. Lines show cross-model means; shaded regions denote the cross-model envelope.

3.3 From Bias Diagnostics to Q-Cal

Based on the above analysis, the commonly used visual prior and early attention signals are biased image-conditioned signals rather than query-conditioned ones. Clearly, bias results in distinct weak zones for both signals when dealing with diverse, multi-round problems. Hybrid sampling signals can resolve this bias with a slightly larger visual token budget, enabling the pruned tokens to incorporate both global and local information. However, this conservative retained set can be much larger

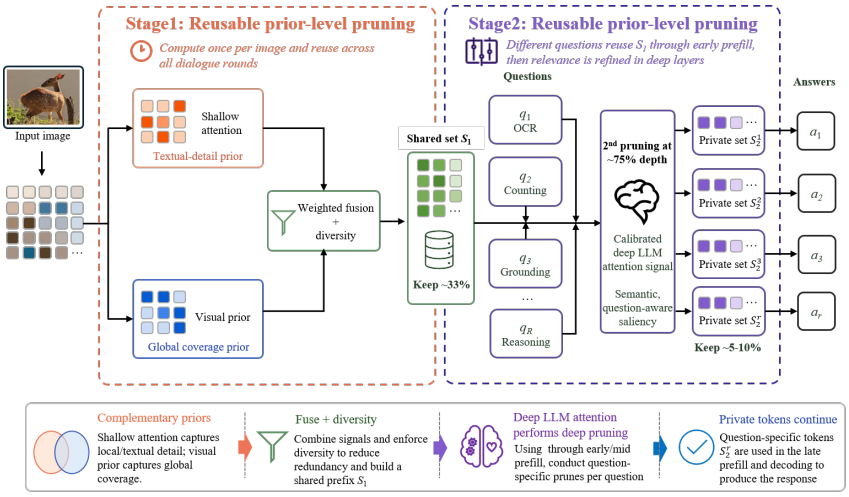


Figure 4: Q-Cal overview. Stage 1 combines complementary early-decoder and Deep-ViT priors to build a reusable image prefix cache. For each question, Stage 2 applies the calibrated deep attention policy π_m inside that prefix and produces a compact question-specific visual-token set.

than the target evidence set $E_r^\star$. It leads to unnecessary overhead in multi-turn tasks, especially those requiring long responses. The question then becomes: Is it possible to approximate $E_r^\star$ at low cost?

Based on this, we propose a two-stage dual-signal pruning method, Q-Cal: Stage 1. We combine two low-cost complementary signals—the visual prior and early attention—to retain a conservative visual subset, thereby preserving candidate evidence for possible questions. Stage 2. After the conservative visual token set is fed into the LM, it remains accessible to each new question q_r in every round at the early and mid-level (top 75%) attention layers. Subsequently, based on deep attention scores, visual tokens are further pruned toward an approximation of $E_r^\star$, which becomes the only visual evidence accessible in the deeper attention layers and the subsequent answer generation process. An overview of this pruning method is shown in Figure 4.

3.4 Implementation Details of Q-Cal

3.4.1 Stage-1 Shared Prefix Selection

Stage 1 uses complementary but question-invariant visual priors to build a broad, reusable visual prefix before the per-question round begins. Let $a_0(i)$ be the early local textual/detail prior for visual token i, and let $g(i)$ be a global coverage prior. A generic stage-1 score is:

$$s_1(i) = \lambda \cdot \mathrm{norm}(a_0(i)) + (1-\lambda) \cdot \mathrm{norm}(g(i)). \tag{3}$$

The fused score is then passed through a diversity-aware selector rather than a plain top-k rule. Starting from an empty set, Q-Cal repeatedly selects the highest remaining token under the current score and suppresses nearby tokens in visual-embedding space:

$$s_1^{(t+1)}(j) = \max\Big(0,\ s_1^{(t)}(j) - s_1^{(t)}(i_t)\exp\Big[-\gamma\big(1-\cos(z_j, z_{i_t})\big)^2\Big]\Big), \qquad i_t = \arg\max_j s_1^{(t)}(j), \tag{4}$$

where z_i denotes the visual-token embedding and already selected tokens are excluded from later steps. This iterative selection stops when $|S_1| = \rho_1 n$, producing the conservative visual subset $S_1 \subseteq \{1, \ldots, n\}$. In this work, $\lambda = 0.5$, $\rho_1 = 33\%$, and $\gamma = 20$.

3.5 Stage-2 Deep Attention Calibration

Stage 2 refines the shared prefix S_1 into a question-specific subset $S_2^{(r)} \subseteq S_1$ that approximates $E_r^\star$ under a lower final retention rate ρ_2. Instead of choosing one decoder layer by hand, Q-Cal calibrates a sparse mixture of deeper text-visual attention layers:

$$s_2(i \mid q_r) = \sum_{\ell \in \mathcal{L}} \alpha_\ell A_\ell(i \mid I, q_r), \tag{5}$$

where A_ℓ is text-visual attention at language-model layer ℓ and $\alpha_\ell \geq 0$, $\sum_\ell \alpha_\ell = 1$.

The calibration pipeline has three steps. First, it builds a visually grounded sample set spanning eight broad capabilities: object and scene recognition, multi-image visual comparison, scene-text reading, document understanding, chart reasoning, diagrammatic reasoning, knowledge-grounded visual reasoning, and hallucination-sensitive verification. Second, it derives budget-specific oracle token sets from gradient sensitivity under teacher forcing. Third, it fits the layer weights so that the top-ranked tokens under s_2 overlap with these oracle sets, with a cost penalty that discourages unnecessarily late second-stage pruning. At inference time, Q-Cal applies the calibrated score only inside S_1 and keeps the top $\rho_2 n$ visual tokens for the remaining layers and answer generation. Appendix A gives the full calibration objective, filtering criteria, and cost proxy.

4 Experiments

4.1 Models and Benchmarks

We evaluate six Qwen-family VLM backbones (Wang et al., 2024; Bai et al., 2025b;a): Qwen2-VL-2B, Qwen2-VL-7B, Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B, and Qwen3-VL-8B. The evaluation matrix includes general visual question answering and commonsense recognition (GQA, VQAv2, Visual7W, VisualGenomeQA) (Hudson & Manning, 2019; Goyal et al., 2017; Zhu et al., 2016; Krishna et al., 2017), object-existence and hallucination-sensitive verification (POPE) (Li et al., 2023), compositional and counting-oriented reasoning (TallyQA) (Acharya et al., 2019), explicit spatial grounding (GQA-Grounding) (Hudson & Manning, 2019), and multi-turn image-grounded dialogue (InViG) (Zhang et al., 2024a). Each image is paired with up to 15 questions as a multi-round task. Each backbone apply the calibrated policy weights obtained from deep attention calibration. The calibrated policy weights are reported in Appendix B. The ρ_2 set as 5% and 10%.

4.2 Baselines

We compare against representative pruning families: FastV (Chen et al., 2024), ID-Selection (Huang et al., 2026), ZSPAPrune (Zhang et al., 2025a), SVD-Prune (Apedo et al., 2026), DivPrune (Alvar et al., 2025), VisPruner and FasterVLM (Zhang et al., 2025b; 2024b), AgilePruner (Baek et al., 2026), PACT (Dhouib et al., 2025), PTP (Liang et al., 2025), and SparseVILA (Khaki et al., 2025).

4.3 Main Results

We first evaluate methods under strict matching of the computation cost C defined in Section 3.1. This is the most direct comparison of pruning quality: each baseline is rerun with the same C as our method for the corresponding model, benchmark, and final budget. Table 2 reports the model-level matrix across the six backbones, with full-token baselines shown in the model header rows. Under the same cost, all baseline families fall below our method.

Then, we release the computation cost restriction from both pre-filling and decoding to pre-filling only. Figure 5 summarizes the resulting speedup–fidelity trade-off averaged over the six backbones: the left strips encode token-layer speedup and the right strips encode average fidelity, where longer is better. In this setting, several baselines can approach Q-Cal in fidelity, but they have substantially higher cost; Q-Cal therefore retains the same accuracy regime with a clear cost advantage.

4.4 Ablation Experiments

4.4.1 Stage-1 Budget

To assess whether the Stage-1 prefix budget of 33% is sufficient to retain evidence for diverse downstream questions, we evaluated the pruning effect on the capacity-balanced evaluation set across models. From Figure 6a, the cross-model mean accuracy rises sharply when the budget is expanded from 5% to 33% (0.426 to 0.563), indicating that lower budgets remove evidence needed by questions. Past 33%, the marginal gain is small: 50% retention reaches 0.577, only 0.014 above 33%. Therefore, 33% is used as the shared Stage-1 prefix budget: it preserves most of the evidence needed for diverse queries while keeping the downstream token-layer cost low.

Table 2: Strict C-matched comparison. Method rows report paired 5% / 10% accuracy. The TL/Ret column gives the final retained-token for Q-Cal and the equivalent retention used for each baseline.

Method	TL/Ret. (%)	GQA	POPE	VQAv2	V7W	VG-QA	TallyQA	GQA-G	InViG	Avg. Fid. (%)
Qwen2-2B		0.625	0.878	0.724	0.466	0.475	0.851	0.702	0.073	-/-
AgilePruner	12.28 / 15.68	0.544 / 0.555	0.765 / 0.779	0.623 / 0.646	0.377 / 0.397	0.393 / 0.413	0.733 / 0.758	0.592 / 0.613	0.055 / 0.054	85.1 / 88.0
DivPrune	12.27 / 15.63	0.525 / 0.541	0.733 / 0.750	0.585 / 0.609	0.359 / 0.379	0.376 / 0.391	0.707 / 0.726	0.575 / 0.592	0.050 / 0.049	81.6 / 84.3
FastV	12.23 / 15.62	0.527 / 0.550	0.739 / 0.756	0.582 / 0.613	0.355 / 0.378	0.371 / 0.395	0.702 / 0.736	0.570 / 0.591	0.054 / 0.056	81.5 / 85.0
FasterVLM	12.71 / 15.78	0.564 / 0.581	0.797 / 0.808	0.651 / 0.666	0.408 / 0.419	0.412 / 0.429	0.820 / 0.832	0.645 / 0.658	0.052 / 0.051	90.8 / 92.8
ID-Selection	12.72 / 15.31	0.594 / 0.600	0.805 / 0.813	0.682 / 0.690	0.422 / 0.431	0.446 / 0.456	0.812 / 0.820	0.664 / 0.666	0.060 / 0.060	93.5 / 94.7
PACT	12.28 / 15.69	0.585 / 0.597	0.810 / 0.821	0.667 / 0.676	0.415 / 0.429	0.442 / 0.452	0.832 / 0.847	0.660 / 0.667	0.059 / 0.061	93.3 / 95.0
PTP	12.26 / 15.63	0.581 / 0.594	0.806 / 0.812	0.663 / 0.676	0.410 / 0.430	0.447 / 0.456	0.801 / 0.810	0.658 / 0.670	0.058 / 0.059	92.3 / 94.2
SVD-Prune	12.20 / 15.60	0.554 / 0.576	0.734 / 0.761	0.628 / 0.652	0.388 / 0.400	0.393 / 0.418	0.756 / 0.784	0.631 / 0.648	0.056 / 0.058	86.5 / 89.6
SparseVILA	12.60 / 16.06	0.312 / 0.332	0.362 / 0.363	0.469 / 0.490	0.326 / 0.339	0.329 / 0.346	0.648 / 0.832	0.645 / 0.658	0.084 / 0.084	66.3 / 72.0
VisPruner	12.34 / 15.80	0.567 / 0.582	0.800 / 0.812	0.653 / 0.668	0.410 / 0.422	0.434 / 0.448	0.823 / 0.838	0.646 / 0.661	0.050 / 0.052	91.5 / 93.7
ZSPAPrune	12.28 / 15.65	0.551 / 0.577	0.738 / 0.752	0.635 / 0.657	0.390 / 0.404	0.396 / 0.406	0.720 / 0.753	0.612 / 0.632	0.054 / 0.054	85.5 / 88.3
Ours	5% / 10%	**0.610 / 0.615**	**0.858 / 0.857**	**0.712 / 0.714**	**0.458 / 0.456**	**0.460 / 0.465**	**0.850 / 0.850**	**0.675 / 0.675**	**0.101 / 0.098**	**98.5 / 98.7**
Qwen2-7B		0.646	0.884	0.753	0.492	0.507	0.881	0.718	0.110	-/-
AgilePruner	11.72 / 14.99	0.567 / 0.589	0.845 / 0.855	0.656 / 0.679	0.393 / 0.413	0.412 / 0.428	0.767 / 0.792	0.625 / 0.640	0.100 / 0.104	87.5 / 90.1
DivPrune	11.66 / 14.91	0.534 / 0.556	0.845 / 0.857	0.590 / 0.612	0.366 / 0.386	0.394 / 0.411	0.750 / 0.769	0.602 / 0.623	0.100 / 0.102	83.7 / 86.5
FastV	11.68 / 14.95	0.543 / 0.558	0.812 / 0.832	0.643 / 0.665	0.380 / 0.397	0.392 / 0.412	0.738 / 0.765	0.592 / 0.613	0.100 / 0.102	84.1 / 87.0
FasterVLM	12.21 / 14.87	0.604 / 0.607	0.871 / 0.872	0.687 / 0.695	0.433 / 0.462	0.453 / 0.461	0.857 / 0.875	0.679 / 0.695	0.104 / 0.106	93.9 / 95.7
ID-Selection	12.41 / 15.81	0.613 / 0.619	0.861 / 0.869	0.693 / 0.709	0.455 / 0.469	0.449 / 0.461	0.843 / 0.857	0.681 / 0.693	0.106 / 0.109	94.2 / 95.8
PACT	12.24 / 14.86	0.600 / 0.619	0.861 / 0.864	0.687 / 0.700	0.429 / 0.440	0.445 / 0.455	0.848 / 0.857	0.670 / 0.679	0.106 / 0.107	93.1 / 94.6
PTP	11.67 / 14.15	0.594 / 0.610	0.867 / 0.871	0.682 / 0.689	0.429 / 0.439	0.455 / 0.463	0.849 / 0.862	0.673 / 0.686	0.101 / 0.103	93.1 / 94.6
SVD-Prune	12.25 / 15.62	0.605 / 0.618	0.838 / 0.857	0.664 / 0.680	0.400 / 0.423	0.421 / 0.440	0.810 / 0.830	0.667 / 0.686	0.104 / 0.107	90.4 / 92.9
SparseVILA	12.63 / 16.10	0.429 / 0.434	0.871 / 0.872	0.592 / 0.600	0.365 / 0.462	0.371 / 0.377	0.857 / 0.875	0.481 / 0.695	0.073 / 0.071	80.9 / 87.8
VisPruner	12.22 / 14.47	0.602 / 0.615	**0.873 / 0.872**	0.687 / 0.694	0.433 / 0.462	0.450 / 0.463	0.861 / **0.877**	0.680 / 0.696	0.105 / 0.107	93.9 / 95.8
ZSPAPrune	11.62 / 14.88	0.588 / 0.593	0.855 / 0.866	0.665 / 0.684	0.411 / 0.426	0.430 / 0.442	0.800 / 0.828	0.640 / 0.664	0.108 / 0.111	90.1 / 92.5
Ours	5% / 10%	**0.629 / 0.633**	0.870 / 0.869	**0.740 / 0.744**	**0.481 / 0.483**	**0.501 / 0.503**	**0.863** / 0.860	**0.699 / 0.703**	**0.153 / 0.156**	**98.9 / 99.2**
Qwen2.5-3B		0.616	0.877	0.723	0.467	0.473	0.861	0.680	0.085	-/-
AgilePruner	13.37 / 17.37	0.415 / 0.427	0.730 / 0.748	0.512 / 0.533	0.302 / 0.319	0.306 / 0.324	0.773 / 0.807	0.485 / 0.509	0.018 / 0.019	74.1 / 77.1
DivPrune	14.05 / 18.04	0.407 / 0.418	0.666 / 0.692	0.485 / 0.507	0.291 / 0.305	0.301 / 0.315	0.763 / 0.787	0.467 / 0.488	0.015 / 0.016	70.9 / 73.7
FastV	13.38 / 17.31	0.408 / 0.417	0.702 / 0.718	0.498 / 0.524	0.286 / 0.313	0.295 / 0.316	0.739 / 0.762	0.463 / 0.484	0.016 / 0.019	71.2 / 74.2
FasterVLM	13.56 / 17.52	0.439 / 0.451	0.717 / 0.730	0.530 / 0.550	0.326 / 0.333	0.336 / 0.348	0.808 / 0.826	0.514 / 0.524	0.020 / 0.022	77.1 / 79.1
ID-Selection	12.35 / 15.67	0.576 / 0.585	0.743 / 0.752	0.678 / 0.560	0.416 / 0.433	0.425 / 0.336	0.837 / 0.850	0.515 / 0.527	0.019 / 0.020	88.0 / 84.9
PACT	13.27 / 17.21	0.439 / 0.446	0.736 / 0.750	0.554 / 0.565	0.416 / 0.327	0.336 / 0.348	**0.848 / 0.854**	0.518 / 0.527	0.018 / 0.018	80.8 / 80.3
PTP	13.45 / 17.31	0.410 / 0.421	0.702 / 0.728	0.505 / 0.527	0.286 / 0.308	0.298 / 0.318	0.743 / 0.775	0.465 / 0.491	0.017 / 0.019	71.6 / 74.9
SVD-Prune	13.56 / 17.29	0.405 / 0.417	0.670 / 0.692	0.454 / 0.477	0.263 / 0.285	0.284 / 0.306	0.791 / 0.798	0.455 / 0.478	0.014 / 0.016	69.7 / 72.6
SparseVILA	12.08 / 15.63	0.523 / 0.526	0.827 / 0.840	0.607 / 0.619	0.377 / 0.389	0.380 / 0.391	0.779 / 0.793	0.586 / 0.600	0.090 / 0.088	87.1 / 88.8
VisPruner	13.60 / 17.49	0.439 / 0.450	0.716 / 0.731	0.531 / 0.550	0.327 / 0.333	0.336 / 0.348	0.809 / 0.826	0.514 / 0.523	0.019 / 0.021	77.1 / 79.1
ZSPAPrune	13.30 / 18.02	0.428 / 0.436	0.700 / 0.716	0.523 / 0.540	0.324 / 0.329	0.325 / 0.336	0.813 / 0.819	0.495 / 0.509	0.017 / 0.018	75.8 / 77.4
Ours	5% / 10%	**0.590 / 0.597**	**0.875 / 0.877**	**0.697 / 0.706**	**0.438 / 0.443**	**0.444 / 0.453**	0.846 / 0.850	**0.660 / 0.663**	**0.142 / 0.138**	**98.2 / 98.8**
Qwen2.5-7B		0.616	0.875	0.742	0.475	0.483	0.866	0.680	0.066	-/-
AgilePruner	12.62 / 16.16	0.453 / 0.468	0.801 / 0.812	0.587 / 0.609	0.337 / 0.352	0.336 / 0.352	0.811 / 0.822	0.511 / 0.531	0.030 / 0.034	80.5 / 82.8
DivPrune	12.44 / 15.89	0.440 / 0.456	0.703 / 0.721	0.544 / 0.566	0.314 / 0.329	0.326 / 0.342	0.746 / 0.776	0.467 / 0.489	0.028 / 0.028	74.3 / 77.2
FastV	12.58 / 16.05	0.445 / 0.463	0.787 / 0.802	0.565 / 0.590	0.325 / 0.341	0.331 / 0.341	0.788 / 0.800	0.485 / 0.515	0.030 / 0.032	78.2 / 81.0
FasterVLM	12.40 / 15.91	0.504 / 0.511	0.791 / 0.800	0.609 / 0.619	0.371 / 0.378	0.373 / 0.379	0.840 / 0.841	0.546 / 0.566	0.034 / 0.036	84.7 / 86.0
ID-Selection	12.31 / 15.78	0.588 / 0.599	0.819 / 0.831	0.621 / 0.626	0.372 / 0.388	0.373 / 0.385	0.836 / 0.855	0.650 / 0.662	0.033 / 0.036	89.3 / 91.3
PACT	12.51 / 16.00	0.491 / 0.502	0.817 / 0.829	0.616 / 0.626	0.369 / 0.378	0.369 / 0.379	0.824 / 0.840	0.550 / 0.562	0.033 / 0.035	84.8 / 86.5
PTP	12.56 / 16.10	0.454 / 0.472	0.794 / 0.812	0.579 / 0.598	0.334 / 0.349	0.334 / 0.347	0.795 / 0.803	0.503 / 0.527	0.030 / 0.030	79.7 / 82.0
SVD-Prune	12.79 / 15.87	0.604 / 0.514	0.813 / 0.822	0.592 / 0.609	0.355 / 0.369	0.365 / 0.374	0.831 / 0.837	0.659 / 0.669	0.034 / 0.034	88.5 / 88.2
SparseVILA	12.60 / 16.06	0.394 / 0.399	0.773 / 0.754	0.532 / 0.548	0.328 / 0.340	0.337 / 0.345	0.736 / 0.736	0.433 / 0.437	0.031 / 0.034	74.3 / 74.8
VisPruner	12.41 / 15.85	0.503 / 0.512	0.792 / 0.800	0.608 / 0.620	0.370 / 0.378	0.373 / 0.380	0.838 / 0.840	0.547 / 0.562	0.035 / 0.037	84.7 / 86.0
ZSPAPrune	12.42 / 15.81	0.494 / 0.504	0.800 / 0.808	0.621 / 0.626	0.365 / 0.374	0.364 / 0.376	0.821 / 0.833	0.544 / 0.555	0.035 / 0.038	84.2 / 85.7
Ours	5% / 10%	**0.609 / 0.610**	**0.873 / 0.873**	**0.725 / 0.726**	**0.470 / 0.467**	**0.480 / 0.483**	**0.864 / 0.864**	**0.672 / 0.673**	**0.104 / 0.104**	**99.8 / 100.0**
Qwen3-4B		0.639	0.901	0.744	0.482	0.490	0.887	0.716	0.069	-/-
AgilePruner	13.33 / 17.01	0.518 / 0.541	0.833 / 0.848	0.598 / 0.628	0.358 / 0.381	0.386 / 0.409	0.719 / 0.771	0.574 / 0.605	0.035 / 0.034	81.7 / 85.6
DivPrune	13.37 / 16.97	0.510 / 0.522	0.756 / 0.775	0.579 / 0.605	0.328 / 0.350	0.362 / 0.383	0.692 / 0.723	0.556 / 0.573	0.032 / 0.034	77.4 / 80.5
FastV	13.18 / 16.82	0.536 / 0.550	0.848 / 0.861	0.626 / 0.645	0.391 / 0.407	0.409 / 0.421	0.784 / 0.809	0.587 / 0.604	0.039 / 0.041	85.7 / 88.0
FasterVLM	13.34 / 16.97	0.572 / 0.585	0.875 / 0.883	0.656 / 0.675	0.404 / 0.419	0.433 / 0.444	0.810 / 0.832	0.650 / 0.658	0.036 / 0.040	89.9 / 92.0
ID-Selection	13.87 / 17.12	0.584 / 0.592	0.893 / 0.905	0.673 / 0.679	0.426 / 0.435	0.445 / 0.451	0.843 / 0.856	0.660 / 0.667	0.044 / 0.046	92.7 / 94.0
PACT	13.38 / 16.27	0.574 / 0.582	0.891 / 0.903	0.658 / 0.677	0.415 / 0.436	0.442 / 0.446	0.838 / 0.852	0.667 / 0.640	0.044 / 0.045	91.9 / 93.0
PTP	13.29 / 16.89	0.547 / 0.560	0.854 / 0.857	0.641 / 0.656	0.400 / 0.409	0.411 / 0.426	0.793 / 0.809	0.603 / 0.614	0.039 / 0.041	87.0 / 88.6
SVD-Prune	13.89 / 17.05	0.548 / 0.567	0.890 / 0.901	0.623 / 0.654	0.390 / 0.416	0.420 / 0.434	0.822 / 0.849	0.629 / 0.646	0.043 / 0.045	88.6 / 91.6
SparseVILA	12.99 / 16.43	0.592 / 0.600	0.884 / 0.891	0.675 / 0.688	0.403 / 0.414	0.430 / 0.442	0.797 / 0.824	0.663 / 0.673	0.082 / 0.078	91.9 / 93.7
VisPruner	13.32 / 17.00	0.572 / 0.585	0.879 / 0.884	0.658 / 0.675	0.406 / 0.419	0.433 / 0.444	0.812 / 0.837	0.652 / 0.660	0.037 / 0.041	90.3 / 92.2
ZSPAPrune	13.54 / 17.24	0.564 / 0.577	0.870 / 0.875	0.642 / 0.657	0.398 / 0.408	0.422 / 0.433	0.802 / 0.818	0.630 / 0.638	0.039 / 0.038	88.6 / 90.3
Ours	5% / 10%	**0.628 / 0.630**	**0.906 / 0.906**	**0.729 / 0.731**	**0.475 / 0.473**	**0.489 / 0.488**	**0.882 / 0.880**	**0.699 / 0.703**	**0.111 / 0.114**	**99.8 / 100.0**
Qwen3-8B		0.626	0.903	0.758	0.475	0.499	0.892	0.713	0.087	-/-
AgilePruner	11.97 / 15.73	0.453 / 0.460	0.836 / 0.842	0.536 / 0.542	0.322 / 0.329	0.326 / 0.334	0.786 / 0.805	0.546 / 0.554	0.065 / 0.063	78.2 / 79.3
DivPrune	12.06 / 15.70	0.437 / 0.455	0.798 / 0.815	0.484 / 0.497	0.295 / 0.316	0.301 / 0.321	0.748 / 0.775	0.522 / 0.541	0.065 / 0.061	73.7 / 76.4
FastV	12.00 / 15.67	0.459 / 0.465	0.836 / 0.847	0.510 / 0.526	0.312 / 0.327	0.317 / 0.327	0.784 / 0.801	0.545 / 0.560	0.063 / 0.063	77.2 / 79.0
FasterVLM	12.33 / 15.55	0.601 / 0.610	0.889 / 0.896	0.564 / 0.722	0.455 / 0.461	0.354 / 0.473	0.867 / **0.875**	0.685 / 0.695	0.074 / 0.076	96.3 / 91.4
ID-Selection	12.54 / 15.81	0.488 / 0.604	0.898 / 0.898	0.551 / 0.559	0.347 / 0.350	0.349 / 0.358	0.833 / 0.841	0.676 / 0.690	0.071 / 0.073	87.4 / 85.9
PACT	12.33 / 15.64	0.592 / 0.606	0.897 / **0.900**	0.558 / 0.556	0.344 / 0.445	0.351 / 0.357	0.850 / 0.854	0.593 / 0.600	0.068 / 0.071	87.9 / 86.6
PTP	12.03 / 15.70	0.465 / 0.468	0.849 / 0.848	0.542 / 0.542	0.323 / 0.328	0.325 / 0.332	0.796 / 0.809	0.564 / 0.572	0.058 / 0.061	79.2 / 80.0
SVD-Prune	13.23 / 15.55	0.480 / 0.600	0.887 / 0.897	0.529 / 0.544	0.428 / 0.445	0.458 / 0.473	0.810 / 0.845	0.601 / 0.689	0.071 / 0.074	90.3 / 88.0
SparseVILA	11.91 / 15.55	0.601 / 0.610	0.896 / 0.896	0.707 / 0.722	0.455 / 0.462	0.472 / 0.477	0.867 / **0.875**	0.685 / 0.695	0.088 / 0.090	96.3 / 97.4
VisPruner	12.33 / 15.55	0.601 / 0.610	0.889 / 0.895	0.720 / 0.564	0.454 / 0.461	0.475 / 0.477	0.865 / 0.873	0.685 / **0.696**	0.075 / 0.077	96.3 / 94.0
ZSPAPrune	12.00 / 15.69	0.472 / 0.481	0.868 / 0.884	0.545 / 0.554	0.338 / 0.345	0.350 / 0.359	0.801 / 0.821	0.574 / 0.585	0.063 / 0.068	80.9 / 82.7
Ours	5% / 10%	**0.611 / 0.611**	**0.899** / 0.899	**0.738 / 0.748**	**0.458 / 0.464**	**0.476 / 0.481**	**0.870** / 0.871	**0.691** / 0.693	**0.132 / 0.133**	**98.4 / 98.9**

Note. For one-shot pruning baselines, TL/Ret. reports the equivalent visual-token retention rate used to match Q-Cal's C. For SparseVILA, it reports the equivalent prefill-stage visual-token retention rate.

4.4.2 Calibration Source

To assess whether the learned layer mixture depends on a particular calibration dataset, we evaluated the performance of a series of policies calibrated with varying calibration sources. Each source is replaced by another benchmark targeting the same capability family listed in Appendix A Table 3. Figure 6b reports all eight single-source substitutions and the two-, three-, and four-source replacements under the same validation test. Across budgets, most settings remain within an absolute deviation of 0.01 from the balanced policy, with a maximum observed deviation of 0.016. This suggests that calibration benefits from covering diverse visual-evidence regimes rather than memorizing specific calibration datasets.

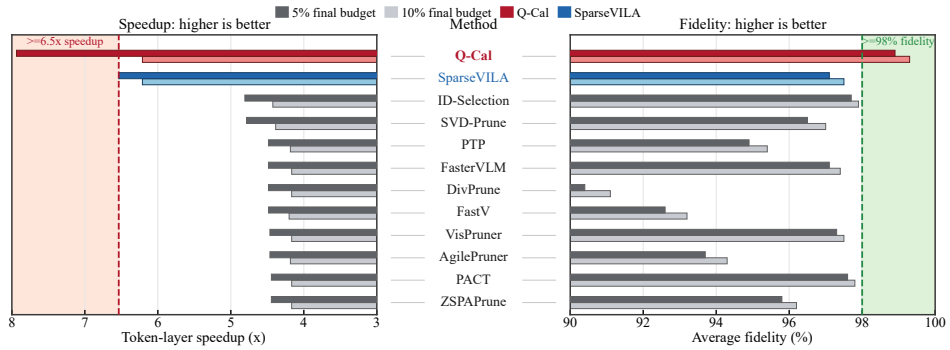


Figure 5: Speedup–fidelity comparison under the prefill-equivalent setting. Dark and light strips denote 5% and 10% final budgets, respectively. The red band marks the region above 6.5x token-layer speedup, and the green band marks the fidelity region above 98%.

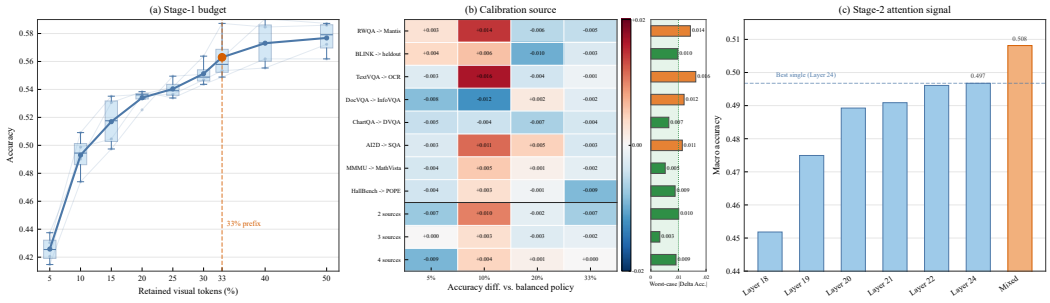


Figure 6: Ablation experiments. (a) Stage-1 budget sweep, showing cross-model accuracy distributions and the mean trend as the prefix budget increases. (b) Calibration-source substitution, showing accuracy changes relative to the balanced calibration policy across token budgets; the side bars report the worst-case absolute deviation. (c) Stage-2 attention signal comparison, showing aggregate accuracy for single-layer controls and the calibrated mixed policy under the same setting.

4.4.3 Stage-2 Attention Signal

To test whether the second pruning stage benefits from a calibrated multi-layer attention signal, we compare it with a layer family in which Stage-2 uses only one language-model layer's text-visual attention after the same Stage-1 candidate pool. Figure 6c reports the aggregate result on Qwen2.5-VL-3B at 10% final visual-token retention, using the same samples and pruning pipeline.

The best single-layer control is Layer 24, reaching 0.497 macro accuracy. The calibrated multi-layer Stage-2 signal reaches 0.508, improving the aggregate score by 0.011 under the same retained-token budget. This indicates that the calibrated Stage-2 policy captures complementary token-ranking information distributed across depth rather than merely selecting a lucky attention layer.

5 Conclusion

This work provides a diagnostic account of what commonly used visual-token pruning signals actually preserve and when reliable query-specific signal occurs. The results show that early text-visual attention is largely prompt-invariant and biased toward local details, whereas visual priors preserve broader spatial evidence; query-discriminative attention emerges only deeper in the language model; and diversity-aware selection reduces redundancy but cannot correct the bias. Leveraging the complementary preferences of early signal families, Q-Cal is designed with two pruning stages. Stage 1 keeps a reusable prefix from complementary early priors, then stage 2 applies calibrated deep attention for query-specific refinement at each round. Across six Qwen-family models and eight benchmarks, Q-Cal improves the cost–accuracy frontier at 5–10% final visual-token retention in multi-round vision question-answer tasks.

AI USE STATEMENT

In this work, we used generative AI tools for (i) writing polish, including improving the clarity, grammar, and structure of the manuscript; (ii) retrieval and discovery, including locating and summarizing related work and identifying relevant methods and benchmarks; and (iii) research ideation and execution, including implementing experiments and debugging code. The authors are responsible for the final methods, experiments, citations, figures, and claims in this paper.

ETHICS STATEMENT

The 240-task diagnostic pool was selected from existing, publicly released vision-language benchmarks; we did not collect new images or questions. Two annotators added labels for the visual-evidence retrieval pattern of each selected task. Because the pool was filtered for visual dependence and baseline correctness, it is not representative of the source benchmarks. The original datasets' licenses and privacy restrictions continue to govern use and redistribution of their images and questions.

REPRODUCIBILITY STATEMENT

Section 3 specifies the diagnostic and pruning procedures, and Section 4 reports the evaluation settings and metrics. The appendices provide calibration details, expanded signal diagnostics, baseline reproduction settings, and examples from the annotated task pool. The linked anonymized repository provides the code for implementation and evaluation.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025. doi: 10.1109/CVPR52734.2025.00877.

Yvon Apedo, Martyna Poreba, Michal Szczepanski, and Samia Bouchafa. Beyond attention scores: Svd-based vision token pruning for efficient vision-language models. *arXiv preprint arXiv:2604.11530*, 2026.

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.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

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*, 2024.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14582–14592, 2025. doi: 10.1109/CVPR52734.2025.01359.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

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*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2024.

Zhaohong Huang, Wenjing Liu, Yuxin Zhang, Fei Chao, and Rongrong Ji. Id-selection: Importance-diversity based visual token selection for efficient lvlm inference. *arXiv preprint arXiv:2604.05601*, 2026.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *arXiv preprint arXiv:1902.09506*, 2019.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, 2016.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. Sparsevila: Decoupling visual sparsity for efficient vlm inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23784–23794, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Yuxuan Liang, Xu Li, Xiaolei Chen, Yi Zheng, Haotian Chen, Bin Li, and Xiangyang Xue. Pyramid token pruning for high-resolution large vision-language models via region, token, and instruction-guided importance. *arXiv preprint arXiv:2509.15704*, 2025.

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

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023b.

Pan Lu, Swaroop Mishra, Tony 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. In *Advances in Neural Information Processing Systems*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyang Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

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*, 2019.

Rinyoichi Takezoe, Yaqian Li, Zihao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learnpruner: Rethinking attention-based token pruning in vision language models. In *International Conference on Learning Representations*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Ziyang Wang, Mengwei Li, Hao Yin, Wenhao Liu, and Zilei Wang. Posprune: Visual token pruning with positional bias correction for efficient large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10485–10493, 2026.

xAI. Grok-1.5 Vision Preview. `https://x.ai/news/grok-1.5v`, 2024. Introduces the RealWorldQA benchmark.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14593–14603, 2025.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025. doi: 10.1609/aaai.v39i21.34366.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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*, 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. doi: 10.1609/aaai.v40i15.38234.

Hanbo Zhang, Jie Xu, Yuchen Mo, and Tao Kong. Invig: Benchmarking open-ended interactive visual grounding with 500k dialogues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5508–5518, 2024a.

Pu Zhang, Yuwei Li, Xingyuan Xian, and Guoming Tang. Zspaprune: Zero-shot prompt-aware token pruning for vision-language models. *arXiv preprint arXiv:2510.17197*, 2025a. doi: 10.48550/arXiv.2510.17197.

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 preprint arXiv:2412.01818v1*, 2024b.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025b.

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

Qihui Zhu, Tao Zhang, Yuchen Wang, Shuangwu Chen, Xiaobin Tan, Jian Yang, Yang Liu, and Yinfei Pan. Hawk: Head importance-aware visual token pruning in multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39583–39592, 2026.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

A Q-CAL CALIBRATION DETAILS

To obtain the sampling weights for the multi-layer cross-modal attention signals, we calibrate across eight general visual capabilities. Table 3 lists the capability families, primary calibration sources, and the same-family replacement sources used in the calibration-source ablation. Effective samples are filtered from the primary sources and kept only if the full-token baseline obtains a nonzero task score and the answer genuinely depends on the visual input rather than question-only language priors. This avoids calibrating on model failures or knowledge-dominant examples that provide little evidence about visual-token selection.

Table 3: Calibration capability coverage and representative sources.

Capability family	Primary source	Replacement source
Object and scene recognition	RealWorldQA (xAI, 2024)	Mantis (Jiang et al., 2024)
Multi-image visual comparison	BLINK (Fu et al., 2024)	BLINK-heldout (Fu et al., 2024)
Scene-text reading	TextVQA (Singh et al., 2019)	OCRBench (Liu et al., 2023b)
Document understanding	DocVQA (Mathew et al., 2021)	InfographicVQA (Mathew et al., 2022)
Chart reasoning	ChartQA (Masry et al., 2022)	DVQA (Kafle et al., 2018)
Diagrammatic reasoning	AI2D (Kembhavi et al., 2016)	ScienceQA (Lu et al., 2022)
Knowledge-grounded visual reasoning	MMMU (Yue et al., 2024)	MathVista (Lu et al., 2024)
Hallucination-sensitive verification	HallBench (Guan et al., 2024)	POPE (Li et al., 2023)

For each retained calibration sample $x = (I, q)$, the full-token model first generates an answer $y^\star$. We treat this answer as the teacher-forcing target and backpropagate its negative log-likelihood to the visual-token embeddings, giving a token-level oracle score

$$o_i(x) = \left\| \nabla_{v_i} [-\log p(y^\star \mid I, q)] \right\|_2 \left\| v_i \right\|_2 , \qquad i \in \mathcal{V}_x. \tag{6}$$

For every target budget $\rho \in \mathcal{B}$, the oracle target set is

$$\mathcal{T}_{x,\rho} = \mathrm{TopK}_{i \in \mathcal{V}_x}(o_i(x), k_\rho), \qquad k_\rho = \lceil \rho |\mathcal{V}_x| \rceil. \tag{7}$$

Calibration fits a policy for sampling and weighting multiple text-visual attention layers. For each task and candidate layer ℓ, let $A_\ell(i \mid x)$ be the min-max normalized attention score over visual tokens. Given weights w, the policy score is

$$s_w(i \mid x) = \sum_{\ell \in \mathcal{L}_{75}} w_\ell A_\ell(i \mid x), \qquad w_\ell \geq 0, \quad \sum_{\ell \in \mathcal{L}_{75}} w_\ell = 1, \tag{8}$$

where $\mathcal{L}_{75}$ denotes candidate layers in the first 75% of language-model depth. For each target budget ρ, the retained set is

$$\widehat{\mathcal{T}}_{x,\rho}(w) = \mathrm{TopK}_{i \in \mathcal{V}_x}\left(s_w(i \mid x), \lceil \rho |\mathcal{V}_x| \rceil\right). \tag{9}$$

The recovery term measures overlap with the gradient-oracle targets:

$$R(w) = \frac{1}{|\mathcal{D}_{cal}||\mathcal{B}|} \sum_{x \in \mathcal{D}_{cal}} \sum_{\rho \in \mathcal{B}} \frac{|\widehat{\mathcal{T}}_{x,\rho}(w) \cap \mathcal{T}_{x,\rho}|}{|\mathcal{T}_{x,\rho}|}. \tag{10}$$

We also include a cost term for the layer where second-stage pruning occurs. Let $S(w)$ be the deepest active layer,

$$S(w) = \max_{\ell \in \mathcal{L}_{75} : w_\ell \geq \epsilon} (\ell + 1), \tag{11}$$

where $\epsilon = 0.05$ treats smaller weights as inactive. The token-layer cost proxy is

$$C(w) = \frac{\rho_1 S(w) + \rho_2 (L - S(w))}{L}. \tag{12}$$

The final policy is obtained by maximizing oracle-set recovery while penalizing delayed second-stage pruning:

$$w^\star = \arg \max_{w_\ell \geq 0,\ \sum_{\ell \in \mathcal{L}_{75}} w_\ell = 1} \left[R(w) - \lambda_c C(w)\right]. \tag{13}$$

B Stage-2 Policy Weights

The calibrated stage-2 policy is instantiated separately for each backbone because Qwen-family models have different decoder depths. Table 4 reports the resulting layer mixtures after applying the 75% depth cap used in the main experiments. A layer is treated as active only when its normalized weight ratio is larger than 0.05; layers below this threshold are not used in the stage-2 score computation.

Table 4: Stage-2 policies used for the six-model main experiment under the bounded-depth setting. Each row reports dominant normalized layer weights ($w > 0.05$) in the model's native layer indices; lower-weight layers are inactive for stage-2 scoring and are summarized as other mass.

Model	LM layers	Depth cap	Dominant layer weights	Other mass
Qwen2-VL-2B	28	L0–L20	L0:0.070, L8:0.120, L14:0.482, L15:0.056, L17:0.114, L20:0.141	0.017
Qwen2-VL-7B	28	L0–L20	L0:0.078, L1:0.055, L8:0.210, L17:0.242, L19:0.226, L20:0.107	0.082
Qwen2.5-VL-3B	36	L0–L26	L9:0.181, L10:0.212, L17:0.163, L23:0.123, L24:0.067, L26:0.174	0.080
Qwen2.5-VL-7B	28	L0–L20	L1:0.072, L2:0.281, L19:0.262, L20:0.365	0.019
Qwen3-VL-4B	36	L0–L26	L8:0.219, L16:0.232, L20:0.082, L24:0.421	0.046
Qwen3-VL-8B	36	L0–L26	L6:0.103, L9:0.356, L17:0.287, L21:0.157, L23:0.076	0.020

C Full Signal Diagnostics

Table 5 expands the prompt-invariance diagnostic with supplementary signals for both groups. Figure 7 reports the pairwise agreement among representative signal families, and Figure 8 visualizes their retained-token patterns on the same image and five native questions. The full set follows the same pattern as the representative signals in the main text: early attention, prompt-guided scores, and visual priors remain highly stable across prompts, while deeper text-visual attention is substantially query-discriminative and has low agreement with either early signal family.

Table 5: Full prompt-invariance diagnostics across four VLMs on 100 VQAv2 validation images with five questions per image. The visual-encoder CLS-attention row is measured on LLaVA-1.5-7B, whose CLIP visual encoder exposes CLS-to-patch attention; Qwen visual backbones do not expose a compatible CLS-attention map.

Signal family		Signal detail	$J_{10\%}$ mean	$J_{10\%}$ var.	$J_{20\%}$ mean	$J_{20\%}$ var.	ρ mean	ρ var.
Query-conditioned signal	Early text–visual attention	Decoder layer 0	0.960	0.004883	0.957	0.005518	0.989	0.000370
		Decoder layer 1	0.924	0.002811	0.939	0.000946	0.990	0.000134
		PosPrune PBC early attention	0.987	0.000167	0.985	0.000148	1.000	0.000000
	Prompt-guided pruning scores	ZSPAPrune prompt–visual cosine	0.604	0.025142	0.649	0.022680	0.885	0.012114
		IVTP pooled-instruction signal	0.876	0.008327	0.893	0.006264	0.989	0.000098
Deep text–visual attention		Mid-depth decoder attention	0.428	0.013579	0.503	0.011231	0.775	0.010403
		75%-depth decoder attention	0.349	0.003954	0.442	0.002396	0.714	0.001090
Visual representation prior		ViT received-attention saliency	1.000	0.000000	1.000	0.000000	1.000	0.000000
		SVD-Prune leverage saliency	1.000	0.000000	1.000	0.000000	1.000	0.000000
		OTPrune embedding-OT selector	1.000	0.000000	1.000	0.000000	1.000	0.000000
		Visual-encoder CLS attention (LLaVA-only)	1.000	0.000000	1.000	0.000000	1.000	0.000000

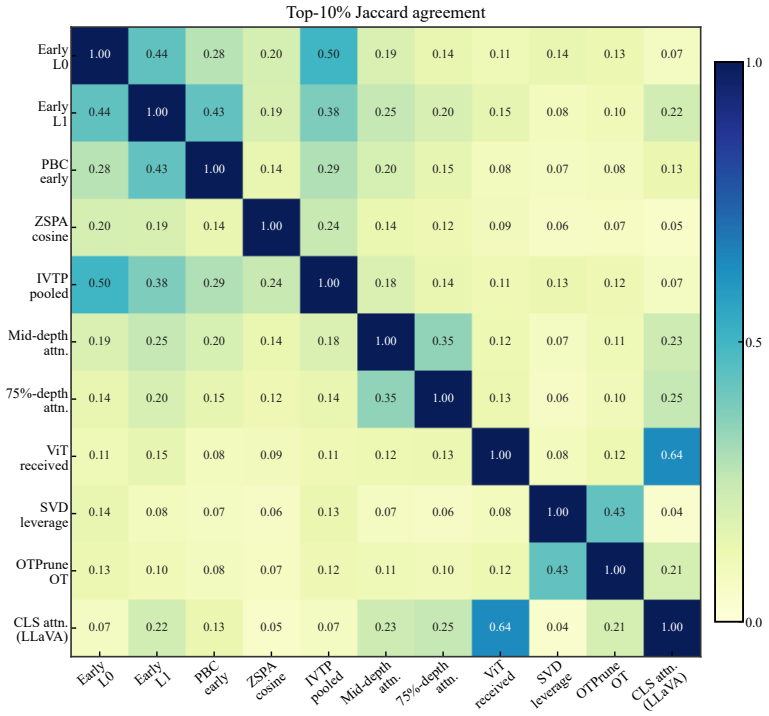


Figure 7: Cross-signal agreement measured by top-10% Jaccard overlap on the same 100-image, five-question VQAv2 diagnostic. Rows and columns include the Table 5 signals; CLS-attention entries are measured on LLaVA-1.5-7B only, and the other entries are averaged over the four-model diagnostic set.

D Selection-Behavior Diagnostics

Table 6 provides the complete numerical results behind the selection-behavior diagnostic summarized in Figure 2. The table reports cross-model accuracy on the 240 labelled tasks, grouped by the visual-evidence regime introduced in Section 3.2. The "Full" column gives the full-token reference for each task type; the remaining columns compare visual-prior, early-attention, dualsignal, and deep-attention scores under top-k selection and diversity-suppressed selection at 10%, 20%, and 33% visual-token budgets.

E Baseline Reproduction

Before including an external pruning method in the main comparison, we audit whether the method can be executed under its intended model family, pruning budget, and benchmark setting. Table 7 records the original paper value when a matched table entry is available and reports our reproduced value under the corresponding setting.

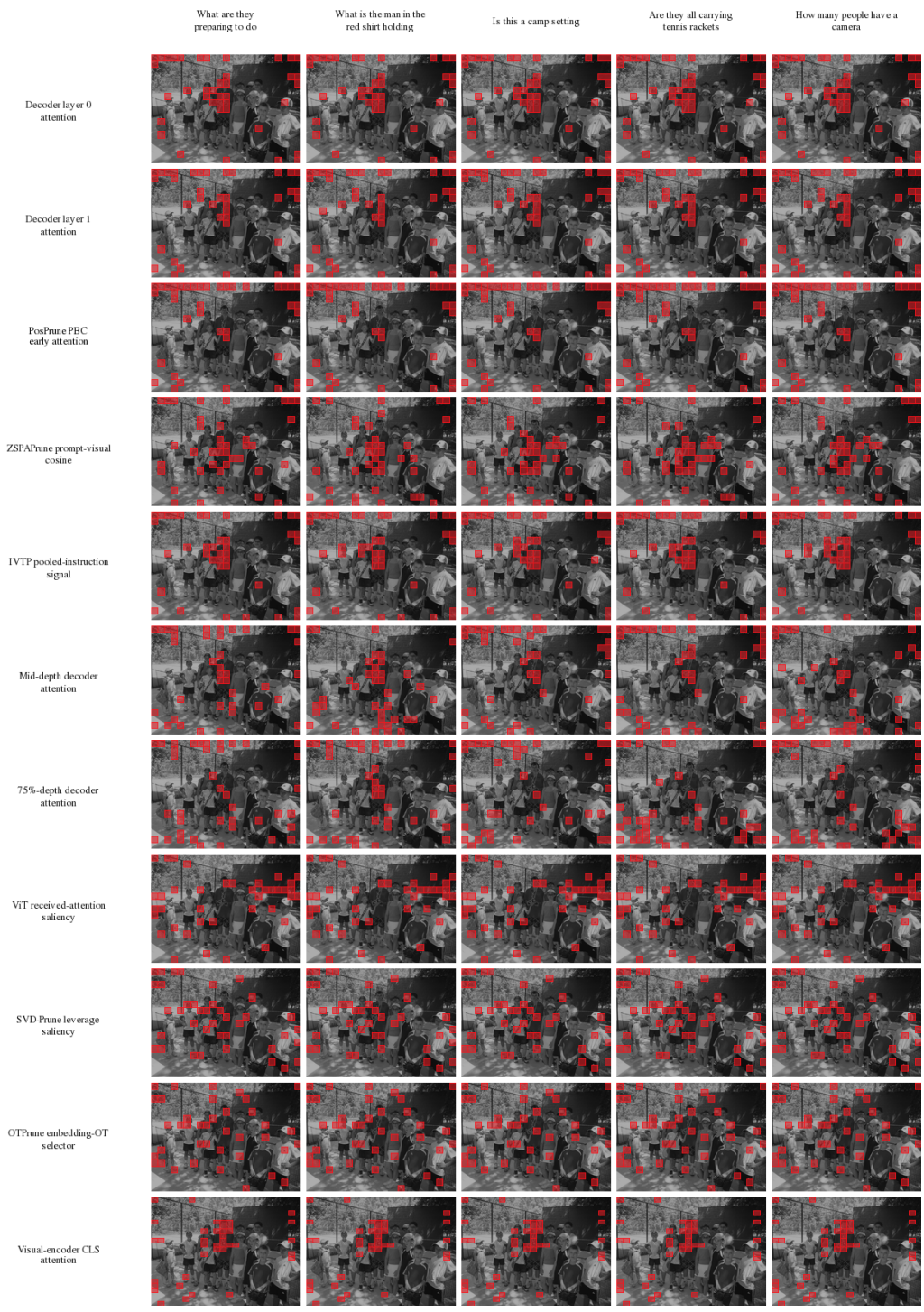


Figure 8: Full qualitative prompt-invariance visualization for the diagnostic signals. Columns are the five VQAv2 questions paired with the same image, and each row highlights the same number of top-ranked visual tokens for one measured signal.

F Selection-Behavior Task Pool

The 240-task pool is constructed with 80 global-like, 80 local-like, and 80 no-difference tasks. Candidate tasks are first filtered by basic model competence and visual dependence: the full-token

Table 6: Cross-model signal-bias diagnostic on 240 labelled tasks under three visual-token budgets.

			Visual prior		Early attn.		DualSignal		Deep attn.	
Task type	Full	Budget	Top-k	+Div.	Top-k	+Div.	Top-k	+Div.	Top-k	+Div.
Overall	0.901	10%	0.576	0.664	0.616	0.727	0.681	0.775	**0.862**	0.856
		20%	0.711	0.749	0.752	0.826	0.825	0.851	**0.871**	0.866
		33%	0.798	0.781	0.817	0.864	0.860	0.873	**0.883**	0.869
Global-like	0.831	10%	**0.802**	0.723	0.053	0.436	0.455	0.651	0.766	0.771
		20%	**0.801**	0.758	0.454	0.694	0.689	0.750	0.789	0.785
		33%	0.810	0.761	0.629	0.766	0.754	0.790	**0.816**	0.803
Local-like	0.913	10%	0.022	0.354	**0.894**	0.829	0.689	0.758	0.880	0.861
		20%	0.427	0.573	0.871	0.855	0.858	0.865	**0.889**	0.879
		33%	0.652	0.655	0.883	0.881	0.882	0.883	**0.892**	0.871
No-difference	0.960	10%	0.903	0.914	0.903	0.917	0.899	0.918	**0.939**	0.936
		20%	0.906	0.917	0.929	0.929	0.926	**0.937**	0.935	0.935
		33%	0.932	0.926	0.937	0.946	0.944	**0.947**	0.942	0.933

Table 7: Baseline reproduction under original or paper-aligned settings.

Method	Test model	Benchmark	Paper baseline	Paper acc.	Repro acc.
AgilePruner	Qwen2.5-VL-7B	TextVQA	AgilePruner 512 tokens	79.80	84.38
AgilePruner	Qwen2.5-VL-7B	AI2D	AgilePruner 512 tokens	82.10	81.25
FastV	LLaVA-1.5-7B	MMMU	FastV K=2, R=50%	34.40	34.33
FastV	LLaVA-1.5-7B	A-OKVQA	FastV K=2, R=50%	77.00	84.33
FasterVLM	LLaVA-1.5-7B	TextVQA	FasterVLM 58 tokens	55.28	58.03
FasterVLM	LLaVA-1.5-7B	POPE	FasterVLM 58 tokens	75.85	77.06
VisPruner	LLaVA-1.5-7B	TextVQA	VisPruner 128 tokens	57.00	58.53
VisPruner	LLaVA-1.5-7B	POPE	VisPruner 128 tokens	84.60	85.90
DivPrune	LLaVA-1.5-7B	POPE	DivPrune, layer 0	86.02	85.43
DivPrune	LLaVA-1.5-7B	MME	DivPrune, layer 0	1328.00	1318.34
SVD-Prune	LLaVA-1.5-7B	GQA	SVD-Prune 32 tokens	53.52	51.67
SVD-Prune	LLaVA-1.5-7B	TextVQA	SVD-Prune 32 tokens	54.81	52.33
PyramidDrop	LLaVA-1.5-7B	TextVQA	PDrop schedule [8,16,24]	57.60	58.80
PyramidDrop	LLaVA-1.5-7B	POPE	PDrop schedule [8,16,24]	84.80	86.00
FitPrune	LLaVA-1.5-7B	TextVQA	FitPrune 60% reduction	58.20	60.57
FitPrune	LLaVA-1.5-7B	POPE	FitPrune 60% reduction	86.50	86.33
HAWK	Qwen2.5-VL-7B	TextVQA	HAWK fixed-resolution, 128 tokens	79.20	79.17
PTP	InternVL2-8B	TextVQA	PTP r=0.5, α=0.5	76.40	75.00
PTP	InternVL2-8B	POPE	PTP r=0.5, α=0.5	87.20	89.88
ID-Selection	Qwen2.5-VL-7B	TextVQA	ID-Selection 88.9% reduction	73.80	75.33
ID-Selection	Qwen2.5-VL-7B	POPE	ID-Selection 88.9% reduction	78.10	83.77
D2Pruner	Qwen2.5-VL-7B	TextVQA	D2Pruner 10% retained	70.54	71.88
D2Pruner	Qwen2.5-VL-7B	POPE	D2Pruner 10% retained	78.34	65.62
ZSPAPrune	Qwen2.5-VL-7B	POPE	ZSPAPrune 10% retained	69.00	77.11
ZSPAPrune	Qwen2.5-VL-7B	TextVQA	ZSPAPrune 10% retained	54.90	56.25
SparseVLM	LLaVA-1.5-7B	TextVQA	SparseVLM 128 tokens	56.70	60.13
SparseVLM	LLaVA-1.5-7B	POPE	SparseVLM 128 tokens	85.00	85.04

model must answer the task correctly, while removing the image or replacing it with an unrelated image must lead to an incorrect answer. The annotation criterion is how the visual evidence needed for answering is retrieved: global-like tasks require broad scene or spatial-coverage evidence; local-like tasks depend on a small localized detail; no-difference tasks are those where either retrieval mode can support the same answer or where the visual evidence for an alternative answer is absent. For example, in a negative existence or relation question, if neither local detail nor global spatial layout provides evidence for answering "yes", the task is labelled no-difference. Two annotators independently annotate the remaining visually grounded tasks. When the two annotations disagree, the annotators discuss the case to obtain a unified label; cases that cannot be resolved are discarded. There are examples of three types of tasks.

F.1 GLOBAL-LIKE TASKS

Example 1.



Question: What is the name of the floor pattern?
Answer: checkered.
Preference reason: The answer depends on a floor pattern spread across the room; broad spatial layout is more useful than a single small patch.

Example 2.



Question: What is the weather like?
Answer: rainy.
Preference reason: Rain is supported by scene-wide cues, including wet road surface, umbrellas, and the overall outdoor context.

Example 3.



Question: How many levels are on this bus?
Answer: 2.
Preference reason: Counting bus levels requires retaining the whole bus silhouette and its vertical structure.

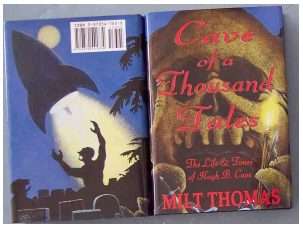
F.2 LOCAL-LIKE TASKS

Example 1.



Question: What letters are on the tail of the plane?
Answer: vh-bar.
Preference reason: The answer is a small alphanumeric mark on the airplane tail; dropping that local region removes the evidence.

Example 2.



Question: Who is the author of this book?
Answer: Milt Thomas.
Preference reason: The answer comes from small text printed on the book cover.

Example 3.



Question: What number route?
Answer: 151.
Preference reason: The route number is a tiny display on the front of the bus.

F.3 NO-DIFFERENCE TASKS

Example 1.



Question: Is this photo taken in a place of coffee shop? Please answer yes or no.
Answer: No.
Preference reason: The image lacks visual evidence for a coffee shop; neither local details nor global scene layout supports a yes answer.

Example 2.



Question: Is the pineapple on the right of the pot in the image? Please answer yes or no.
Answer: No.
Preference reason: The queried object-relation is absent; local inspection and broader layout both reject the yes hypothesis.

Example 3.



Question: Is the boy playing tee ball?
Answer: no.
Preference reason: The relevant tee-ball evidence is absent; broad action context and local equipment cues lead to the same negative answer.