OUTPUT-ALIGNED VISUAL TOKEN ROUTING FOR EFFICIENT VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token pruning accelerates Vision-Language Models (VLMs), yet existing methods predominantly rely on intermediate proxies such as attention weights or feature diversity to rank tokens. Visual elements typically exhibit strong spatial and semantic interdependencies. Consequently, ranking tokens in isolation fails to preserve the joint context required for generation, a challenge we term proxy metric misalignment. We propose Output-Aligned Routing (OAR), which supervises visual token selection directly via the inherent generation behavior of a frozen base VLM. In the offline stage, the base model uses unlabeled calibration data to evaluate retained visual subsets through the output KL divergence between unpruned and pruned generation distributions. We distill these divergence signals into soft inclusion targets to train a 1.91M-parameter Router. At inference time, the Router selects and gathers the top-B tokens in a single, query-aware $O(N)$ pass prior to language-model prefill, requiring zero online intervention from the base model. Across demanding benchmarks spanning charts, documents, infographics, and scene text, OAR preserves 92–95% of full-model accuracy on ChartQA and TextVQA while retaining only 80 visual tokens, an approximately 85% reduction in the visual sequence. Crucially, it achieves up to a 2.94× end-to-end speedup and a 4.8× reduction in KV cache memory, with an online selection overhead of under 1 ms.

1 INTRODUCTION

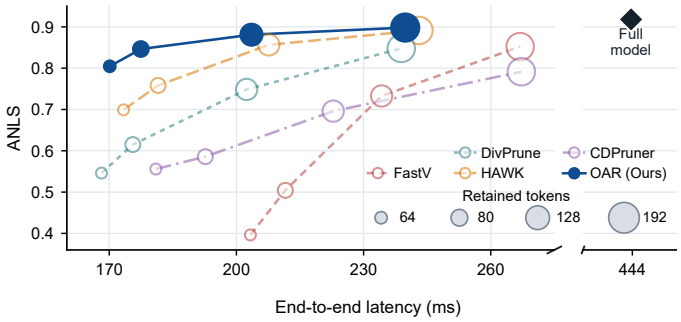


Figure 1: End-to-end latency and open-ended ChartQA Average Normalized Levenshtein Similarity (ANLS) at 64, 80, 128, and 192 retained visual tokens. Marker area denotes the token budget; the broken axis separates the full model.

Vision-Language Models (VLMs) achieve strong performance in visual question answering, document understanding, and multimodal reasoning (Bai et al., 2025; Zhu et al., 2025). Their accuracy, however, comes with a substantial inference cost. A high-resolution image or document page can produce hundreds or thousands of visual tokens, lengthening language-model prefill and expanding the KV cache (Takezoe et al., 2026; Zhu et al., 2026). This computational burden limits the practical deployment of VLMs in resource-constrained environments and real-time applications.

Visual token pruning addresses this cost by retaining a small set of informative tokens before or within the language model. Existing methods predominantly estimate token importance from heuristic-

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.

driven proxies, such as attention weights (Chen et al., 2024a; Zhu et al., 2026), or reduce redundancy using feature diversity, similarity, and coverage (Alvar et al., 2025; Yang et al., 2025; Wen et al., 2025; Zhang et al., 2025a). Their success depends on whether these intermediate signals successfully identify the visual evidence actually needed for generation.

We find that this assumption is often insufficient. High-attention tokens can omit a necessary relation, while a diverse set can cover the image yet break non-additive local dependencies (such as the relationship between a chart legend, its curve, and the corresponding axis value). The selected tokens may therefore look useful in isolation but fall short of preserving the joint context required for generation. We term this key challenge proxy metric misalignment. Our diagnostic study shows that feature-diverse subsets frequently change the base VLM's answer, whereas subsets chosen by their effect on the model's output preserve its generation more reliably.

Based on this observation, we propose OAR (Output-Aligned Routing), which shifts the paradigm from heuristic-driven pruning to output-aligned selection. During an offline, label-free calibration phase, the frozen base VLM evaluates retained token subsets using the KL divergence between full and pruned generation distributions. We convert these divergence scores into token inclusion probabilities, which provide soft targets for a 1.91M-parameter Router. By decoupling subset evaluation from online deployment, OAR distills the base model's intrinsic generation behavior into a lightweight selector (Hinton et al., 2015). At inference time, the Router scores all visual tokens once, retains the top B, and gathers them into a compact tensor before language-model prefill.

Across demanding visual reasoning tasks involving charts, documents, infographics, and scene text, OAR consistently preserves 92–95% of full-model accuracy on ChartQA and TextVQA while retaining only 80 visual tokens, an approximately 85% reduction in the visual sequence. It achieves up to a 2.5× end-to-end speedup and reduces KV cache memory footprint by 4.8×, while the complete online selection path costs less than 1 ms. The learned Router can also be reused across related visual domains without target-domain calibration.

In summary, the main contributions of this work are as follows:

- We identify proxy metric misalignment, a key challenge where pointwise proxy scores struggle to preserve joint visual context. We verify this empirically through controlled interventions, demonstrating that visual token utility is highly non-additive and context-dependent.
- We propose distilling sequence-level KL divergence from joint subset evaluations into a lightweight, geometry- and query-aware Router. To make this large-scale distillation computationally tractable, we introduce a label-free trajectory approximation that captures the base model's dominant output changes without requiring human annotations.
- Across demanding reasoning benchmarks, OAR achieves up to a 2.5× end-to-end speedup and a 4.8× KV cache reduction with sub-millisecond online overhead. Beyond efficiency, it supports zero-shot reuse across related domains, preserves critical visual dependencies, and significantly reduces object hallucinations compared to proxy-based baselines.

2 Related Work

2.1 Efficient Multimodal Large Language Models

Modern multimodal large language models combine vision encoders with pretrained language backbones for visual question answering and document reasoning (Bai et al., 2025; Zhu et al., 2025; Choi et al., 2026). Fine-grained images can produce hundreds or thousands of visual tokens, increasing attention cost during prefill and KV-cache memory during decoding (Vasu et al., 2025; Takezoe et al., 2026; Zhu et al., 2026). Visual token reduction addresses this sequence-length bottleneck for efficient deployment.

2.2 Visual Token Reduction and Selection Policy

Visual token reduction methods fall into training-free heuristics and lightweight trainable selectors. Within the training-free category, attention-based methods such as FastV (Chen et al., 2024a) and HAWK (Zhu et al., 2026) rank tokens using early-layer or head-weighted attention. Progressive schemes such as PyramidDrop and VScan distribute selection across stages to reduce premature

information loss, whereas SparseVLM and PriorTR use text guidance or corrected visual priors to improve question relevance and mitigate attention bias (Xing et al., 2025; Zhang et al., 2025c; 2026; Chen et al., 2026). Another training-free family targets visual redundancy. ToMe, VisionZip, and DART merge similar representations (Bolya et al., 2023; Yang et al., 2025; Wen et al., 2025), while DivPrune, VisPruner, and CDPruner optimize feature diversity or conditional coverage (Alvar et al., 2025; Zhang et al., 2025a;b). Lightweight trainable approaches including ATP-LLaVA and LearnPruner instead learn compact selectors from visual features and task context (Ye et al., 2025; Takezoe et al., 2026).

These methods primarily infer utility from intermediate proxies such as attention or feature distance (Chen et al., 2024a; Zhu et al., 2026; Alvar et al., 2025; Zhang et al., 2025b). Structured reasoning exposes their limitation: a chart legend, curve, and axis value can be useful jointly while appearing weak in isolation. OAR therefore supervises selection with the base VLM's output change under complete retained contexts, directly aligning the Router with generation behavior.

3 METHODOLOGY

Given an image–question input $x = (I, q)$, where I is the image and q is the question, and a retained-token budget B, OAR learns which visual tokens to retain from the output behavior of a frozen base VLM. We first show why independently ranked proxy scores can struggle to preserve the context required for generation. We then formulate joint subset evaluation through output divergence and describe how OAR converts this expensive offline signal into a lightweight online Router.

3.1 MOTIVATION

Limitations of Independent Ranking. To motivate our output-aligned approach, we first analyze the assumptions behind independent token ranking. Most efficient selectors assign one score to each visual token and retain the top B (Chen et al., 2024a; Zhu et al., 2026; Alvar et al., 2025; Zhang et al., 2025a). This strategy assumes that the contribution of a token can be estimated independently of the other retained tokens. Visual evidence is often non-additive. A chart legend is useful only with the corresponding curve and axis value. Likewise, a document field may require both its label and content. Consequently, the utility of a token can change when the surrounding context changes.

Utility Decomposition. Let $\mathcal{R}_x(S)$ denote the output divergence induced by retaining token set S, as formalized in Eq. 3. Let $S_0 \subseteq \mathcal{V}$ be a fixed anchor context shared by all candidates and let $A \subseteq \mathcal{V} \setminus S_0$ contain the $b = B - |S_0|$ additional tokens. We characterize non-additive dependencies through $U_x(A) = \mathcal{R}_x(S_0) - \mathcal{R}_x(S_0 \cup A)$. For every nonempty interaction set $H \subseteq A$, let $m_x(H)$ denote its Möbius coefficient, computed over the auxiliary subsets $Q \subseteq H$:

$$U_x(A) = \sum_{\emptyset \neq H \subseteq A} m_x(H), \qquad m_x(H) = \sum_{Q \subseteq H} (-1)^{|H|-|Q|} U_x(Q), \tag{1}$$

where singleton coefficients describe isolated contributions and terms with $|H| \geq 2$ describe interactions.

Proposition 1 (Additive limit)**.** On a finite token ground set and for a fixed added-token budget b, if all higher-order coefficients $m_x(H)$ with $|H| \geq 2$ vanish, ranking singleton utilities recovers the globally optimal set. More generally, define $e_x(A) = \sum_{H \subseteq A, |H| \geq 2} m_x(H)$. If $|e_x(A)| \leq \epsilon$ for every feasible $|A| = b$, the top-b singleton solution is within 2ϵ utility of the optimum.

The proof is given in Appendix B.1. Proposition 1 explains when independent proxy scores succeed and when they depart from the globally useful context: stronger visual coupling increases the higher-order interaction error ϵ and weakens singleton ranking.

Diagnostic Analysis on ChartQA. At the same 80-token budget, we compare max–min feature diversity with direct evaluation of retained subsets using the base VLM's output difference. Diversity selects the minimum-divergence candidate in only 44.8% of 192 held-out examples. Output-based selection raises ANLS from 0.5178 to 0.6470, exact match from 26.56% to 38.54%, and fidelity from 27.08% to 39.58% at nearly identical latency (Appendix Table 6). Across retained contexts,

88.19% of token-group utilities change sign, and 98.96% of examples reverse their top or bottom group ordering (Figure 2(b)). Feature diversity alone therefore does not predict answer-preserving subsets.

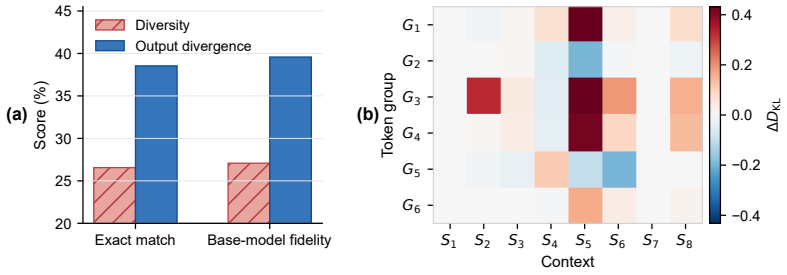


Figure 2: Controlled ChartQA diagnosis at an 80-token budget. (a) Exact match and base-model fidelity. (b) Each row fixes a token group G_i, each column a retained context S_j, and color shows the change in output divergence after adding G_i to S_j. Red indicates reduced divergence, blue indicates increased divergence, and sign changes within a row reveal context-dependent utility.

3.2 Objective Formulation

Output-Aligned Subset Evaluation. To address the aforementioned issues, our work draws inspiration from policy distillation (Rusu et al., 2015) and decision-focused learning (Wilder et al., 2019; Elmachtoub & Grigas, 2022), which advocate that a compact policy should be supervised directly by the quality of the downstream decisions it induces. We bring this principle to visual token routing, formulating the token pruning process as a subset selection problem guided by the model's final generation distribution rather than intermediate proxies. A base VLM f_θ with frozen parameters θ encodes I into N visual tokens $\mathcal{V} = \{v_i\}_{i=1}^{N}$. Given budget B, token pruning retains a subset $\mathcal{S} \subseteq \mathcal{V}$ with $|\mathcal{S}| = B$. Let $y = (y_1, \ldots, y_T)$ denote an output sequence of length T, with t indexing its decoding steps, and let $p_\theta(\cdot \mid \cdot)$ denote the frozen VLM's next-token distribution. The full and pruned visual contexts define the autoregressive distributions

$$p_{\mathcal{V}}(y \mid x) = \prod_t p_\theta(y_t \mid y_{<t}, q, \mathcal{V}), \qquad p_{\mathcal{S}}(y \mid x) = \prod_t p_\theta(y_t \mid y_{<t}, q, \mathcal{S}). \tag{2}$$

We measure the output change caused by pruning with

$$\mathcal{R}_x(\mathcal{S}) = D_{\mathrm{KL}}(p_{\mathcal{V}}(y \mid x) \,\|\, p_{\mathcal{S}}(y \mid x)), \tag{3}$$

and seek $\mathcal{S}_x^* = \arg\min_{|\mathcal{S}|=B} \mathcal{R}_x(\mathcal{S})$. Unlike an attention or diversity score, Eq. 3 evaluates the retained tokens jointly through their effect on generation.

Tractable Trajectory Approximation. Exact sequence-level KL requires summing over the full vocabulary at every possible generation trajectory and is therefore impractical for large-scale target generation. We use a tractable approximation that follows the full model's greedy trajectory $\bar{y}_{1:T}$. In visual-document QA and structured reasoning, the output space is highly directed and dominated by a few high-probability semantic paths. Tracking the base model's greedy trajectory thus provides an exceptionally rich and focused signal, capturing the exact reasoning trace where proxy metric misalignment is most destructive. At decoding step t, let $p_{\mathcal{V},t}$ and $p_{\mathcal{S},t}$ denote the full- and reduced-context next-token distributions under the shared prefix $\bar{y}_{<t}$. Let $A_{K,t} = \mathrm{TopK}(p_{\mathcal{V},t})$. The shared map $c_{A_{K,t}}$ retains the probabilities indexed by $A_{K,t}$ from either distribution and aggregates its complement into one tail bin:

$$\widehat{\mathcal{R}}_x^{(K)}(\mathcal{S}) = \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}\big(c_{A_{K,t}}[p_{\mathcal{V},t}] \,\|\, c_{A_{K,t}}[p_{\mathcal{S},t}]\big), \tag{4}$$

where we default to $K = 128$. Crucially, this objective relies entirely on the base VLM's intrinsic generation behavior, preserving richer reasoning traces than ground-truth answer supervision while operating label-free.

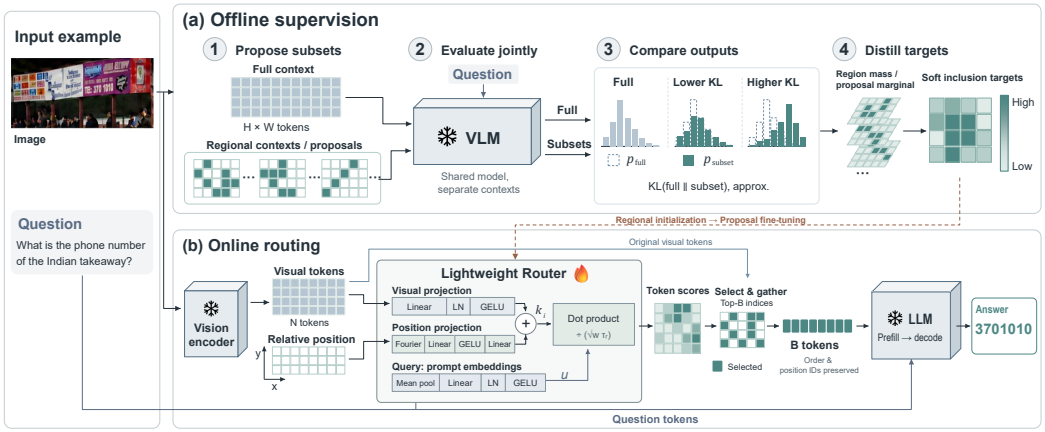


Figure 3: Overview of OAR. Offline, a frozen VLM evaluates candidate token subsets through output divergence, producing soft inclusion targets to train a lightweight Router. Online, the Router uses visual features, question information, and spatial positions to select a compact visual context before language-model prefill.

Corollary 1 (Next-token margin preservation)**.** At a shared prefix $\bar{y}_{<t}$, let $a = \arg\max_b p_{\mathcal{V},t}(b)$ denote the full-context winner, let $\gamma_t = p_{\mathcal{V},t}(a) - \max_{b \neq a} p_{\mathcal{V},t}(b)$ denote its probability margin, and let $R_t(\mathcal{S}) = D_{\mathrm{KL}}(p_{\mathcal{V},t} \| p_{\mathcal{S},t})$ denote the corresponding step-level output divergence. The reduced context preserves the full model's next-token winner whenever $\gamma_t > \sqrt{2R_t(\mathcal{S})}$.

The proof follows from Pinsker's inequality and is given in Appendix B.2. The condition gives a decision-level interpretation to the KL objective: controlling output divergence bounds deviation from the base model's greedy trajectory. Empirically, it certifies 87.88% of decoding steps and 76.11% of complete trajectories at 80 retained tokens, with 100% fidelity on certified cases. The top-128 approximation matches exact-KL coverage (Appendix C.9).

3.3 OAR: Offline Distillation and Online Routing

Framework Overview. Figure 3 summarizes OAR. The offline stage constructs output-aligned token targets from joint subset evaluations. The Router is fixed after offline training and selects a fixed number of visual tokens before language-model prefill.

3.3.1 Offline Target Generation

Coarse-to-Fine Target Construction. Exhaustive evaluation of $\binom{N}{B}$ subsets is computationally intractable. To navigate this combinatorial space, the offline teacher employs an $O(G + M)$ coarse-to-fine intervention strategy.

Stage 1 (Coarse Regional Interventions): To capture the non-additive local dependencies formulated in Section 3.1, we first evaluate macroscopic spatial blocks. We partition each native visual grid into $G = g^2$ relative spatial groups $\{C_j\}_{j=1}^{G}$ ($g = 4$ or 8). For Qwen2.5-VL's native height–width position (h_i, w_i) of token i (Bai et al., 2025), let $h_{\min}, h_{\max}, w_{\min}, w_{\max}$ denote the coordinate extrema. We define its normalized coordinate as

$$z_i = \left(\frac{h_i - h_{\min}}{\max(h_{\max} - h_{\min}, 1)}, \frac{w_i - w_{\min}}{\max(w_{\max} - w_{\min}, 1)} \right) \in [0,1]^2. \tag{5}$$

which maps any native resolution or aspect ratio to a unified continuous grid. Let $\mathcal{A}$ be a fixed set of uniformly spaced background anchors and define the regional candidate $S_j = C_j \cup \mathcal{A}$. Its divergence $r_j = \widehat{\mathcal{R}}_x^{(K)}(S_j)$ evaluates regional evidence within a coherent global layout. With target temperature $\tau > 0$, the regional probability $\pi_x(j)$ and token target t_i are

$$\pi_x(j) = \frac{\exp(-r_j/\tau)}{\sum_{k=1}^{G} \exp(-r_k/\tau)}, \qquad t_i = \frac{\pi_x(j)}{|C_j|} \quad (i \in C_j), \tag{6}$$

where we use $\tau = 0.03$. Distributing each regional mass across its tokens yields dense, layout-aware supervision from at most G reduced-context calls.

Stage 2 (Fine Proposal Refinement): While intermediate proxy metrics (e.g., attention or feature diversity) are insufficient as deterministic deployment filters, they provide complementary complete-set candidates. We form a finite pool $\mathcal{P}_x = \{S_x^{(m)}\}_{m=1}^{M}$ from $M = 3$ regional, conditional-diversity, and head-aware proposals. Their output divergences define a Gibbs distribution over this pool with $\tau_p = 0.01$:

$$\rho_x(m) = \frac{\exp[-\widehat{\mathcal{R}}_x^{(K)}(S_x^{(m)})/\tau_p]}{\sum_{\ell=1}^{M} \exp[-\widehat{\mathcal{R}}_x^{(K)}(S_x^{(\ell)})/\tau_p]}. \tag{7}$$

The final target is the weighted marginal over this finite proposal pool,

$$t_i^{\mathcal{P}} = \sum_{m=1}^{M} \rho_x(m) \frac{\mathbf{1}[i \in S_x^{(m)}]}{|S_x^{(m)}|}. \tag{8}$$

This normalized inclusion weight sums to one over i. When all candidates have size B, $t_i^{\mathcal{P}} = \Pr_{S \sim \rho_x}(i \in S)/B$. Fitting this weighted marginal lets the Router synthesize sets beyond the discrete support pool. Its deployed top-B set exceeds every individual proposal family in Appendix C.8. The cached targets are then used without further base-VLM evaluations.

3.3.2 Online Lightweight Routing

Dynamic Token Selection. During inference, we deploy a lightweight Router to select visual tokens based on the current query and geometric layout. Let $q \in \mathbb{R}^{d_{model}}$ be the mean-pooled input embedding of non-image prompt tokens. Bypassing the language-model Transformer blocks entirely, this pooled query representation captures the global semantic intent of the question (e.g., target entities or analytical dimensions) without incurring the latency overhead of multi-layer cross-attention. Empirically, this lightweight query alignment is highly effective at steering the router to relevant regions while keeping the online overhead strictly sub-millisecond. With hidden width $w = 256$, let W_v and W_q denote learned projections, let $\psi(z_i)$ be the 10-D Fourier encoding of relative position z_i, and let $g_\phi(\psi) = W_{g,2}\mathrm{GELU}(W_{g,1}\psi + b_{g,1}) + b_{g,2}$ be the two-layer geometry branch. With Router temperature $\tau_r > 0$, the token representation k_i, query representation u, and score s_i are

$$\begin{aligned} k_i &= \mathrm{GELU}(\mathrm{LN}(W_v v_i)) + g_\phi(\psi(z_i)), \\ u &= \mathrm{GELU}(\mathrm{LN}(W_q q)), \qquad s_i = \frac{k_i^\top u}{\sqrt{w}\,\tau_r}, \end{aligned} \tag{9}$$

This continuous Fourier encoding provides the Router with robust spatial awareness independent of the dynamic visual token count N. The 1.91M-parameter Router has $O(Nw)$ complexity and is trained with cross-entropy against the regional or proposal target $\widetilde{t}$:

$$\mathcal{L}_{\text{router}} = -\sum_{i=1}^{N} \widetilde{t}_i \log\left(\frac{\exp(s_i)}{\sum_j \exp(s_j)}\right). \tag{10}$$

The Router independently scores tokens for sub-millisecond deployment, while its targets are marginal inclusion probabilities from globally evaluated subsets. Top-B selection is the optimal fixed-size projection of these marginals under expected overlap (Appendix B.4), retaining set-aware supervision without online combinatorial search.

The proposal-refined Router retains the B highest-scoring tokens to form the compact visual context. Let $\hat{S}$ denote the final selected index set and let d be the visual-token feature dimension. We sort $\hat{S}$ and gather the compact tensor $V_{\hat{S}} \in \mathbb{R}^{B \times d}$. For Qwen2.5-VL, the corresponding temporal, height, and width position-ID triples are gathered in the same order. Text position IDs remain unchanged. Only these tokens enter language-model prefill and the resulting KV cache.

Table 1: **Accuracy and latency at matched token budgets.** ChartQA and DocVQA report ANLS; AI2D and TextVQA report accuracy. Latency is measured end to end. Original images average approximately 550 visual tokens, so retaining 80 tokens removes about 85% of the visual sequence.

Method	ChartQA (Open)		DocVQA		AI2D		TextVQA	
	ANLS ↑	E2E (ms) ↓	ANLS ↑	E2E (ms) ↓	Acc (%) ↑	E2E (ms) ↓	Acc (%) ↑	E2E (ms) ↓
Full model	0.9180	443.77	0.9110	502.26	84.10	305.2	82.58	308.50
128 Retained Tokens								
FastV	0.7327	234.26	0.5840	247.92	73.15	140.5	78.40	145.2
HAWK	0.8565	207.65	0.6120	241.45	78.60	141.2	81.30	146.0
DivPrune	0.7486	202.41	0.5512	242.52	70.44	152.6	74.20	158.4
CDPruner	0.6960	222.83	0.6288	264.87	79.10	160.1	82.50	165.7
DART	0.5899	173.47	0.3487	216.27	75.58	101.40	63.83	158.49
VisionZip	0.7895	272.15	0.4670	329.75	78.95	137.43	63.37	260.37
PriorTR	0.7713	237.86	0.4775	268.70	80.73	114.96	76.06	188.65
ToMe	0.6797	240.84	0.5161	284.59	80.57	122.14	71.14	203.46
OAR (Ours)	**0.8810**	203.55	**0.7950**	253.27	**83.05**	138.8	**84.50**	143.5
80 Retained Tokens								
FastV	0.5043	211.54	0.3447	222.81	64.20	106.5	71.10	110.2
HAWK	0.7582	181.52	0.4437	213.79	68.30	107.0	75.74	158.55
DivPrune	0.6150	175.56	0.4028	216.76	59.80	118.4	64.68	151.79
CDPruner	0.5858	192.68	0.3057	233.48	66.50	126.1	44.56	164.68
DART	0.4430	170.19	0.2532	209.32	71.92	96.32	54.29	152.96
VisionZip	0.6602	268.01	0.3347	320.53	77.01	131.67	55.99	254.74
PriorTR	0.6514	216.85	0.3733	237.45	78.47	104.71	71.85	175.99
ToMe	0.5482	220.10	0.3948	260.46	78.34	115.47	64.41	194.01
OAR (Ours)	**0.8463**	177.46	**0.6897**	230.32	**79.55**	103.8	**78.43**	153.63

4 Experiments

We evaluate OAR on structured reasoning, fine-grained perception, and hallucination benchmarks. The experiments compare task quality at matched token budgets and end-to-end latency, then analyze the contribution of output-aligned supervision and the lightweight Router.

4.1 Experimental Setup

We evaluate dense structural reasoning on ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), InfographicVQA (Mathew et al., 2022), and AI2D (Kembhavi et al., 2016); fine-grained perception on TextVQA (Singh et al., 2019); and visual dependency and hallucination on MMStar (Chen et al., 2024b), POPE (Li et al., 2023), and CHAIR (Rohrbach et al., 2018).

We use Qwen2.5-VL-7B (Bai et al., 2025) as the primary frozen base VLM and evaluate InternVL3-8B (Zhu et al., 2025) on ChartQA. The comparison includes attention-based FastV (Chen et al., 2024a), HAWK (Zhu et al., 2026), and PriorTR (Chen et al., 2026); diversity- or merging-based DivPrune (Alvar et al., 2025), ToMe (Bolya et al., 2023), VisionZip (Yang et al., 2025), and DART (Wen et al., 2025); and the conditional-diversity method CDPruner (Zhang et al., 2025b). All methods use the same actual token budgets and generation settings.

The Router is calibrated on unlabeled image-question pairs from each dataset's training split to distill the base model's inherent generation behavior, requiring no human annotations (Appendix A). Unless identified as an ablation, OAR uses regional initialization followed by proposal refinement. This one-time offline target generation is highly efficient (e.g., 2.63 GPU-hours for ChartQA; Appendix E).

All benchmarks and end-to-end latency measurements (covering image preparation, routing, language-model prefill, and decoding) are conducted on a single NVIDIA H20 GPU at batch size 1.

4.2 Main Results

Table 1 shows that OAR advances the latency–quality frontier at both budgets. At 128 tokens it attains the highest quality on all four benchmarks; at 80 tokens it remains strongest while removing about 85% of the visual sequence. Figure 1 confirms the same trend across 64, 80, 128, and 192 tokens on open-ended ChartQA.

Table 2: InternVL3-8B adaptability on open-ended ChartQA at 80 retained tokens.

Method	ANLS ↑	E2E (ms) ↓
Full model	0.8024	354.35
Uniform	0.3931	187.48
DivPrune	0.2997	194.26
CDPruner	0.3273	197.56
Query relevance	0.1260	193.95
OAR (Ours)	**0.5487**	189.95

Table 3: Offline-target ablations on open-ended ChartQA at 80 retained tokens. All variants use the same Router, training pairs, optimization, token budget, and decoding settings. Only the offline target changes.

Target	ANLS ↑	EM (%) ↑	E2E (ms) ↓
CDPruner-set imitation	0.5143	32.78	177.49
HAWK-set imitation	0.7762	69.07	177.01
Output-divergence proposals	**0.8463**	**76.85**	177.46

The largest gains occur when answers require linked visual elements. At 80 tokens, OAR exceeds HAWK by 24.6 ANLS points on DocVQA and 8.8 on ChartQA by retaining table headers with cells and legends with curves. It also extends to scene text, reaching 78.43% on TextVQA with a 2.01× end-to-end speedup.

Adaptability to other base VLMs. We fit the same 1.91M-parameter Router to InternVL3-8B using 4,000 unlabeled pairs and 1.22 GPU-hours of target generation. At 80 tokens, Table 2 shows gains of 0.2214 ANLS over CDPruner and 0.2491 over DivPrune with a 1.87× speedup, demonstrating efficient adaptation across vision encoders and language backbones.

Zero-shot routing. A single Router trained on TextVQA is evaluated unchanged on DocVQA and InfographicVQA. Without target-domain calibration, it obtains 0.5145 and 0.3455 ANLS, exceeding the strongest training-free baselines by 0.0708 and 0.0150, respectively. The complete source–target matrix is provided in Appendix 6.

Deployment efficiency. On open-ended ChartQA, the full model requires 151.96 ms to first token and 72.80 ms per subsequent decoding step. Including preparation and token selection, OAR reduces these costs to 100.73/20.36 ms with 80 retained tokens and 112.25/32.23 ms with 192 retained tokens. At 80 tokens, OAR further reduces prefill allocation by 4.90× and KV-cache memory by 4.84×; complete memory measurements are provided in Appendix E.1.

4.3 Impact on Object Hallucination and Visual Dependency

We evaluate decision-critical evidence on POPE, CHAIR, and MMStar. At the same 80-token budget, OAR retains 97.6% of full-model POPE accuracy with a 1.75× speedup. Table 4 further shows lower CHAIR hallucination and stronger MMStar visual dependency than diversity- and merging-based baselines.

Table 4: **Hallucination and visual dependency at 80 tokens.** All selectors use the same token budget and generation protocol. $C_S \downarrow$ and $C_I \downarrow$ indicate lower object hallucination.

Method	**CHAIR**		**POPE**	**MMStar**
	C_S (%) ↓	C_I (%) ↓	Acc (%) ↑	Acc (%) ↑
Full model	44.2	11.5	86.77	58.6
FastV	48.6	13.8	84.49	37.2
HAWK	50.1	14.1	83.02	39.5
DivPrune	61.2	19.3	83.80	45.1
ToMe	59.8	18.7	77.9	44.8
OAR(Ours)	**46.5**	**12.4**	**84.67**	**54.2**

Figure 4 provides a representative example of the retained evidence underlying the quantitative gains.

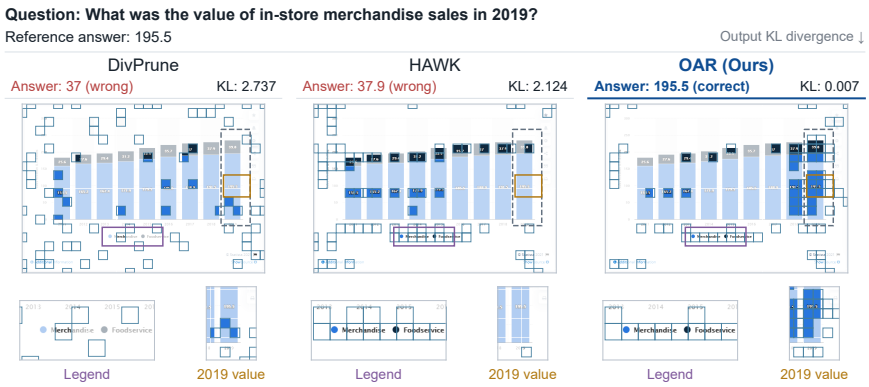


Figure 4: Retained evidence on ChartQA at 80 tokens. Zoomed views show the legend and queried 2019 bar. Each panel reports the answer and output KL divergence from the full model (lower is better). OAR preserves the complete legend–bar–value chain.

4.4 Ablation Study

We evaluate the core components of OAR on open-ended ChartQA at 80 tokens, reporting both ANLS and base-model fidelity (the percentage of greedy generation trajectories that match the unpruned base model).

Table 5 shows that regional structure and proposal refinement are complementary, together raising ANLS by nearly 20 points over uniform supervision. With the same Router capacity, data, and latency, output-divergence targets improve ANLS by 7.0 points and exact match by 7.8% over the strongest proxy target (Table 3). Figure 5 reaches full-data ANLS with approximately 2,200 unlabeled pairs and remains effective below 100 pairs.

Table 5: Component ablations on ChartQA at 80 retained tokens.

Variant	Fidelity ↑	ANLS ↑
OAR (Full)	**92.83%**	**0.8463**
Target Identity (Offline Construction)		
w/o Regional Interventions	81.40%	0.7215
w/o Proposal Refinement	88.25%	0.7960
Replace w/ Uniform Targets	74.12%	0.6540
Router Architecture (Online)		
w/o Text Condition q	79.55%	0.6988
w/o Fourier Coord $\psi(z_i)$	85.30%	0.7512

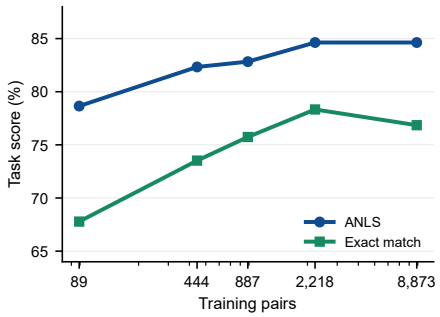


Figure 5: Data efficiency of output-aligned Router training.

The Router's architectural inputs independently contribute to this performance: removing the question embedding q lowers ANLS to 0.6988, while removing the continuous Fourier coordinates $\psi(z_i)$ lowers it to 0.7512 (Table 5), validating the necessity of both query-awareness and robust spatial grounding.

5 Conclusion and Limitations

We presented OAR, which distills output changes from joint subset interventions into a lightweight 1.91M-parameter Router. Its single $O(N)$ pass selects compact visual contexts with sub-millisecond overhead before language-model prefill, advancing the accuracy–efficiency frontier across charts, documents, and scene text.

While transferring OAR to a different base VLM requires regenerating offline targets and retraining the Router, this adaptation is lightweight and relies entirely on unlabeled data. Extending this output-aligned routing principle to multi-turn interactions, multi-image contexts, and video streams represents a promising direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to draft portions of the manuscript and edit the paper for clarity and style. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *CVPR*, 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *ICLR*, 2023. URL https://github.com/facebookresearch/ToMe.

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 *ECCV*, pp. 19–35, 2024a. doi: 10.1007/978-3-031-73004-7_2.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024b.

Zengjie Chen, Yuxiang Cai, Jingcai Guo, Taotao Cai, Jianwei Yin, and Zhi Chen. Accelerating multimodal large language models with prior-corrected token reduction. In *European Conference on Computer Vision*, 2026. URL https://github.com/CodeChildCZJ/PriorTR.

Joonmyung Choi, Sanghyeok Lee, Jongha Kim, Sehyung Kim, Dohwan Ko, Jihyung Kil, and Hyunwoo J. Kim. Docprune: Efficient document question answering via background, question, and comprehension-aware token pruning. In *CVPR*, pp. 3543–3552, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Choi_DocPrune_Efficient_Document_Question_Answering_via_Background_Question_and_Comprehension-aware_CVPR_2026_paper.html.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pp. 235–251, 2016.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. 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*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, pp. 2200–2209, 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*, pp. 1697–1706, 2022. doi: 10.1109/WACV51458.2022.00176.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pp. 4035–4045, 2018. doi: 10.18653/v1/D18-1437.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pp. 8317–8326, 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. URL https://openreview.net/forum?id=Dxb6gBJHby.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. FastVLM: Efficient vision encoding for vision language models. In *CVPR*, pp. 19769–19780, 2025.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. In *EMNLP*, pp. 9961–9980, 2025. doi: 10.18653/v1/2025.emnlp-main.505. URL https://aclanthology.org/2025.emnlp-main.505/.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *AAAI*, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

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 *CVPR*, pp. 14593–14603, 2025.

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 *CVPR*, pp. 19792–19802, 2025.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *CVPR*, pp. 24972–24982, 2025.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=KZYhyilFnt.

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 *ICCV*, pp. 20857–20867, 2025a.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *NeurIPS*, 2025b. URL https://github.com/Theia-4869/CDPruner.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparse-VLM: Visual token sparsification for efficient vision-language model inference. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857, 2025c. URL https://proceedings.mlr.press/v267/zhang25s.html.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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 *CVPR*, pp. 39583–39592, 2026.

A Implementation and Setup Details

Target construction. The frozen base VLM is evaluated in inference mode. The regional initialization uses a relative $g \times g$ grid, $|\mathcal{A}| = 16$, $K = 128$, and a validation-selected softmax temperature. The proposal pool uses $M = 3$ complete 80-token sets (regional, conditional-diversity, and head-aware) with $\tau_p = .01$. The shared full-model trajectory is cached once per example and reused across all reduced-context calls. Regional and proposal construction require respectively $1 + G$ and $1 + M$ base-model executions per example. For every shared prefix, the cache stores the top-128 probabilities and one aggregated tail bin together with candidate masks, margins, and model outputs.

Dynamic-resolution coordinates. Qwen2.5-VL supplies native temporal, height, and width positions for every visual token. We normalize the two spatial axes independently according to Eq. 5, apply the $g \times g$ partition in this relative coordinate system, and sample $\mathcal{A}$ uniformly over the example's native raster order. Consequently, neither the Fourier branch nor the intervention grid assumes a fixed N, image size, or aspect ratio. The online computation remains one $O(Nw)$ scoring pass, and the selected tokens retain their native 3-D MRoPE triples during gathering.

Split provenance. Offline targets and Router fitting use training partitions only. ChartQA uses 8,873 fit examples and a disjoint 398-example validation split, with the 540-example test untouched; DocVQA and InfographicVQA use 8,000 fit and 1,000 validation examples before evaluation on their disjoint official sets. Every cache row stores the dataset identifier and split membership, and checkpoint selection is performed only on the validation partition.

Question state and composition. The question vector is the mean of frozen input-embedding vectors at attended non-image prompt positions; no language-model Transformer block is executed. Main proposal-refined results apply Router top-B directly. The 64-routed plus 16-residual policy is retained only as a controlled regional-target variant.

Router architecture and training. Both visual and question projections map 3,584 dimensions to width 256 using Linear–LayerNorm–GELU blocks. The geometry branch maps the 10-dimensional Fourier encoding through $10 \to 256 \to 256$ with GELU, and a learned scalar initializes $\tau_r = 0.2$. The Router has 1,905,153 parameters. Regional initialization uses AdamW for 15 epochs with learning rate 3×10^{-4}, weight decay 0.01, gradient-norm clipping at 1.0, and batch size 8. Proposal targets fine-tune the same architecture for five epochs at 10^{-4}, with checkpoint selection using validation teacher cross-entropy.

Token gathering and timing. Selected indices are sorted by raster position and gathered once. Their native Qwen2.5-VL temporal/height/width position-ID triples are gathered in the same order; non-visual position IDs remain those returned by the base model's native `get_rope_index`. The budget counts only visual tokens entering the language model. CUDA synchronization brackets every component, and we account for $T_{\mathrm{E2E}} = T_{\mathrm{prep}} + T_{\mathrm{router}} + T_{\mathrm{gather/remap}} + T_{\mathrm{prefill}} + T_{\mathrm{decode}}$, together with generated-token count, time per decoded token, and TTFT. Offline teacher calls are accounted for separately. All main latency numbers use one NVIDIA H20 (96GB), batch size one, bf16, and identical deterministic decoding settings.

B Mathematical Proofs and Theoretical Analysis

B.1 Proof of Proposition 1

Write $U_x(A) = w_x(A) + e_x(A)$, where $w_x(A) = \sum_{i \in A} m_x(\{i\})$. Let P be the top-b singleton set and A^* an optimal size-b set. Since P maximizes w_x,

$$U_x(A^*) - U_x(P) = w_x(A^*) - w_x(P) + e_x(A^*) - e_x(P) \leq 2\epsilon. \tag{11}$$

If every higher-order coefficient vanishes, then $e_x(A) = 0$ for all A, so the bound is zero and singleton ranking is exact.

B.2 Proof of Corollary 1

Let $p^{\mathrm{f}} = p_{\mathcal{V},t}$, $p^{\mathrm{r}} = p_{\mathcal{S},t}$, and $d_{\mathrm{TV}} = \mathrm{TV}(p^{\mathrm{f}}, p^{\mathrm{r}})$. For the full-model winner a and any $b \neq a$,

$$p^{\mathrm{r}}(a) - p^{\mathrm{r}}(b) \geq \gamma_t - 2d_{\mathrm{TV}} \geq \gamma_t - \sqrt{2R_t(\mathcal{S})}, \tag{12}$$

where the second inequality is Pinsker's inequality. The stated condition makes the difference positive for every b, preserving the next-token winner. Applying the same argument at every shared prefix preserves the full greedy trajectory.

B.3 Softmax target as an entropy-regularized optimum

For a distribution ρ over the finite candidate pool $\mathcal{C}_x$, consider

$$\mathcal{J}(\rho) = \mathbb{E}_{S\sim\rho}[\widehat{\mathcal{R}}_x^{(K)}(S)] - \tau H(\rho). \tag{13}$$

Substitution of Eq. 6 yields $\mathcal{J}(\rho) = \tau D_{\mathrm{KL}}(\rho\|\pi_x) - \tau \log Z_x$, where $Z_x = \sum_{S\in\mathcal{C}_x} \exp[-\widehat{\mathcal{R}}_x^{(K)}(S)/\tau]$. Non-negativity and strict convexity of KL on the simplex make π_x the unique minimizer.

B.4 Marginal projection of set-level supervision

For a teacher posterior ρ_x over size-B subsets, let $\bar{\mu}_{x,i} = \Pr_{S\sim\rho_x}(i \in S) = Bt_i^{\mathcal{P}}$. For any deterministic size-B set A,

$$\mathbb{E}_{S\sim\rho_x}|A \cap S| = \sum_{i\in A} \bar{\mu}_{x,i}. \tag{14}$$

Thus selecting the top-B marginals is the Bayes-optimal fixed-size projection under expected overlap. The controlled target comparison in Table 3 holds the Router, training pairs, budget, and online path fixed: output-divergence supervision reaches 0.8463 ANLS, compared with 0.7762 for the strongest proxy-imitation target. The factorized online scorer therefore amortizes the marginal preferences induced by complete-set evaluations rather than recomputing interactions at deployment.

B.5 Why diversity does not guarantee answer preservation

Consider candidate groups a, b, c and budget two. Let a and b be arbitrarily close in feature space and c arbitrarily far from both. Suppose the decision requires the conjunction $\{a, b\}$, so $U(\{a, b\}) = 1$ and every selected pair containing c has utility zero. A max–min diversity objective selects c and one of a, b, although its decision utility is lower. Feature distances can be rescaled without changing these utilities, so dispersion alone cannot guarantee preservation of the model's decision.

Table 6: ChartQA diagnosis at an 80-token budget.

Selector	ANLS ↑	Fidelity (%) ↑	Gen. (ms) ↓
Diversity proxy	0.5178	27.08	98.83
Output divergence	**0.6470**	**39.58**	98.53
Full model	0.9306	100.0	–

C Extended Quantitative Results and Ablations

C.1 Complete Controlled Budget Matrix

Table 7 complements the open-ended evaluation with the controlled binary setting. Reporting quality and latency together exposes how each selector responds to a larger retained-token budget.

Table 7: **Controlled binary ChartQA across token budgets.** Entries are accuracy (%)/median end-to-end latency (ms). VisionZip uses its released Eager implementation. † denotes the full-data group-context Router with 64 routed tokens and 16 residual anchors.

Method	64 tokens	80 tokens	128 tokens	192 tokens
DART	74.81/104.14	79.81/106.79	85.74/111.00	90.93/118.01
VisionZip	88.52/190.55	93.15/190.87	96.30/196.90	96.48/205.10
PriorTR	94.44/107.02	94.26/107.81	95.37/112.41	95.74/118.41
FastV	92.22/105.38	93.33/108.68	94.81/111.55	95.93/117.79
OAR	90.74/96.60	92.41/99.54	94.26/104.40	95.37/111.88
OAR†	–	95.19/103.81	–	–

C.2 Sensitivity and Structural Ablations

Table 8 isolates target identity, regional granularity, residual anchors, and intervention context while keeping the online Router fixed. These controls separate offline design choices from deployment capacity.

Table 8: **Structural design choices on ChartQA.** The online Router is fixed.

Factor	Variants	Accuracy (%)
Teacher target	uniform / shuffled / matched KL	80.37 / 81.48 / **94.63**
Group grid	2×2 / 4×4 / 8×8	86.30 / 92.41 / **94.63**
Residual anchors at an 80-token budget	0 / 16	91.67 / **92.41**
Teacher context	conditional marginal / fixed anchor	93.70 / **95.19**

The vocabulary approximation is changed without changing candidate masks. Moving from a binary answer-candidate KL to top-128-plus-tail increases the mean number of unique divergence values per validation example from 10.74 to 64.98 and reduces the tied-pair rate from 12.25% to 0.016%.

Table 9: **Sensitivity to target and Router settings at an 80-token budget.** Accuracy and base-model fidelity are percentages; output divergence is measured on the selected tensor.

Factor	Setting	Accuracy ↑	Fidelity ↑	Divergence ↓
Softmax temperature	$\tau = .03$	**95.56**	95.19	**.06305**
	$\tau = .05$	95.00	94.63	.06583
	$\tau = .10$	95.00	95.00	.06927
	$\tau = .20$	94.07	94.44	.08608
Candidate pool	6×6, 16 anchors	**95.37**	**95.37**	.08182
	8×8, 0 anchors	95.00	95.00	.07804
	8×8, 8 anchors	95.19	95.19	.07532
	8×8, 32 anchors	94.44	94.44	**.07258**
Router width	128	94.81	95.56	**.07235**
	192	**95.56**	**96.30**	.07394
	256	95.19	95.56	.07850
	512	**95.56**	95.56	.07752

A sharp softmax target produces the strongest executed-set result, reaching 95.56% accuracy and 0.06305 output divergence at $\tau = .03$. Candidate interventions provide complementary operating points: the 6×6 grid with 16 anchors maximizes accuracy, while additional anchors reduce divergence. Router performance saturates by width 192; the 256-width, 1.91M-parameter configuration remains on the same high-accuracy plateau.

Table 10: Proposal-target sensitivity on open-ended ChartQA at 80 retained tokens.

Factor	Setting	ANLS ↑	EM (%) ↑
τ_p	.01	**.8463**	**76.85**
	.03	.8358	76.11
	.10	.8238	74.26
Support points	Regional + HAWK	.8280	76.11
	Regional + CDPruner	.6604	52.04
	Regional + HAWK + CDPruner	**.8463**	**76.85**

Table 11: Top-K probability-aggregation sensitivity on open-ended ChartQA at 80 retained tokens.

K	Step coverage (%)	Sequence coverage (%)	Precision (%)
32	87.88	76.11	100.0
64	87.88	76.11	100.0
128	87.88	76.11	100.0
256	87.88	76.11	100.0

The default sharp proposal target and the complete three-family support set achieve the strongest open-ended result. The certificate statistics are unchanged from top-32 through top-256 aggregation, showing that the retained probability mass already captures the decision-relevant output distribution at small K.

Table 12: **Multi-budget offline-target ablation on open-ended ChartQA.** For each row, one frozen Router checkpoint is evaluated at every retained-token budget; each entry reports ANLS / EM (%).

Target	64 tokens	80 tokens	128 tokens	192 tokens
CDPruner-set imitation	0.4532 / 24.63	0.5143 / 32.78	0.6592 / 52.41	0.7560 / 65.00
HAWK-set imitation	0.6938 / 58.89	0.7762 / 69.07	0.8531 / 79.26	0.8879 / 83.33
Regional-set imitation	0.6489 / 46.67	0.6649 / 51.30	0.7376 / 61.67	0.7955 / 70.19
Output-divergence proposals	**0.8045 / 71.11**	**0.8463 / 76.85**	**0.8810 / 83.52**	**0.8983 / 84.81**

Table 12 shows that output-divergence supervision is strongest at every budget. Its advantage over the best single-family imitation is largest at 64 retained tokens (0.8045 versus 0.6938 ANLS) and remains at 192 tokens (0.8983 versus 0.8879), showing that the target-quality gain persists from aggressive to relaxed compression.

C.3 Matched-Training Controls

The matched-capacity target-quality comparison is reported in the main paper (Table 3). This appendix complements it with the training data-efficiency curve.

Table 13: **Router data efficiency at an 80-token budget.** Training subsets are nested and use no answer labels; all rows share the same 540 evaluation examples.

Fit pairs	ANLS ↑	EM (%) ↑
89	0.7865	67.78
444	0.8233	73.52
887	0.8283	75.74
2,218	**0.8463**	**78.33**
8,873	**0.8463**	76.85

The Router is effective in the few-example regime and reaches its full-data ANLS level with one quarter of the available pairs. A small unlabeled calibration set is therefore sufficient to train a single-pass selection policy.

C.4 Paired Statistical Analysis

The main paper reports effect sizes and trends, while paired significance is recorded here. On ChartQA, OAR improves ANLS over matched HAWK by 0.0881, with a paired bootstrap 95% confidence interval of $[0.0576, 0.1126]$. Exact-match outcomes contain 76 wins and 22 losses, giving $p = 3.90 \times 10^{-8}$. In the controlled ChartQA diagnosis, complete-subset output evaluation improves exact match over diversity with 23 wins and no losses ($p = 2.38 \times 10^{-7}$), and improves base-model fidelity with 24 wins and no losses ($p = 1.19 \times 10^{-7}$).

C.5 Cross-domain Zero-shot Routing

Figure 6 reports every source–target direction without regenerating targets or updating Router parameters.

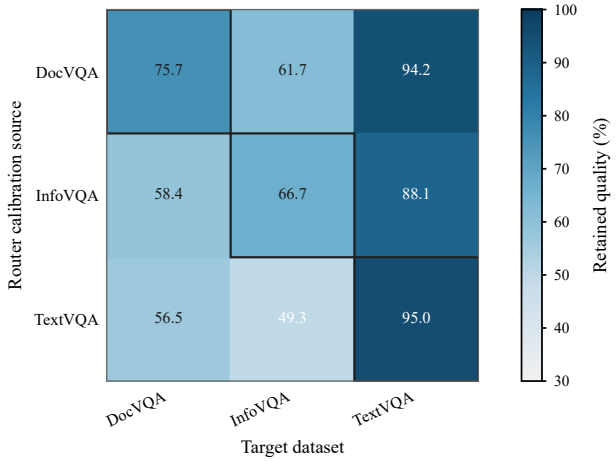


Figure 6: **Cross-domain routing at 80 retained tokens.** Values are target-task quality retained relative to the unpruned model. Diagonal cells are in-domain references and off-diagonal cells are zero-shot transfers.

C.6 Dynamic-Resolution Robustness

Table 14 groups examples by aspect ratio to test routing under different native image layouts. The visual-token budget remains fixed at 80 across all strata.

Table 14: **Robustness to dynamic resolution at an 80-token budget.** Aspect strata use the official DocVQA and InfographicVQA test images; N is the visual-token count before routing.

Dataset	Aspect stratum	Examples	Mean N	Uniform ANLS	DivPrune ANLS	OAR ANLS
DocVQA	portrait (< 0.5)	57	549.3	0.2142	0.2981	**0.4533**
	ordinary (0.5–2)	5,171	558.0	0.2957	0.4013	**0.6900**
	landscape (> 2)	121	531.7	0.4015	0.5172	**0.7895**
InfographicVQA	portrait (< 0.5)	1,529	544.8	0.2891	0.2793	**0.4530**
	ordinary (0.5–2)	1,266	547.3	0.3040	0.3127	**0.4856**

OAR exceeds both baselines in the reported aspect strata. The DocVQA portrait and landscape strata contain 57 and 121 examples, respectively, so their estimates have less sample support than the ordinary-aspect stratum. The comparisons use the native aspect-ratio strata listed in the table. On InfographicVQA, the lowest and highest token-count quartiles obtain 0.4716 and 0.4715 ANLS; on DocVQA the corresponding values are 0.6838 and 0.7045. These comparisons describe performance across the observed native token counts.

C.7 Permutation-based Conditional Marginal Targets

For a random group permutation σ, let $P_{\sigma,j}$ contain the groups preceding C_j. The conditional-marginal variant assigns

$$\phi_x(j) = \mathbb{E}_\sigma \left[\widehat{\mathcal{R}}_x^{(K)}(A \cup P_{\sigma,j}) - \widehat{\mathcal{R}}_x^{(K)}(A \cup P_{\sigma,j} \cup C_j) \right]. \tag{15}$$

The expectation is approximated with four fixed permutations per example. This variant expands the intervention context but requires substantially more base-model calls. Table 8 compares it with the canonical group-plus-anchor target through the same online Router and budget.

C.8 Evaluation of Router-selected Subsets

Let S_x^* minimize exact output divergence over all size-B sets, let $S_{\mathcal{C}}^*$ minimize it within the finite target pool $\mathcal{C}_x$, and let $\widetilde{S}_{\mathcal{C}}$ minimize the practical trajectory surrogate within that pool. For the Router output S_θ,

$$\mathcal{R}_x(S_\theta) - \mathcal{R}_x(S_x^*) = [\mathcal{R}_x(S_\theta) - \mathcal{R}_x(\widetilde{S}_{\mathcal{C}})] + [\mathcal{R}_x(\widetilde{S}_{\mathcal{C}}) - \mathcal{R}_x(S_{\mathcal{C}}^*)] \tag{16}$$

$$+ [\mathcal{R}_x(S_{\mathcal{C}}^*) - \mathcal{R}_x(S_x^*)]. \tag{17}$$

The first term is measured by re-executing the deployed top-B set, the second by changing the top-K probability aggregation while holding candidate masks fixed, and the third by expanding the regional and heterogeneous-proposal pools. This decomposition attributes approximation error to a specific stage without treating the finite candidate pool as the unknown global optimum.

We evaluate the Router's own top-B set rather than selecting the best cached proposal. At an 80-token budget on ChartQA, this set reaches 0.8463 ANLS, while the strongest individual proposal family, HAWK, reaches 0.7582 under the same base VLM and evaluation protocol. The gain beyond every discrete support family shows that the weighted marginal is not restricted to reproducing one cached proposal mask.

C.9 Empirical Margin and Certificate Distributions

Figure 7 shows the quantities underlying Table 15. The margin distribution is concentrated above the divergence thresholds, and increasing the retained budget shifts the ratio $\gamma_t/\sqrt{2R_t(\mathcal{S})}$ toward the certified side of the boundary. The exact and top-128-plus-tail conditions produce identical aggregate coverage in this experiment, while the certificate remains selective: it certifies 76.11% and 83.33% of full trajectories at 80 and 128 tokens rather than inferring fidelity for uncovered cases.

Table 15: Empirical next-token certificate on open-ended ChartQA. Step coverage reports exact/top-128-plus-tail KL.

Tokens	Examples	Step coverage	Precision	Seq. coverage	Fidelity
80	All	87.88/87.88	100.00	76.11	100.00
80	Correct	90.82/90.82	100.00	79.45	100.00
128	All	91.00/91.00	100.00	83.33	100.00
128	Correct	93.65/93.65	100.00	86.69	100.00

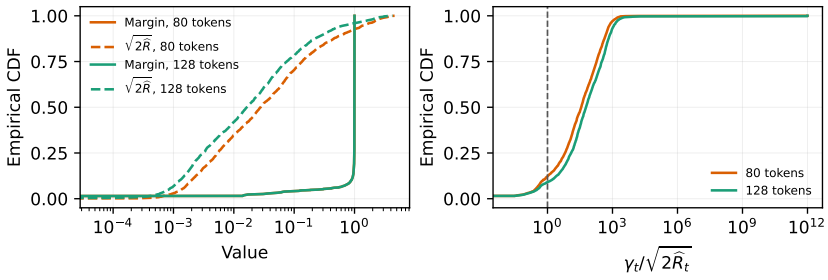
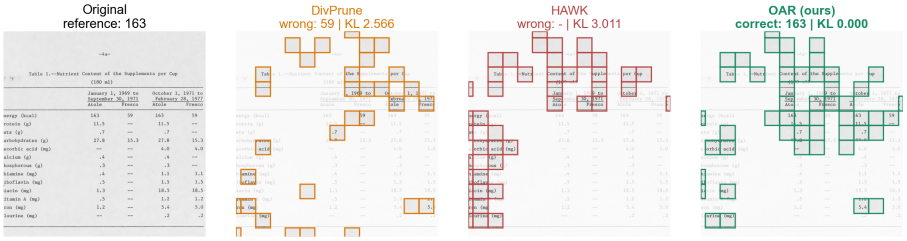


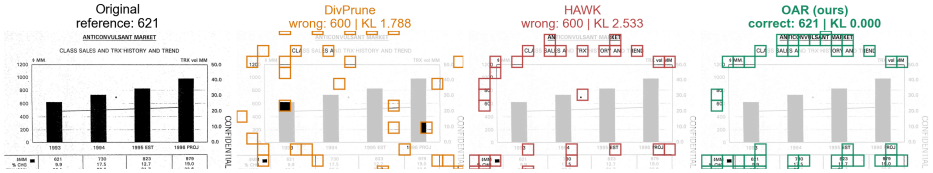
Figure 7: Empirical margin and certificate distributions on open-ended ChartQA at 80 and 128 retained tokens. The dashed vertical line marks the certificate boundary.

D QUALITATIVE RESULTS AND VISUALIZATIONS

Figures 8 and 9 extend the retained-evidence analysis across documents and infographics. On DocVQA, OAR jointly preserves the queried nutrient row, period header, and value in Figure 8(a). In Figure 8(b), it retains the 1993 label, the sales series, and the corresponding value 621, while DivPrune and HAWK return 600.



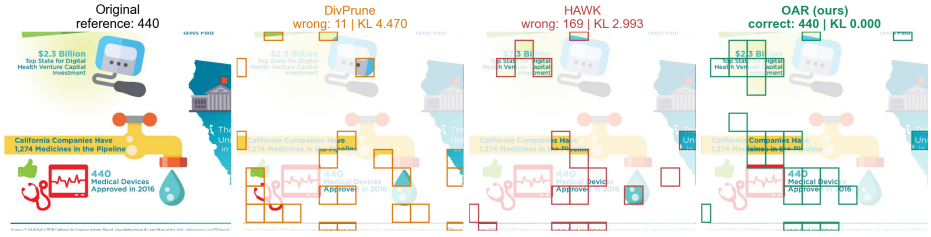
(a) What is the value of Atole for "Energy" during the time period January 1, 1969 to September 30,1971?



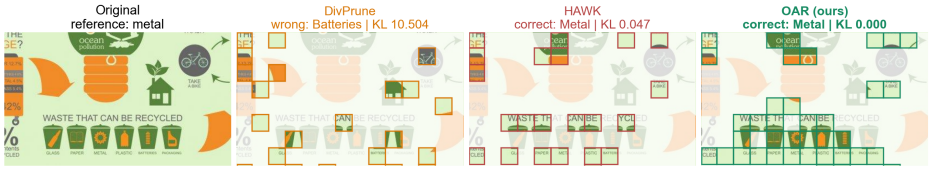
(b) What was the Sales ($MM) in 1993?

Figure 8: Additional DocVQA evidence at an 80-token budget. Each row shows the same enlarged region across methods; boxes denote retained tokens.

On InfographicVQA, Figure 9(a) shows that OAR concentrates evidence around the "440 medical devices" statement instead of unrelated prominent numbers. Figure 9(b) requires reading an ordered list: OAR retains the recyclable-item sequence and recovers its third entry, Metal.



(a) What is the number of medical devices approved in 2016?



(b) Which is listed third among the types of waste that can be recycled?

Figure 9: Additional InfographicVQA evidence at an 80-token budget. Each row shows the same enlarged region across methods; boxes denote retained tokens.

TextVQA further tests fine-grained associations among nearby text instances. Figure 10 contains several phone numbers; OAR retains the business name and its corresponding number, 370 1010, whereas DivPrune and HAWK combine digits from unrelated signs.

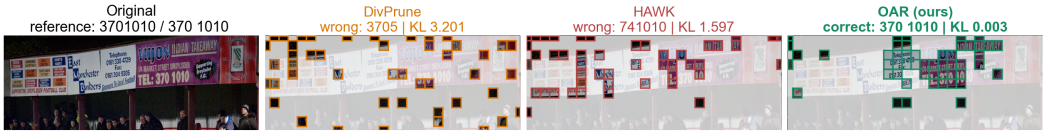


Figure 10: Additional TextVQA evidence at an 80-token budget. Outlined patches are selected; dimmed patches are discarded.

E Efficiency and Cost Accounting

E.1 Sequence-memory measurements

Table 16 separates sequence-dependent allocations from total peak GPU memory. At 80 tokens, OAR reduces prefill allocation and KV-cache memory by 4.90× and 4.84×, respectively; total peak memory also includes the frozen model and other allocations.

Table 16: **Memory efficiency on open-ended ChartQA at 80 retained tokens.** Measurements use one H20, batch size one, bf16, SDPA, and 540 examples. Parentheses report reduction relative to the full model.

Method	Visual Tokens	Prefill Alloc. (MiB)	KV Cache (MiB)	Peak GPU (MiB)
Full model	559.2	113.27 (1.00×)	33.086 (1.00×)	16025.72
FastV	80	119.35 (0.95×)	6.836 (4.84×)	16031.80
HAWK	80	23.64 (4.79×)	6.836 (4.84×)	**16019.01**
DivPrune	80	23.64 (4.79×)	6.836 (4.84×)	**16019.01**
CDPruner	80	23.64 (4.79×)	6.836 (4.84×)	**16019.01**
OAR (Ours)	80	**23.13 (4.90×)**	**6.836 (4.84×)**	16026.28

E.2 Offline target-generation cost

Table 17 reports the one-time cost of constructing the training targets. Regional and proposal costs are listed separately: together they require 2.63 GPU-hours on ChartQA. No teacher calls are made during deployment.

Table 17: **One-time offline target-generation cost on one H20.** Counts include the full-model trajectory and all reduced-context evaluations.

Dataset / target	Budget	Examples	Calls / example	GPU-hours
ChartQA regional ($G = 16$, 16 anchors)	80	9,271	17	1.13
ChartQA proposal ($M = 3$)	80	9,271	4	1.50
DocVQA regional ($G = 64$, 16 anchors)	80	9,000	65	4.51
DocVQA proposal ($M = 3$)	80	9,000	4	1.94
InfographicVQA regional ($G = 64$, 16 anchors)	80	9,000	65	4.02
InfographicVQA proposal ($M = 3$)	80	9,000	4	1.51
TextVQA regional ($G = 16$, 16 anchors)	80	9,000	17	1.88
TextVQA proposal ($M = 3$)	80	9,000	4	1.34
Deployment teacher	any	–	0	0