BORROWING DIFFUSION'S GAZE FOR VLM VISUAL TOKEN PRUNING

Anonymous authors
Paper under double-blind review

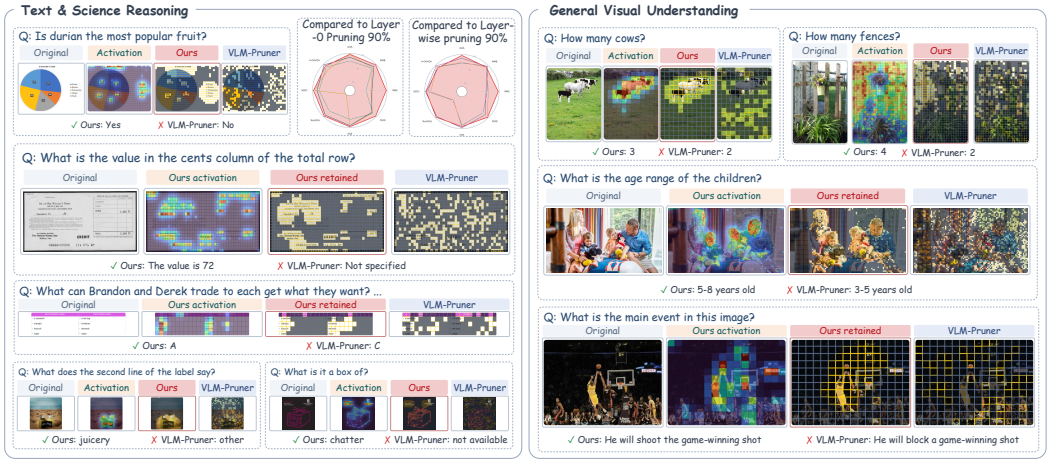


Figure 1: Our DiffScorer preserves query-relevant visual evidence tokens across general visual understanding and text/science reasoning tasks, enabling more accurate answers while operating entirely at Layer 0.

ABSTRACT

Vision–language models (VLMs) encode images into many visual tokens, most of which are query-irrelevant yet computationally expensive. Existing methods perform one-shot or layer-wise pruning based on visual sensitivity, feature redundancy, or text–visual attention. However, these signals are largely derived from the target VLM itself and may be affected by structural biases such as attention sinks, limiting their reliability as indicators of query-relevant visual evidence. We introduce a Layer-0 pruning framework that draws on the strong text–image alignment of pretrained diffusion models to improve query-relevant visual evidence selection. Specifically, we distill diffusion text–image attention into **DiffScorer**, a lightweight scorer that provides an external signal for selecting query-relevant visual tokens before LLM prefill. To preserve this strong distilled prior while correcting its residual attention biases, **counterfactual Swap-GRPO** explores budget-preserving token exchanges around DiffScorer's deterministic selection and optimizes the swap trajectories with answer rewards from a frozen VLM. Only DiffScorer is updated in both stages. Extensive experiments across VLM architectures, pruning ratios, and visual understanding benchmarks demonstrate that DiffScorer consistently outperforms competitive one-shot, layer-wise, and learned pruning methods. Notably, for high-resolution inputs containing over 10K visual tokens, DiffScorer achieves a $1.50\times$ improvement in end-to-end generation throughput while preserving 99.5% of the unpruned accuracy.

1 INTRODUCTION

Vision–language models (VLMs) have made substantial progress in visual question answering and multimodal reasoning (Liu et al., 2024a; Bai et al., 2025c;b; Li et al., 2025), yet their inference cost

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.

grows rapidly with visual input resolution. A single high-resolution image can produce thousands or even tens of thousands of visual tokens (Chen et al., 2024b), increasing language-model prefill cost, memory consumption, and subsequent decoding overhead. Many of these tokens, however, are redundant for a given query. Retaining sufficient visual evidence to support accurate answers under a limited token budget is therefore a central challenge in efficient VLM inference.

Existing methods explore different paradigms for visual token pruning. One line of work performs one-shot pruning before the language model or within its early layers, allowing discarded tokens to bypass most or all of the computationally expensive LLM backbone (Wen et al., 2025; Chen et al., 2024a; Alvar et al., 2025; Yang et al., 2025b). Another performs layer-wise pruning across language-model layers, exploiting richer multimodal context to mitigate the risk of prematurely removing critical information (Ye et al., 2025; Liu et al., 2024b; Xing et al., 2024; Zhang et al., 2025b). These strategies reflect a trade-off between early compression and informed selection: earlier pruning offers greater computational savings but requires identifying relevant evidence with less contextual information. This challenge is particularly acute at Layer 0, where mistakenly discarded evidence may be difficult to recover through subsequent reasoning.

A key challenge is thus to obtain reliable importance signals for early token selection. Existing approaches typically rely on token sensitivity, feature redundancy, or text-to-visual attention from shallow or deep VLM layers (Kim et al., 2026; Chen et al., 2026; Fang et al., 2026; Takezoe et al., 2026). Despite their differences, these approaches largely share a common premise: signals produced by the VLM itself can identify the visual regions needed to answer the query. However, recent studies of cross-modal attention in VLMs suggest that these internal signals can exhibit attention sinks, attention shifts, and other structural biases (Yang et al., 2025a; Zhang et al., 2025a; Kang et al., 2025; Lin et al., 2025). Such distortions can propagate to token selection, causing pruning methods to overlook critical visual evidence (Bai et al., 2025a; Endo et al., 2025).

To address this issue, we turn to visual generative models as a source of external supervision for visual evidence selection. Text-conditioned diffusion models are trained to synthesize visual content from language descriptions (Rombach et al., 2022; Peebles & Xie, 2023), encouraging alignment between text tokens and image regions. Motivated by this property, we use cross-modal attention from a pretrained diffusion model as an external visual grounding prior. Specifically, we extract semantic phrases of relevant visual evidence from the VQA sample, obtain the corresponding text–image attention maps from the diffusion model, and distill them into **DiffScorer**, a lightweight, query-conditioned visual token scorer. DiffScorer operates directly on the VLM's Layer-0 visual features and predicts their relevance to the input query prior to LLM prefill. As illustrated in Fig. 1, this external prior yields more complete query-relevant evidence and more accurate answers across general VQA and text/science reasoning than representative VLM-internal selectors.

Although diffusion attention provides a useful spatial prior, it may retain residual attention biases from diffusion teacher. Therefore, we employ reinforcement learning to address these limitations. Crucially, however, DiffScorer's deterministic selection already covers most useful regions after distillation. Sampling arbitrary masks would discard this strong initialization and produce many uninformative rollouts, whereas the remaining errors often lie near the keep–discard boundary: answer-critical tokens can retain locally high scores yet lose out under the fixed budget. We therefore introduce **counterfactual Swap-GRPO**, which explores locally around the deterministic mask by exchanging retained and discarded tokens without replacement, creating controlled counterfactual comparisons that better exploit the diffusion prior. Answer rewards from the frozen VLM and group-relative advantages then optimize each swap trajectory, suppressing semantically related but answer-irrelevant tokens while recovering omitted evidence.

We conduct extensive experiments across multiple VLM architectures, pruning ratios, and visual understanding benchmarks. Compared with representative one-shot and layer-wise pruning methods, our framework consistently achieves a better accuracy–efficiency trade-off. On Qwen3-VL, it retains 92.04% of the unpruned model's average performance while pruning 90% of visual tokens. These benefits extend to high-resolution inputs containing over 10K visual tokens: on HR-Bench 8K, pruning 80% of visual tokens achieves $1.50\times$ the end-to-end generation throughput of the unpruned model while retaining 99.5% of its performance. We further compare DiffScorer with strong learnable pruning baselines trained using VLM-internal signals in our ablation studies, supporting the effectiveness of our diffusion-distilled scorer.

Our main contributions are as follows:

- We introduce **DiffScorer**, a lightweight, query-conditioned scorer that distills text–image alignment knowledge from a pretrained diffusion model, providing an external visual evidence grounding prior for Layer-0 visual token pruning.
- We propose **counterfactual Swap-GRPO**, which constructs budget-preserving token swaps through sampling without replacement around a deterministic retained set, and refines the scorer using answer rewards and swap-trajectory probabilities.
- We evaluate our framework across multiple VLM architectures, token budgets, and visual understanding benchmarks, demonstrating strong performance retention and improvements in end-to-end generation throughput.

2 Related Work

2.1 Visual Token Reduction

Visual-token reduction is widely used to improve VLM inference efficiency. One-shot methods remove tokens before or within shallow language-model layers using text–vision attention, visual saliency, redundancy, or diversity criteria (Chen et al., 2024a; Zhang et al., 2025a; Wen et al., 2025; Alvar et al., 2025; Chen et al., 2026; Wu et al., 2026). Layer-wise methods instead prune or merge tokens across multiple layers to exploit increasingly informative multimodal representations (Zhang et al., 2025b; Xing et al., 2024; Hu et al., 2026; Singh et al., 2026). The former offers greater acceleration but must identify critical evidence with limited cross-modal context, whereas the latter retains more information at the cost of additional early-layer computation. To combine the advantages of both, recent approaches have further explored learned token-selection mechanisms to enable effective pruning at Layer 0 through differentiable selection and downstream supervision (Shao et al., 2025; Sun et al., 2026; Huang et al., 2025; Zhu et al., 2026; Takezoe et al., 2026). TPRL and GRIP-VLM further optimize discrete selection policies using reinforcement learning (Cao et al., 2026; Huang et al., 2026). Nevertheless, their supervision is largely derived from the target VLM's attention, representations, or outputs, and may therefore inherit its internal biases.

2.2 Reliability of VLM-Internal Signals

Recent studies have questioned whether VLM-internal signals reliably indicate task-relevant visual evidence. Tokens receiving low text-to-vision attention can still substantially affect model predictions (Yang et al., 2025a), while early pruning may produce spatially biased selections and discard fine-grained local evidence (Endo et al., 2025; Bai et al., 2025a). Attention sinks and position- or layer-dependent shifts further distort attention-based importance estimates (Kang et al., 2025; Qi et al., 2026; Sun et al., 2025). In contrast, text-conditioned diffusion models exhibit strong spatial alignment between linguistic concepts and image regions, with their cross-attention maps supporting visual attribution and grounding (Tang et al., 2023; Kondapaneni et al., 2024; Liu et al., 2024c). Instead of relying on the target VLM's internal attention or hidden states, we distill the diffusion attention prior into a query-conditioned Layer-0 scorer, DiffScorer. To correct residual biases in this prior, counterfactual Swap-GRPO explores budget-preserving token exchanges around the deterministic selection and optimizes the scorer using answer rewards from a frozen VLM.

3 Method

For an image–question pair (I, q), DiffScorer scores the frozen VLM's visual and question features before the first LLM block:

$$(\mathbf{V}, \mathbf{T}) = \Phi_0(I, q), \qquad \mathbf{s} = f_\theta(\mathbf{V}, \mathbf{T}) \in \mathbb{R}^N. \tag{1}$$

Here, $\mathbf{V} \in \mathbb{R}^{N\times d}$ and $\mathbf{T} \in \mathbb{R}^{L\times d}$ are visual and question tokens. For retention ratio ρ, a deterministic mask $\mathbf{m}$ retains $K = \mathrm{round}(\rho N)$ visual tokens for LLM prefill. Our two-stage training paradigm first distills diffusion visual evidence attention into DiffScorer, then refines it with frozen-VLM answer rewards through counterfactual Swap-GRPO. Both stages update only DiffScorer.

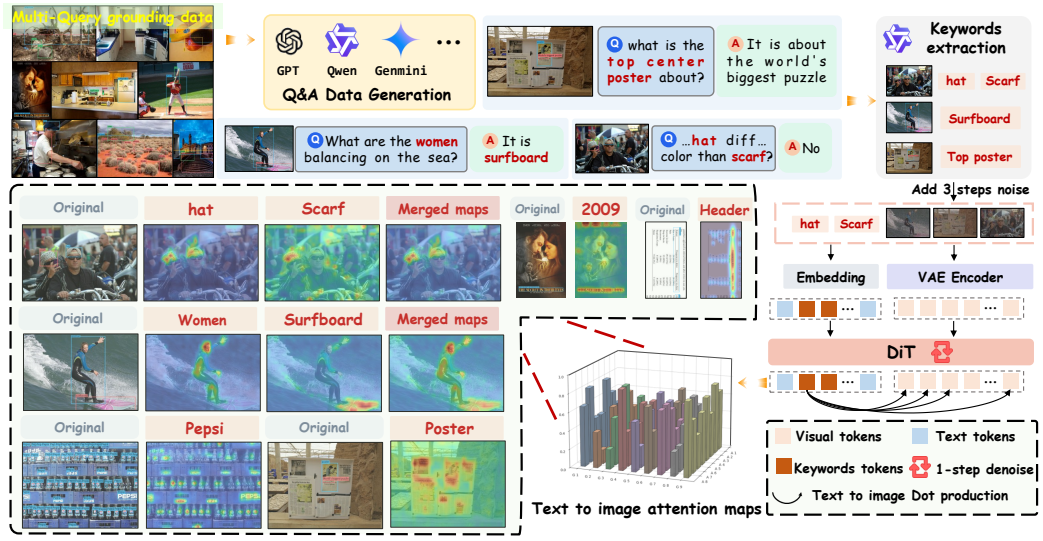


Figure 2: Overview of our diffusion-grounded supervision data generation pipeline.

3.1 Diffusion Grounded Evidence Supervision Data Generation

3.1.1 Grounding-Aware VQA Construction

Figure 2 shows the supervision data generation pipeline. Given multi-object grounding annotations $\mathcal{G} = \{(o_r, \mathbf{b}_r)\}_{r=1}^{R}$, with entity o_r and region $\mathbf{b}_r$, a frozen VLM generates multiple question–answer pairs about categories, attributes, counts, locations, and relations, while extracting $\mathcal{E}_m$ from q_m or y_m as lexical spans that refer to the visual evidence needed to answer q_m:

$$\{(q_m, y_m, \mathcal{E}_m)\}_{m=1}^{M} = \mathcal{F}_{\text{gen}}(I, \mathcal{G}), \qquad \mathcal{E}_m = \{e_{m,j}\}_{j=1}^{J_m}. \tag{2}$$

Here, q_m, y_m, and $\mathcal{E}_m$ denote the question, answer, and answer-relevant evidence terms.

3.1.2 Diffusion Visual Evidence Attention Generation

For each generated VQA sample, we first encode the image into the diffusion latent space and add the noise at timestep t_3 to the image. We then form a text condition x_m from the question, answer, and evidence terms, while recording each term's token span. A frozen diffusion model then performs one denoising step at the timestep t_3 with noisy image latent z_{t_3} and the text condition x_m as inputs:

$$\mathbf{z}_0 = \mathcal{E}_{\text{VAE}}(I), \qquad \mathbf{z}_{t_3} = \alpha_{t_3}\mathbf{z}_0 + \sigma_{t_3}\boldsymbol{\epsilon}, \qquad \widehat{\mathbf{z}} = \mathcal{D}_\phi(\mathbf{z}_{t_3}, t_3, x_m), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

Scheduler coefficients α_{t_3} and σ_{t_3} set the corruption level. Light noising and one-step denoising activate text–image correspondence while preserving input structure. In the teacher's joint attention over text and image streams, we extract the off-diagonal text-query-to-image-key block and average over all sites and heads, denoted by $\mathcal{C}$:

$$\mathbf{A}_{T\to I}^{\ell,h} = \left[\text{Softmax}\left(\frac{\mathbf{Q}^{\ell,h}(\mathbf{K}^{\ell,h})^\top}{\sqrt{d_h}}\right)\right]_{\mathcal{T},\mathcal{I}}, \qquad \overline{\mathbf{A}}_{T\to I} = \frac{1}{|\mathcal{C}|}\sum_{(\ell,h)\in\mathcal{C}} \mathbf{A}_{T\to I}^{\ell,h}. \tag{4}$$

For evidence term $e_{m,j}$ with text-token span $\mathcal{S}_{m,j}$, we average attention within the span and combine evidence maps by spatial confidence to form a final merged teacher distribution $\mathbf{p}_m^D$:

$$\mathbf{a}_{m,j} = \mathcal{N}\left(\frac{1}{|\mathcal{S}_{m,j}|}\sum_{t\in\mathcal{S}_{m,j}} \overline{\mathbf{A}}_{T\to I}[t,:]\right), \qquad \mathbf{p}_m^D = \mathcal{N}\left(\sum_{j=1}^{J_m}\omega_{m,j}\mathbf{a}_{m,j}\right), \tag{5}$$

where $\mathcal{N}(\cdot)$ denotes sum normalization over image positions and $\omega_{m,j}$ measures the spatial confidence of the corresponding keyword attention. The resulting teacher distribution $\mathbf{p}_m^D$ is resized to the Layer-0 visual-token grid of the target VLM.

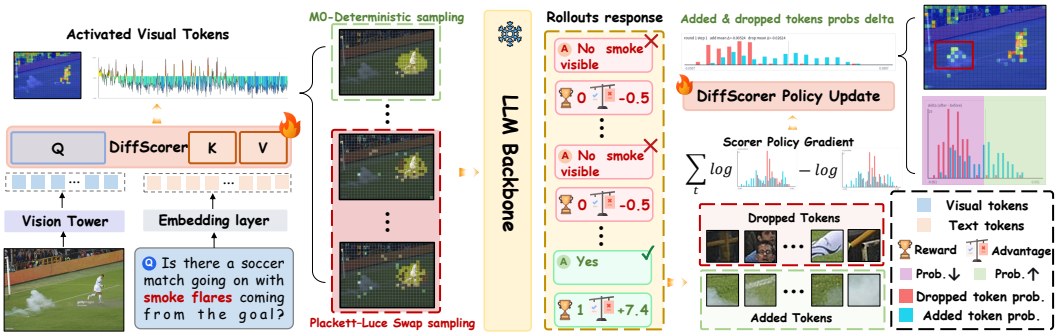


Figure 3: Counterfactual Swap-GRPO: sample budget-preserving swaps around a deterministic anchor and refine DiffScorer using frozen-VLM answer rewards.

3.2 Query-Conditioned DiffScorer

DiffScorer uses a small stack of cross-attention blocks with question text queries and visual keys and values. A lightweight network aggregates the resulting vision-aware text tokens into one query vector, whose normalized similarity to each projected visual token yields $\mathbf{s}$. We train the student distribution $\mathbf{p}_m^\theta = \mathrm{Softmax}(\mathbf{s}_m)$ with the KL objective:

$$\mathcal{L}_{\text{distill}} = D_{\text{KL}}\left(\mathbf{p}_m^D \,\|\, \mathbf{p}_m^\theta\right) = \sum_{i=1}^{N} p_{m,i}^D \log \frac{p_{m,i}^D}{p_{m,i}^\theta}. \tag{6}$$

3.3 Counterfactual Swap-GRPO

3.3.1 Why Local Swap Exploration?

After diffusion distillation, DiffScorer's deterministic mask already covers most answer-relevant regions, so sampling arbitrary masks risks discarding useful evidence and producing low-value rollouts. Yet critical tokens may still receive moderate or locally high scores and be excluded by budget competition, spatial aggregation, or dominant regions. We therefore explore local swaps around the deterministic mask to recover such evidence while preserving the distilled prior, as shown in Fig. 3.

3.3.2 Budget-Preserving Swap Policy

The deterministic anchor $\mathbf{m}_0$ retains K tokens and stays fixed within each rollout group; its construction is excluded from the action probability. We standardize valid-token scores as $z_i = (s_i - \mu_{\mathbf{s}})/(\sigma_{\mathbf{s}} + \epsilon)$. At swap step t, the drop set $\mathcal{D}_t$ contains the remaining tokens from $\mathbf{m}_0$, while the add set $\mathcal{A}_t$ contains the remaining tokens from its initial complement. We preferentially remove low-scoring retained tokens and recover high-scoring discarded tokens:

$$\pi_\theta^-(d_t \mid \mathcal{D}_t) = (1-\eta)\frac{\exp(-z_{d_t}/\tau_-)}{\sum_{i\in\mathcal{D}_t}\exp(-z_i/\tau_-)} + \frac{\eta}{|\mathcal{D}_t|}, \qquad \pi_\theta^+(a_t \mid \mathcal{A}_t) = (1-\eta)\frac{\exp(z_{a_t}/\tau_+)}{\sum_{i\in\mathcal{A}_t}\exp(z_i/\tau_+)} + \frac{\eta}{|\mathcal{A}_t|}. \tag{7}$$

A sampled token is removed from its candidate set, resulting in Plackett–Luce sampling without replacement. After r swaps(set by swap fraction), the final mask and trajectory likelihood are

$$\mathbf{m}_g = (\mathbf{m}_0 \setminus \{d_1,\ldots,d_r\}) \cup \{a_1,\ldots,a_r\}, \quad \log \pi_\theta(\tau_g \mid \mathbf{m}_0) = \sum_{t=1}^{r}\left[\log \pi_\theta^-(d_t \mid \mathcal{D}_t) + \log \pi_\theta^+(a_t \mid \mathcal{A}_t)\right]. \tag{8}$$

Each trajectory retains exactly K tokens. Rollouts differ only in a few exchanged tokens, providing controlled counterfactual comparisons at fixed image, question, and budget.

3.3.3 Policy Optimization

For each sample, the group contains one deterministic anchor and $G-1$ independent swap trajectories. The frozen VLM answers with each selected visual tokens to obtain trajectory level reward. Let $\overline{R}_{-g}$ denote the mean reward excluding rollout g and deterministic anchor $\mathbf{m}_0$, we define

$$\widehat{y}_g = F_{\text{VLM}}(I, q; \mathbf{m}_g), \quad R_g = R(\widehat{y}_g, y), \quad A_g = \frac{R_g - \frac{1}{2}R_0 - \frac{1}{2}\overline{R}_{-g}}{\widehat{\sigma}_\rho + \epsilon}. \tag{9}$$

Here, R_0 is the anchor reward and $\widehat{\sigma}_\rho$ is an exponential moving average of within-group reward standard deviation for retention ratio ρ. The advantage compares each swap with both the anchor and the other rollouts. Let $\varrho_g(\theta) = \exp[\log \pi_\theta(\tau_g) - \log \pi_{\theta_{\text{old}}}(\tau_g)]$ be the trajectory-level importance ratio. The complete objective combines clipped policy optimization, entropy regularization, and a constraint toward the diffusion-distilled initialization θ_0:

$$\mathcal{L}_{\text{RL}} = -\frac{1}{G-1}\sum_{g=1}^{G-1} \min\left(\varrho_g A_g,\ \text{clip}(\varrho_g, 1-\epsilon_c, 1+\epsilon_c)A_g\right) - \beta_H \overline{\mathcal{H}} + \beta_{\text{KL}} D_{\text{KL}}. \tag{10}$$

The deterministic rollout serves only as a counterfactual reference and is excluded from the policy gradient; only swap-trajectory probabilities backpropagate to DiffScorer. This refinement suppresses answer-irrelevant selections and recovers missed evidence (Fig. 4). At inference, DiffScorer deterministically produces an exact-budget Layer-0 mask; see Appendix A.

4 Experiments

4.1 Experimental Setup

Training data. We sample 100K multi-query grounding instances from Locate Anything (Wang et al., 2026), Qwen3.5-27B then generates 140K VQA pairs with answer-relevant evidence terms. Qwen-Image-2512 (Wu et al., 2025) supplies merged attention maps for distillation (Sec. 3.1); a 12K subset is further used for Swap-GRPO. Both stages' data are disjoint from evaluation benchmarks.

Backbones and baselines. We primarily evaluate on Qwen3-VL-8B-Instruct (Bai et al., 2025b) and validate cross-architecture generalization on LLaVA-1.5-7B (Liu et al., 2024a). For one-shot pruning, we compare against DivPrune, ZOO-Prune, PruneSId, OTPrune, and VLM-Pruner (Alvar et al., 2025; Kim et al., 2026; Fang et al., 2026; Chen et al., 2026; Wu et al., 2026), which cover diversity-, redundancy-, transport-, and query-aware selection criteria. We further compare our one-shot pruner with layer-wise methods SparseVLM, iLLaVA, and DUET-VLM (Zhang et al., 2025b; Hu et al., 2026; Singh et al., 2026), using the same final visual-token budgets. For the scorer ablation, we retrain learned selectors Dynamic-LLaVA, VisionSelector, and TwigVLM (Huang et al., 2025; Zhu et al., 2026; Shao et al., 2025) using the same questions, reference answers, and evidence terms $(q, y, \mathcal{E})$ as DiffScorer, while retaining their native VLM-derived objectives. A GRIP-VLM-style policy (Huang et al., 2026) is used to compare rollout strategies.

Benchmarks and metrics. We evaluate general visual understanding on GQA, MMBench, MME, POPE, ScienceQA-IMG, TextVQA, the image split of SEED-Bench, and A-OKVQA (Hudson & Manning, 2019; Liu et al., 2024d; Fu et al., 2025; Li et al., 2023b; Lu et al., 2022; Singh et al., 2019; Li et al., 2023a; Schwenk et al., 2022). We report accuracy or F1 as appropriate and perception-plus-cognition scores for MME. Avg. is the mean per-benchmark score divided by its unpruned score, expressed as a percentage; bold/underlined values indicate best/second-best results within each block. On HR-Bench 4K/8K (Wang et al., 2025), we report Fine-grained Single-instance Perception (FSP), Fine-grained Cross-instance Perception (FCP), and overall accuracy. Efficiency is measured by batch-1 end-to-end throughput uses identical generation settings and includes visual encoding, DiffScorer, post-processing, prefill, and decoding, excluding model/data loading.

Implementation details. Both training stages freeze the VLM backbone and update only DiffScorer. Supervised diffusion distillation is trained for four epochs, while Swap-GRPO is run for two epochs. The default RL configuration uses $G = 8$ and swap fraction 0.15; the visual retention ratio is sampled from $\{0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1\}$ during training. All experiments are conducted on one node with $8\times$ NVIDIA A100 80GB GPUs. The supervised stage takes approximately one hour, and the RL stage takes approximately 40 minutes. Detailed DiffScorer architectural specifications and training hyperparameters are reported in Appendix B.4 Table 9.

4.2 Main Results

Layer-0 pruning on Qwen3-VL. In Table 1, DiffScorer retains 98.04%, 95.30%, and 92.04% of unpruned performance at 70%, 80%, and 90% pruning ratio. At 90%, it leads on seven of eight benchmarks, supporting our central hypothesis that an external diffusion-grounded prior is particularly valuable when Layer-0 selection errors cannot be repaired by subsequent VLM layers.

Table 1: Comparative experiments of Layer-0 visual token pruning on Qwen3-VL-8B-Instruct. The best result is in **bold** and the second best is underlined.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 30% visual tokens (↓ 70%)									
DivPrune	61.96	86.92	2224	88.12	86.76	79.52	75.24	85.33	95.74%
ZOO-Prune	62.72	87.84	2162	88.48	88.20	70.65	76.47	85.24	94.78%
PruneSId	62.73	87.16	2114	87.58	88.15	71.85	76.84	84.10	94.38%
OTPrune	61.87	87.55	2217	88.65	88.15	75.73	76.44	84.89	95.59%
VLM-Pruner	62.47	87.75	2237	88.56	88.59	80.10	75.94	86.03	96.63%
Ours	**62.86**	**88.61**	**2304**	**88.98**	**91.13**	**81.48**	**77.47**	**86.20**	**98.04%**
Retain 20% visual tokens (↓ 80%)									
DivPrune	60.37	85.05	2063	87.41	84.63	75.38	73.61	83.49	92.80%
ZOO-Prune	60.32	85.08	2044	87.76	85.97	61.02	75.51	83.76	91.11%
PruneSId	60.69	85.38	2106	87.15	86.66	64.11	75.23	82.36	91.77%
OTPrune	61.08	85.31	2088	87.22	86.91	62.87	75.50	82.97	91.73%
VLM-Pruner	**61.62**	86.23	2140	87.53	86.67	**77.80**	74.93	**85.33**	94.73%
Ours	61.09	**87.41**	**2233**	**88.51**	**88.01**	75.72	**75.66**	84.63	**95.30%**
Retain 10% visual tokens (↓ 90%)									
DivPrune	55.76	80.80	1859	85.58	80.61	68.21	70.70	80.09	87.43%
ZOO-Prune	56.87	80.53	1856	83.55	83.94	43.73	72.58	77.38	83.98%
PruneSId	55.95	82.84	2004	82.24	82.75	51.76	72.23	79.91	86.07%
OTPrune	58.28	81.15	1935	85.04	84.32	40.97	73.28	78.43	84.87%
VLM-Pruner	57.26	82.30	1950	84.52	84.33	71.11	71.77	**81.92**	89.61%
Ours	**58.52**	**83.76**	**2169**	**87.83**	**85.02**	**71.88**	**73.58**	80.96	**92.04%**

Comparison with layer-wise pruning. Table 2 compares our Layer-0 selector with methods that defer part of the selection process to later LLM layers. Despite pruning before the first LLM block, DiffScorer still exceeds the strongest layer-wise baseline DUET-VLM by 0.46 and 1.12 average performance retention points at 70% and 90% pruning ratio. Thus, DiffScorer preserves more task-relevant evidence while allowing discarded tokens to bypass the entire LLM backbone.

Table 2: Comparisons of our Layer-0 selector with Layer-wise visual token pruning methods on Qwen3-VL-8B-Instruct.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 30% visual tokens (↓ 70%)									
SparseVLM	61.42	87.92	2273	88.13	89.74	78.11	77.13	85.68	96.57%
iLLaVA	61.79	87.87	2210	88.32	90.68	77.81	76.16	85.50	96.23%
DUET-VLM	**63.10**	88.49	**2309**	87.40	90.53	78.47	**77.85**	**87.42**	97.58%
Ours	62.86	**88.61**	2304	**88.98**	**91.13**	**81.48**	77.47	86.20	**98.04%**
Retain 10% visual tokens (↓ 90%)									
SparseVLM	55.32	**84.80**	2001	83.89	84.28	59.72	72.77	81.66	88.17%
iLLaVA	55.63	83.92	2018	87.56	85.17	59.56	71.28	81.74	88.58%
DUET-VLM	57.36	84.40	1983	85.70	**85.72**	71.42	**73.97**	**82.44**	90.92%
Ours	**58.52**	83.76	**2169**	**87.83**	85.02	**71.88**	73.58	80.96	**92.04%**

Generalization across VLM architectures. We instantiate the same framework on LLaVA-1.5-7B. As shown in Table 3, DiffScorer achieves the highest aggregate performance retention at all three budgets: 98.97%, 97.58%, and 95.57% (Table 3). The consistent behavior on Qwen3-VL and LLaVA indicates that the benefit is not tied to a particular vision encoder or language backbone. Additional DivPrune and PruneSId results appear in Appendix C.2.

Table 3: Comparative experiments on Layer-0 visual-token pruning on LLaVA-1.5-7B.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
LLaVA-1.5-7B	63.79	73.96	1785.12	85.94	69.46	45.97	66.16	80.87	100.00%
Retain 30% visual tokens (↓ 70%)									
ZOO-Prune	61.40	73.08	**1781**	83.44	68.81	43.71	63.91	78.86	97.52%
OTPrune	61.28	73.45	1771	85.08	68.41	42.06	63.54	79.30	97.21%
VLM-Pruner	**62.95**	73.15	1758	85.01	70.35	43.96	**65.18**	79.56	98.60%
Ours	61.57	**74.36**	1764	**85.82**	**70.45**	**44.89**	64.76	**80.09**	**98.97%**
Retain 20% visual tokens (↓ 80%)									
ZOO-Prune	60.09	72.27	1745	85.05	68.17	42.76	62.58	**79.21**	96.54%
OTPrune	60.20	73.11	1752	85.64	68.12	39.19	61.98	78.68	95.66%
VLM-Pruner	**62.40**	72.83	1716	85.95	**70.00**	43.21	63.05	78.95	97.52%
Ours	60.31	**73.55**	**1754**	**86.04**	69.45	**43.51**	**63.44**	79.04	**97.58%**
Retain 10% visual tokens (↓ 90%)									
ZOO-Prune	59.51	71.74	1731	85.57	68.71	40.56	60.51	76.76	95.04%
OTPrune	58.05	72.33	1673	84.59	68.46	32.47	59.12	**78.51**	92.07%
VLM-Pruner	**60.39**	72.51	1705	85.01	68.45	40.22	61.07	77.03	95.09%
Ours	58.31	**72.52**	**1756**	**85.78**	**69.06**	**40.66**	**61.87**	77.24	**95.57%**

High-Resolution Accuracy and Efficiency. Table 4 measures the practical benefit of DiffScorer. On HR4K, 30% retention (13,746 tokens → 4,128 tokens) raises overall accuracy from 78.13 to 79.00 and throughput from 16.72 to 21.03 tokens/s (1.26×). On HR8K, 20% retention (16,224 tokens → 3,244 tokens) preserves 99.5% of unpruned accuracy (74.25 versus 74.63) with 1.50× throughput. DUET-VLM reaches higher accuracy (75.50) but only 1.08× throughput. DiffScorer gives the largest measured speedup at both resolutions; additional baselines comparisons and their computational complexity analysis appear in Appendix C.3 and Appendix C.4 respectively.

Table 4: HR-Bench accuracy and end-to-end generation throughput (tokens/s).

Method	*HR4K: 13,746 → 4,128; retain 30% (↓ 70%)*				*HR8K: 16,224 → 3,244; retain 20% (↓ 80%)*			
	FSP Acc.↑	FCP Acc.↑	Overall Acc.↑	Throughput tok/s ↑	FSP Acc.↑	FCP Acc.↑	Overall Acc.↑	Throughput tok/s ↑
Qwen3-VL-8B	92.25	64.00	78.13	16.72 (1.00×)	85.25	64.00	74.63	14.23 (1.00×)
iLLaVA	90.00	63.50	76.75	17.62 (1.05×)	84.75	61.50	73.13	15.91 (1.12×)
DUET-VLM	**92.50**	64.50	78.50	15.55 (0.93×)	**86.00**	**65.00**	**75.50**	15.44 (1.08×)
ZOO-Prune	91.00	63.75	77.38	18.53 (1.11×)	83.25	59.50	71.38	16.97 (1.19×)
OTPrune	91.50	65.25	78.38	18.61 (1.11×)	82.25	62.00	72.13	17.69 (1.24×)
VLM-Pruner	91.25	61.25	76.25	20.22 (1.21×)	85.00	57.50	71.25	19.53 (1.37×)
Ours	91.75	**66.25**	**79.00**	**21.03 (1.26×)**	85.50	63.00	74.25	**21.29 (1.50×)**

4.3 Ablation Studies

Does diffusion supervision improve Layer-0 selection? Beyond the Layer-0 pruning methods compared in Table 1, we further compare DiffScorer with the learned selectors in Table 5. DiffScorer without RL retains 96.98% on average, exceeding Dynamic-LLaVA (94.70%), VisionSelector (95.78%), and TwigVLM (96.02%), supporting the value of an external diffusion prior. Figure 4 corroborates this result qualitatively: DiffScorer covers a more complete set of answer-relevant regions. At 20% and 10% retention, DiffScorer without RL also surpasses the strongest learned baselines, as shown in Table 11 in Appendix C.1. Because evidence boxes annotations provide a direct alternative source of supervision, we construct a controlled Box+SAM baseline to directly assess whether diffusion-derived supervision provides a more effective training signal for Layer-0 evidence selection. Specifically, we derive soft targets from annotation boxes and SAM masks while keeping training setting fixed. At 30% retention, diffusion supervision improves performance retention from 95.72% to 96.98%, supporting the advantage of its query-conditioned evidence targets. Construction details, lower-budget results, and qualitative comparisons are provided in Appendix B.1.

Table 5: Supervision and rollout ablations on Qwen3-VL-8B at 70% pruning.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	SEED$^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 30% visual tokens (↓ 70%) Supervision source									
Dynamic-LLaVA	59.46	86.28	2238	86.36	87.56	76.20	76.58	84.54	94.70%
VisionSelector	60.21	87.11	2235	86.61	**90.63**	**80.13**	75.24	84.63	95.78%
TwigVLM	61.15	86.63	2283	86.47	90.28	78.92	75.31	**85.41**	96.02%
Box+SAM supervision	60.54	86.37	**2300**	87.95	87.55	78.98	75.84	84.19	95.72%
DiffScorer w/o RL	**62.00**	**87.13**	2292	**88.37**	90.08	79.97	**77.07**	**85.41**	**96.98%**
Rollout strategy with the diffusion-distilled scorer									
GRIP-VLM-style rollouts	62.37	88.08	2290	88.63	90.13	80.72	77.03	85.94	97.40%
Ours (full)	**62.86**	**88.61**	**2304**	**88.98**	**91.13**	**81.48**	**77.47**	**86.20**	**98.04%**

Effect of counterfactual Swap-GRPO. Swap-GRPO raises average performance retention from 96.98% to 98.04% and improves all eight benchmarks as shown in Table 5. Figure 4 shows how Swap-GRPO suppresses distractors and recovers evidence omitted by the distilled scorer. With the same distilled scorer and budget, GRIP-VLM-style rollouts achieve 97.40%, versus 98.04% for counterfactual swaps, supporting our motivation of local exploration in Sec. 3.3. More ablations about group size, swap fraction, pruning position, architecture, and post-processing appear in Appendix B.

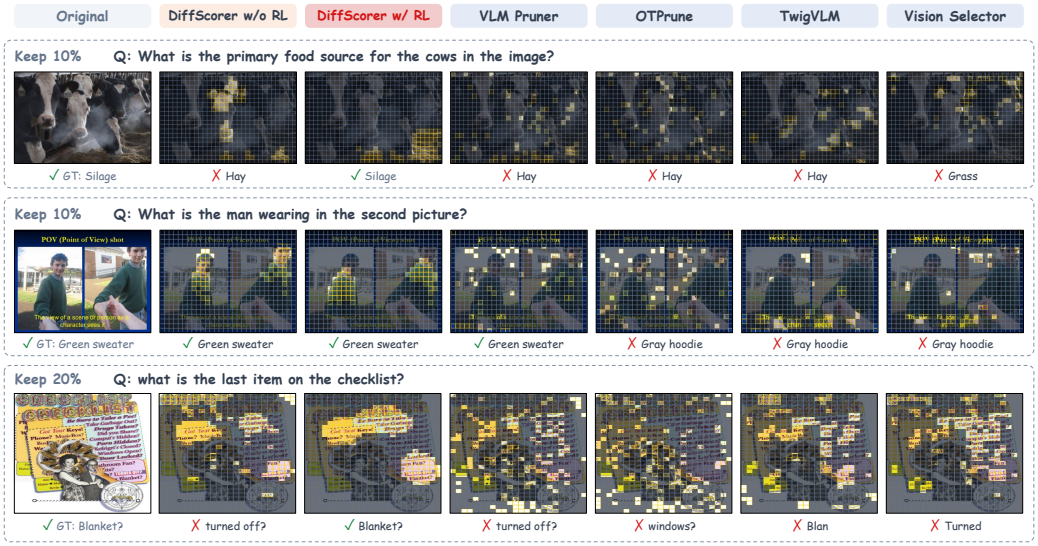


Figure 4: Qualitative effect of counterfactual Swap-GRPO.

5 Conclusion

We present DiffScorer, a Layer-0 visual-token pruning framework that distills the strong text–image alignment capability of a pretrained diffusion model, instead of relying on the potentially biased attention or hidden representations of the target VLM. To further correct residual attention biases without erasing this distilled evidence-selection prior, we introduce counterfactual Swap-GRPO, which optimizes local, budget-preserving token exchanges around DiffScorer's deterministic selection using answer rewards from a frozen VLM. Extensive experiments and ablations support this design: at 70% pruning ratio, diffusion supervision alone achieves 96.98% average performance retention, outperforming the strongest learned selector and one-shot pruning method, while Swap-GRPO further raises performance retention to 98.04% and improves all benchmarks.

AI USE STATEMENT

AI-assisted tools were used in the preparation of this manuscript for two purposes. First, they were used to improve the clarity, grammar, and overall presentation of the writing. These tools served only as editorial aids and did not determine the scientific claims, methodology, experimental design, results, or conclusions of the work. Second, AI-assisted tools were used to support literature retrieval and the discovery of potentially relevant related work. All retrieved references and associated information were independently verified against their original sources and carefully reviewed by the authors before being cited or incorporated into the manuscript. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 9392–9401. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00877. URL https://openaccess.thecvf.com/content/CVPR2025/html/Alvar_DivPrune_Diversity-based_Visual_Token_Pruning_for_Large_Multimodal_Models_CVPR_2025_paper.html.

Bizhe Bai, Jianjian Cao, Yadan Luo, and Tao Chen. Local information matters: Inference acceleration for grounded conversation generation models through adaptive local-aware token pruning. *arXiv preprint arXiv:2503.23959*, 2025a.

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, 2025b. URL https://arxiv.org/abs/2511.21631.

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, 2025c. URL https://arxiv.org/abs/2502.13923.

Sihan Cao, Jianwei Zhang, Pengcheng Zheng, Jiaxin Yan, Caiyan Qin, Yalan Ye, Wei Dong, Peng Wang, Yang Yang, and Chaoning Zhang. Language-guided token compression with reinforcement learning in large vision-language models. *arXiv preprint arXiv:2603.13394*, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pp. 19–35. Springer, 2024a. doi: 10.1007/978-3-031-73004-7_2. URL https://doi.org/10.1007/978-3-031-73004-7_2.

Xiwen Chen, Wenhui Zhu, Gen Li, Xuanzhao Dong, Yujian Xiong, Hao Wang, Peijie Qiu, Qingquan Song, Zhipeng Wang, Shao Tang, Yalin Wang, and Abolfazl Razi. Otprune: Distribution-aligned visual token pruning via optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 5849–5859, June 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl:

Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024b.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22826–22835. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02119. URL https://doi.org/10.1109/ICCV51701.2025.02119.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 51976–51999, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/555479a201da27c97aaeed842d16ca49-Paper-Conference.pdf.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

Lianyu Hu, Liqing Gao, Fanhua Shang, Liang Wan, and Wei Feng. illava: An image is worth fewer than 1/3 input tokens in large multimodal models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 73724–73737, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/77c87a15bbf0aad017c53995b832cf84-Paper-Conference.pdf.

Mingzhe Huang, Weijun Wang, Xin Ding, Liang Mi, Hao Wen, Yuanchun Li, Lichen Pang, Shansong Yang, Yunxin Liu, and Ting Cao. Grip-vlm: Group-relative importance pruning for efficient vision-language models. *arXiv preprint arXiv:2605.13375*, 2026.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, and Shaohui Lin. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 69927–69955, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/aeafb73dfed3007ec5299be1604d6f99-Paper-Conference.pdf.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogério Guimarães, and Pietro Perona. Text-image alignment for diffusion-based perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13883–13893. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01317. URL https://doi.org/10.1109/CVPR52733.2024.01317.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 5334–5342. AAAI Press, 2025. doi: 10.1609/AAAI.V39I5.32567. URL https://doi.org/10.1609/aaai.v39i5.32567.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024a.

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

Xuyang Liu, Siteng Huang, Yachen Kang, Honggang Chen, and Donglin Wang. VGDIFFZERO: text-to-image diffusion models can be zero-shot visual grounders. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 2765–2769. IEEE, 2024c. doi: 10.1109/ICASSP48485.2024.10445945. URL https://doi.org/10.1109/ICASSP48485.2024.10445945.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024d. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Zheng Qi, Chao Shang, Evangelia Spiliopoulou, and Nikolaos Pappas. Capturing gaze shifts for guidance: Cross-modal fusion enhancement for VLM hallucination mitigation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Nyk5TF8H5x.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pp. 146–162. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Zhenwei Shao, Mingyang Wang, Zhou Yu, Wenwen Pan, Yan Yang, Tao Wei, Hongyuan Zhang, Ning Mao, Wei Chen, and Jun Yu. Growing a twig to accelerate large vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20064–20074. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01866. URL https://doi.org/10.1109/ICCV51701.2025.01866.

Aditya Kumar Singh, Hitesh Kandala, Pratik Prabhanjan Brahma, Zicheng Liu, and Emad Barsoum. Duet-vlm: Dual stage unified efficient token reduction for vlm training and inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17651–17660, June 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Fengyuan Sun, Leqi Shen, Hui Chen, Sicheng Zhao, Jungong Han, and Guiguang Ding. AdaTP: Attention-debiased token pruning for video large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3273–3286, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.174. URL https://aclanthology.org/2025.findings-emnlp.174/.

Yuyao Sun, Tao Deng, Shuang Li, Deqing Wang, Hao Geng, and Minjun Yu. Learning to predict middle-layer attention in mllms for visual token pruning, 2026. URL https://arxiv.org/abs/2608.06411.

Rinyoichi Takezoe, Yaqian Li, Zi-Hao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learnpruner: Rethinking attention-based token pruning in vision language models. In

C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 66381–66400, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6cf7a37e761f55b642cf0939b4c64bb8-Paper-Conference.pdf.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5644–5659, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.310. URL https://aclanthology.org/2023.acl-long.310/.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision - ECCV 2026 - 19th European Conference, Malmö, Sweden, September 8-12, 2026, Proceedings, Part LXVI*, volume 17066 of *Lecture Notes in Computer Science*, pp. 336–357. Springer, 2026. doi: 10.1007/978-3-032-37718-0_20. URL https://doi.org/10.1007/978-3-032-37718-0_20.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 7907–7915. AAAI Press, 2025. doi: 10.1609/AAAI.V39I8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9961–9980, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.505. URL https://aclanthology.org/2025.emnlp-main.505/.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

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

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

Dingchen Yang, Bowen Cao, Anran Zhang, Weibo Gu, Winston Hu, and Guang Chen. Beyond intermediate states: Explaining visual redundancy through language. *arXiv preprint arXiv:2503.20540*, 2025a.

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 *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025, pp. 19792–19802. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.01843. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_VisionZip_Longer_is_Better_but_Not_Necessary_in_Vision_Language_CVPR_2025_paper.html.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24972–24982. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02325. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_ATP-LLaVA_Adaptive_Token_Pruning_for_Large_Vision_Language_Models_CVPR_2025_paper.html.

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 *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20857–20867. IEEE, 2025a. doi: 10.1109/ICCV51701.2025.01939. URL https://doi.org/10.1109/ICCV51701.2025.01939.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/zhang25s.html.

Jiaying Zhu, Yurui Zhu, Xin Lu, Wenrui Yan, Dong Li, Kunlin Liu, Xueyang Fu, and Zheng-Jun Zha. Visionselector: End-to-end learnable visual token compression for efficient multimodal llms, 2026. URL https://arxiv.org/abs/2510.16598.

A Inference Setup

A.1 Motivation: Multi-Peak Visual Evidence

After diffusion distillation and Swap-GRPO refinement, DiffScorer can assign high scores to several complementary evidence regions. Relational questions may depend on spatially separated objects, while fine-grained questions may jointly require a subject, an attribute region, and local context. As illustrated in Figure 5, these cases produce multi-peak score fields. Native token-wise Top-K can concentrate the budget on a dominant mode and miss a weaker, spatially separated mode. We therefore use a deterministic entropy-guided, cluster-aware post-processor that preserves the local structure of positive responses while increasing candidate coverage when the score field is dispersed.

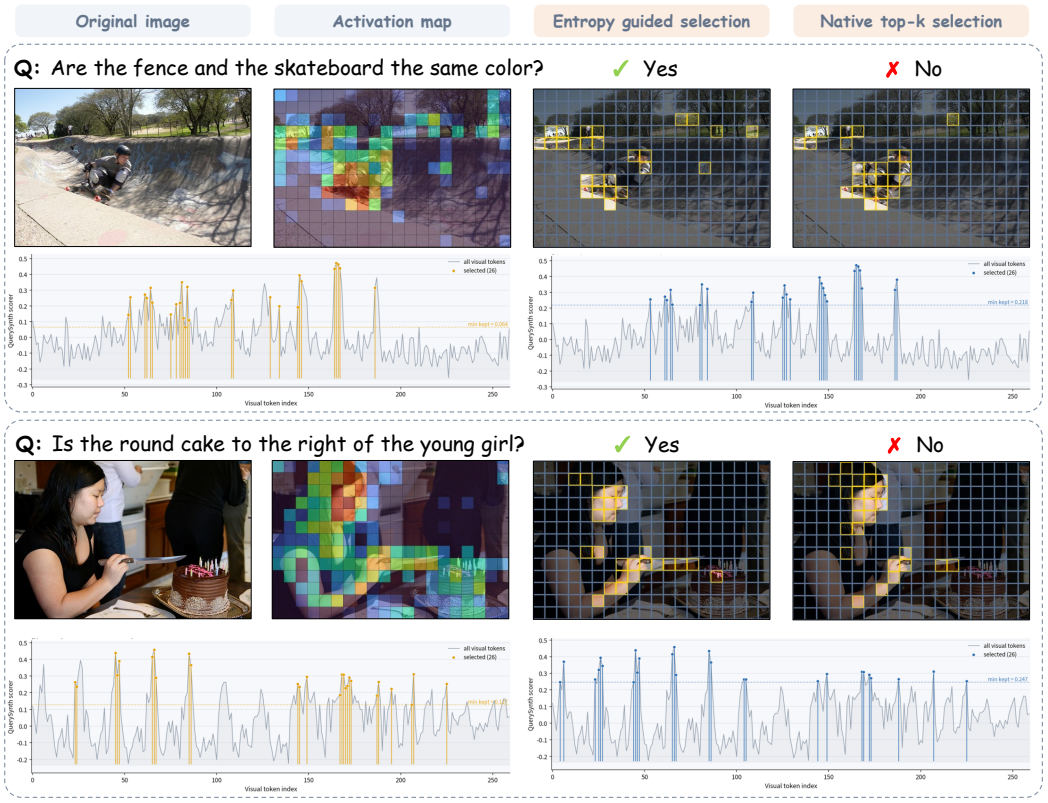


Figure 5: Multi-peak visual evidence and entropy-guided selection. DiffScorer can produce spatially separated modes for complementary evidence. Max-preserving smoothing retains local maxima and expands their spatial support; normalized score entropy enlarges the candidate set for dispersed fields before cluster-aware allocation. This reduces the risk that native Top-K spends the budget on only the dominant mode.

A.2 Entropy-Guided Cluster-Aware Selection

Let $\mathbf{h} \in \mathbb{R}^{H\times W}$ be the DiffScorer field over N valid visual tokens, and let $K = \mathrm{round}(\rho N)$ be the final retention budget. The post-processor has three deterministic stages.

Adaptive max-preserving smoothing. We first rectify the score field,

$$h_i^+ = \max(h_i, 0). \tag{11}$$

Let $s = \min(H, W)$ denote the short side of the token grid. Using $\sigma_0 = 1.25$ as the reference bandwidth at $s = 24$, we scale the Gaussian bandwidth with resolution:

$$\begin{aligned} \sigma_{\text{eff}} &= \min\left\{ \mathrm{clip}\left(\sigma_0 \frac{s}{24}, 0.5, 4\right), \frac{s}{12} \right\}, \\ k &= 2\left\lceil 3\sigma_{\text{eff}} \right\rceil + 1. \end{aligned} \tag{12}$$

For the normalized $k \times k$ Gaussian kernel G and reflect-padded convolution,

$$G(u,v) = \frac{1}{Z_G} \exp\left(-\frac{u^2+v^2}{2\sigma_{\text{eff}}^2}\right), \qquad \widetilde{\mathbf{h}} = G * \mathbf{h}^+, \qquad h_i^s = \max\left(h_i^+, \widetilde{h}_i\right). \tag{13}$$

The pointwise maximum prevents smoothing from attenuating an isolated positive maximum, while the convolution supplies nearby support so that a genuine local mode can form a spatial component.

Entropy-adaptive candidate budget. Because $\mathbf{h}^s$ is non-negative by construction, no additional score shift is required. If $Z_s = \sum_j h_j^s = 0$, the score field has no clustering semantics and we directly return

$$E = \mathrm{TopK}_K(\mathbf{h}^s), \qquad K = \mathrm{round}(\rho N). \tag{14}$$

All ties are broken stably by token index (equivalently, raster order), making this degenerate case deterministic. In the remainder, assume $Z_s > 0$ and define

$$p_i = \frac{h_i^s}{Z_s}, \qquad \mathcal{H}(\mathbf{p}) = -\sum_{i=1}^{N} p_i \log p_i, \qquad \overline{\mathcal{H}} = \frac{\mathcal{H}(\mathbf{p})}{\log N}. \tag{15}$$

For any requested retention ratio ρ, the final and maximum candidate budgets are

$$K = \mathrm{round}(\rho N), \qquad M_{\max} = \max\{K, \mathrm{round}(\rho_{\max} N)\}, \tag{16}$$

and the entropy-adaptive candidate budget and pool are

$$M = K + \mathrm{round}\left[(M_{\max} - K)\overline{\mathcal{H}}\right], \qquad \mathcal{C} = \mathrm{TopK}_M(\mathbf{h}^s). \tag{17}$$

Since $0 \le \overline{\mathcal{H}} \le 1$ and $M_{\max} \ge K$, this construction guarantees

$$K \le M \le M_{\max}. \tag{18}$$

For illustration, under the 10% retention setting, $\rho = 0.10$, so $K = \mathrm{round}(0.10N)$ and $M_{\max} = \mathrm{round}(0.40N)$. The same allocation rule is applied for any requested retention ratio ρ; in our evaluations, $\rho \in \{0.30, 0.20, 0.10\}$. We set $\rho_{\max} = 0.40$, which exceeds the largest evaluated retention ratio of 0.30. A concentrated field uses close to the minimum K candidates, whereas a flatter or multi-modal field exposes more secondary responses to the next stage.

Spatial clustering and exact-budget allocation. We run DBSCAN on the grid coordinates $\mathbf{x}_i = (r_i, c_i)$ of candidates $i \in \mathcal{C}$, with $\varepsilon = 2.5$ grid cells and $\mathrm{minPts} = 2$:

$$\mathcal{N}_\varepsilon(i) = \{j \in \mathcal{C} : \|\mathbf{x}_i - \mathbf{x}_j\|_2 \le \varepsilon\}. \tag{19}$$

A candidate is a core point when $|\mathcal{N}_\varepsilon(i)| \ge \mathrm{minPts}$, and clusters are expanded from core points by standard DBSCAN. Let the non-noise clusters and the DBSCAN noise set be

$$\{c_1, \ldots, c_C\}, \qquad \mathcal{O} = \mathcal{C} \setminus \bigcup_{c=1}^{C} c. \tag{20}$$

Noise points do not participate in the preferred inter-cluster budget allocation, but they remain in $\mathcal{C}$ for the final budget fill. If DBSCAN produces no cluster, i.e., $C = 0$, we directly return

$$E = \mathrm{TopK}_K(\mathbf{h}^s \mid \mathcal{C}). \tag{21}$$

This avoids creating a one-point fallback cluster that cannot carry a budget larger than one.

For $C > 0$, define the total clustered capacity and the budget assigned preferentially to genuine clusters as

$$S = \sum_{c=1}^{C} |c|, \qquad K_{\text{cl}} = \min(K, S). \tag{22}$$

When $K_{\text{cl}} \ge C$, every cluster first receives one token and the remaining budget is

$$R = K_{\text{cl}} - C. \tag{23}$$

If $S > C$, allocate this remainder in proportion to the remaining capacities:

$$q_c = R\frac{|c|-1}{S-C}, \qquad K_c = 1 + \lfloor q_c \rfloor. \tag{24}$$

The unassigned slots are then given, one at a time, in descending order of $q_c - \lfloor q_c \rfloor$ until $\sum_c K_c = K_{\mathrm{cl}}$; ties are broken by the deterministic DBSCAN cluster label. Because $R \le S - C$, this largest-remainder allocation satisfies

$$1 \le K_c \le |c|. \tag{25}$$

If $S = C$ within this branch, all clusters are singletons and necessarily $K_{\mathrm{cl}} = C$, so we set $K_c = 1$ for every cluster. The clustered selection is then

$$E_{\mathrm{cl}} = \bigcup_{c=1}^{C} \mathrm{TopK}_{K_c}(\mathbf{h}^s \mid c). \tag{26}$$

When $K_{\mathrm{cl}} < C$, let r_c be the highest-scoring representative of cluster c and instead take

$$E_{\mathrm{cl}} = \mathrm{TopK}_{K_{\mathrm{cl}}}(\mathbf{h}^s \mid \{r_1, \ldots, r_C\}). \tag{27}$$

Thus, in either branch, $|E_{\mathrm{cl}}| = K_{\mathrm{cl}}$. Finally, we deterministically fill any remaining slots from all unselected candidates:

$$E = E_{\mathrm{cl}} \cup \mathrm{TopK}_{K-|E_{\mathrm{cl}}|}(\mathbf{h}^s \mid \mathcal{C} \setminus E_{\mathrm{cl}}), \tag{28}$$

where $\mathrm{TopK}_0 = \emptyset$. This fill can select high-scoring noise points as well as unselected points from genuine clusters. Since $|\mathcal{C}| = M \ge K$, it follows strictly that

$$|E| = K. \tag{29}$$

All score ties in the clustering and fill stages are broken stably by token index or raster order. Keeping noise outside the preferred cluster allocation preserves the meaning of DBSCAN: treating every noise point as a singleton cluster could otherwise force a slot for each isolated point whenever the cluster budget is at least the resulting cluster count.

This post-processor has no learnable parameters and does not change DiffScorer's role as the token-importance scorer. The diffusion teacher and stochastic swap sampling are absent at inference; only the selected visual tokens enter LLM prefill. The end-to-end throughput measurements in Table 4 include this deterministic post-processing.

B Additional Ablation Studies

B.1 Box-SAM versus Diffusion Supervision

Evidence-box annotations offer a direct way to extract the labeled objects and use their spatial support as supervision. To isolate the contribution of our diffusion-based target construction, we therefore build an additional annotation-box+SAM (Kirillov et al., 2023) baseline while holding the scorer architecture, training data and optimization schedule fixed. Specifically, both variants use the same 140K VQA pairs, Qwen3-VL-8B-Instruct backbone, DiffScorer, training schedule, and evaluation protocol, and neither uses Swap-GRPO. For the box+SAM variant, each annotated evidence box prompts SAM to produce an instance mask. When an instance contains multiple evidence boxes, we merge their masks into a single supervision target. We then soften this target so that locations near each mask region's geometric center receive larger weights and the weights decrease toward the boundary. The diffusion variant instead uses the merged, query-conditioned diffusion attention maps described in Sec. 3.1. Thus, the comparison changes only how the 140K supervision targets are constructed.

Spatial confidence for diffusion targets. Each keyword map M_t is first sum-normalized into a probability map $\widehat{M}_t$. We measure its spatial support as

$$A_t = \left| \left\{ x : \widehat{M}_t(x) \ge 0.25 \cdot \max_x \widehat{M}_t(x) \right\} \right|. \tag{30}$$

The spatial weight, implemented as `area_factor`, is

$$\alpha_t = \mathrm{clip}\left(\left(\frac{\mathrm{median}(A)}{A_t} \right)^{\gamma}, 0.55, 1.75 \right). \tag{31}$$

Table 6: Supervision-source ablation on Qwen3-VL-8B-Instruct at 20% and 10% visual-token retention. Both variants use the same 140K VQA pairs and exclude answer-level RL.

Supervision	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	SEED$^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 20% visual tokens (↓ 80%)									
Box+SAM supervision	58.94	85.24	2198	87.14	85.97	74.68	72.95	81.83	92.95%
Diffusion supervision	**60.55**	**85.59**	**2224**	**87.72**	**88.45**	**77.52**	**75.09**	**84.10**	**94.95%**
Retain 10% visual tokens (↓ 90%)									
Box+SAM supervision	56.94	81.43	2004	85.53	82.99	71.12	69.27	79.30	88.91%
Diffusion supervision	**57.31**	**82.63**	**2095**	**86.81**	**84.43**	**72.45**	**72.29**	**79.83**	**90.76%**

This weighting downweights spatially broad maps and upweights compact maps before keyword-level evidence is merged; it instantiates the spatial-confidence factor $\omega_{m,j}$ in Eq. 5.

Table 6 shows that diffusion supervision improves all eight benchmarks at both lower retention budgets. It raises average retention from 92.95% to 94.95% at 20% retention and from 88.91% to 90.76% at 10% retention, absolute gains of 2.00 and 1.85 points. Together with the 1.26-point gain at 30% retention in Table 5, the advantage persists as the token budget becomes tighter.

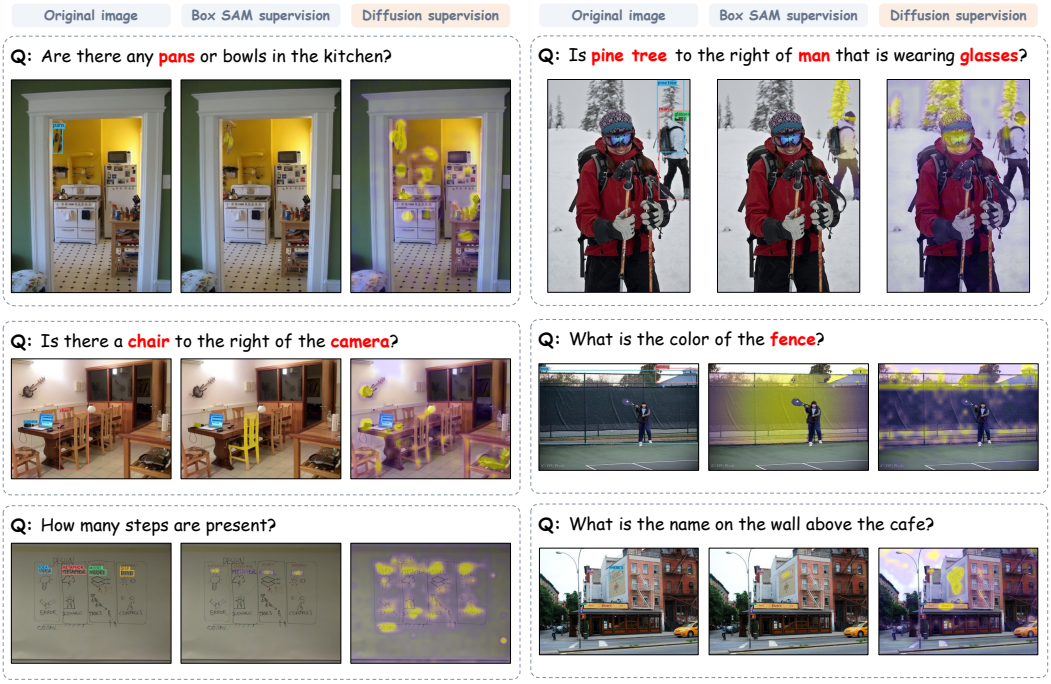


Figure 6: Qualitative comparison of box+SAM and diffusion supervision. Box-derived masks mainly cover the interiors of annotated objects, whereas diffusion attention can preserve query-conditioned evidence beyond the boxes, including complementary objects, relations, contextual regions, and text–object associations.

Figure 6 explains the quantitative gap. In the top-left kitchen example, box+SAM supervision remains concentrated on the annotated pan, while diffusion supervision also responds to other semantically relevant cookware and the surrounding kitchen regions needed to assess whether pans or bowls are present. In the top-right spatial-relation example, diffusion attends not only to the boxed pine tree but also to the relevant people, the glasses, and additional tree context needed to resolve the relation. The two bottom-row examples show the same behavior in counting and text-centric settings. For the diagram, diffusion identifies the step-bearing regions across the board rather than

only isolated annotated fragments. For the cafe example, it precisely localizes the target name while retaining the facade and sign context that links the text to the cafe.

The benefit is therefore not merely sharper segmentation. Box+SAM targets inherit the spatial support and coverage choices of the annotations, encouraging the scorer to equate evidence with labeled object interiors. This geometric coverage bias can underweight relational partners, scene context, unboxed instances, and text–object association cues. Diffusion attention is query-conditioned and can distribute supervision across the complete set of semantically relevant regions. Distilling this signal better matches the downstream selection objective: preserve all visual evidence required to answer under a fixed token budget, rather than reproduce the extent of the available boxes. The weaker box+SAM results are consequently consistent with target-induced bias rather than differences in data, architecture, or answer-level training.

B.2 Rollout Group Size and Swap Fraction

Swap-GRPO performs local exploration around DiffScorer's deterministic selection, so the rollout group size G and swap fraction f_{swap} jointly control the breadth and radius of the counterfactual neighborhood. Let $\mathbf{m}_0$ be the deterministic anchor mask containing K retained tokens. For each sampled trajectory, the swap fraction determines the number of exchanges as

$$r = \text{round}(f_{\text{swap}} K). \tag{32}$$

The trajectory drops r distinct tokens from $\mathbf{m}_0$ and adds the same number of distinct tokens from its complement, thereby preserving the exact budget $|\mathbf{m}_g| = K$. Thus, $f_{\text{swap}} = r/K$ denotes the fraction of anchor-retained tokens that are replaced, rather than the fraction of all visual tokens.

Increasing G exposes each query to more independently sampled, budget-preserving swap trajectories and therefore broadens group-level coverage. By contrast, increasing f_{swap} replaces more distinct anchor tokens in each trajectory and expands the perturbation radius around $\mathbf{m}_0$. The two parameters must therefore provide sufficient diversity for learning while preserving the strong diffusion-distilled prior.

Table 7: Ablation of Swap-GRPO group size and swap fraction on Qwen3-VL-8B-Instruct with 70% visual-token pruning.

Configuration	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
					Rollout group size G				
$G = 4$	62.24	87.69	2307	88.70	**91.22**	81.06	76.13	85.76	97.44%
$G = 8$	**62.86**	**88.61**	2304	**88.98**	91.13	81.48	**77.47**	**86.20**	**98.04%**
$G = 16$	62.82	88.10	2311	88.95	89.64	80.37	77.26	85.59	97.51%
$G = 32$	62.55	88.22	**2341**	88.76	91.08	**81.62**	76.60	85.50	97.87%
					swap fraction				
0.15	**62.86**	**88.61**	2304	88.98	**91.13**	**81.48**	**77.47**	**86.20**	**98.04%**
0.20	62.58	87.46	2306	**89.08**	90.18	81.06	76.65	85.68	97.46%
0.25	62.20	87.92	2287	88.44	90.38	80.61	77.13	86.11	97.36%
0.30	**62.86**	87.11	**2307**	88.88	90.53	80.96	77.10	**86.20**	97.62%

We vary the group size over $G \in \{4, 8, 16, 32\}$. The corresponding average performance retentions are 97.44%, 98.04%, 97.51%, and 97.87%, making $G = 8$ the best accuracy–cost choice. We also vary the swap fraction from 0.15 to 0.30. The default value of 0.15 achieves 98.04%, whereas fractions of 0.20, 0.25, and 0.30 obtain 97.46%, 97.36%, and 97.62%, respectively. The non-monotonic group-size trend and the degradation under larger swap fractions show that local exploration is not improved by simply enlarging the neighborhood. Too little exploration may provide insufficient counterfactual diversity, whereas excessive breadth or perturbation can introduce trajectories that move unnecessarily far from the reliable diffusion prior. The best setting therefore favors a moderate number of local alternatives with a small swap radius.

B.3 Pruning Position and Inference Post-Processing

The pruning position determines both the contextual information available to token selection and the point at which computational savings begin. Although the vision encoder has already contextualized the image tokens, Layer-0 pruning operates on projected visual features before any LLM block, whereas later pruning can use representations that have undergone additional LLM-side cross-token fusion. This extra context may potentially benefit selection, but every preceding block must process the full visual sequence. We therefore evaluate whether postponing pruning provides sufficient accuracy gains to justify its additional cost, together with the contribution of deterministic inference post-processing.

Table 8: Ablation of the pruning position and deterministic inference post-processing on Qwen3-VL-8B-Instruct with 70% visual-token pruning.

Configuration	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Pruning position									
Layer 0	**62.86**	**88.61**	2304	88.98	**91.13**	**81.48**	**77.47**	**86.20**	**98.04%**
Layer 1	62.35	88.38	**2315**	89.07	90.73	81.11	77.43	85.50	97.77%
Layer 2	62.35	88.33	2312	**89.11**	90.78	80.93	77.31	85.76	97.75%
Inference post-processing									
Full	**62.86**	88.61	2304	**88.98**	**91.13**	**81.48**	**77.47**	86.20	**98.04%**
w/o spatial smoothing	62.60	**88.63**	**2338**	88.66	90.48	80.92	76.98	**86.29**	97.89%
w/o entropy-guided clustering	62.51	87.69	2302	88.89	90.63	80.94	77.35	86.11	97.65%

Pruning position and inference post-processing. Layer-0 pruning achieves 98.04% average retention, compared with 97.77% and 97.75% when postponing the same selection to Layers 1 and 2. The later positions retain the full visual sequence for additional computation but do not improve selection, indicating that diffusion-grounded scores already identify the required evidence before the first LLM block.

The full post-processor improves average retention by 0.15 points over removing spatial smoothing (97.89%) and by 0.39 points over removing entropy-guided clustering (97.65%). The latter result is consistent with Figure 5: adaptive candidate expansion and spatial grouping help expose weaker complementary modes during discretization. The gains are deliberately modest because the underlying scores are already strong. Even the larger component ablation retains 97.65%, exceeding the strongest competing one-shot method (VLM-Pruner, 96.63%), layer-wise method (DUET-VLM, 97.58%), and learned selector (TwigVLM, 96.02%) at the same 30% retention budget. Thus, the post-processor improves multi-peak coverage, while the main advantage remains robustly attributable to diffusion distillation and Swap-GRPO.

B.4 DiffScorer Architecture

DiffScorer uses the same backbone-independent scorer configuration in the LLaVA-1.5-7B and Qwen3-VL-8B experiments. The Vision-Q/Text-KV ablation also keeps the 4096-to-1024 projection and two-block cross-attention trunk unchanged; it differs only at the final readout, where visual-side outputs are scored token-wise by an MLP and softmax instead of aggregating text-side outputs into a query and comparing it with visual keys. Table 9 summarizes the shared architecture and the SFT and Swap-GRPO optimization settings.

We compare the proposed text-query/visual-key–value design with a reversed cross-attention variant that uses visual queries and text keys and values, followed by an MLP and softmax over visual tokens. Both variants use the same diffusion supervision and exclude Swap-GRPO.

The two directions perform similarly across budgets: the proposed design obtains 96.98% versus 96.88% average retention at 30%, 94.95% versus 94.47% at 20%, and a rounded tie of 90.76% at 10%. Both directions form an $L \times N$ attention matrix and thus have the same leading $\mathcal{O}(LNd)$ attention cost. The efficiency distinction lies in the scoring head: Text-Q/Vision-KV aggregates L vision-aware text tokens once into a single query and scores N visual tokens by normalized similarity, whereas the reversed variant retains N cross-attention outputs and applies an MLP independently

Table 9: DiffScorer architecture and training hyperparameters. The architecture is shared by the LLaVA and Qwen experiments; the Vision-Q/Text-KV ablation changes only the final readout.

Stage	Setting	Value
Architecture		
Architecture	Input / internal width	4096 / 1024 (`hidden_size` / `query_dim`)
Architecture	Cross-attention	2 layers, 8 heads, and 128 dimensions per head
Architecture	FFN hidden width	2048 (`ffn_mult`=2.0)
Architecture	Dropout	0
Architecture	Trainable parameters	26,245,121 (≈26.25M)
Architecture	Output readout	Text-Q/Vision-KV; the Vision-Q/Text-KV ablation shares the same trunk and changes only the terminal aggregation and scoring head
Supervised fine-tuning (SFT)		
SFT	Optimizer	AdamW with PyTorch defaults: $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$
SFT	Learning rate / weight decay	2×10^{-5} / 0.01
SFT	Global effective batch size	256 samples per optimizer step
SFT	Trainable module	DiffScorer only; the VLM, including the LLM, is frozen
SFT	Score temperature / logit scale	1.0 / disabled (`LEARNABLE_LOGIT_SCALE=false`)
Swap-GRPO refinement		
RL	Scorer optimizer	AdamW with default $\beta = (0.9, 0.999)$; only the scorer is updated
RL	Scorer LR / weight decay	5×10^{-6} / 0.01
RL	Batch / group / accumulation	64 / 8 / 1
RL	Training duration	2 epochs
RL	Scorer temperature	1.0
RL	Text rollout	Temperature 0.0 with `do_sample=false`
RL	Answer judge / reward	Qwen3.5-2B; binary correctness $R \in \{0, 1\}$
RL	Swap temperature	$\tau_{\text{drop}} = \tau_{\text{add}} = 1.0$
RL	Swap fraction	0.15; each trajectory drops and replaces $r = \text{round}(0.15K)$ distinct tokens from the deterministic K-token anchor mask $\mathbf{m}_0$
RL	Exploration coefficient	$\eta = 0.03$
RL	Entropy / KL coefficients	0.001 / 0.01; KL compares the current scorer with the frozen initial scorer over visual-token distributions
RL	PPO clip / epochs	0.1 / 1

Table 10: DiffScorer architecture ablation on Qwen3-VL-8B-Instruct. Text-Q/Vision-KV is the proposed scorer; Vision-Q/Text-KV applies a dense MLP scoring head to the visual-side cross-attention outputs. Neither variant uses Swap-GRPO.

Architecture	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 30% visual tokens (↓ 70%)									
Vision-Q/Text-KV	61.91	**87.71**	2284	88.05	89.24	79.73	**77.26**	**85.68**	96.88%
Text-Q/Vision-KV	**62.00**	87.13	**2292**	**88.37**	**90.08**	**79.97**	77.07	85.41	**96.98%**
Retain 20% visual tokens (↓ 80%)									
Vision-Q/Text-KV	60.50	**86.09**	2191	87.36	86.12	77.18	75.06	**84.45**	94.47%
Text-Q/Vision-KV	**60.55**	85.59	**2224**	**87.72**	**88.45**	**77.52**	**75.09**	84.10	**94.95%**
Retain 10% visual tokens (↓ 90%)									
Vision-Q/Text-KV	**57.59**	82.17	2073	86.34	83.04	**73.71**	72.23	**81.22**	**90.76%**
Text-Q/Vision-KV	57.31	**82.63**	**2095**	**86.81**	**84.43**	72.45	**72.29**	79.83	**90.76%**

to each. The proposed design therefore reduces dense per-visual-token head computation while preserving accuracy; we do not attribute a reduction in the shared attention FLOPs to the attention direction itself.

C Additional Comparison Results

C.1 Learned Pruners at Lower Retention Rates

Table 11 extends the learned-scorer comparison in Table 5 to 20% and 10% visual-token retention. Because these baselines were originally trained with different backbones and datasets, comparing their released checkpoints would confound pruning quality with model and data differences. We

therefore retrain Dynamic-LLaVA, VisionSelector, and TwigVLM using their official implementations, with a common Qwen3-VL-8B-Instruct backbone and the same 140K VQA records. Their native objectives already expose the learned selectors to both the question q and reference answer y: Dynamic-LLaVA minimizes the downstream autoregressive generation loss plus keep-rate regularization, propagating gradients through its discrete decisions with Gumbel–Softmax and a straight-through estimator (Huang et al., 2025); VisionSelector minimizes $\mathcal{L}_{\text{CE}} + \lambda_t \mathcal{L}_{\text{constraint}}$ (Zhu et al., 2026); and TwigVLM updates only the twig block using the conventional autoregressive loss (Shao et al., 2025). Our evidence terms are not an independent annotation source: $\mathcal{E} = g(q, y)$ is deterministically extracted from the same question–answer pair. Thus, although DiffScorer explicitly materializes $\mathcal{E}$ when constructing its diffusion teacher, it receives no answer or evidence information unavailable to the baselines. At inference, all methods use only the image and question and are evaluated under identical retention budgets. Under this information-matched setting, DiffScorer without RL performs best on all eight benchmarks at both budgets, achieving 94.95% and 90.76% average retention, respectively.

Table 11: Comparison with learned visual-token pruners on Qwen3-VL-8B-Instruct at 20% and 10% retention. DiffScorer w/o RL uses diffusion distillation without answer-level RL.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
Qwen3-VL-8B	64.12	89.74	2374	88.91	94.49	83.30	78.84	88.03	100.00%
Retain 20% visual tokens (↓ 80%)									
Dynamic-LLaVA	57.33	84.64	2114	85.23	87.06	72.75	74.71	82.71	92.10%
VisionSelector	57.32	84.78	2189	84.77	87.56	75.14	72.92	82.18	92.52%
TwigVLM	59.61	84.45	2100	82.47	86.32	73.35	73.26	83.23	91.90%
DiffScorer w/o RL	**60.55**	**85.59**	**2224**	**87.72**	**88.45**	**77.52**	**75.09**	**84.10**	**94.95%**
Retain 10% visual tokens (↓ 90%)									
Dynamic-LLaVA	52.90	79.74	1996	81.10	84.04	65.15	70.43	76.86	86.31%
VisionSelector	52.44	79.56	2002	79.71	83.19	68.10	68.12	76.68	85.97%
TwigVLM	56.01	80.46	1926	77.17	82.15	68.64	67.74	78.78	86.21%
DiffScorer w/o RL	**57.31**	**82.63**	**2095**	**86.81**	**84.43**	**72.45**	**72.29**	**79.83**	**90.76%**

C.2 Additional LLaVA Baselines

Table 12 reports DivPrune and PruneSId on LLaVA-1.5-7B alongside our method at all three budgets. The main comparison is in Table 3.

Table 12: Additional Layer-0 baselines on LLaVA-1.5-7B.

Method	GQA Acc.	MMB Acc.	MME P+C	POPE F1	SQA Acc.	TextVQA Acc.	$\text{SEED}^{\text{Image}}$ Acc.	A-OKVQA Acc.	Avg.
LLaVA-1.5-7B	63.79	73.96	1785.12	85.94	69.46	45.97	66.16	80.87	100.00%
Retain 30% visual tokens (↓ 70%)									
DivPrune	61.38	72.55	**1792**	83.46	68.81	42.77	63.68	79.21	97.26%
PruneSId	60.21	72.99	1708	83.36	68.26	40.50	62.81	78.68	95.54%
Ours	**61.57**	**74.36**	1764	**85.82**	**70.45**	**44.89**	**64.76**	**80.09**	**98.97%**
Retain 20% visual tokens (↓ 80%)									
DivPrune	**60.69**	72.26	1722	84.89	68.56	41.26	62.16	78.51	95.95%
PruneSId	59.44	72.37	1720	84.83	67.92	39.44	61.28	77.37	94.75%
Ours	60.31	**73.55**	**1754**	**86.04**	**69.45**	**43.51**	**63.44**	**79.04**	**97.58%**
Retain 10% visual tokens (↓ 90%)									
DivPrune	**59.07**	71.19	1678	85.77	68.91	38.23	60.00	76.76	93.83%
PruneSId	58.43	71.81	1700	84.62	68.17	38.06	59.97	76.50	93.57%
Ours	58.31	**72.52**	**1756**	**85.78**	**69.06**	**40.66**	**61.87**	**77.24**	**95.57%**

C.3 Additional HR-Bench Baselines

Table 13 reports DivPrune, SparseVLM, and PruneSId on HR4K and HR8K alongside the unpruned backbone and our method. The remaining baselines are compared in Table 4.

Table 13: Additional HR-Bench baselines: accuracy and end-to-end generation throughput (tokens/s).

Method	*HR4K: 13,746 → 4,128; retain 30% (↓ 70%)*				*HR8K: 16,224 → 3,244; retain 20% (↓ 80%)*			
	FSP Acc.↑	FCP Acc.↑	Overall Acc.↑	Throughput tok/s ↑	FSP Acc.↑	FCP Acc.↑	Overall Acc.↑	Throughput tok/s ↑
Qwen3-VL-8B	92.25	64.00	78.13	16.72 (1.00×)	85.25	64.00	74.63	14.23 (1.00×)
DivPrune	90.50	59.50	75.00	20.56 (1.23×)	82.00	57.75	69.88	19.97 (1.40×)
SparseVLM	**92.50**	64.25	78.38	19.78 (1.18×)	85.25	62.00	73.63	18.66 (1.31×)
PruneSId	91.00	63.25	77.13	19.23 (1.15×)	83.75	60.50	72.13	17.28 (1.21×)
Ours	91.75	**66.25**	**79.00**	**21.03 (1.26×)**	**85.50**	**63.00**	**74.25**	**21.29 (1.50×)**

C.4 Computational Complexity and Throughput Analysis

Tables 4 and 13 show that DiffScorer provides the largest measured end-to-end speedup at both high-resolution settings. We explain this result by separating the cost of token selection from the cost that remains inside the LLM.

Cost accumulated across LLM layers. Let V be the initial number of visual tokens, T the number of text tokens, P the number of LLM layers, d the LLM width, and V_ℓ the visual-token count entering layer ℓ. Ignoring constants, prefill costs

$$\mathcal{C}_{\text{prefill}} = \mathcal{C}_{\text{select}} + \sum_{\ell=1}^{P} \mathcal{O}\big((T+V_\ell)d^2 + (T+V_\ell)^2 d\big)\,, \tag{33}$$

where the first term inside the sum covers projections and feed-forward layers, and the second covers attention score computation and value aggregation. During autoregressive decoding, each generated token also attends to a KV cache of length approximately $T + V_\ell$, so early visual-token removal reduces both prefill and every subsequent decoding step.

DiffScorer fixes $V_\ell = \rho V$ for all LLM layers because selection occurs before Layer 0. Its inference-time scoring cost is

$$\mathcal{C}_{\text{DiffScorer}} = \mathcal{O}(TVd_s) + \mathcal{C}_{\text{post}}(V), \tag{34}$$

where $d_s \ll d$ is the lightweight scorer width and $\mathcal{C}_{\text{post}}$ denotes smoothing, candidate ranking, and clustering on the two-dimensional score field. The scorer uses text–visual interactions rather than a $V \times V$ token-relation matrix and is evaluated only once. Consequently, the visual-dominated quadratic attention term is reduced approximately by ρ^2 in every LLM layer (to 9% at $\rho = 0.30$ and 4% at $\rho = 0.20$), while the linear projection and feed-forward term scales approximately with ρ. Vision encoding is still included in the reported throughput, so these values are conservative relative to LLM-prefill-only speedups.

Delayed and layer-wise reduction. A method that first prunes after q layers must pay for q full-token blocks:

$$\mathcal{C}_{\text{delayed}} \;\geq\; q\,\mathcal{C}_{\text{block}}(T+V) + (P-q)\,\mathcal{C}_{\text{block}}(T+\rho V) + \mathcal{C}_{\text{select}}. \tag{35}$$

VLM-Pruner (Wu et al., 2026), for example, relies on token keys from the preceding layer and therefore prunes at Layer 2; it additionally constructs semantic and spatial token relations, performs buffered greedy selection, and fuses discarded information. It consequently pays for two full-token LLM blocks before realizing any reduction. For genuinely layer-wise methods, $V_1 \geq V_2 \geq \cdots \geq V_P = \rho V$, and Equation 33 contains the larger intermediate budgets rather than the final budget at every layer. SparseVLM (Zhang et al., 2025b) repeatedly derives text-guided importance, estimates the rank of its priority matrix, and recycles removed tokens; its reported rank computation scales as $\mathcal{O}(TV_\ell \min(T, V_\ell))$ at each sparsified layer. DUET-VLM (Singh et al., 2026) combines vision-side redundancy clustering with progressive text-guided dropping inside the LLM; its reference schedule

retains substantial visual context through the middle layers and reaches the budget only in later layers. iLLaVA (Hu et al., 2026) reduces computation in both the vision encoder and LLM, but reaches its final count through repeated token-merging and recycling modules across blocks. These progressive designs favor information preservation, yet their early and middle layers still process more than the final matched budget and their repeated selection or merging adds runtime overhead. This explains why a matched final token count does not imply a matched end-to-end speedup.

One-shot selection overhead. Layer-0 baselines avoid full-token LLM blocks, but their selectors can become costly when V exceeds 10K. DivPrune (Alvar et al., 2025) solves a max–min diversity problem using global token distances, which requires a pairwise similarity structure and greedy subset construction. OTPrune (Chen et al., 2026) likewise forms a token kernel for distribution alignment and uses incremental submodular selection, with its greedy update scaling as $\mathcal{O}(K^2V)$ for retained budget K. PruneSID (Fang et al., 2026) performs principal semantic component analysis, semantic grouping, and intra-group non-maximum suppression, introducing global decomposition and within-group comparison costs. ZOO-Prune (Kim et al., 2026) estimates sensitivity using repeated zeroth-order perturbations at the projector and then performs diversity-aware subset selection. Although these operations are cheaper than full LLM reasoning, their cost grows with the unpruned token set and is paid before the LLM savings begin. Pairwise $V \times V$ analysis is particularly visible at HR4K/HR8K, where V is 13,746/16,224.

Connection to measured throughput. DiffScorer moves the expensive supervision needed to learn query relevance into offline training. At inference, it performs one lightweight, query-conditioned Layer-0 pass, applies the deterministic score-field post-processor, and presents the exact final budget to every LLM layer. It therefore avoids (i) full-token LLM warm-up layers, (ii) repeated layer-wise ranking and merging, and (iii) pairwise visual-token subset optimization. This structure matches the measurements: at HR4K, DiffScorer reaches $1.26\times$ throughput versus the strongest baseline speedup of $1.23\times$; at HR8K, the gap widens to $1.50\times$ versus $1.40\times$. The widening at the longer visual sequence is expected because the saved LLM attention grows quadratically with sequence length, whereas DiffScorer's cross-modal scoring grows with TV and uses small scorer width d_s.

C.5 Additional Qualitative Results

Figure 7 and 8 provides additional comparisons of retained visual evidence and answers across DiffScorer before and after Swap-GRPO and representative pruning baselines. Across questions that require object attributes, spatial relations, and localized text or objects, DiffScorer preserves complementary answer-relevant regions more consistently. The examples also show the role of answer-level refinement: Swap-GRPO adjusts the distilled selection when semantically related distractors compete with the evidence required for the final answer.

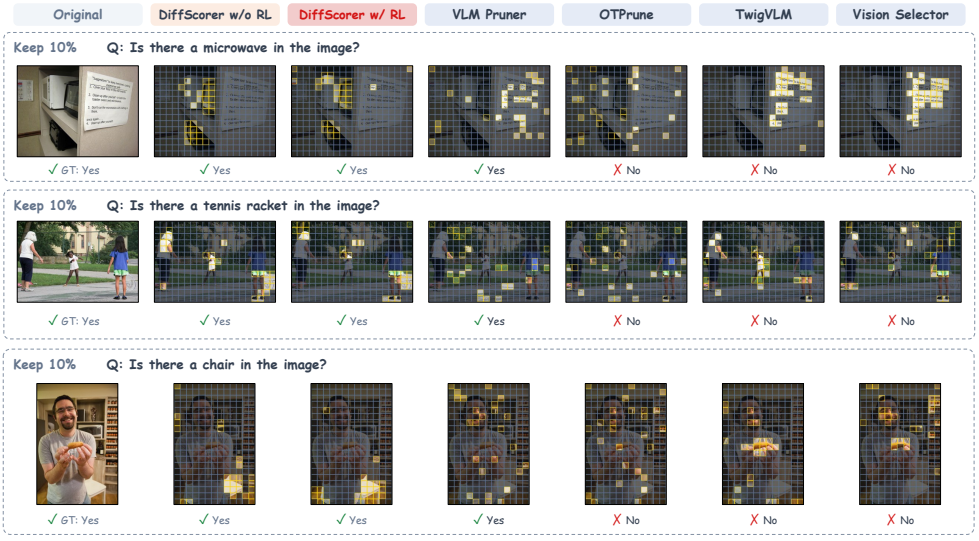


Figure 7: Additional qualitative comparisons of visual-token selection part1.

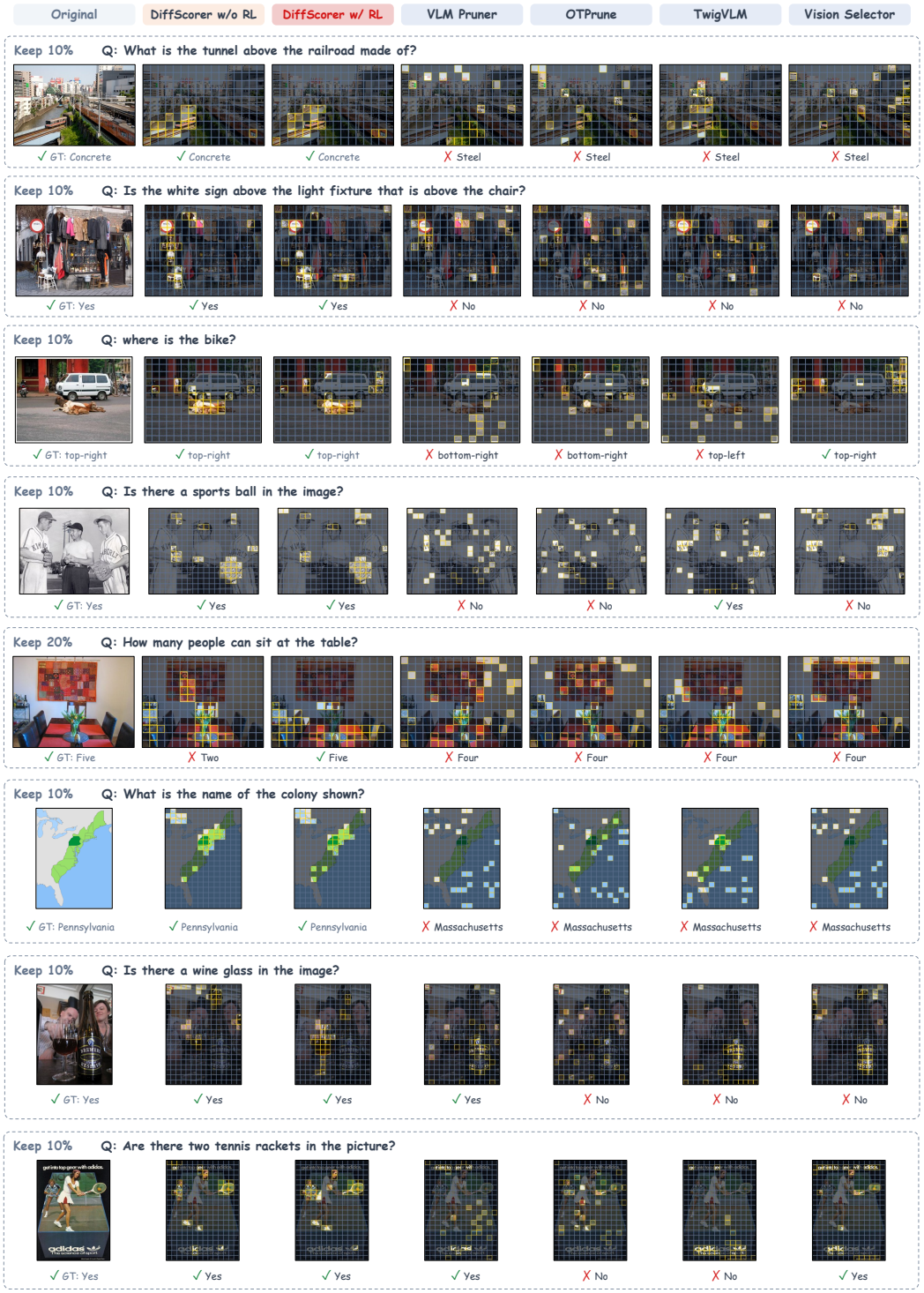


Figure 8: Additional qualitative comparisons of visual-token selection part2. Highlighted cells indicate retained evidence for DiffScorer before and after Swap-GRPO and for representative pruning baselines; the predicted answer is shown beneath each selection. Swap-GRPO preserves the strong diffusion-distilled prior while correcting residual selections that favor related distractors over answer-critical evidence.