MVP: Multi-View-Guided Token Pruning for Vision-Language Models

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) achieve strong multimodal understanding by encoding images as sequences of visual tokens, yet cost hundreds and thousands of tokens per image. Existing pruning methods often rely on a single criterion, capturing only a partial view of token necessity. Controlled experiments reveal that token necessity is multi-faceted, requiring the joint exploitation of heterogeneous pruning perspectives. To address this, we propose a novel Multi-View-Guided Token Pruning (MVP-Token) method to learn token necessity from multiple complementary pruning perspectives. We first establish a dataset **MVP-PruneSet**, which encompasses diverse pruning criteria by running orthogonal pruning methods, including a novel question-guided method **GRASP** at the encoder side. We then propose **MVP-Pruner** to capture such multi-view supervision signals. Specifically, a lightweight dual-attention module predicts soft token-importance labels under multi-view guidance. Extensive experiments demonstrate the effectiveness of MVP-Token across different model architectures. On LLaVA-1.5-7B, our method can surpass the single view baseline with a clear margin. On Qwem-2.5-VL-7B, our method achieves a 1.33× speedup and a 24.6% latency reduction under dymatic resolution while surpassing VisPruner by 2.1% of the average performance.

1 Introduction

By processing images as visual token sequences alongside text, Vision-Language Models (VLMs) (Liu et al., 2024; Bai et al., 2025b) excel across various multimodal tasks, such as visual question answering (Antol et al., 2015; Goyal et al., 2017), visual reasoning (Hudson and Manning, 2019) and image captioning (Chen et al., 2015). Despite their strong multimodal reasoning capability, modern VLMs inevitably process hundreds or even thousands of visual tokens per image. This dense representation incurs heavy computational overhead due to the quadratic complexity of attention mechanisms. Therefore, pruning irrelevant visual tokens without sacrificing performance offers a principled path to inference efficiency.

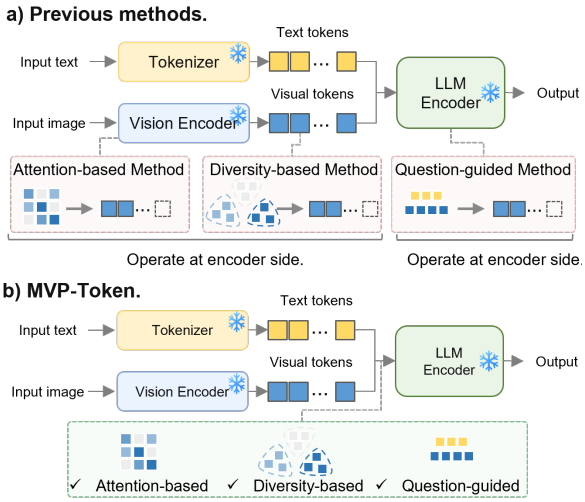


Figure 1: (a) Previous methods exploit only patial cues at different model stages. (b) **MVP-Token** uses single post-encoder pruner models multi cues.

Numerous token pruning methods have been proposed, which can be broadly categorized into four families. **Attention-based** methods leverage internal attention weights or saliency scores to identify and retain tokens that receive high response values (Zhang et al., 2025a; Zou et al., 2026). **Diversity-based** methods focus on maintaining a spatially and semantically diverse set of tokens to ensure broad visual coverage (Deng et al., 2026; Yang et al., 2025; Baek et al., 2026). **Question-guided** methods incorporate the query text into the pruning process, selecting tokens most relevant to answering the specific question (Zhang et al., 2025b; Sun et al., 2025). **Training-based** methods fine-tune the model on large instruction-following corpora in an end-to-end manner, relying on the model to implicitly discover token importance as a

byproduct of next-token prediction (Takezoe et al., 2026; Ye et al., 2025). However, as shown in figure 1, each of these criteria only captures a partial view of token importance, and single-view pruning signals are often biased and insufficient to characterize the overall token necessity.

For different pruning methods, there are two key points. (1) **Different pruning behaviors.** Different methods preserve markedly different visual regions, reflecting distinct inductive biases over token relevance. (2) **Complementary pruning capabilities.** When evaluated on identical samples, the success–failure patterns of different methods are interleaved, meaning each method can solve specific examples missed by others. This inherent complementarity suggests that token pruning should not merely rely on a single heuristic mask. Instead, heterogeneous pruning outcomes should be treated as structured supervision, where different methods provide partial yet complementary evidence about visual token importance. ***Which tokens are jointly necessary across multiple complementary pruning perspectives? How to exploit the contrastive structure that emerges when different methods disagree on the same sample?***

To address these questions, we propose **MVP-Token** (Multi-View-Guided Token Pruning), a two-stage framework leveraging multi-view token pruning for efficient VLMs. In the first stage, we construct a multi-method aggregated dataset **MVP-PruneSet** to provide diverse perspectives. In particular, as many question-guided methods prune inside the decoder, making them misaligned with encoder-side methods. We design a novel question-guided method **GRASP** (Grounding-Responsive Aware Semantic Pruning) to ground the input query onto the object-level token grid at the encoder side. For each method and sample, we record whether the pruned representation suffices to answer the question correctly, yielding completeness annotations from question-guided, attention-based, and diversity-based perspectives. In the second stage, to efficiently fuse the multi-view pruning information, we design a lightweight adapter **MVP-Pruner** trained with **SFF-Label** (Success–Failure Family Label) supervision and a rank loss. Specifically, SFF-Label is a soft labeling strategy that aggregates heterogeneous pruning signals into per-token importance labels. The rank loss directly supervises the relative importance of token pairs at both coarse and fine-grained levels. In this way, MVP-Pruner learns complementary multi-view evidence into a single, unified importance score for each visual token, enabling efficient and accurate pruning at inference time.

We evaluate our method on LLaVA-1.5-7B (Liu et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025b) across vision-language benchmarks. MVP-Token outperforms existing pruning methods. Specifically, it exceeds VisPruner by 3.35% and 1.73% in average accuracy evaluating Qwen2.5-VL-7B at 20% and 30% retention respectively, while achieving a lower per-sample latency (89.22 ms vs. 98.16 ms) and a higher inference speedup (1.33× vs. 1.21×) over the uncompressed baseline. Additionally, the visualization of retained tokens shows that MVP-Token possesses a high degree of interpretability, effectively balancing question-conditioned relevance and global contextual coverage.

2 Related Work

Vision-language models. VLMs achieve remarkable success in multimodal understanding and instruction following. To align modalities, they typically employ a visual encoder such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023) to extract high-dimensional features, which are flattened into visual token sequences. These tokens are then projected into a language decoder via linear projections (Liu et al., 2023), abstractors (Bai et al., 2023), or cross-attention layers (Alayrac et al., 2022) for autoregressive generation. However, modern architectures increasingly adopt higher input resolutions or dynamic scaling to preserve fine-grained details (Liu et al., 2024; Bai et al., 2025b). This trend explodes the sequence length to hundreds or thousands of tokens, imposing a heavy computational burden due to the quadratic complexity of attention. Optimizing token efficiency via pruning is thus essential to accelerate VLM inference.

Visual token pruning. Existing visual token reduction methods can be broadly divided into four families according to their selection criteria. (1) **Attention-based methods** leverage attention maps to identify salient regions. For example, VisPruner (Zhang et al., 2025a) uses self-attention scores from the vision encoder to dynamically evaluate token significance, while HoloV (Zou et al., 2026) utilizes global dependencies to preserve critical foreground objects. Though training-free, they tend to overconcentrate on highly attended regions, potentially overlooking low-saliency yet necessary

text or context. (2) **Diversity-based methods** mitigate informational redundancy by ensuring spatial and semantic coverage through clustering or sampling. SCOPE (Deng et al., 2026) maximizes both semantic distinctiveness and spatial coverage. Some recent works incorporate mixed objectives: VisionZip (Yang et al., 2025) combines spatial downsampling with attention-guided aggregation, and AgilePruner (Baek et al., 2026) dynamically balances attention and feature diversity by employing an image-complexity adaptive threshold to remove redundant tokens. (3) **Question-guided methods** incorporate textual instructions to anchor task-relevant regions. SparseVLM (Zhang et al., 2025b) filters irrelevant tokens layer-by-layer within the decoder via cross-modal attention matrices, while LVPruning (Sun et al., 2025) extracts query-specific regions early to prevent background patches from entering the decoder. Nevertheless, their efficacy heavily relies on cross-modal alignment estimators, which can degrade into noise when text comprehension or grounding is weak. (4) **Training-based methods** parameterize the selection module and optimize token distributions via data-driven supervision. ATP-LLaVA (Ye et al., 2025) incorporates learnable routing parameters supervised by multi-task instruction loss, and LearnPruner (Takezoe et al., 2026) designs a plug-and-play network to learn instance-level token values from the next-token prediction objective. Regrettably, aligning selection with implicit generation losses without modeling fine-grained, token-level necessity directly often induces biased or unstable compression under tight budgets.

In summary, existing literature predominantly relies on a singular pruning dimension, which introduces heuristic biases and fails to fully characterize token importance. This multi-faceted nature of token necessity motivates a holistic framework that synthesizes these complementary perspectives.

3 Preliminary Study

Token pruning methods for VLMs have been proposed from several distinct perspectives, each guided by a different hypothesis about what makes a visual token necessary. To investigate the relationship among pruning strategies, we conduct a preliminary study compare three representative methods embodying diverse philosophies: SCOPE (diversity-based), VisPruner (attention-based), and SparseVLM (question-guided).

3.1 Spatial Distribution of Retained Tokens

Observation 1: Different methods attend to fundamentally different visual regions.

Figure 2 visualizes the tokens retained by each method on identical image–question pairs. Despite operating under identical budgets on identical inputs, the spatial distributions of retained tokens diverge substantially.

Diversity-based pruning (SCOPE) scatters selections across distinct spatial areas to maintain global coverage. In contrast, Attention-based pruning (VisPruner) concentrates on visually salient foreground regions. Yet, question-guided pruning (SparseVLM) anchors retention to regions semantically grounded in the query text. These qualitative differences are not incidental, as each method's core hypothesis is about what makes a token necessary. No single method captures all aspects of visual relevance simultaneously, motivating us to unify their diverse signals.

3.2 Inference Correctness Across Methods

Observation 2: Different methods succeed and fail on different samples, revealing complementary capabilities.

To further investigate whether this spatial divergence translates into behavioral differences, we analyze the inference correctness of each method on GQA (Hudson and Manning, 2019).

Correctness patterns are interleaved, not aligned. Figure 3 reports the accuracies of SparseVLM (53.5%), VisPruner (55.9%), and SCOPE (58.4%) at a 64-token budget. In addition, the accuracy that any method answers correctly under pruning, namely the union (65.3%) exceed **+6.9%** over the strongest individual method. This gap reveals a substantial complementary correctness space, where samples that lie beyond the reach of any single method yet remain collectively solvable.

To characterize the structure of this complementarity, Figure 3 reports the pairwise Jaccard similarity between per-sample correctness sets over 12,578 GQA samples. Though similarities are substantial (0.767–0.830) on simple samples, the imperfect overlap shows that methods often uniquely succeed where others fail.

3.3 Takeaway

The observations reveal that ***token necessity is inherently multi-faceted and cannot be faithfully characterized by any single pruning criterion.***

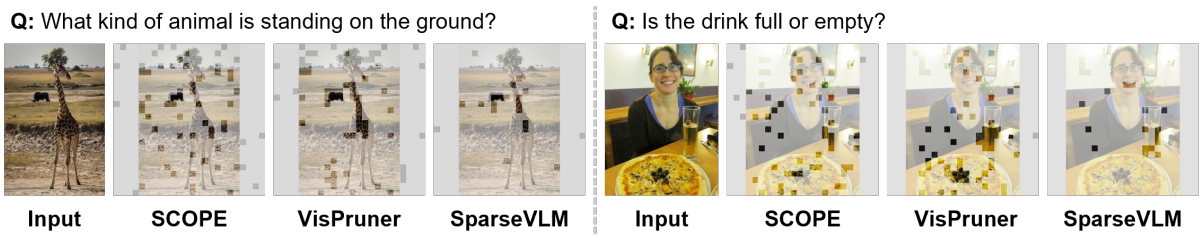


Figure 2: Visualization of retained visual tokens (highlighted patches) under three pruning methods at a 64-token budget. Each ethod preserves different spatial regions, reflecting distinct inductive biases over token relevance.

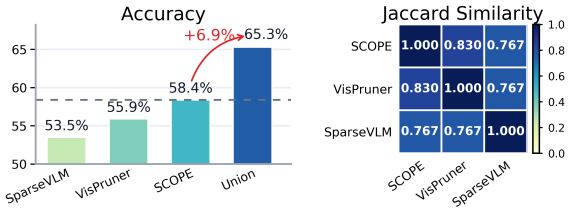


Figure 3: Left: Accuracy of each pruning method and their union under a 64-token budget on GQA. Right: Jaccard similarity between per-method correctness sets.

This directly motivates our **MVP-Token** to pursue multi-view-guided token pruning.

4 Method

Based on discussions above, we propose **MVP-Token**, a two-stage framework that prunes visual tokens by learning the complementary evidence from multiple pruning perspectives. (1) **MVP-PruneSet** collects multi-view token masks and per-sample correctness annotations from three diverse pruning families. (2) **SFF-Label** translates complementary success–failure evidence into relative importance labels. (3) **MVP-Pruner**, a compact question-aware scorer trained with an objective that directly matches top-K selection.

4.1 MVP-PruneSet

As mentioned, each pruning family embodies a distinct inductive bias making them complementary rather than redundant. MVP-PruneSet is designed to preserve the diversity of pruning evidence.

Mask collection. For each training sample we collect masks from question-guided (GRASP), attention-based (VisPruner), and diversity-based (SCOPE). The proposed GRASP produces a variable-size mask determined by the question content, while SCOPE and VisPruner each contribute two fixed budgets ($K = 64$ and 128) to enrich the evidence range, giving five masks in total:

$$\mathcal{M} = \{M_{\mathrm{g}}, M_{\mathrm{v},K}, M_{\mathrm{s},K}\}, \tag{1}$$

where $M_m \in \{0,1\}^N$ and N is the number of visual tokens $\{\mathbf{v}_i\}$ encoded from the image.

Question-guided mask. Existing question-guided methods prune inside the LLM decoder, after all visual tokens have already entered the key-value cache. Their masks are therefore unsuitable as supervision for a pre-projector pruner. We introduce **GRASP** (Grounding-Aware Semantic Pruning) to supply a **question-guided mask** as well as at the **encoder side.** Specially, a grounding model (Qwen3-VL(Bai et al., 2025a)) extracts the minimal set of visually relevant targets for the question. And then, the targets are localized by a text-prompted segmentation model (SAM3(Carion et al., 2025)) as object masks, which are projected onto the VLM's 24×24 (LLaVA) patch grid with identical preprocessing, yielding $M_{\mathrm{m}} \in \{0,1\}^{576}$ that is aligned with encoder-side methods.

Moreover, each mask is applied into the frozen VLM to record whether the pruned representation still answers the question correctly. This turns each mask into structured evidence as successful masks confirm retained tokens are jointly sufficient and failed masks reveal the retained sets are missing something necessary.

4.2 MVP-Pruner

We design MVP-Pruner to learn the token importance labels from multiple views. Specially, visual and question tokens are projected into a shared latent space. Each pruner block applies visual self-attention (spatial dependencies), cross-attention to question tokens (query context), and a SwiGLU FFN. A linear head outputs token-level, importance logit $s_i = g_\theta(\mathbf{v}_i, \mathbf{T})$ per visual token, where $\mathbf{T}$ denotes the question token sequence. At training, the backbone is frozen and only the pruner is updated; at inference, only the top-K tokens by s_i are

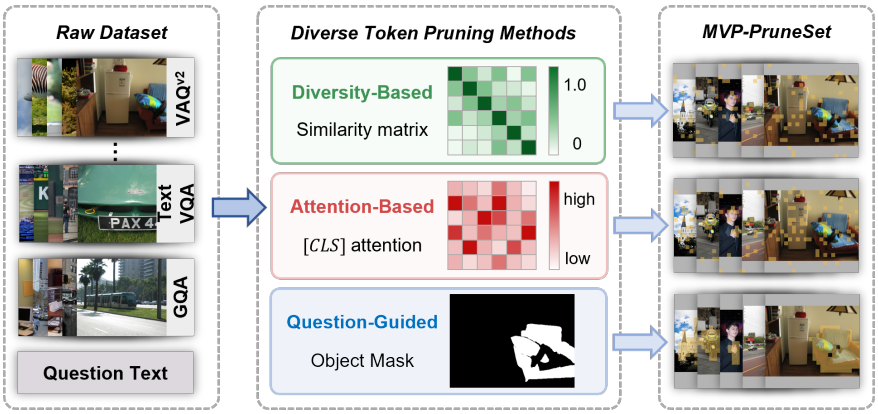


Figure 4: Overview of MVP-PruneSet Construction. Three complementary pruning families annotate each training sample with success–failure evidence to ensure complete multi-view coverage.

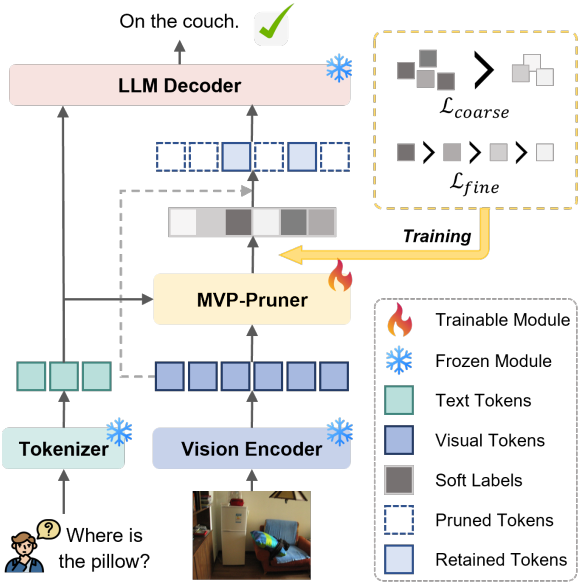


Figure 5: Overview of MVP-Pruner, supervised via a ranking-centered objective.

forwarded to the frozen VLM. Specifically, **MVP-Pruner** is trained with **SFF-Label** as the supervision signal and a ranking-centered objective that directly supervises the relative importance of token pairs at both coarse and fine-grained levels.

SSF-Label. A token selected by multiple independent views carries strictly stronger evidence of necessity than one selected by only one. SFF-Label operationalizes this intuition as a continuous importance label $y_i \in [0, 1]$. It lets tokens naturally stratify by the breadth of cross-view evidence supporting them, preserving the multi-faceted structure of token necessity.

Reliability-weighted. Not all views are equally trustworthy. We measure each mask type's evidence efficiency $E_m = \mathrm{Acc}_m / R_m$, which rewards high accuracy at low retention rates. Mask-level weights α_m are normalized within each family; family-level weights β_f are normalized across families so that GRASP's single mask is compared fairly against multi-budget families.

Evidence fusion. Masks are split by correctness into correct set $\mathcal{C}^+$ and wrong set $\mathcal{C}^-$. Within each family f, positive support $p_{f,i}$ and failure selection strength $q_{f,i}$ are defined as reliability-weighted retention rates over $\mathcal{C}^+ \cap f$ and $\mathcal{C}^- \cap f$. Cross-family aggregation then yields global signals:

$$P_i = \sum_{f \in \mathcal{A}^+} \bar{\beta}_f \, p_{f,i}, \tag{2}$$

$$Q_i = \sum_{f \in \mathcal{A}^-} \bar{\beta}_f \, q_{f,i}, \tag{3}$$

where $\bar{\beta}_f$ denotes β_f normalized over families with at least one correct ($\mathcal{A}^+$) or failed ($\mathcal{A}^-$) mask. P_i is high only when multiple successful perspectives across *different* families all retain token i. And Q_i captures how often failed masks select i without successful backing.

Soft label. A failed mask is not purely negative evidence as it may retain genuinely necessary tokens while also including redundant or distracting ones that cause the prediction to fail. Treating Q_i as a hard suppressor would therefore penalize necessary tokens that happen to co-occur in failed masks. We instead combine P_i and Q_i via a Noisy-OR formulation (Pearl, 1989):

$$y_i = P_i \big[1 - Q_i(1 - P_i)\big]. \tag{4}$$

The suppression term $Q_i(1 - P_i)$ acts only on the positive-evidence *deficit* $(1 - P_i)$: tokens with strong multi-view positive support are largely unaffected, while tokens absent from successful masks

yet persistent in failed ones are progressively down-weighted. The resulting label stratifies tokens by the convergence of independent pruning evidence, preserving the multi-faceted structure of token necessity identified in Section 3.

Rank loss. Pruning is actually a ranking problem. At inference, MVP-Pruner selects the top-K tokens by score, so only the relative order of s_i matters, not its absolute value. Fitting $s_i \approx y_i$ numerically would impose an unnecessary constraint: a pruner that correctly ranks all tokens is equally good regardless of its absolute output scale. We therefore train with a purely order-based objective:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{budget}}. \tag{5}$$

For valid pairs $\mathcal{P} = \{(i,j) \mid \Delta_{ij} = y_i - y_j > \delta\}$, we penalize ordering violations via a soft-margin penalty:

$$\ell(i,j;\, m) = \text{softplus}\Big(\tfrac{m-(s_i-s_j)}{\tau}\Big), \tag{6}$$

where m is a target margin. The rank loss is a weighted sum of this penalty over $\mathcal{P}$, using two complementary weighting strategies:

$$\mathcal{L}_{\text{rank}} = \lambda_c\, \mathcal{L}_{\text{coarse}} + \lambda_f\, \mathcal{L}_{\text{fine}}. \tag{7}$$

Coarse term: weight each pair by Δ_{ij} and set margin $m = \gamma\Delta_{ij}$. Pairs with large label gaps receive both higher weight and a proportionally larger required margin, pushing clearly necessary tokens far from clearly unnecessary ones.

Fine term: weight each pair by Δ_{ij}^{-1} and use a fixed small margin $m = \mu$. This inverts the coarse strategy: near-boundary pairs with small Δ_{ij} are up-weighted, concentrating the loss on the tokens near the top-K cutoff where a single rank swap changes which tokens get retained. The higher weight on the fine term reflects that near-boundary ordering is the primary failure mode in practice.

BCE loss. After multi-view aggregation, SFF-Label provides relative token importance rather than calibrated probabilities, we do not directly regress its raw values. Instead, for a target token budget K, we construct a binary budget label z^K by assigning 1 to the top-K tokens in the aggregated soft label y and 0 to the remaining tokens. Given the predicted token scores s, we optimize the budget-level BCE objective:

$$\mathcal{L}_{\text{budget}} = \text{BCEWithLogits}(s, z^K). \tag{8}$$

This loss encourages the pruner to learn the coarse important/redundant token distribution.

5 Experiments

5.1 Experimental Setup

Implementation details. We evaluate MVP-Token on LLaVA-1.5-7B and Qwen2.5-VL-7B which are from different VLM families. The backbone VLM is kept frozen throughout. Only MVP-Pruner is trained. Rank loss hyperparameters follow the values ($\delta = 0.02$, $\lambda_c = 0.3$, $\lambda_f = 0.7$, $\gamma = 4.0$, $\mu = 0.1$, $\tau = 1.0$). At inference, MVP-Pruner scores all N visual tokens and the top-K are forwarded to the frozen VLM.

Training data. We train MVP-Pruner on the subsets of LLaVA-665K (Liu et al., 2023). We retain the GQA, OCR-VQA, and TextVQA splits, discarding long-dialogue, referring-expression, and purely descriptive subsets, as these either lack a definitive correct answer or do not stress visual token selection. The TextVQA subset within LLaVA-665K is descriptive, so we replace it with the TextVQA training split. To mitigate the memorization effect that LLaVA-665K is LLaVA-1.5's own fine-tuning corpus, we additionally exclude samples that the VLM answers correctly without visual input.

Evaluation benchmarks. We evaluate on eight vision-language benchmarks covering complementary reasoning dimensions: **VQAv2** (Goyal et al., 2017) and **GQA** (Hudson and Manning, 2019) for general visual question answering and compositional reasoning; **TextVQA** (Singh et al., 2019) for text recognition in natural images; **VizWiz** (Gurari et al., 2018) for robustness on real-world low-quality images; **POPE** (Li et al., 2023b) for object hallucination evaluation; **ScienceQA** (Lu et al., 2022) for multi-hop knowledge reasoning; **SEED-Bench** (Li et al., 2023a) for general multimodal understanding; and **MME** (Fu et al., 2026) for comprehensive multi-task evaluation. These benchmarks stress both spatial understanding and fine-grained token discrimination, providing a thorough evaluation of pruning quality.

Comparison methods. We compare against representative methods spanning all pruning families. For training-free baselines, we include the attention-based methods **FastV** (Chen et al., 2024) and **VisPruner** (Zhang et al., 2025a); the diversity-based methods **VisionZip** (Yang et al., 2025); and the question-guided method **SparseVLM** (Zhang et al., 2025b). We additionally compare against the training-based variant **VisionZip** 🔥, which fur-

Method	GQA	MME	POPE	SQA	VQA^{Text}	SEED	VizWiz	VQA^{V2}	Avg.
Total 576 Tokens									
Vanilla	**61.9**	**1862.0**	**85.9**	**69.5**	**58.2**	**58.6**	**50.0**	**78.5**	**100%**
Retain 192 Tokens (↓ 67%)									
FastV	52.7	1612.0	64.8	67.3	52.5	57.1	50.8	67.1	89.84%
SparseVLM	57.6	1721.0	83.6	**69.1**	56.1	55.8	50.5	75.6	96.39%
VisionZip	59.3	1782.6	85.3	68.9	57.3	56.4	51.6	76.8	98.21%
VisPruner	59.5	1779.0	**86.7**	68.5	57.4	58.5	51.5	77.0	98.84%
VisionZip 🔥	**60.1**	**1834.0**	84.9	68.2	**57.8**	57.1	**52.6**	**77.4**	**99.14%**
MVP-Token 🔥	**60.1**	1795.2	86.0	68.7	56.3	**58.7**	52.2	77.3	99.03%
Retain 128 Tokens (↓ 78%)									
FastV	49.6	1490.0	59.6	60.2	50.6	55.9	51.3	61.8	84.98%
SparseVLM	56.0	1696.0	80.5	67.1	54.9	53.4	51.4	73.8	94.26%
VisionZip	57.6	1761.7	83.2	68.9	56.8	54.9	52.0	75.6	96.91%
VisPruner	58.2	1743.0	84.6	**69.1**	57.0	57.0	**52.6**	75.8	97.84%
VisionZip 🔥	58.9	**1823.0**	83.7	68.7	56.1	55.8	49.9	**76.6**	97.28%
MVP-Token 🔥	**59.3**	**1760.0**	**85.6**	68.5	**57.2**	**57.4**	52.5	76.4	**98.39%**
Retain 64 Tokens (↓ 89%)									
FastV	46.1	1256.0	48.0	51.1	47.8	51.9	50.8	55.0	76.71%
SparseVLM	52.7	1505.0	75.0	62.2	51.8	51.1	53.1	68.2	89.01%
VisionZip	55.1	1690.0	77.0	69.0	55.5	52.2	52.9	72.4	93.90%
VisPruner	55.4	1691.0	80.4	69.2	55.8	53.7	53.3	72.7	95.03%
VisionZip 🔥	**57.0**	**1756.0**	80.9	68.8	56.0	53.4	46.8	**74.2**	94.38%
MVP-Token 🔥	56.8	1709.0	**82.1**	68.8	**56.3**	**55.2**	**53.8**	74.1	**96.38%**

Table 1: **Main results on LLaVA-1.5-7B.** We compare MVP-Token with representative visual token pruning methods under different retained-token budgets. The average score is normalized to the vanilla model with 576 visual tokens. The best results are highlighted in bold, and the second-best results are underlined. Higher values indicate better performance. 🔥denotes trainable methods, while the others are training-free.

ther fine-tunes the selection module on instruction-following data, and **DivPrune** (Alvar et al., 2025) on Qwen2.5-VL experiments.

5.2 Main Results

Results on LLaVA-1.5-7b. LLaVA-1.5-7B is a widely used VLM benchmark, which turns images into fixed 576 visual tokens. Table 1 reports the performance comparison on LLaVA-1.5-7B under different visual token budgets. Overall, MVP-Token achieves consistently strong performance across 192, 128, and 64 retained tokens, demonstrating its effectiveness under both moderate and aggressive pruning settings. As reducing 77.8% of visual tokens, MVP-Token preserves 99.03% of the vanilla performance, achieving competitive or superior results on most benchmarks. When the token budget is reduced to 128, MVP-Token further maintains 98.39% of the original performance and obtains the best average score among all compared methods. Under the most aggressive 64-token setting, MVP-Token still retains 96.38% of the vanilla performance, outperforming existing training-free pruning methods by a clear margin in average performance and also surpassing the fine-tuned baselines. The results show that MVP-Token learning token necessity from multi-view pruning supervision preserve task-relevant visual information more effectively than single-criterion pruning methods.

Method	Retain 30%	Retain 20%
VisPruner	94.15%	88.39%
MVP-Token 🔥	**95.88%**	**91.74%**

Table 2: **Summary results on Qwen-2.5-VL-7B.** We report the normalized average score under different retained-token ratios. Detailed benchmark results are provided in Appendix Table 3.

Results on Qwen-2.5-VL-7B. Qwen-2.5-VL-7B is a recent VLM encoding images into dynamic visual tokens. Table 2 reports results under dynamic-resolution settings across two retention ratios. MVP-Token consistently outperforms VisPruner under all retain ratio. Notably, the performance gap widens at the aggressive 20% budget (+3.35% vs. +1.73%), suggesting that multi-view supervision provides more robust importance estimation. Moreover, the detailed results is in appendix3

5.3 Ablation Study

Figure 6 validate the effectiveness of SFF-Label and Rank loss. The proposed soft label naturally combines the advantages of intersection and union labels: tokens repeatedly selected by successful

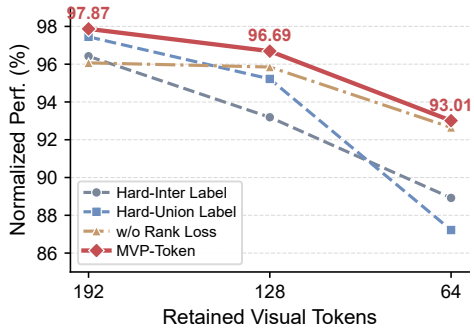


Figure 6: **Ablation study on key components.** Results are reported on LLaVA-1.5-7B using GQA, MME, and POPE with the budget of 64-token, 128-token, 192-token.

pruning views receive higher scores, while less frequently selected tokens are still assigned lower confidence. By contrast, the hard-union label contains too many positive tokens on average, reducing token discrimination and causing a large drop under the 64-token budget. The hard-intersection label is overly conservative, resulting in fewer positive tokens and weaker performance at larger budgets. Removing rank loss still allows the model to learn a coarse importance distribution with BCE, but weakens fine-grained token ordering, especially at the 192-token budget. MVP-Token consistently performs best across all budgets, showing the necessity of both soft multi-view labels and ranking supervision.

5.4 Visualization Results

Interpretability analysis. Figure 7 visualizes the pruning behavior of MVP-Token, and the heatmaps are directly produced by the learned pruner. The learned importance maps show that MVP-Token combines question-specific relevance, local discriminative evidence, and global contextual coverage. This behavior is consistent with our multi-view supervision design, where complementary pruning views provide success–failure evidence for learning token necessity. These results suggest that MVP-Token learns an interpretable and balanced pruning policy. Detailed case studies are provided in Appendix D.

Question-conditioned visualization. Figure 8 further visualizes MVP-Token on the same image with different questions. The retained tokens differ across questions. This comparison shows that MVP-Token adapts token selection according to the visual evidence required by text inputs.

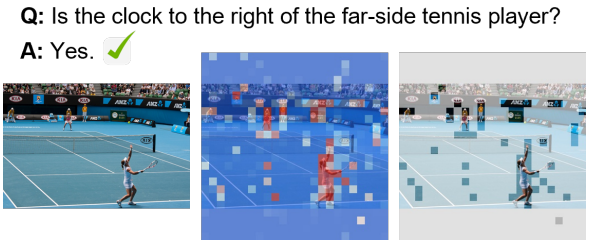


Figure 7: **Qualitative visualization of MVP-Token.** From left to right: original image, learned token-importance heatmap (**red** indicates high score, while **blue** indicates low score), and retained top-64 visual tokens. MVP-Token assigns high importance to question-relevant regions – clock and tennis players, while preserving sparse contextual tokens for spatial reasoning.

Figure 8: **Question-conditioned pruning visualization.** For the same image, MVP-Token retains different visual tokens according to different questions. For the heatmaps, **red** indicates high importance, while **blue** indicates low importance.

6 Conclusion

In this work, we propose **MVP-Token**, a multi-view token pruning framework that reframes visual token pruning as a multi-view learning problem. By aggregating complementary success–failure evidence from diverse pruning perspectives, MVP-Pruner learns a balanced pruning policy that integrates question-aware relevance, local discriminability, and global spatial coverage simultaneously. Extensive experiments on two representative VLMs across eight benchmarks demonstrate that MVP-Token consistently outperforms both training-free and fine-tuned baselines under aggressive token budgets.

Limitations

MVP-Pruner currently requires a pre-specified token budget K at inference time. Since MVP-Pruner is trained to learn relative token importance rather than absolute token necessity, it cannot adaptively determine how many tokens a given sample requires. As a result, visually simple samples may still retain redundant tokens under a fixed budget, as illustrated by qualitative examples in Appendix D. Developing a budget-free variant that predicts instance-level token counts remains an avenue for future work.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *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*.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. 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*, pages 9392–9401. Computer Vision Foundation / IEEE.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966v3*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv: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, Ming-Hsuan 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. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

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

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Judea Pearl. 1989. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

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

Rinyoichi Takezoe, Yaqian Li, Zihao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. 2026. Learnpruner: Rethinking attention-based token pruning in vision language models. *arXiv preprint arXiv:2604.23950*.

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

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24972–24982.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

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

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. 2025b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2026. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832.

A Evaluation Setup and Benchmark Details

A.1 Preliminary Study Setup

All experiments areconducted on LLaVA-1.5-7B with pruning applied strictly between the visual encoder and the multimodal projector, evaluated on GQA under a uniform budget of 64 retained tokens.

A.2 Evaluation Benchmarks

In this section, we provide a brief introduction to the evaluation benchmarks used in our experiments.

VQA-v2 (Goyal et al., 2017). VQA-v2 is a large-scale visual question answering benchmark built on images from MS-COCO (Chen et al., 2015). Each image is paired with open-ended questions and ten human-provided ground-truth answers. The dataset is carefully balanced to reduce language priors, requiring models to genuinely rely on visual evidence to answer correctly. And it's denoted as $\textbf{VQA}^{\textbf{V2}}$ in our experimental tables.

GQA (Hudson and Manning, 2019). GQA focuses on compositional visual reasoning over scene graph structures derived from Visual Genome (Krishna et al., 2017). Questions involve multi-step spatial relations, object attributes, and logical operations, making it a challenging testbed for fine-grained visual understanding.

TextVQA (Singh et al., 2019). TextVQA evaluates a model's ability to read and reason about text appearing within natural images. Answering questions may require direct text recognition or contextual reasoning over both visual and textual cues in the image. It is denoted as $\textbf{VQA}^{\textbf{Text}}$ in our experimental tables.

VizWiz (Gurari et al., 2018). VizWiz is collected from real-world photos taken by visually impaired users, featuring low image quality, unusual viewpoints, and ambiguous content. It evaluates model robustness under challenging, out-of-distribution visual conditions that are rarely present in standard benchmarks.

POPE (Li et al., 2023b). POPE targets object hallucination in VLMs through binary yes/no questions about the presence of objects in images. It provideds a comprehensive measure of hallucination robustness.

ScienceQA (Lu et al., 2022). ScienceQA consists of multiple-choice science questions spanning natural science, language science, and social science. It denoted as **SQA** in our experimental tables.

SEED-Bench (Li et al., 2023a). SEED-Bench is a comprehensive multiple-choice benchmark covering 12 evaluation dimensions of multimodal understanding, including scene understanding, instance attributes, spatial relations, and instance interaction.

MME (Fu et al., 2026). MME measures both perception and cognition abilities of VLMs across 14 subtasks. Perception subtasks cover coarse-grained recognition (existence, count, position, color) and fine-grained recognition (scenes, landmarks, artworks). Cognition subtasks include commonsense reasoning, numerical calculation, text translation, and code reasoning.

B Complete Empircal Results

Table 3 presents the comprehensive empirical comparison between MVP-Token and other baseline methods on Qwen-2.5-VL-7B-Instruct.

C Efficiency

We evaluate inference efficiency under the dynamic-resolution setting (MinPix $= 256 \times 28 \times 28$, MaxPix $= 2048 \times 28 \times 28$), comparing vanilla Qwen2.5-VL-7B, VisPruner, and MVP-Token at a 20% token retention ratio. As shown in Table 4, retaining only 20% of visual tokens reduces the average token count from 867.59 to $\sim$173, yielding a **1.33**$\times$ speedup and a 24.6% reduction in end-to-end latency for MVP-Token, compared to **1.21**$\times$ for VisPruner.

We note that Memory (GB) reflects peak GPU memory during inference rather than the marginal cost of visual tokens alone. Since model weights and resident caches dominate GPU memory, token pruning yields limited memory savings but clear benefits in prefill-stage FLOPs and latency.

The latency advantage of MVP-Token over VisPruner stems from their selection mechanisms: MVP-Token applies a lightweight learned scorer and selects tokens via a single top-K operation, whereas VisPruner performs iterative importance-and-diversity-based selection, incurring higher overhead under variable-length dynamic-resolution inputs. This suggests that in large-scale multimodal

Method	GQA	MME	SEED	VQA^{Text}	SQA	Avg.
Original						
Vanilla	60.50	2327.00	74.20	77.70	87.50	100%
Retain 30% Tokens						
VisPruner	56.80	2161.70	71.40	70.30	85.10	94.15%
MVP-Token 🔥	**58.40**	**2172.80**	**71.60**	**73.70**	**85.90**	**95.88%**
Retain 20% Tokens						
VisPruner	54.10	2066.00	68.90	58.60	83.54	88.39%
MVP-Token 🔥	**57.00**	**2115.00**	**69.80**	**64.00**	**85.00**	**91.74%**

Table 3: **Detailed results on Qwen-2.5-VL-7B.** The experiments are conducted under the dynamic-resolution setting (MinPix = $256 \times 28 \times 28$, MaxPix = $2048 \times 28 \times 28$). The average score is normalized by the original model.

Method	MME Infer T (s)	GPU Util.	Memory (GB)	Avg. Token	Latency (ms)	Speedup (x)
NVIDIA RTX PRO 6000 Single GPU Performance						
Qwen2.5-VL-7B	281.10	52.62	18.20	867.59	118.41	1.00
VisPruner	233.04	46.49	18.17	173.52	98.16	1.21
MVP-Token	211.81	45.67	18.17	173.85	89.22	1.33

Table 4: **Comparison of Inference Efficiency.** All results are reproduced on a single NVIDIA RTX PRO 6000 GPU under the same hardware and evaluation settings.

models, the primary efficiency gain from visual token pruning lies in shortening the computation path rather than reducing peak memory consumption.

D Case Studies

D.1 Diverse Question Types

We present qualitative examples spanning four representative question types. In each case, MVP-Token selects semantically relevant tokens and produces the correct answer.

D.2 Question-Conditioned Pruning on the Same Image

To further illustrate the question-awareness of MVP-Token, we apply it to the same image under two different questions. The retained tokens differ substantially, confirming that MVP-Pruner adapts its selection to the specific visual evidence required by each query.

D.3 Fixed-Budget Behaviour on Visually Simple Samples

As discussed in Section 6, MVP-Pruner learns *relative* token importance rather than an absolute necessity threshold. On visually simple samples where few patches are truly informative, the importance heatmap exhibits no strongly peaked responses, and the fixed budget forces the model to retain tokens that are collectively low-importance. Figure 11 illustrates representative cases where this behaviour is most pronounced.

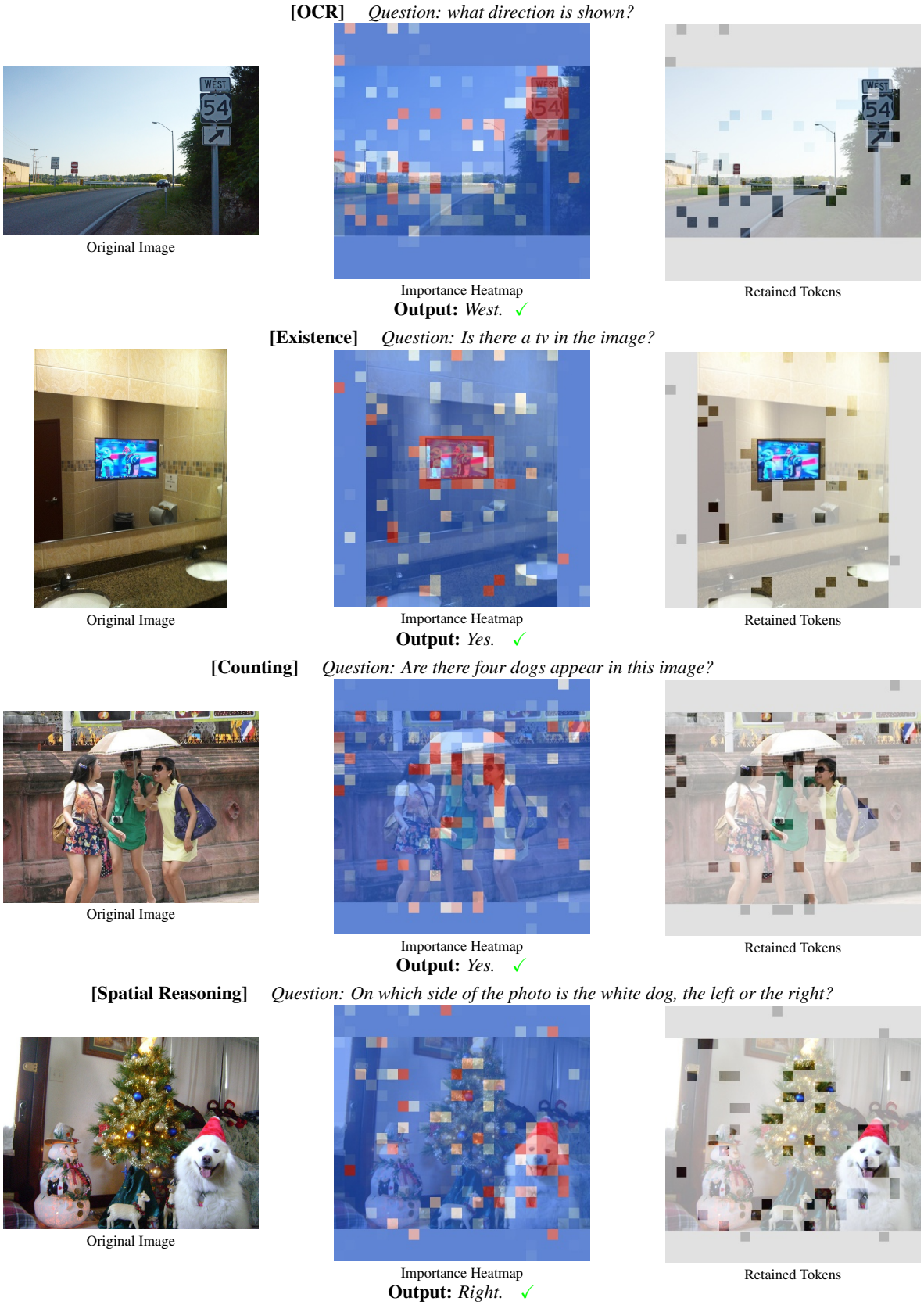


Figure 9: **MVP-Token on diverse question types.** For each example, we show the original image, the learned importance heatmap (**red**: high, **blue**: low), the retained top-64 tokens, and the model output. MVP-Token consistently focuses on question-relevant regions across different visual reasoning demands.

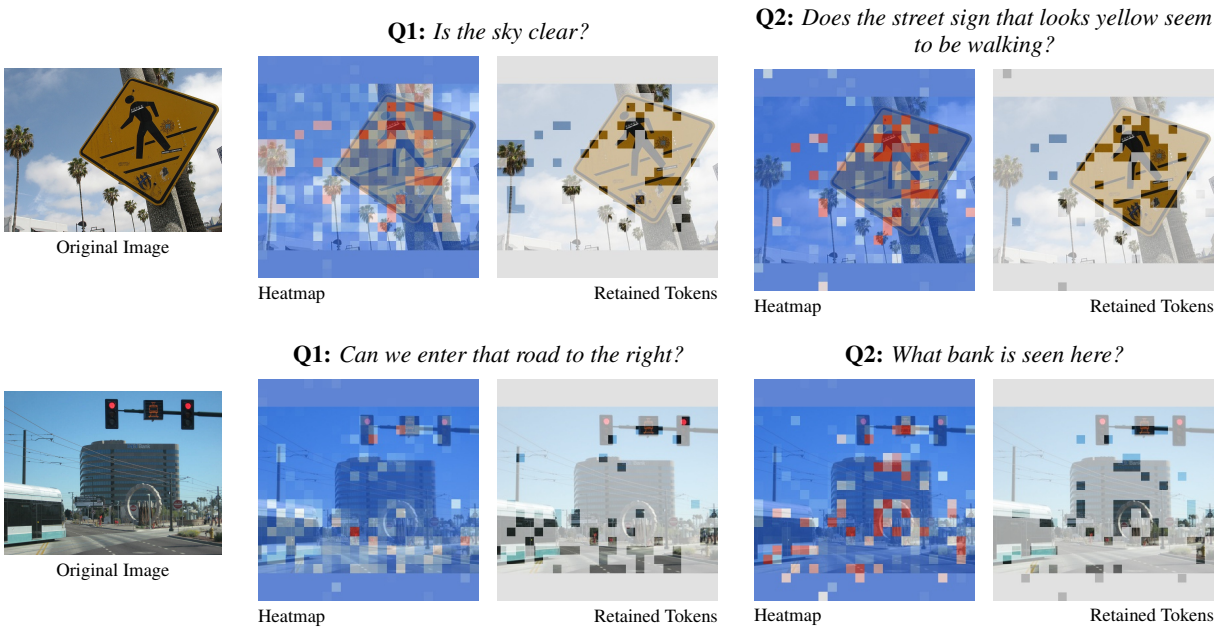
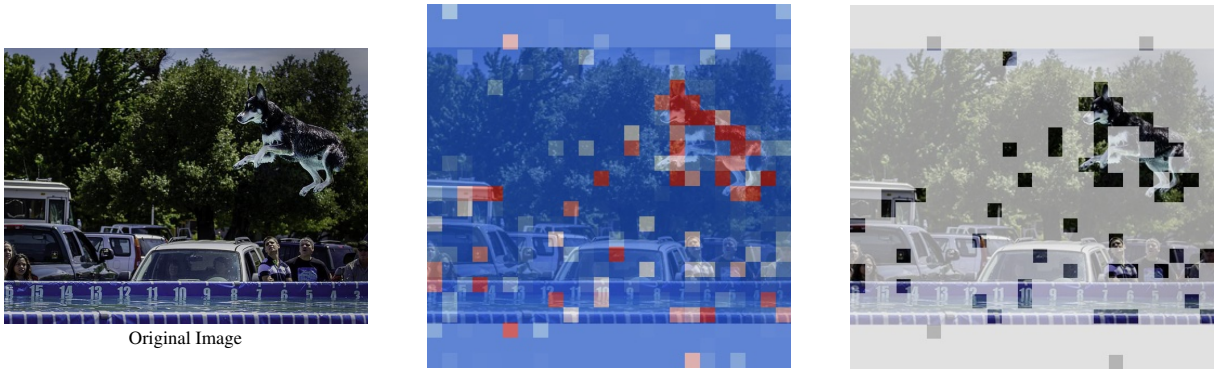


Figure 10: **Question-conditioned pruning on the same image.** Different questions elicit markedly different retention patterns, demonstrating that MVP-Pruner dynamically adapts token selection to the specific visual evidence demanded by each query.

Question: *Is there only one dog in the image?*



Original Image | Importance Heatmap | Retained Tokens

Output: *Yes.* ✓ (answered correctly despite token redundancy)

Question: *Is there a person in the image?*

Original Image | Importance Heatmap | Retained Tokens

Output: *Yes.* ✓ (answered correctly despite token redundancy)

Figure 11: **Fixed-budget behaviour on visually simple samples.** When a sample requires few discriminative tokens, the learned importance heatmap shows no strongly peaked response. Under a fixed budget K, the pruner retains dispersed low-importance tokens to fill the quota, revealing the limitation of relative-importance training discussed in Section 6.