QUEST: QUERY-GUIDED EVIDENCE SELECTION FOR VISUAL TOKEN REDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long visual token sequences increase the inference cost of the large language model (LLM) in vision-language models (VLMs). Effective visual token reduction requires preserving the visual information needed to answer the textual query. An external vision-language alignment model (VLAM) can provide relevance between image regions and the query. However, directly mapping this relevance to the visual tokens of the target VLM is difficult because the VLAM and the target VLM may use different vision encoders and visual representations. To address this issue, we propose QUEST, a training-free Pre-LLM visual token selection method. QUEST uses the relevance from an external VLAM to identify query-relevant Evidence Seed units and uses the corresponding vision encoder tokens as Evidence Seeds. It constructs Evidence Attention from the attention between the Evidence Seeds and other visual tokens at an intermediate layer of the target VLM's vision encoder. QUEST also selects an Evidence Source Layer based on layer-wise attention distance. Evidence Attention is combined with Visual Guidance, followed by diversity-aware selection to choose the final visual tokens under a given token budget. We evaluate QUEST on LLaVA-NeXT-7B and Qwen2.5-VL-7B across multiple VLAMs, token budgets, and nine benchmarks. The results show high average performance retention relative to the full-token VLMs across the evaluated token budgets.

1 INTRODUCTION

A typical vision-language model (VLM) consists of a vision encoder, a vision-language connector, and a large language model (LLM) (Liu et al., 2023; Li et al., 2023a; Zhu et al., 2025; Bai et al., 2025) (Fig. 1a). The input image and textual query are converted into visual tokens and text tokens, respectively, and jointly fed into the LLM. Recent VLMs use hundreds to thousands of visual tokens to process high-resolution images. This increases the sequence length, attention computation, and memory consumption of the LLM. Visual representations, however, contain substantial redundancy (Yang et al., 2025), and the information needed for a query is often concentrated in only part of the image (Guo et al., 2025). Reducing the number of visual tokens while preserving task-relevant visual information is therefore important for efficient VLM inference.

Visual token reduction can be broadly divided into token compression, which merges multiple tokens into fewer representations (Bolya et al., 2023; Li et al., 2025), and token selection or pruning, which retains only a subset of important tokens (Chen et al., 2024a; Zhang et al., 2025c;a). In this work, we focus on training-free token selection that can be applied to existing VLMs without model retraining (Chen et al., 2024a; Zhang et al., 2025c; Xing et al., 2025; Zhang et al., 2025a). We further categorize these methods as Inner-LLM or Pre-LLM according to where token reduction is performed (Guo et al., 2025).

Inner-LLM methods (Fig. 1b) feed all visual tokens into the LLM and remove a subset of them at early or intermediate layers (Chen et al., 2024a; Zhang et al., 2025c; Xing et al., 2025). These methods can estimate token importance after cross-modal interaction. However, they cannot reduce LLM computation before pruning, and text-visual attention may not accurately reflect visual relevance (Zhang et al., 2025a). Pre-LLM methods (Fig. 1c) remove visual tokens before they enter the LLM, directly reducing LLM computation (Yang et al., 2025; Zhang et al., 2025a). They must identify query-relevant visual tokens without relying on interactions inside the LLM.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

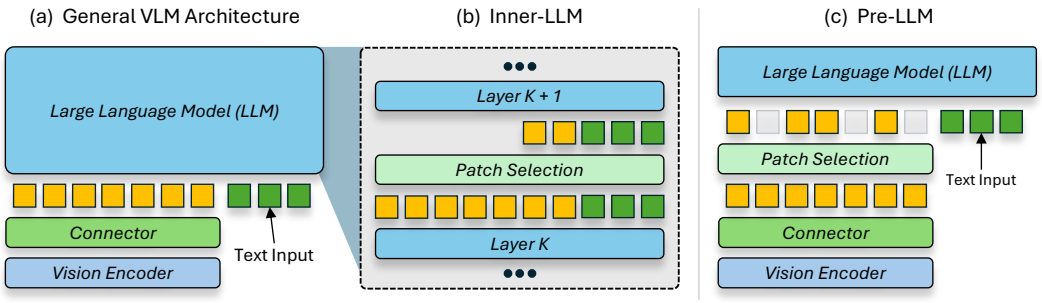

Figure 1: Comparison of Inner-LLM and Pre-LLM visual token selection. (a) A standard VLM architecture. (b) Inner-LLM visual token selection. (c) Pre-LLM visual token selection.

Existing Pre-LLM token selection methods use visual signals such as vision encoder attention, visual diversity, and sensitivity (Yang et al., 2025; Alvar et al., 2025; Kim et al., 2026). Visually important regions, however, do not always correspond to the regions needed to answer the textual query. To incorporate query information, CROP uses a separate lightweight VLM to detect query-relevant regions (Guo et al., 2025), while CDPruner, CrisPrune, and SFPruner use the textual query during token selection (Zhang et al., 2025b; Liu et al., 2026c; Song et al., 2026).

Methods that rely on vision-language alignment inside the target VLM may not apply equally well across different VLM architectures. The LLaVA family can exploit patch-level alignment from CLIP-based vision encoders (Liu et al., 2023; 2024a;b; Radford et al., 2021; Zhou et al., 2022). In contrast, Qwen2.5-VL, InternVL3, and BLIP-2 use different vision encoders and multimodal architectures (Bai et al., 2025; Zhu et al., 2025; Li et al., 2023a). An external vision-language alignment model (VLAM) can instead provide relevance between image regions and the textual query (Zhou et al., 2022; Wang et al., 2024; Lan et al., 2024; Zhai et al., 2023).

Directly mapping query relevance from an external VLAM to the visual tokens of a target VLM is not straightforward. VLAMs and VLMs may differ in input resolution, patch size, preprocessing, and feature representation. As a result, the same image region can be tokenized and represented differently (Dosovitskiy et al., 2021). Visual tokens in a Vision Transformer also become contextualized as they pass through self-attention layers. Spatial matching of the VLAM relevance map alone cannot fully capture how query-relevant information is represented inside the target VLM.

To address this problem, we propose Query-Guided Evidence Selection for Visual Token Reduction (QUEST). Figure 2 shows the overall pipeline. QUEST selects the Top-K Evidence Seed units based on VLAM relevance and uses the corresponding vision encoder tokens as Evidence Seeds. It then constructs Evidence Attention from the attention between the Evidence Seeds and other visual tokens at an intermediate layer of the target VLM's vision encoder. This step expands query-relevant evidence to related visual tokens. QUEST combines Evidence Attention with Visual Guidance and applies an existing diversity-aware selection method to select the final visual tokens under a given token budget. VLAM relevance is thus used to initialize query-relevant evidence rather than directly determine the final token importance score. This differs from existing methods that directly combine text relevance with visual saliency or diversity scores (Liu et al., 2026c; Song et al., 2026).

Attention characteristics in Vision Transformers can vary across layers (Liu et al., 2026b), while several existing Pre-LLM methods use information from the final layer (Shang et al., 2025; Ma et al., 2026). QUEST analyzes layer-wise attention patterns using attention distance (Liu et al., 2026a) and selects an Evidence Source Layer from which to extract Evidence Attention. Our ablation study shows that the selected intermediate layer achieves higher performance than the final layer.

Our main contributions are as follows.

- We propose a query-aware Pre-LLM visual token selection method that uses an external vision-language alignment model without relying on the target VLM's native vision-language alignment.

- We introduce Evidence Attention, which expands VLAM-identified Evidence Seed units into related visual evidence using the target VLM's internal attention. We also select the Evidence Source Layer based on layer-wise attention characteristics.
- We evaluate QUEST on VLMs with different vision encoders and connector architectures, and analyze performance retention and inference cost across multiple benchmarks, VLAMs, and token budgets.

2 Related Work

2.1 Large Vision-Language Models and Inference Inefficiency

A vision-language model (VLM) typically combines a vision encoder, a vision-language connector, and a large language model (LLM) to process visual information (Li et al., 2023a; Liu et al., 2023). Recent VLMs such as LLaVA-NeXT (Liu et al., 2024b), LLaVA-OneVision (Li et al., 2024), and Qwen2.5-VL (Bai et al., 2025) support various visual inputs, including high-resolution images, multiple images, and videos. Their long visual token sequences increase the computation and memory usage of the LLM (Vaswani et al., 2017). To reduce this cost, ToMe (Bolya et al., 2023), TRIPS (Jiang et al., 2022), and AIM (Zhong et al., 2025) use token merging or pruning.

2.2 Training-Free Visual Token Selection

Training-free visual token selection can be divided into Inner-LLM and Pre-LLM methods based on where tokens are removed. Inner-LLM methods such as FastV (Chen et al., 2024a), PyramidDrop (Xing et al., 2025), and SparseVLM (Zhang et al., 2025c) use token representations or attention scores after the visual tokens enter the LLM. These methods can use query-image interactions. However, they cannot reduce LLM computation before pruning.

Pre-LLM methods remove a subset of visual tokens before they enter the LLM. HiRED (Arif et al., 2025), VisionZip (Yang et al., 2025), and VisPruner (Zhang et al., 2025a) use vision encoder attention or token redundancy. DART (Wen et al., 2025) and DivPrune (Alvar et al., 2025) use visual diversity and redundancy. ZOO-Prune (Kim et al., 2026) estimates token importance using the sensitivity of the connector output. These visual signals do not directly reflect the textual query. Therefore, CDPruner (Zhang et al., 2025b), CrisPrune (Liu et al., 2026c), and SFPruner (Song et al., 2026) use text-visual relevance for token selection. PinPoint (Kwon et al., 2026) instead trains a lightweight instruction-region alignment module to identify a contiguous image region related to the query.

2.3 Vision-Language Alignment Models

CLIP (Radford et al., 2021) aligns images and text in a shared representation space through contrastive learning. However, CLIP is trained mainly with image-level representations, which limits fine-grained region-text alignment. MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), and ClearCLIP (Lan et al., 2024) modify pretrained CLIP representations to improve patch-level semantic alignment. Implementation details are provided in Appendix C.2. Prior work focuses on obtaining localized vision-language representations. QUEST uses this local alignment to identify query-relevant evidence and select visual tokens.

3 Method

QUEST consists of three stages: Semantic Evidence Mining, Guidance Fusion, and Diversity-Aware Selection. The overall framework is shown in Figure 2. QUEST first identifies Evidence Seed units using an external VLAM. It then expands query-relevant evidence using internal attention from the target VLM and selects the final visual tokens by considering visual saliency and diversity.

3.1 Preliminaries

VLM Architecture. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the VLM vision encoder $\mathrm{VE}_{\texttt{VLM}}$ produces visual tokens:

$$V = \mathrm{VE}_{\texttt{VLM}}(I) \in \mathbb{R}^{N_v \times d_{\mathrm{VE}}}. \tag{1}$$

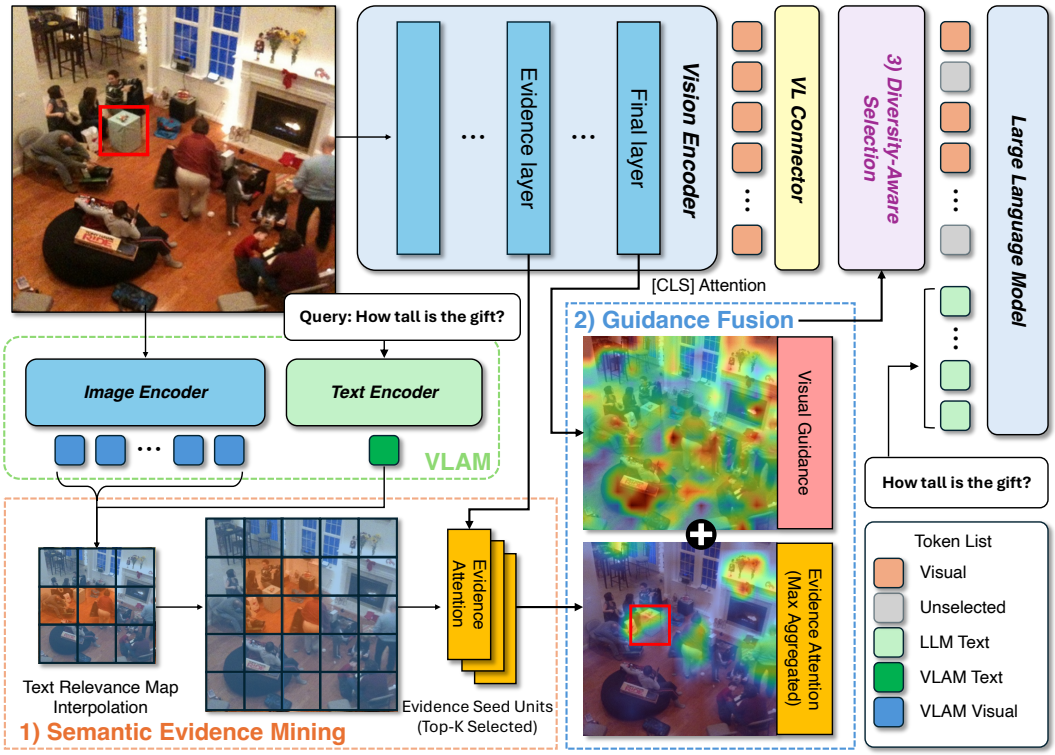

Figure 2: An overview of QUEST. An external VLAM identifies Evidence Seed units, and the corresponding vision encoder tokens are used to compute Evidence Attention. QUEST combines Evidence Attention with Visual Guidance and selects the final visual tokens using Diversity-Aware Selection.

The vision-language connector Conn maps them to the visual representations fed into the LLM:

$$\hat{V} = \mathrm{Conn}(V) \in \mathbb{R}^{\hat{N}_v \times d_{\mathrm{LLM}}}. \tag{2}$$

Here, N_v and $\hat{N}_v$ denote the numbers of visual tokens before and after the connector, and d_{VE} and d_{LLM} denote the hidden dimensions of the vision encoder and LLM. For token-wise projection, $\hat{N}_v = N_v$. For connectors with spatial merging, $\hat{N}_v < N_v$ may hold. QUEST computes guidance in the vision encoder token space and selects B tokens from the $\hat{N}_v$ connector-output tokens, where $B < \hat{N}_v$.

3.2 Semantic Evidence Mining

Evidence Seed. Given an input image I and query Q, the VLAM computes a patch-level text relevance map R:

$$R = \mathrm{VLAM}(I, Q) \in \mathbb{R}^{N \times N}, \tag{3}$$

where $R_{i,j}$ represents the semantic relevance between the image region at (i, j) and the query. The VLAM and target VLM may have different spatial resolutions. Therefore, we interpolate R to a grid that matches the visual token structure of the target VLM:

$$\tilde{R} = \mathrm{Interp}\left(R, M_r, M_c\right) \in \mathbb{R}^{M_r \times M_c}. \tag{4}$$

QUEST selects the K Evidence Seed units with the highest relevance in $\tilde{R}$. We define $E \subseteq \{1, \ldots, N_v\}$ as the set of vision encoder token indices corresponding to these units. When token resolution is preserved, each unit corresponds to one vision encoder token. With spatial merging, one unit may correspond to multiple tokens. Model-specific mappings are described in Appendix C.

Evidence Attention. Evidence Seeds correspond to regions with high query relevance in the VLAM. However, query-relevant information may also exist in other visual tokens. QUEST uses attention between the Evidence Seeds and other visual tokens to expand the query-relevant evidence. At vision encoder layer ℓ, let H_ℓ denote the number of attention heads and $A_V^{(\ell)} \in \mathbb{R}^{H_\ell \times N_v \times N_v}$ denote the self-attention submatrix whose query and key positions both correspond to visual tokens. The Evidence Attention score for visual token j is computed as follows:

$$G_j^E = \max_{i \in E} \left(\frac{1}{H_\ell} \sum_{h=1}^{H_\ell} A_{V,hij}^{(\ell)} \right), \qquad j = 1, \ldots, N_v. \tag{5}$$

We first average the attention from each Evidence Seed to token j across heads and then take the maximum across Evidence Seeds. An ablation of the aggregation method is provided in Appendix D.1.

Evidence Source Layer. Evidence Attention can vary with the depth of the vision encoder (Liu et al., 2026b). QUEST uses attention distance (Liu et al., 2026a) to select the Evidence Source Layer. Let $p_i = (x_i, y_i)$ be the 2D patch-grid coordinate of visual token V_i, and define $d(V_i, V_j) = \|p_i - p_j\|_2$. We average the visual attention at layer ℓ over the heads:

$$\bar{A}_{ij}^{(\ell)} = \frac{1}{H_\ell} \sum_{h=1}^{H_\ell} \left(A_V^{(\ell)} \right)_{h,i,j}. \tag{6}$$

Using this matrix, we define the maximum attention distance as

$$D_{\max}^{(\ell)} = \frac{1}{N_v} \sum_{i=1}^{N_v} d\left(V_i, V_{\arg\max_j \bar{A}_{ij}^{(\ell)}} \right). \tag{7}$$

For each query token, we measure the spatial distance to the visual token receiving its strongest attention. $D_{\max}^{(\ell)}$ is the average of this distance over all query tokens. For comparison, we also compute $D_{\text{avg}}^{(\ell)}$ as the average attention-weighted spatial distance over all query-key pairs.

For each model, we compute $D_{\max}$ using only images from the GQA validation set and average it across samples. We select the layer with the largest increase in the average $D_{\max}$ over the previous measured layer as the Evidence Source Layer. Textual queries and annotations are not used in this process. The analysis is presented in Sec. 4.3.

3.3 Guidance Fusion

Visual Guidance. Evidence Attention captures query-relevant evidence. However, visual information that is not directly related to the query may still be important for understanding the image. To capture this information, we compute Visual Guidance $G^V \in \mathbb{R}^{N_v}$ from the final-layer attention of the vision encoder. Let L denote the final vision encoder layer, and let $A^{(L)}$ denote its self-attention matrix with H_L attention heads. For vision encoders with a `[CLS]` token, we average the `[CLS]`-to-token attention across heads. For encoders without a `[CLS]` token, we average the attention over the visual query rows:

$$G_j^V = \begin{cases} \frac{1}{H_L} \sum_{h=1}^{H_L} A_{h,\texttt{CLS},j}^{(L)}, & \text{with } \texttt{[CLS]}, \\ \frac{1}{H_L N_v} \sum_{h=1}^{H_L} \sum_{i=1}^{N_v} A_{h,i,j}^{(L)}, & \text{without } \texttt{[CLS]}, \end{cases} \qquad j = 1, \ldots, N_v. \tag{8}$$

Fused Guidance. Evidence Attention G^E and Visual Guidance G^V may have different scales. We therefore apply min-max normalization to each guidance independently:

$$\tilde{G}_i^E = \frac{G_i^E - \min_j G_j^E}{\max_j G_j^E - \min_j G_j^E + \epsilon}, \qquad \tilde{G}_i^V = \frac{G_i^V - \min_j G_j^V}{\max_j G_j^V - \min_j G_j^V + \epsilon}, \tag{9}$$

where ϵ is a small constant for numerical stability. We compute the Fused Guidance G using a balance factor $\alpha \in [0, 1]$:

$$G = \alpha \tilde{G}^E + (1 - \alpha) \tilde{G}^V \in \mathbb{R}^{N_v}. \tag{10}$$

Table 1: Performance comparison on LLaVA-NeXT-7B. CLIP† uses negated image-text cosine similarity.

Method	VLAM	GQA	MMBench	MME	POPE	SQA-IMG	TextVQA	Avg.
Total 2880 ***Tokens*** (100.0%)								
LLaVA-NeXT-7B		62.50	65.80	1511.80	86.80	67.50	60.30	100.00%
Retain 640 ***Tokens*** (22.2%)								
VisionZip	-	61.30	66.30	**1493.40**	86.30	68.10	**59.90**	99.55%
DivPrune	-	61.58	65.38	1469.70	85.51	67.82	57.00	98.10%
CDPruner	CLIP†	62.60	66.30	1474.20	87.30	67.90	58.40	99.41%
SFPruner	CLIP	62.70	**67.10**	1480.30	87.10	67.60	59.00	99.76%
QUEST	CLIP†	**63.64**	65.00	1490.73	**87.71**	**68.17**	59.23	**99.91%**
Retain 320 ***Tokens*** (11.1%)								
VisionZip	-	59.30	63.10	1397.10	82.10	67.30	**58.80**	95.83%
DivPrune	-	59.63	63.66	1423.30	83.47	67.82	56.20	96.02%
CDPruner	CLIP†	61.60	**65.50**	1453.10	87.20	67.80	57.40	98.39%
SFPruner	CLIP	61.50	65.10	**1507.60**	86.70	67.80	58.30	**99.01%**
QUEST	CLIP†	**61.63**	**65.50**	1474.50	**87.50**	**68.42**	58.02	**99.01%**

The setting of α is described in Appendix D.2. When the connector changes the token resolution, we map G^E and G^V to the connector-output token space. We then apply the same normalization and fusion to obtain $\hat{G} \in \mathbb{R}^{\hat{N}_v}$. When the token resolution is preserved, $\hat{G} = G$. The model-specific mappings are described in Appendix C.

3.4 Diversity-Aware Selection

Using only the Fused Guidance $\hat{G}$ for selection may retain similar tokens. To reduce this redundancy, QUEST uses the Semantics-Guided Ridge Leverage Score (SG-RLS) and Directional Masking (DM) from SFPruner (Song et al., 2026). We compute SG-RLS using the connector-output features $\hat{V}$ and Fused Guidance $\hat{G}$. We then apply cosine-similarity-based DM to obtain the final score:

$$S^{\text{final}} = \text{DM}\left(\text{SG-RLS}(\hat{V}, \hat{G}), \hat{V}\right), \qquad \hat{P} = \text{TopK}(S^{\text{final}}, B). \tag{11}$$

Details of SG-RLS and DM are provided in Appendix A.

4 Results

4.1 Experimental Setup

VLM Backbones and Benchmarks We evaluate LLaVA-NeXT-7B (Liu et al., 2024b) and Qwen2.5-VL-7B (Bai et al., 2025). We refer to the alignment model used for visual token selection as a vision-language alignment model (VLAM). We use nine benchmarks in total. GQA (Hudson & Manning, 2019), POPE (Li et al., 2023b), MME (Fu et al., 2025), MMBench (Liu et al., 2024c), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), AI2D (Kembhavi et al., 2016), and MMStar (Chen et al., 2024b) are evaluated with LMMs-Eval (Zhang et al., 2024). HBench (Guan et al., 2024) is evaluated with VLMEvalKit (Duan et al., 2024). We ran all inference experiments in FP16 on a single RTX PRO 6000 Blackwell Max-Q GPU. Details of the benchmarks are provided in Appendix B.

Hyperparameters We set the balance factor α to 0.5 based on the experiments in Appendix D.2. For the ridge parameter λ, we follow the setting of Song et al. (2026).

Table 2: Performance comparison on Qwen2.5-VL-7B. CLIP† uses negated image-text cosine similarity.

Method	VLAM	AI2D	HBench	MMBench	MME	MMStar	POPE	TextVQA	Avg.
Average 614.39 ***Tokens*** (100.0%)									
Qwen2.5-VL-7B		82.90	66.50	83.00	2312.70	62.90	86.50	78.90	100.00%
Average 184.32 ***Tokens*** (30.0%)									
Random	-	78.27	61.62	80.07	2263.71	55.90	83.26	71.18	93.82%
DivPrune	-	80.15	62.78	80.93	2225.76	56.34	84.79	74.28	95.22%
CDPruner	CLIP†	80.41	61.62	80.93	2298.06	59.35	**87.20**	76.26	96.90%
SFPruner	CLIP†	**81.54**	62.36	**82.22**	**2351.04**	58.37	87.02	**76.66**	97.63%
QUEST	CLIP†	81.12	**63.83**	81.96	2323.72	**60.10**	87.19	76.55	**98.06%**
Average 122.88 ***Tokens*** (20.0%)									
Random	-	73.87	55.73	78.18	2194.62	52.25	80.85	66.62	88.99%
DivPrune	-	77.01	59.40	80.15	2182.20	52.73	83.70	71.60	92.07%
CDPruner	CLIP†	79.02	59.94	79.81	**2282.25**	56.10	**87.23**	73.95	94.86%
SFPruner	CLIP†	**79.76**	60.67	79.90	2274.92	56.79	86.56	74.44	95.25%
QUEST	CLIP†	79.21	**61.41**	**80.33**	2249.97	**58.32**	86.89	**74.54**	**95.66%**
Average 61.44 ***Tokens*** (10.0%)									
Random	-	69.62	54.78	74.91	2044.69	47.76	75.88	66.62	84.73%
DivPrune	-	70.98	54.89	77.15	2099.20	45.93	78.94	65.61	85.62%
CDPruner	CLIP†	74.97	57.94	77.75	**2157.05**	50.02	**87.41**	68.58	90.28%
SFPruner	CLIP†	74.77	56.99	76.98	2121.47	**51.84**	86.06	68.97	89.95%
QUEST	CLIP†	**75.00**	**58.04**	**78.26**	2078.12	**51.84**	86.00	**69.73**	**90.30%**

4.2 Performance Comparison

LLaVA-NeXT-7B. Table 1 compares performance on LLaVA-NeXT-7B. For existing methods, we use the reported results and their original VLAM settings. We evaluate QUEST directly under the same token budgets. Avg denotes the average performance retention, with the full-token performance normalized to 100%. QUEST achieves the highest Avg of 99.91% with 640 tokens (22.2% retain). With 320 tokens (11.1% retain), QUEST achieves 99.01%, matching SFPruner.

Qwen2.5-VL-7B. Table 2 compares performance on Qwen2.5-VL-7B. Previous studies use different evaluation settings, so we reproduce all baselines using the setting of SFPruner (Song et al., 2026). Qwen2.5-VL-7B does not provide a paired text encoder. For baselines that use text relevance, we therefore use the same CLIP† as QUEST. We specify the token budget by retention ratio to account for dynamic resolution. QUEST achieves the highest Avg at all retention ratios of 30%, 20%, and 10%.

4.3 Ablation Studies

Evidence Source Layer Selection. Figure 3 shows the layer-wise changes in $D_{\max}$ and D_{avg} for the CLIP and Qwen2.5-VL vision encoders. $D_{\max}$ increases most at layer 12 in CLIP and layer 15 in Qwen2.5-VL. We select these layers as the Evidence Source Layers.

We perform a source-layer ablation on HBench, MMStar, and POPE. For both VLMs, the layer selected by $D_{\max}$ gives the highest average performance as shown in Table 3. For Qwen2.5-VL-7B, we evaluate all global-attention layers. In both VLMs, the selected intermediate layer also performs better on average than the final layer.

VLAM Choice. Table 4 shows that CLIP† performs best on LLaVA-NeXT-7B, while SCLIP performs best on Qwen2.5-VL-7B. This result suggests that the appropriate VLAM may depend on

Table 3: Performance across Evidence Source Layers on LLaVA-NeXT-7B and Qwen2.5-VL-7B. All results use SG-RLS with $K = 20$ Evidence Seed units. CLIP† uses negated image-text cosine similarity. Source indices are zero-based. Values in parentheses show changes relative to source layers 12 and 15 for LLaVA-NeXT-7B and Qwen2.5-VL-7B, respectively.

VLAM	Source Index	Retain	HBench	MMStar	POPE	Avg.
LLaVA-NeXT-7B		100%	54.57	37.09	86.80	100.00%
CLIP†	8	11.1%	53.63 (-2.10%)	**35.62** (+0.11%)	87.37 (-0.15%)	98.32% (-0.73%)
CLIP†	**12**	11.1%	**54.78**	35.58	**87.50**	**99.04%**
CLIP†	23	11.1%	53.94 (-1.53%)	34.88 (-1.97%)	86.92 (-0.66%)	97.67% (-1.38%)
Qwen2.5-VL-7B		100%	66.50	62.90	86.50	100.00%
CLIP†	7	20%	**62.25** (+1.37%)	56.86 (-2.50%)	**87.16** (+0.31%)	94.92% (-0.26%)
CLIP†	**15**	20%	61.41	**58.32**	86.89	**95.17%**
CLIP†	23	20%	60.88 (-0.86%)	58.07 (-0.43%)	86.52 (-0.43%)	94.63% (-0.57%)
CLIP†	31	20%	61.20 (-0.34%)	57.39 (-1.59%)	86.89	94.57% (-0.63%)

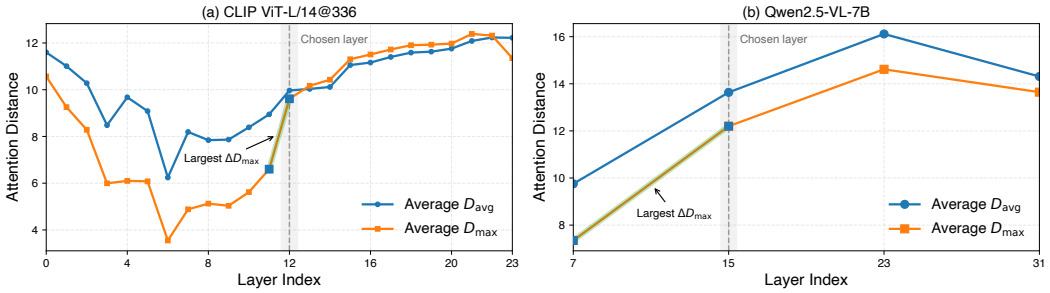

Figure 3: Layer-wise attention distance analysis using only images from the GQA validation set. (a) $D_{\max}$ increases most at CLIP layer 12. (b) $D_{\max}$ increases most at Qwen2.5-VL layer 15. For Qwen2.5-VL, only global-attention layers are evaluated. D_{avg} is shown for comparison.

Table 4: Performance comparison across VLAMs on Qwen2.5-VL-7B and LLaVA-NeXT-7B. All results use SG-RLS with $K = 20$ Evidence Seed units. The Evidence Source Layers are 15 and 12, respectively. CLIP† uses negated image-text cosine similarity. Values in parentheses show changes relative to CLIP† for each VLM.

VLAM	HBench	MMBench	MME	POPE	Avg.
Average 614.39 ***Tokens*** (100.0%)					
Qwen2.5-VL-7B	66.5	83.0	2312.7	86.5	100.00%
Average 122.88 ***Tokens*** (20.0%)					
CLIP†	61.4	80.3	2250.0	86.9	96.72%
MaskCLIP	61.9 (+0.75%)	80.4 (+0.12%)	**2310.1** (+2.60%)	**87.5** (+0.69%)	97.75% (+1.03%)
SCLIP	**63.9** (+3.76%)	**80.8** (+0.60%)	2261.0 (+0.48%)	87.0 (+0.12%)	**97.94%** (+1.22%)
ClearCLIP	61.4	80.3	2271.3 (+0.92%)	86.4 (-0.58%)	96.82% (+0.10%)
SigLIP2	59.5 (-2.86%)	79.5 (-0.96%)	2199.2 (-2.20%)	83.0 (-4.51%)	94.07% (-2.65%)
2880 ***Tokens*** (100.0%)					
LLaVA-NeXT-7B	54.57	65.80	1511.80	86.80	100.00%
320 ***Tokens*** (11.1%)					
CLIP†	54.78	**65.50**	1474.50	87.50	**99.57%**
MaskCLIP	53.63 (-2.10%)	63.40 (-3.21%)	1475.87 (+0.09%)	**87.54** (+0.05%)	98.27% (-1.30%)
SCLIP	54.36 (-0.77%)	63.49 (-3.07%)	**1487.46** (+0.88%)	86.84 (-0.75%)	98.63% (-0.94%)
ClearCLIP	54.47 (-0.57%)	63.75 (-2.67%)	1472.10 (-0.16%)	86.55 (-1.09%)	98.44% (-1.13%)
SigLIP2	**54.89** (+0.20%)	62.29 (-4.90%)	1384.09 (-6.13%)	83.09 (-5.04%)	95.63% (-3.94%)

Table 5: Performance with different numbers of Evidence Seed units on Qwen2.5-VL-7B. All results use SG-RLS with Evidence Source Layer 15. CLIP† uses negated image-text cosine similarity. Values in parentheses show changes relative to $K = 20$.

VLAM	Seed Units (K)	Retain	HBench	MMStar	POPE	Avg.
-	-	100%	66.50	62.90	86.50	100.00%
CLIP†	15	20%	60.99 (-0.68%)	**58.60** (+0.48%)	86.82 (-0.08%)	95.08% (-0.09%)
CLIP†	**20**	20%	**61.41**	58.32	**86.89**	**95.17%**
CLIP†	25	20%	61.09 (-0.52%)	58.14 (-0.31%)	86.83 (-0.07%)	94.89% (-0.29%)

Table 6: Inference efficiency on Qwen2.5-VL-7B with the input resolution fixed at 896×896. CLIP† uses negated image-text cosine similarity. E2E prefilling includes the runtime of the vision encoder and token selection. Values in parentheses show ratios relative to the full model.

Method	VLAM	FLOPs (T)	Memory (GB)	E2E Prefilling (ms)	Decoding (ms/token)
Total 1024 ***Tokens*** (100.0%)					
Qwen2.5-VL-7B		40.6	15.8	198.9	197.1
Retain 410 ***Tokens*** (40.0%)					
CDPruner	-	**11.6** (×0.29)	**15.7** (×0.99)	290.3 (×1.46)	**26.0** (×0.13)
SFPruner	-	11.7 (×0.29)	16.2 (×1.03)	**169.4** (×0.85)	**26.0** (×0.13)
QUEST	CLIP†	12.1 (×0.30)	17.0 (×1.08)	194.1 (×0.98)	**26.0** (×0.13)
Retain 205 ***Tokens*** (20.0%)					
CDPruner	-	**8.7** (×0.21)	**15.7** (×0.99)	219.8 (×1.11)	**24.0** (×0.12)
SFPruner	-	8.8 (×0.22)	16.2 (×1.03)	**153.5** (×0.77)	**24.0** (×0.12)
QUEST	CLIP†	9.2 (×0.23)	17.0 (×1.08)	183.7 (×0.92)	**24.0** (×0.12)

the target VLM. We use CLIP† as the default VLAM because it is simple to implement and allows comparison with existing query-aware baselines under the same VLAM setting.

Evidence Seed Count. We compare $K \in \{15, 20, 25\}$ to study the effect of the number of Evidence Seed units. Table 5 shows the highest average performance at $K = 20$. Performance decreases by 0.09 and 0.28 percentage points at $K = 15$ and $K = 25$, respectively. Therefore, we use $K = 20$ in all main experiments.

4.4 Inference Efficiency

Table 6 compares FLOPs, memory, E2E prefilling latency, and decoding latency on Qwen2.5-VL-7B under the same visual token budgets. We fix the input resolution to 896×896 for all experiments. QUEST uses an external CLIP† VLAM. Therefore, it requires slightly more FLOPs and memory than CDPruner and SFPruner. Its end-to-end prefilling latency is lower than CDPruner but higher than SFPruner. Decoding latency is the same across all three methods.

5 Concluding Remarks

We propose QUEST, a training-free Pre-LLM method for visual token selection. QUEST identifies query-relevant Evidence Seed units using an external vision-language alignment model (VLAM) and expands the related visual evidence using internal attention from the target VLM's vision encoder. It also selects an Evidence Source Layer based on layer-wise attention characteristics. QUEST then combines Evidence Attention with Visual Guidance and applies diversity-aware selection to choose the final visual tokens. Across different VLMs, VLAMs, benchmarks, and token budgets, QUEST shows high performance retention while reducing the number of visual tokens. The results suggest that external vision-language alignment can provide initial query evidence, which can then be expanded using the target VLM's internal attention for query-aware visual token selection.

AI USE STATEMENT

We used generative AI tools to assist with grammar correction and LaTeX formatting. We reviewed and revised all AI-assisted content. We independently verified all numerical results, technical claims, and citations. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only publicly available models and benchmarks and does not involve human subjects or the collection of personal data.

REPRODUCIBILITY STATEMENT

We describe the proposed method, implementation details, experimental settings, hyperparameters, and ablation configurations in the main paper and appendices. We report the evaluated models, benchmarks, token budgets, and source-layer configurations needed to reproduce the experiments. We will make the source code publicly available upon publication.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. HiRED: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

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

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024b. URL https://arxiv.org/pdf/2403.20330.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

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 *Advances in Neural Information Processing Systems*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.pdf.

Jiawei Guo, Feifei Zhai, Pu Jian, Qianrun Wei, and Yu Zhou. CROP: Contextual region-oriented visual token pruning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Chaoya Jiang, Haiyang Xu, Chenliang Li, Ming Yan, Wei Ye, Shikun Zhang, Bin Bi, and Songfang Huang. TRIPS: Efficient vision-and-language pre-training with text-relevant image patch selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4084–4096, 2022. URL https://aclanthology.org/2022.emnlp-main.273/.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pp. 235–251, 2016. URL https://arxiv.org/pdf/1603.07396.

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*, 2026.

Mincheol Kwon, Minseung Lee, Seonga Choi, Miso Choi, Kyeong-Jin Oh, Hyunyoung Lee, Cheonyoung Park, Yongho Song, Seunghyun Park, and Jinkyu Kim. Focus, don't prune: Identifying instruction-relevant regions for information-rich image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 126–142, 2024. URL https://arxiv.org/pdf/2407.12442.

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, 2024. URL https://arxiv.org/abs/2408.03326. arXiv preprint arXiv:2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023a.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal LLM. *International Journal of Computer Vision*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jinming Liu, Zhaoyang Jia, Jiahao Li, Bin Li, Xin Jin, Wenjun Zeng, and Yan Lu. When MLLMs meet compression distortion: A coding paradigm tailored to MLLMs. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2509.24258.

Jizhihui Liu, Guangdao Zhu, Feiyi Du, Niu Lian, Jun Li, Bin Chen, Weili Guan, and Yaowei Wang. HiPrune: Hierarchical attention for efficient token pruning in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3274–3291. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.162. URL https://aclanthology.org/2026.findings-acl.162/.

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 *European Conference on Computer Vision*, 2024c.

Ziniu Liu, Shuheng Zhou, Mingqing Liu, Hao Deng, and Huijia Zhu. CrisPrune: Combining contextual relevance and intrinsic saliency for efficient visual token pruning in MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13546–13564. Association for Computational Linguistics, 2026c. doi: 10.18653/v1/2026.findings-acl.663. URL https://aclanthology.org/2026.findings-acl.663/.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Kexin Ma, Jing Xiao, Chaofeng Chen, Geyong Min, Guibo Zhu, Jinqiao Wang, and Liang Liao. Decoupled similarity for task-aware token pruning in large vision-language models. In *Proceedings of the ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2604.11240. Accepted at ACM Multimedia 2026.

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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jouwon Song, Woohyeong Kim, and Kyeongbo Kong. SFPruner: Structured redundancy modeling for efficient visual token pruning in high-resolution MLLMs. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.23046. Accepted to ECCV 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, 2024.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

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

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2506.10967.

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

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. AIM: Adaptive inference of multi-modal LLMs via token merging and pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479. arXiv preprint arXiv:2504.10479.

A SG-RLS & DIRECTIONAL MASKING

Semantics-Guided Ridge Leverage Score. The Fused Guidance $\hat{G}$ obtained in Section 3.3 captures both query relevance and visual saliency. However, similar tokens can receive high guidance scores simultaneously, leading to redundant selections. To reduce this redundancy, QUEST adopts the Semantics-Guided Ridge Leverage Score (SG-RLS) proposed in SFPruner (Song et al., 2026). Let $\bar{V}_i$ denote the L_2-normalized feature of $\hat{V}_i$ defined in Eq. 2. We stack these features row-wise to form $\bar{V} \in \mathbb{R}^{\hat{N}_v \times d\mathrm{LLM}}$. Based on this representation, we then compute the Gram matrix C of the visual token features and the Ridge Leverage Score (RLS) ℓ_i for each token as follows:

$$\begin{aligned} C &= \bar{V}^\top \bar{V} \in \mathbb{R}^{d_{\mathrm{LLM}} \times d_{\mathrm{LLM}}}, \\ \ell_i &= \bar{V}_i \left(C + \lambda I_{d_{\mathrm{LLM}}}\right)^{-1} \bar{V}_i^\top, \qquad i = 1, \ldots, \hat{N}_v. \end{aligned} \tag{12}$$

Here, $\lambda > 0$ is a regularization constant, and the RLS assigns high scores to relatively distinctive features; in contrast, it assigns low scores to redundant tokens in the feature space. QUEST calculates the SG-RLS score S_i by combining the RLS and Fused Guidance $\hat{G}$:

$$S_i = \ell_i \hat{G}_i, \qquad i = 1, \ldots, \hat{N}_v, \quad S \in \mathbb{R}^{\hat{N}_v}. \tag{13}$$

Thus, a high S_i indicates that visual token i has both a high Fused Guidance score and high feature-space distinctiveness.

Ranking-Based Directional Masking. SG-RLS reduces global redundancy, but redundancy remains between highly similar tokens. To reduce similarity among these tokens, we adapt Directional Masking, which leverages cosine similarities between visual tokens and the SG-RLS ranking. $\bar{V}$ is L_2 normalized, so cosine similarity matrix C_{sim} can be calculated as follows.

$$C_{\mathrm{sim}} = \bar{V}\bar{V}^\top \in \mathbb{R}^{\hat{N}_v \times \hat{N}_v}. \tag{14}$$

Using C_{sim}, we define the ranking mask M so that only tokens $\hat{V}_j$ with higher SG-RLS scores can suppress $\hat{V}_i$. We then compute the diversity factor m_i and the final score.

$$\begin{aligned} M_{ij} &= \mathbb{I}(S_j > S_i), \qquad M_{ii} = 0, \\ m_i &= 1 - \max_j \left(M_{ij} C_{\mathrm{sim},ij}\right), \\ S_i^{\mathrm{final}} &= S_i m_i, \qquad i = 1, \ldots, \hat{N}_v. \end{aligned} \tag{15}$$

Since $M_{ii} = 0$, the max operation always includes 0. Given the range of cosine similarity, this ensures $0 \le m_i \le 1$. In addition, tokens with the same SG-RLS score do not suppress each other. Finally, we select the indices of the Top-B visual tokens according to S^{final}.

$$\hat{P} = \underset{P \subseteq \{1,\ldots,\hat{N}_v\}, |P|=B}{\arg\max} \sum_{i \in P} S_i^{\mathrm{final}}. \tag{16}$$

Here, $\hat{P}$ denotes the visual token index set selected by QUEST. The entire selection process uses matrix operations, enabling diversity-aware visual token selection without iterative greedy algorithms.

B DATASET

B.1 GENERAL VISUAL UNDERSTANDING AND REASONING

GQA. GQA is a VQA dataset for compositional visual reasoning over real-world images. Built from Visual Genome scene graphs, it evaluates multi-step reasoning over object attributes, spatial locations, and inter-object relationships.

MME. MME evaluates both perception and cognition across 14 subtasks, including object recognition, counting, position, OCR, commonsense reasoning, numerical calculation, translation, and code reasoning. It covers a broad range of capabilities from basic visual perception to high-level multimodal reasoning.

MMBench^{En}. MMBench^{En} contains approximately 3,000 multiple-choice questions covering coarse and fine-grained perception, attribute and relation reasoning, and logical reasoning.

B.2 OCR and Text-Centric Visual Understanding

TextVQA. TextVQA evaluates the ability to recognize and reason about text embedded in natural images. It contains 28,408 images and 45,336 questions requiring joint understanding of OCR content, surrounding visual context, and natural language queries.

B.3 Hallucination and Visual Grounding

POPE. POPE evaluates object hallucination by asking whether specific objects are present in an image. It measures whether model responses are grounded in visual evidence rather than language priors.

HallusionBench. HallusionBench evaluates hallucination and visual illusion using 346 images and 1,129 questions. It focuses on cases where visual evidence may conflict with language priors.

B.4 Scientific and Diagram Reasoning

AI2D. AI2D evaluates understanding and reasoning over scientific diagrams. It contains more than 5,000 diagrams and about 15,000 multiple-choice questions that require joint interpretation of graphical elements, labels, and their relationships.

ScienceQA^{img}. ScienceQA is a multimodal science question answering dataset spanning natural science, social science, and language science. We use its single-image subset to evaluate reasoning that combines visual information, scientific knowledge, and language understanding.

B.5 Comprehensive Visually Grounded Reasoning

MMStar. MMStar contains 1,500 visually dependent samples and evaluates six capabilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics. It emphasizes reasoning grounded in visual evidence rather than language priors.

C Implementation Details

C.1 QUEST Adaptation

LLaVA-NeXT. LLaVA-NeXT represents an AnyRes input using one base tile and multiple spatial tiles that capture local regions. Since the number of spatial tiles varies with the aspect ratio of the input image I, we fix the spatial layout to 2×2 for compatibility with the Vision-Language Alignment Model (VLAM). Each spatial tile contains $24 \times 24 = 576$ patches, yielding a merged 48×48 patch grid over I. Because LLaVA-NeXT uses the same CLIP vision encoder as the VLAM, we do not compute Evidence Attention and Visual Guidance independently for each spatial tile. Instead, we encode the entire image I with the VLAM to obtain $G^E_{\mathrm{global}} \in \mathbb{R}^{576}$ from Eq. 5 and $G^V_{\mathrm{global}} \in \mathbb{R}^{576}$ from Eq. 8, and then map them onto the merged spatial patch grid as follows.

$$\begin{aligned} G^E_{\mathrm{spatial}} &= \mathrm{Flatten}(\mathrm{Interp}(\mathrm{Reshape}(G^E_{\mathrm{global}}, 24, 24), 48, 48)) \in \mathbb{R}^{2304}, \\ G^V_{\mathrm{spatial}} &= \mathrm{Flatten}(\mathrm{Interp}(\mathrm{Reshape}(G^V_{\mathrm{global}}, 24, 24), 48, 48)) \in \mathbb{R}^{2304} \end{aligned} \tag{17}$$

The resulting G^E_{spatial} and G^V_{spatial}, aligned with each patch in the spatial tiles, are combined using the balance factor $\alpha \in [0, 1]$ as in Eq. 10.

$$G = \alpha\, \mathrm{MinMax}(G^E_{\mathrm{spatial}}) + (1 - \alpha)\, \mathrm{MinMax}(G^V_{\mathrm{spatial}}) \tag{18}$$

Qwen2.5-VL. QUEST resizes the input image I to the resolution required by the VLAM and computes the patch-level text relevance map R. When applying QUEST to Qwen2.5-VL, the vision-language connector includes the spatial merging operation. Specifically, Qwen2.5-VL merges each 2×2 group of neighboring patches from the vision encoder into a single visual token before feeding them to the large language model (LLM). Directly interpolating R from Eq. 3 to the patch grid V in Eq. 1, followed by 2×2 average pooling to assign scores to $\hat{V}$ in Eq. 2, can reduce score differences across patches due to the smoothing effects of interpolation and averaging. To preserve the relative relevance distribution, we first normalize R using $\mathrm{Softmax}(\tau R)$ with scaling factor τ, then interpolate it to the grid of V and apply 2×2 average pooling as in the spatial merging operation. Accordingly, the text relevance R^{merge} corresponding to $\hat{V}$ is defined as follows:

$$R^{\mathrm{merge}} = \mathrm{AvgPool}_{2\times2}\left(\mathrm{Interp}\left(\mathrm{Softmax}(\tau R), M_r, M_c\right)\right). \tag{19}$$

We select the Top-K merged-token units with the highest relevance scores from R_{merge} as Evidence Seed units. Each selected unit corresponds to a 2×2 group of pre-merging vision encoder tokens, which are used as Evidence Seeds. Thus, $K = 20$ Evidence Seed units correspond to $|E| = 80$ Evidence Seed tokens. The aggregated G^E and G^V are then processed separately by 2×2 average pooling followed by MinMax scaling, as in Eq. 9. Finally, they are combined using the balance factor α to obtain the Fused Guidance $\hat{G}$.

C.2 Vision-Language Alignment Model

CLIP-Inverse (CLIP†). As noted in MaskCLIP (Zhou et al., 2022), CLIP inherently exhibits patch-level vision-language alignment, but directly exploiting this property is not straightforward. Following the implementation of CDPruner (`https://github.com/Theia-4869/CDPruner`), we therefore invert the relevance map R in Eq. 3 as $R = -R$. For CLIP patch features, higher text relevance can correspond to lower cosine similarity, leading to a counterintuitive interpretation of similarity scores. QUEST adopts the same inversion strategy when using CLIP as the VLAM.

MaskCLIP. MaskCLIP (Zhou et al., 2022) observes that although CLIP is trained to learn image-level global representations, its internal visual features already contain localized semantic information. In standard CLIP, the final attention layer aggregates spatial features using query and key embeddings, which mixes information across patches. MaskCLIP removes the query and key projections from the final attention layer and directly uses the value representations of visual tokens. This preserves local semantics and produces patch-level features that can be compared with text embeddings for dense prediction without additional fine-tuning.

SCLIP. SCLIP (Wang et al., 2024) notes that standard query-key self-attention in CLIP is effective for global recognition but may not preserve spatial consistency well enough for dense prediction. To address this, SCLIP introduces Correlative Self-Attention, which replaces query-key cross-similarity with query-query and key-key similarity. By measuring similarities between visual tokens within the same representation space, it promotes more consistent connections among semantically related regions and improves spatial coherence for dense prediction. This modification can also be applied to pretrained CLIP without additional training.

ClearCLIP. ClearCLIP (Lan et al., 2024) analyzes that the degradation of CLIP's dense representations is caused not only by attention but also by the residual connection and Feed-Forward Network (FFN). In particular, image-text contrastive training emphasizes global representations, which can reduce local discriminability, while the residual connection propagates this global bias to dense features. ClearCLIP therefore removes the residual connection and FFN in the final transformer layer and applies self-self attention. This reduces global bias and produces more discriminative, text-aligned local visual representations.

SigLIP2. SigLIP2 is a vision-language alignment model pretrained with a sigmoid-based contrastive objective instead of the softmax-based contrastive learning used in CLIP. It provides strong image-text alignment and has been widely adopted across various vision-language tasks. To use SigLIP2 as a VLAM, we follow MaskCLIP (Zhou et al., 2022) and remove the attention pooling

Algorithm 1 Text Relevance Computation with MaskCLIP and SCLIP in QUEST

Input: Visual features x, text T, CLIP vision encoder E_{v}, text encoder E_{t}, visual projection P_{v}
Output: Text relevance map $R \in \mathbb{R}^{N \times N}$

text_feat $\leftarrow \mathrm{F.normalize}(E_{\mathrm{t}}(T), \mathrm{dim} = -1) \in \mathbb{R}^{d_{\mathrm{VLAM}}}$
hidden $\leftarrow [x_{\mathrm{cls}}; x] \in \mathbb{R}^{(1+N^2) \times d_{\mathrm{VLAM}}}$
for layer $\in E_{\mathrm{v}}$.encoder.layers$[:-1]$ **do**
    hidden $\leftarrow$ layer(hidden)
end for
last_layer $\leftarrow E_{\mathrm{v}}$.encoder.layers$[-1]$
normed $\leftarrow$ last_layer.layer_norm1(hidden)
if $M = \mathrm{MaskCLIP}$ **then**
    $v \leftarrow$ last_layer.self_attn.v_proj(normed)
    hidden $\leftarrow$ last_layer.self_attn.out_proj(v)
else if $M = \mathrm{SCLIP}$ **then**
    $q, k, v \leftarrow$ last_layer.self_attn(normed)
    $A_q \leftarrow \mathrm{softmax}(qq^{\top} \cdot \mathrm{scale})$
    $A_k \leftarrow \mathrm{softmax}(kk^{\top} \cdot \mathrm{scale})$
    $A \leftarrow A_q + A_k$
    hidden $\leftarrow Av$
    hidden $\leftarrow$ last_layer.self_attn.out_proj(hidden)
end if
% Residual connection and FFN are omitted.
visual_feats $\leftarrow P_{\mathrm{v}}$(last_layer.layer_norm_post(hidden))
visual_feats $\leftarrow \mathrm{F.normalize}(\mathrm{visual_feats}, \mathrm{dim} = -1)$
% w/o [CLS]
$R \leftarrow \mathrm{CosSim}(\mathrm{visual_feats}[1:,:], \mathrm{text_feat}).\,\mathrm{reshape}(N, N)$
return R

applied after the final layer of the vision encoder. We directly compare the resulting patch-level features $h_i^{(L)}$ with the text embedding t using cosine similarity. We then compute the query relevance score using the learned logit scale s and logit bias b of SigLIP2 as $R = \mathrm{Sigmoid}(s\,\mathrm{Sim}(h_i^{(L)}, t) + b)$.

Our Modifications. When initially evaluating different VLAMs, we visualized the text relevance heatmaps produced by MaskCLIP and SCLIP and found that they did not clearly distinguish query-relevant patches from irrelevant ones. Following the analysis in ClearCLIP (Lan et al., 2024), We therefore remove the residual connection and FFN from the final layer of MaskCLIP and SCLIP before using them as VLAMs. Algorithm 1 describes the resulting procedure.

C.3 Alternative Attention Distance

In the main method, we use $D_{\max}$ to select the Evidence Source Layer. As an alternative metric that reflects the overall spatial range of the attention distribution, we use the average attention distance D_{avg} (Liu et al., 2026a), defined as follows:

$$D_{\mathrm{avg}}^{(\ell)} = \frac{1}{N_v} \sum_{i=1}^{N_v} \sum_{j=1}^{N_v} \bar{A}_{ij}^{(\ell)}\, d(V_i, V_j). \tag{20}$$

While $D_{\max}$ measures the spatial distance from each query token to the token receiving the strongest attention, D_{avg} measures the spatial interaction range using the full attention distribution.

D Additional Experiments

D.1 Evidence Attention Aggregation

This section analyzes the Evidence Attention aggregation for different Vision-Language Alignment Models (VLAMs).

Table 7: Evidence Attention aggregation ablation in Qwen2.5-VL-7B. We compare different aggregation strategies for Evidence Attention in Section 3.2, with the Source Index fixed to 15 and the Evidence Seed Unit count fixed to 20. Colored values indicate the change in performance retention relative to max aggregation for the same VLAM.

VLAM	Agg	Retain	POPE	MMBench	MMStar	Avg.
Qwen2.5-VL-7B	-	100%	86.51	82.99	62.90	100.00%
CLIP†	max	20%	86.89	80.33	58.32	**96.65%**
CLIP†	mean	20%	**87.12** (+0.26%)	79.98 (-0.44%)	57.60 (-1.23%)	96.22% (-0.44%)
MaskCLIP	max	20%	87.04	**80.76**	56.95	96.15%
MaskCLIP	mean	20%	**87.26** (+0.25%)	79.81 (-1.18%)	**57.84** (+1.56%)	**96.33%** (+0.19%)
SCLIP	max	20%	**87.46**	80.41	**57.80**	**96.63%**
SCLIP	mean	20%	86.68 (-0.89%)	**80.58** (+0.21%)	57.33 (-0.81%)	96.15% (-0.50%)
ClearCLIP	max	20%	86.44	**80.33**	**57.27**	**95.92%**
ClearCLIP	mean	20%	**86.48** (+0.05%)	80.15 (-0.22%)	56.18 (-1.90%)	95.29% (-0.66%)
SigLIP2	max	20%	**82.99**	79.47	**56.97**	**94.09%**
SigLIP2	mean	20%	82.96 (-0.04%)	**80.33** (+1.08%)	54.59 (-4.18%)	93.16% (-0.99%)

Table 8: Ablation of balance factor in Qwen2.5-VL-7B. The Source Index is fixed to 15, and the Evidence Seed Unit count is fixed to 20. Colored values indicate the performance change relative to $\alpha = 0.5$ within the same VLAM.

VLAM	α	MMBench	POPE	MMStar	Avg.
CLIP†	0.4	80.41 (+0.10%)	86.77 (-0.14%)	57.47 (-1.46%)	96.18%
CLIP†	0.5	80.33	86.89	58.32	96.65%
CLIP†	0.6	80.67 (+0.42%)	86.91 (+0.02%)	58.25 (-0.12%)	**96.75%**
SCLIP	0.4	80.41 (-0.43%)	86.78 (-0.30%)	56.30 (-2.61%)	95.57%
SCLIP	0.5	80.76	87.04	57.81	**96.61%**
SCLIP	0.6	80.76	86.62 (-0.48%)	57.32 (-0.85%)	96.18%
MaskCLIP	0.4	80.41	87.26 (-0.23%)	57.24 (+0.14%)	96.25%
MaskCLIP	0.5	80.41	87.46	57.16	**96.28%**
MaskCLIP	0.6	80.41	87.30 (-0.18%)	56.95 (-0.37%)	96.11%
ClearCLIP	0.4	80.58 (+0.31%)	86.46 (+0.02%)	56.55 (-1.26%)	95.65%
ClearCLIP	0.5	80.33	86.44	57.27	**95.92%**
ClearCLIP	0.6	80.50 (+0.21%)	86.15 (-0.34%)	56.47 (-1.40%)	95.45%

$$G_j^E = \frac{1}{|E|} \sum_{i \in E} \left(\frac{1}{H_\ell} \sum_{h=1}^{H_\ell} A_{V,hij}^{(\ell)} \right), \qquad j = 1, \ldots, N_v. \tag{21}$$

As shown in Eq. 21, mean aggregation averages the attention responses over all Evidence Seeds, instead of selecting the maximum response as in Eq. 5. Table 7 reports the comparison between max aggregation and mean aggregation. When MaskCLIP is used, mean aggregation improves the overall performance by 0.18% percentage points. However, performance decreases in most other cases. This suggests that mean aggregation may dilute informative responses associated with individual Evidence Seeds.

D.2 Balance Factor Ablation

This section analyzes the effect of the balance factor α, which is used to combine $\tilde{G}_E$ and $\tilde{G}_V$ in Eq. 10.

The performance variations for different VLAMs are reported in Table 8. We evaluate CLIP†, SCLIP, MaskCLIP, and ClearCLIP with α values of 0.4, 0.5, and 0.6. The results show that $\alpha = 0.5$ achieves

the best performance for all VLAMs except CLIP†. Therefore, we set $\alpha = 0.5$ as the final hyperparameter value, as it provides the most consistent performance across different VLAMs.

E VISUALIZATIONS

E.1 LAYER-WISE EVIDENCE ATTENTION

The following visualizations compare Evidence Attention across different layers of the Qwen2.5-VL and CLIP vision encoders, together with the Visual Guidance obtained from the final layer. Figures 4 and 5 show the layer-wise Evidence Attention in the Qwen2.5-VL vision encoder, while Figures 6 and 7 show the corresponding results for the CLIP vision encoder. Evidence Attention expands query-relevant evidence from the Evidence Seeds to related visual tokens through the internal attention of the target VLM. In the early layers, Evidence Attention tends to remain relatively local, whereas intermediate layers progressively incorporate information from the surrounding visual context. In contrast, deeper layers exhibit more broadly distributed attention, resulting in less distinctive attention patterns. These observations suggest that Evidence Attention captures information complementary to Visual Guidance, supporting our choice of an intermediate layer as the Evidence Source Layer.

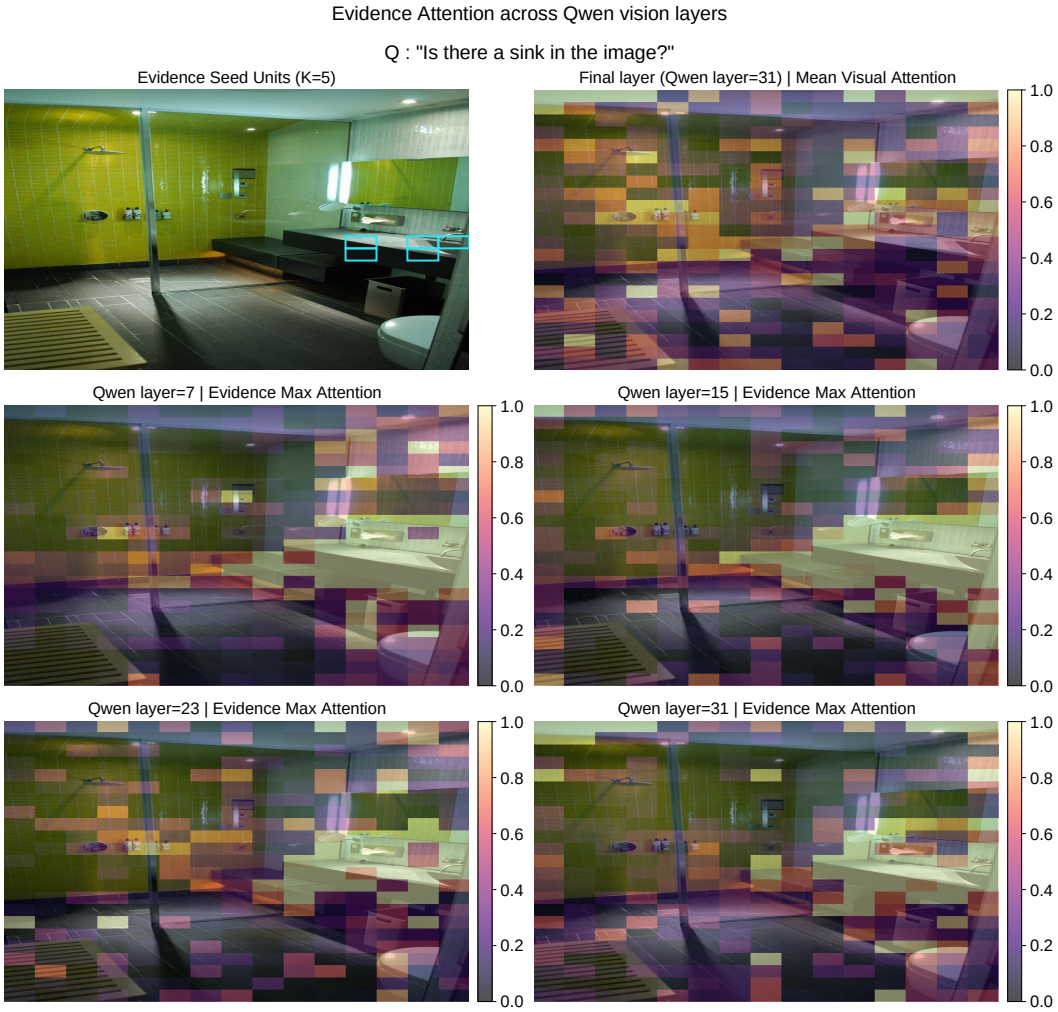

Figure 4: Layer-wise Evidence Attention example in the Qwen2.5-VL vision encoder (1).

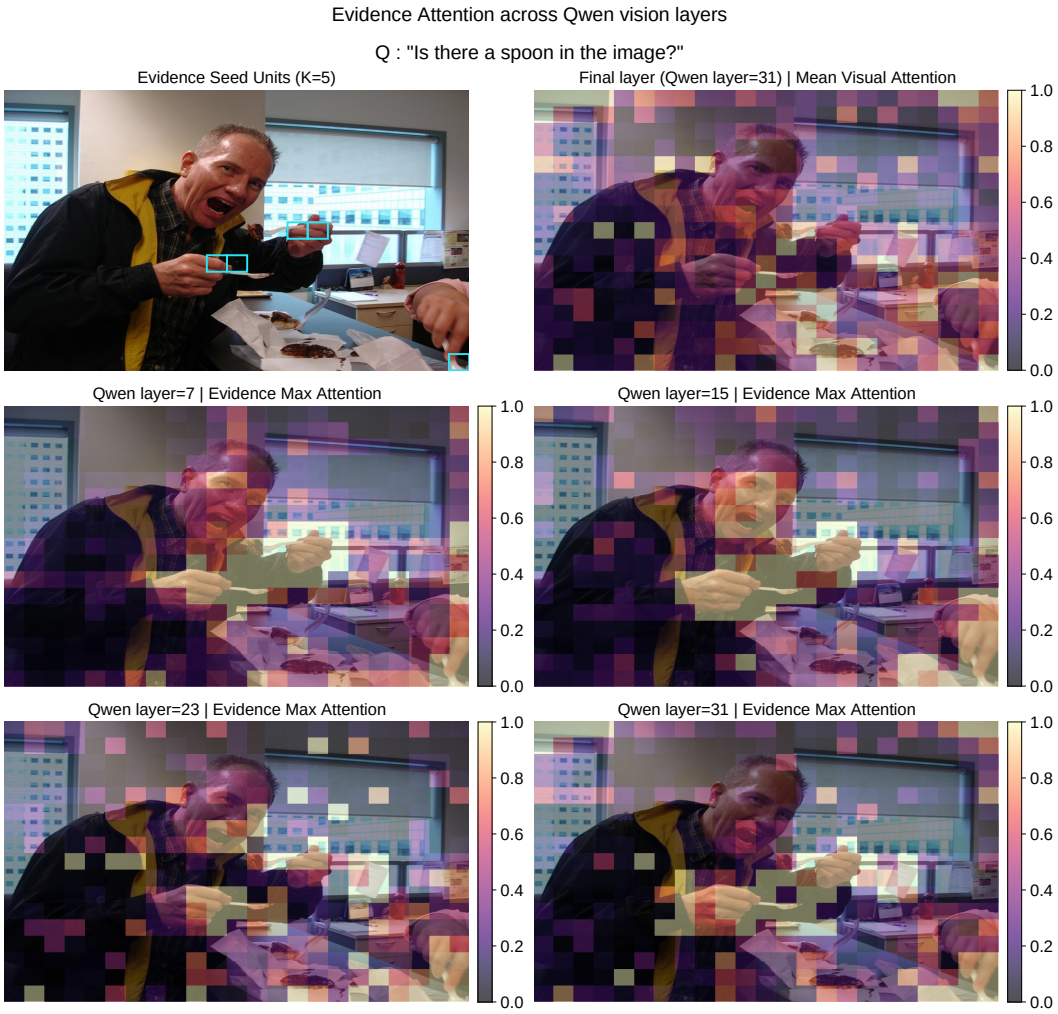

Figure 5: Layer-wise Evidence Attention example in the Qwen2.5-VL vision encoder (2).

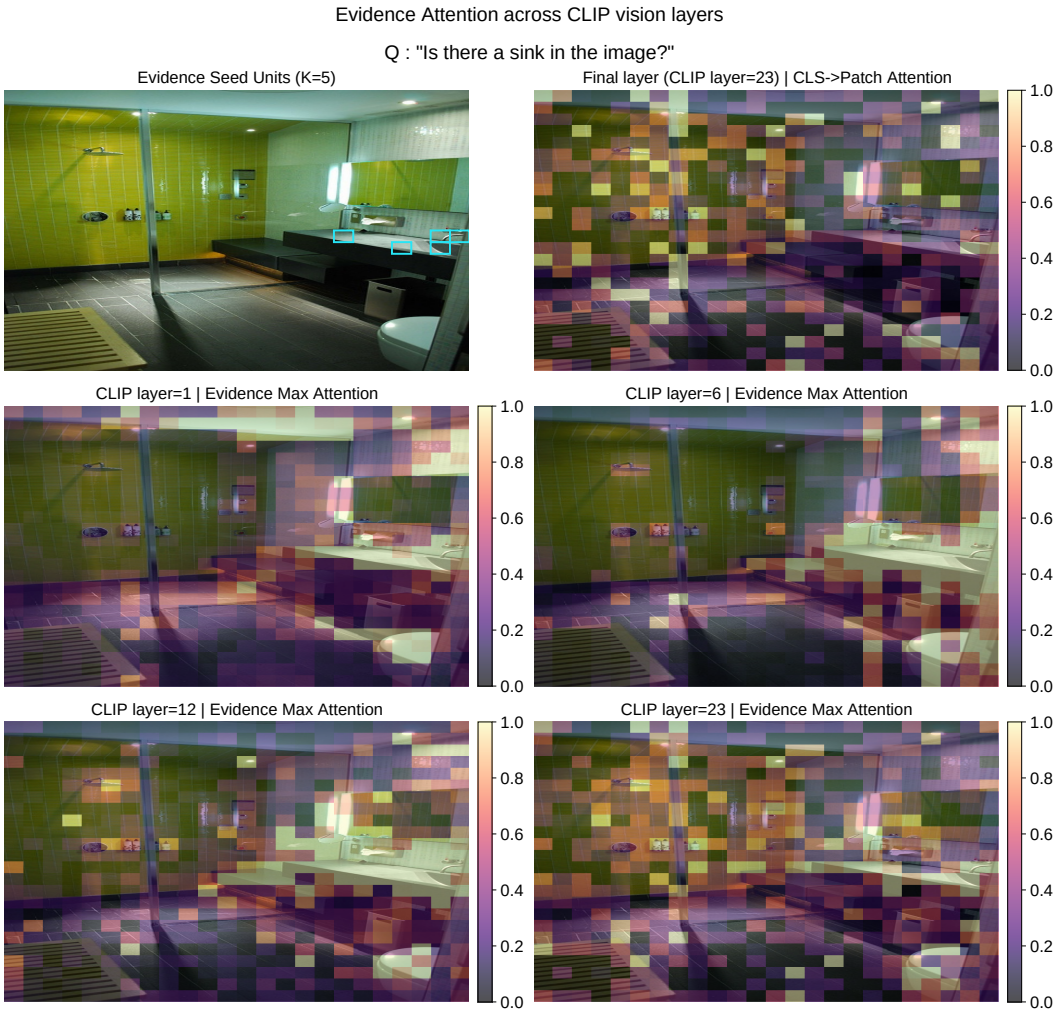

Figure 6: Layer-wise Evidence Attention example in the CLIP vision encoder (1).



Figure 7: Layer-wise Evidence Attention example in the CLIP vision encoder (2).

E.2 VISUAL TOKEN SELECTION RESULTS

The following visualizations compare two selection strategies. For visualization, the selected connector-output tokens are mapped back to the pre-merging vision-encoder token space. (1) selecting visual tokens based only on the Fused Guidance G in Eq. 10, which combines Evidence Attention and Visual Guidance without applying Semantics-Guided Ridge Leverage Score (SG-RLS), and (2) selecting vision encoder tokens with SG-RLS, which additionally incorporates diversity into the selection process. Figure 8 shows that QUEST can identify a substantial portion of the visual information relevant to answering the input query even without explicitly considering diversity through SG-RLS. Furthermore, Figure 9 demonstrates that the diversity introduced by SG-RLS helps reduce visual redundancy while preserving additional visually important regions that may otherwise be overlooked by the Fused Guidance G.

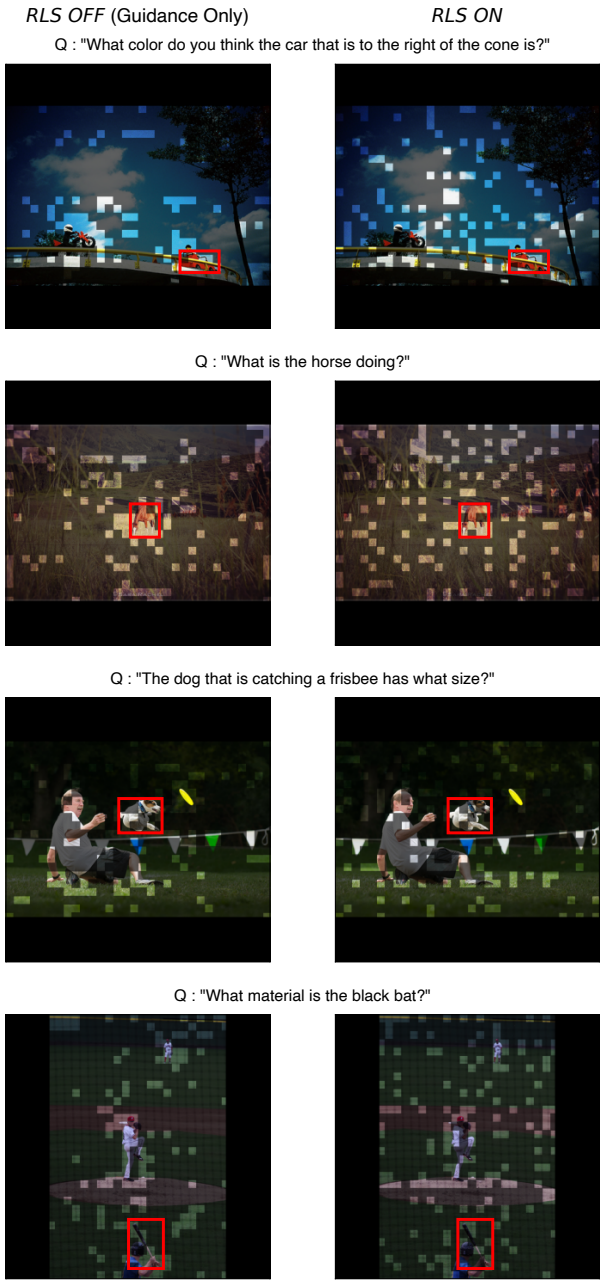

Figure 8: Comparison of visual token selection with and without SG-RLS (1).

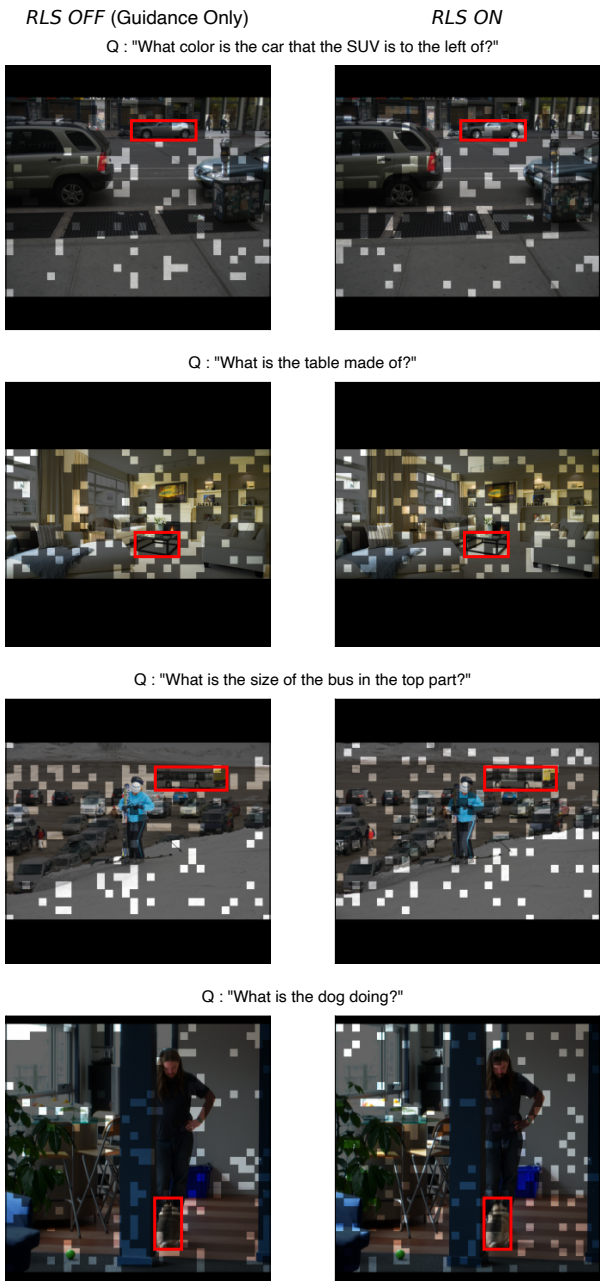

Figure 9: Comparison of visual token selection with and without SG-RLS (2).