READ: QUERY-RELEVANT EVIDENCE RETRIEVAL FOR VISUAL TOKEN COMPRESSION IN EDGE–CLOUD VLM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal foundation models enable edge–cloud collaborative intelligence by transmitting visual tokens from edge devices to a cloud vision–language model for downstream reasoning. However, uploading dense token sequences incurs substantial communication overhead and cloud-side computation and memory costs. Vision-only selectors operate before transmission but cannot adapt to specific user queries. Conversely, query-aware methods typically require executing heavy language or multimodal networks on resource-constrained edge devices, incurring substantial local overhead. We present **READ** (**R**etrieving **E**vidence via **A**ttention **D**istillation), a framework that reformulates pre-transmission token selection as an evidence retrieval task. During offline training, **READ** leverages prompt-to-visual attention from the frozen cloud model to supervise a compact cross-modal retriever that selects query-relevant visual tokens for accurate and efficient collaborative inference. By jointly training the retriever with prefix-ranking and score-calibration objectives, **READ** produces well-calibrated relevance scores that enable edge devices to identify task-critical visual tokens for efficient collaborative inference without running large models locally. Experiments demonstrate that **READ** achieves a competitive efficiency–accuracy trade-off across diverse architectures and transmission budgets, effectively preserving reasoning performance even under aggressive compression with minimal edge overhead.

1 INTRODUCTION

Edge–cloud visual intelligence is shifting from task-specific feature transmission toward unified multimodal foundation models (Bai et al., 2025; Liu et al., 2024a). In this collaborative setup, an edge device tokenizes raw visual inputs and transmits the resulting sequence to a cloud-based vision–language model (VLM) for downstream reasoning. By replacing task-specific edge pipelines with general-purpose representations, this paradigm allows distributed devices to access broad multimodal capabilities hosted in the cloud.

However, transmitting dense visual token sequences introduces two critical operational bottlenecks. On the uplink, transmitting hundreds of high-dimensional embeddings strains constrained network bandwidth. On the cloud, processing long token sequences leads to high prefill computation and large cache memory footprints (Shao et al., 2026). Because natural images contain substantial spatial redundancy, filtering visual tokens prior to transmission can mitigate both bottlenecks. While visual token compression can be achieved through token merging (Zhang et al., 2024a) or token selection (Yuan et al., 2026; Shao et al., 2026), selection is particularly suitable for edge–cloud systems. By directly dropping non-essential or redundant tokens, it reduces communication payload and cloud sequence length without requiring complex token synthesis operations on device.

Existing visual token selection methods generally follow two paradigms. Vision-driven methods evaluate token importance purely from image features without textual context, using training-free heuristics (Yang et al., 2025; Zhang et al., 2025; Alvar et al., 2025; Xu et al., 2026) or lightweight learned scorers (Zhu et al., 2025; Tan et al., 2026). Because these approaches only inspect visual features, they incur minimal edge overhead and reliably preserve prominent objects. However, they are query-agnostic and treat all questions identically. Consequently, they risk discarding subtle,

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.

localized cues, such as small text or fine-grained attributes, that are necessary for answering specific queries. Conversely, query-aware methods incorporate question semantics, but their deployment choices face practical limitations in edge–cloud settings. In-model pruners (Chen et al., 2024; Zhang et al., 2024b) drop tokens dynamically within cloud transformer layers, which reduces computation only after the full sequence has already traversed the transmission network. Proxy-guided methods such as SGP (Zhao et al., 2025) obtain token-importance guidance from an auxiliary small VLM. When adapted to pre-transmission edge selection, however, running the proxy VLM on device introduces substantial memory and latency overhead.

To address these limitations, we introduce READ (**R**etrieving **E**vidence via **A**ttention **D**istillation), a framework that reformulates pre-transmission visual token selection as an evidence retrieval task. Rather than running language-model layers on resource-constrained devices, READ decouples query-aware evidence scoring from online text reasoning: target prompt-to-image attention patterns from a frozen cloud VLM are cached offline and distilled into a compact cross-modal retriever. At inference time, the retriever deployed on the edge device fuses an intrinsic visual salience prior with prompt embeddings in a single forward pass, scoring query-relevant tokens prior to uplink transmission. To maintain reliable token rankings across varying bandwidth constraints without per-budget retraining, READ trains the retriever with dynamic prefix ranking and rank-survival calibration, producing well-calibrated relevance scores that integrate with a lightweight diversity step to avoid redundant selection. We evaluate READ across ten multimodal benchmarks on four representative VLM families: LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), Qwen2.5-VL (Bai et al., 2025), and InternVL2.5 (Chen et al., 2025). Across the evaluated compression budgets, READ achieves competitive accuracy retention against existing baselines while reducing uplink token transmission and lowering end-to-end inference latency.

Our main contributions are summarized as follows:

- We formulate edge–cloud visual token selection as an evidence retrieval task, decoupling query-aware evidence identification from online language-model execution.
- We propose READ, a framework that distills target VLM attention preferences into a compact cross-modal retriever, enabling query-conditioned token selection on edge devices prior to transmission.
- We develop a joint training strategy combining dynamic prefix ranking with rank-survival calibration, producing calibrated relevance scores suitable for diversity-aware coverage across variable budgets.
- We demonstrate across four VLM families and ten benchmarks that READ consistently preserves downstream reasoning accuracy under aggressive compression while reducing cloud memory footprint and computational overheads.

2 Related Work

2.1 Edge–Cloud Collaborative Multimodal Systems

Collaborative intelligence balances computation and communication by partitioning workloads between edge devices and cloud servers (Kang et al., 2017; Matsubara et al., 2022; Gao et al., 2021). In modern vision–language model (VLM) deployments, edge devices typically encode an input image into a sequence of patch tokens and transmit them to a cloud model for downstream reasoning (Bai et al., 2025; Liu et al., 2024a). Prior work on edge–cloud execution has explored hybrid small-and-large model setups (Hao et al., 2024; Jin & Wu, 2024), collaborative inference scheduling (Yuan et al., 2025), speculative decoding across tiers (Jia et al., 2026), input- or token-level routing (She et al., 2025; Šabanović et al., 2026), and semantic communication for vision–language tasks (Song & Kim, 2025; Hsu et al., 2026).

Most of these systems decide where and when to run each component, but treat the visual token sequence itself as fixed. As a result, the edge device still uploads the full sequence of dense patch embeddings over the uplink, and the cloud model must process that entire sequence during prefill. In contrast, READ compresses the visual token sequence directly on the edge prior to transmission.

2.2 Visual Token Compression for Efficient VLM Inference

Visual token compression reduces uplink transmission overhead, KV-cache footprints, and cloud prefill computation by pruning or merging redundant visual tokens before cloud-side execution. Existing

approaches broadly fall into two categories depending on whether token selection is conditioned on the user query.

Query-Agnostic Token Compression. Query-agnostic methods score visual tokens using only image representations, without access to the text prompt. One line of work uses training-free heuristics to prune or merge tokens, relying on feature similarity (Yang et al., 2025; Zhang et al., 2025), encoder self-attention (Shang et al., 2025; Arif et al., 2024), token diversity (Alvar et al., 2025), or salience–coverage trade-offs (Xu et al., 2026). Another line trains lightweight heads to predict patch importance directly from visual features (Zhu et al., 2025; Tan et al., 2026). Because these selectors inspect only visual features, they incur minimal edge overhead and reliably preserve dominant objects. However, being query-agnostic, they risk discarding non-salient or subtle visual cues (such as small text or fine-grained attributes) that are critical for answering specific questions. Trainable patch scorers supervised by generation loss also tend to lack a calibrated score ranking across different token budgets.

Query-Conditioned Token Compression. Query-conditioned methods incorporate the prompt to determine which tokens to retain, yet their deployment strategies introduce practical trade-offs in edge–cloud serving. In-model pruners such as FastV (Chen et al., 2024) and SparseVLM (Zhang et al., 2024b) drop tokens dynamically based on intermediate attention within the cloud model during prefill. While reducing computation in deeper layers, they still require the full token sequence to be uploaded first, offering no reduction in uplink bandwidth or initial cloud prefill. Alternatively, auxiliary schemes such as SGP (Zhao et al., 2025) run a smaller proxy VLM on the edge to identify relevant tokens before transmission. However, executing an autoregressive model on device introduces substantial latency and memory overheads, and capacity discrepancies between the proxy and target models can lead to mismatched token selection. READ avoids running an online language model on the edge by distilling the cloud VLM's prompt-to-image attention offline into a compact retriever, enabling the edge device to select query-relevant tokens prior to transmission.

3 Problem Formulation

3.1 Edge–Cloud Visual Token Compression

Collaborative edge–cloud inference has emerged as a promising paradigm for deploying modern multimodal foundation models. In this setting, the edge device performs front-end tokenization: given an input image I and a textual query Q, an on-device vision encoder E extracts N visual tokens $\boldsymbol{V} = E(I) = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N] \in \mathbb{R}^{N \times d_v}$, while an embedding layer maps the query Q into a sequence of M prompt tokens. Rather than executing heavy multimodal reasoning locally, the edge device transmits these token representations to a cloud-hosted VLM. However, uploading the full, uncompressed visual sequence incurs a substantial N-token uplink transmission payload and an $\mathcal{O}((N+M)^2)$ cloud prefill attention complexity.

To alleviate these operational bottlenecks, the edge device must identify a compact subset of query-relevant evidence tokens $\mathcal{S}_K \subset [N]$ under a constrained budget $K \ll N$ *prior to* transmission, uploading only $\boldsymbol{V}_{\mathcal{S}_K}$. The optimal selection strategy maximizes the downstream response utility:

$$\mathcal{S}_K^* = \arg \max_{\mathcal{S} \subseteq [N],\, |\mathcal{S}|=K} U(\mathcal{S}; I, Q). \tag{1}$$

Because response utility U depends heavily on query semantics, we formulate pre-transmission token selection as an *information retrieval (IR)* problem: downstream reasoning requires isolating the most relevant visual evidence for the query, which emphasizes relative evidence *ranking* over absolute confidence estimation. Crucially, this ranking must be computed without access to cloud-side internal activations or execution of on-device language models, thereby preserving the low-latency advantage of edge offloading.

3.2 Target-Model Attention as Relevance Guidance

Internal prompt-to-visual attention in multimodal foundation models provides a reliable proxy for identifying task-relevant visual evidence (Chen et al., 2024; Zhao et al., 2025). During the prefill stage, cross-modal attention within self-attention layers traces information routing from the textual query

to visual inputs, highlighting the spatial regions the model attends to when interpreting the prompt. Specifically, at layer ℓ and head h, let $\boldsymbol{q}_u^{(\ell,h)}$ and $\boldsymbol{k}_j^{(\ell,h)}$ denote the query and key representations corresponding to prompt position $u \in \mathcal{P}$ and sequence position j, respectively. The normalized attention weight from prompt token u to visual position $i \in \mathcal{V}$ is given by:

$$A_{u,i}^{(\ell,h)} = \frac{\exp\left((\boldsymbol{q}_u^{(\ell,h)})^\top \boldsymbol{k}_i^{(\ell,h)}/\sqrt{d_h}\right)}{\sum_{j\in\mathcal{V}\cup\mathcal{P}_{\le u}} \exp\left((\boldsymbol{q}_u^{(\ell,h)})^\top \boldsymbol{k}_j^{(\ell,h)}/\sqrt{d_h}\right)}, \tag{2}$$

where d_h is the head dimension, and $\mathcal{P}_{\le u}$ denotes the causal context available at prompt position u. By aggregating these attention weights across all L_T layers and H_T heads, we quantify the target model's overall visual evidence preference:

$$a_i^T = \sum_{\ell=1}^{L_T}\sum_{h=1}^{H_T}\sum_{u\in\mathcal{P}} A_{u,i}^{(\ell,h)}, \qquad i \in \mathcal{V}. \tag{3}$$

By defining $\mathcal{P}$ exclusively over input prompt tokens during the prefill phase rather than autoregressively generated tokens, a_i^T naturally satisfies the pre-transmission causal constraint by capturing question-guided visual cues without answer leakage. Moreover, this design enables highly efficient offline caching via a single parallel forward pass, which completely bypasses costly autoregressive decoding.

While a_i^T offers a high-fidelity relevance signal, computing it online on the edge would require executing the full target VLM locally, incurring prohibitive memory and latency overheads. However, because this operational constraint applies only at inference time, the target model's attention can be pre-computed offline. Sorting a_i^T in descending order yields ordinal teacher ranks $r_i^T \in \{1, \dots, N\}$, providing a scale-invariant supervision signal. This decouples complex language-model reasoning from online token selection, framing our objective as learning a lightweight edge retriever supervised by these offline target ranks.

4 READ: Query-Relevant Evidence Distillation

4.1 Overview

Building upon cached teacher attention, READ decouples heavy multimodal reasoning from online token scoring to enable lightweight on-device selection. As depicted in Figure 1, edge inference follows a streamlined three-step pipeline: (i) extract visual tokens $\boldsymbol{V}^{\text{in}}$, prompt embeddings $\boldsymbol{T}^{\text{in}}$, and an intrinsic visual prior $\widetilde{\boldsymbol{\pi}}$ using the local front-end; (ii) predict calibrated patch relevance scores via a compact cross-modal retriever in a single forward pass; and (iii) select a diverse subset of K tokens via relevance-weighted coverage for uplink transmission, saving bandwidth and cloud prefill compute without running on-device language models.

Conceptually, while vision-only methods capture dominant scene structures, they cannot adapt to specific prompts and often miss subtle, task-critical cues. To address this limitation, READ decomposes evidence scoring into an inherent visual prior that anchors dominant scene layout and a query-dependent adjustment from the retriever:

$$s_i = \pi_i \oplus \Delta_{\text{retriever}}(i \mid Q), \tag{4}$$

where π_i provides baseline visual salience, $\Delta_{\text{retriever}}(i \mid Q)$ is the prompt-guided evidence adjustment, and $\oplus$ denotes an adaptive logit combination. By operating directly on evidence ranks, this formulation reliably preserves dominant layout while prioritizing question-relevant fine-grained details under constrained token budgets.

4.2 Visual Salience Prior

While query conditioning is essential, inherent image salience provides a stable, query-independent anchor. Large objects and high-contrast regions often possess baseline relevance, and a strong prior prevents the retriever from drifting when the text prompt is uninformative. From a selected intermediate layer of the frozen vision transformer, we average self-attention across heads and extract

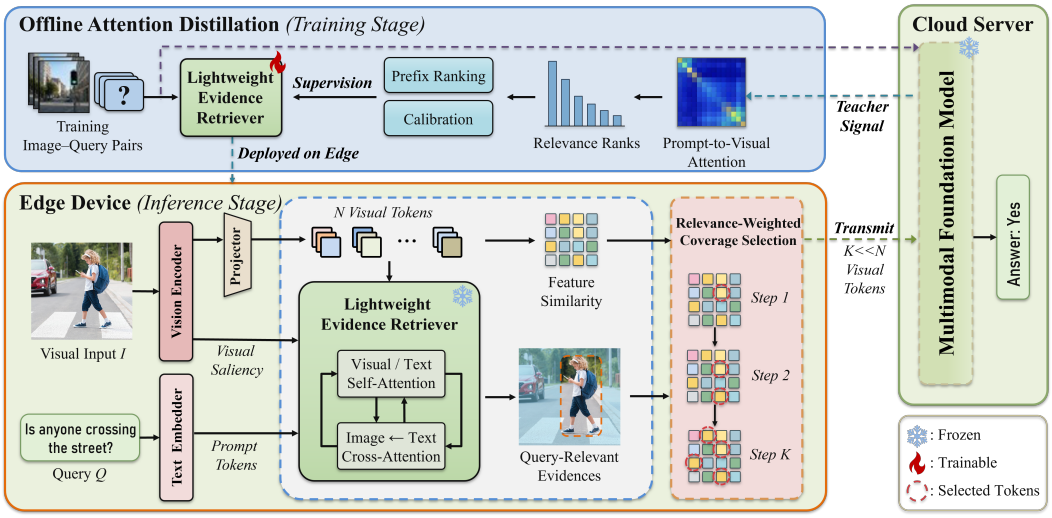


Figure 1: Overview of the READ framework. Offline, prompt-to-visual attention from the frozen cloud VLM is cached and distilled into a compact cross-modal retriever using dynamic prefix ranking and rank-survival calibration. At inference time, the edge device grounds token selection in an intrinsic visual salience prior, updates patch rankings with prompt evidence via the lightweight retriever, and selects a diverse subset of K tokens via relevance-weighted coverage. Transmitting only this compact sequence reduces uplink transmission overhead and cloud computation without running language models on the edge.

the [CLS]-to-patch attention scores. Let $r_i^C \in \{1, \ldots, N\}$ represent its descending one-based rank. Instead of min-max scaling the raw attention scores, which is sensitive to scale and sequence length, we map this rank to the same log-budget survival curve used as our distillation target in Section 4.4:

$$\pi_i = \frac{\log(N/r_i^C)}{\log N}, \qquad \widetilde{\pi}_i = \pi_i - \frac{1}{N}\sum_{j=1}^{N} \pi_j. \tag{5}$$

Under a retention cutoff drawn uniformly in $\log K$, π_i matches the exact analytical probability that patch i survives if selection followed r_i^C. Consequently, the prior is scale-free, adapts automatically to varying sequence lengths, and aligns naturally with our distillation objective. Centering the prior ($\widetilde{\pi}_i$) further removes sample-level additive bias, ensuring that the residual path in Equation 9 does not uniformly shift every logit.

4.3 Cross-Modal Evidence Retriever

To ensure the retriever is both query-aware and sufficiently lightweight for edge deployment, we design a compact cross-modal architecture rather than executing a heavy language-model probe. Specifically, visual tokens from the multimodal projector, $\boldsymbol{V}^{\text{in}} \in \mathbb{R}^{N\times D}$, and prompt token embeddings from the base model, $\boldsymbol{T}^{\text{in}} \in \mathbb{R}^{M\times D}$, are first projected into a shared hidden dimension d:

$$\boldsymbol{V}^{(0)} = W_v\,\mathrm{LN}(\boldsymbol{V}^{\text{in}}), \qquad \boldsymbol{T}^{(0)} = W_t\,\mathrm{LN}(\boldsymbol{T}^{\text{in}}). \tag{6}$$

A lightweight MLP f_π then embeds each scalar survival prior $\widetilde{\pi}_i$ into the hidden dimension d and injects it directly into the visual stream from the very first layer:

$$\boldsymbol{V}^{(0)} \leftarrow \boldsymbol{V}^{(0)} + f_\pi(\widetilde{\boldsymbol{\pi}}). \tag{7}$$

This early injection allows structural salience to condition all subsequent cross-modal feature refinements, rather than merely adjusting the final score.

The retriever then processes these features through a series of cross-modal blocks. Each block contextualizes the text prompt via self-attention, models spatial relationships among visual tokens,

and enables every visual token to actively query the full prompt via image-to-text cross-attention:

$$\widetilde{\boldsymbol{V}}^{(\ell)} = \widehat{\boldsymbol{V}}^{(\ell)} + \mathrm{MCA}_{v\leftarrow t}\big(\mathrm{LN}(\widehat{\boldsymbol{V}}^{(\ell)}), \mathrm{LN}(\boldsymbol{T}^{(\ell)})\big), \tag{8}$$

where $\widehat{\boldsymbol{V}}^{(\ell)}$ denotes intermediate visual features after spatial self-attention. This cross-attention direction ($v \leftarrow t$) enables each visual token to attend to the prompt tokens in a single forward pass, producing query-conditioned representations without autoregressive decoding. Finally, a scoring head projects the refined visual representations into a query-conditioned adjustment Δ_i, yielding the final residual-adjusted logit:

$$z_i = \Delta_i + \gamma\widetilde{\pi}_i, \qquad \gamma = \mathrm{softplus}(\beta) > 0, \tag{9}$$

where β is a learnable parameter. The direct residual connection maintains a short, stable gradient path from image salience, while the unconstrained adjustment Δ_i grants the retriever the capacity to elevate visually inconspicuous patches whenever the user query demands them.

4.4 Prefix-Ranking Distillation and Rank-Survival Calibration

Directly distilling the full attention distribution of a large VLM via conventional divergence losses, such as KL divergence, is impractical for a lightweight edge model, as matching diffuse, long-tailed probabilities across uninformative background patches squanders model capacity. Instead, we tailor our distillation objective directly to the goal of token compression: identifying task-critical visual tokens to answer the prompt. Because downstream selection depends strictly on relative token ordering rather than absolute probabilities, the retriever needs to learn accurate evidence *ranking* rather than full distribution reconstruction. By supervising relative candidate ordering, we concentrate model capacity on top candidate tokens to guarantee dependable ordinal rankings under varying budgets K.

To ensure consistent ranking quality across continuous compression scales without budget-specific retraining, we formulate a dynamic *prefix-ranking* objective by sampling the retention cutoff from a log-uniform distribution: $\log K \sim \mathcal{U}(\log 1, \log N)$. Unlike linear sampling which over-indexes on trivial large budgets, log-uniform sampling assigns equal probability mass across relative compression scales (e.g., $[16, 32]$ is sampled as often as $[64, 128]$). This concentrates capacity on the top prefix where evidence retrieval under tight bandwidth is most critical, while weakly regularizing tail tokens. Consequently, scoring is fully decoupled from any fixed budget, allowing a single trained checkpoint to generalize across diverse compression budgets. For a sampled budget K, a token inside the teacher's top-K prefix ($i^+ \sim \mathcal{U}\{i : r_i^T \le K\}$) must outscore an unselected token ($i^- \sim \mathcal{U}\{i : r_i^T > K\}$):

$$\mathcal{L}_{\mathrm{rank}} = \mathbb{E}_{K,i^+,i^-}\big[\log\big(1 + \exp[-(z_{i^+} - z_{i^-})]\big)\big]. \tag{10}$$

Although ranking establishes relative order, it leaves logit magnitudes unconstrained, which can distort downstream distance-based coverage selection (Section 4.5). We resolve this through *rank-survival calibration*, anchoring logits to the analytical survival probability under the same log-uniform budget law:

$$y_i = \Pr(K \ge r_i^T) = \frac{\log(N/r_i^T)}{\log N}. \tag{11}$$

Here, $y_i \in [0, 1]$ calibrates predicted probabilities to the operational likelihood that a token falls within the teacher's top-K prefix across variable bandwidths. The retriever is supervised via soft-label binary cross-entropy (Appendix A):

$$\mathcal{L}_{\mathrm{cal}} = -\tfrac{1}{N}\sum_{i=1}^{N}\big[y_i \log\sigma(z_i) + (1 - y_i)\log(1 - \sigma(z_i))\big]. \tag{12}$$

The overall training objective is $\mathcal{L}_{\mathrm{READ}} = \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}} + \lambda_{\mathrm{cal}}\mathcal{L}_{\mathrm{cal}}$. During edge inference, $s_i = \sigma(z_i) \in [0, 1]$ provides a rank-survival calibrated score directly compatible with geometric coverage.

4.5 Relevance-Weighted Coverage Selection

Selecting the top-K tokens based on s_i alone risks packing the limited communication budget onto adjacent, highly similar visual patches. Conversely, pure farthest-first coverage selection achieves

Table 1: LLaVA-1.5-7B visual token compression. Best Avg. and Rel. in each block are bold.

Method	VQAv2	GQA	VizWiz	SQA	TextVQA	POPE	MME	MMB	MMB-CN	MM-Vet	Avg.	Rel.
LLaVA-1.5-7B (576)	78.3	61.9	50.2	69.5	58.2	85.3	1514.8	64.3	58.2	29.7	63.1	100.0%
Retain 128 Tokens (↓ 77.8%)												
FastV	69.6	55.8	51.2	69.1	57.0	71.4	1437.4	63.2	56.9	27.2	59.3	94.0%
VisionZip	73.8	57.5	52.0	68.6	56.9	82.6	1439.7	61.4	56.5	31.3	61.3	97.1%
DivPrune	74.3	59.3	53.4	67.8	56.1	86.5	1381.6	60.4	55.1	30.0	61.2	97.0%
VisPruner	73.9	58.5	52.5	69.1	57.0	84.2	1408.6	61.1	56.8	31.9	61.5	97.5%
SparseVLM	75.2	58.5	50.1	69.2	56.6	83.1	1436.3	63.7	58.2	28.9	61.5	97.5%
VisionSelector	74.3	57.9	51.9	68.5	56.9	82.4	1415.4	61.9	57.0	31.5	61.3	97.1%
IDPruner	74.3	59.4	53.0	68.5	57.1	86.2	1398.7	60.7	56.1	32.2	61.7	97.8%
READ (Ours)	75.2	59.7	52.3	69.1	57.2	85.9	1440.5	62.2	56.5	32.6	**62.3**	**98.7%**
Retain 64 Tokens (↓ 88.9%)												
FastV	64.3	51.4	51.1	70.0	55.1	59.8	1238.9	60.1	52.9	25.7	55.2	87.5%
VisionZip	70.6	54.9	52.5	68.9	55.4	77.1	1373.8	59.5	55.3	30.2	59.3	94.0%
DivPrune	72.5	57.9	53.8	67.5	54.8	85.2	1351.8	58.6	51.5	29.1	59.9	94.9%
VisPruner	70.9	55.5	53.3	68.7	55.5	79.6	1353.5	59.1	54.9	29.6	59.5	94.3%
SparseVLM	68.9	53.5	50.1	69.6	53.3	71.6	1298.8	59.7	52.2	25.5	56.9	90.2%
VisionSelector	72.4	55.0	52.9	69.1	55.2	76.0	1347.8	60.5	55.0	29.4	59.3	94.0%
IDPruner	72.4	57.8	54.0	68.0	55.9	85.5	1366.2	59.3	52.7	27.5	60.1	95.2%
READ (Ours)	74.2	58.3	53.2	68.7	56.1	86.4	1450.6	61.3	55.5	29.3	**61.6**	**97.6%**
Retain 32 Tokens (↓ 94.4%)												
FastV	56.0	46.7	49.5	70.1	52.0	38.8	999.0	51.6	44.4	18.8	47.8	75.8%
VisionZip	65.6	51.8	52.9	68.9	53.2	67.9	1245.3	57.1	50.3	25.8	55.6	88.1%
DivPrune	69.4	54.8	53.3	68.8	53.3	82.2	1308.4	56.5	47.4	26.1	57.7	91.4%
VisPruner	66.1	52.3	52.5	68.2	53.8	74.7	1260.8	56.5	49.4	26.8	56.3	89.2%
SparseVLM	62.3	49.9	48.2	70.2	49.6	59.2	1150.0	57.0	46.2	21.2	52.1	82.6%
VisionSelector	69.6	51.7	53.1	69.0	54.0	68.4	1273.8	57.6	52.8	27.5	56.7	89.9%
IDPruner	69.6	55.1	53.7	68.8	54.2	82.0	1293.9	57.0	49.4	27.5	58.2	92.2%
READ (Ours)	72.5	56.7	53.6	68.2	55.3	86.6	1348.1	59.7	52.8	29.2	**60.2**	**95.4%**

broad diversity but frequently retains irrelevant background patches. To strike a balance, we combine both principles. Let $\bar{s}_i = s_i / \max_j s_j$ represent the normalized relevance, and let $\boldsymbol{h}_i$ be the ℓ_2-normalized frozen vision-encoder feature for patch i. We define the cosine distance $d(i,j)$ and the minimum distance (novelty) relative to the currently selected set $\mathcal{S}$ as:

$$d(i,j) = \max\big(0, 1 - \boldsymbol{h}_i^\top \boldsymbol{h}_j\big), \qquad \delta(i \mid \mathcal{S}) = \min_{j \in \mathcal{S}} d(i,j). \tag{13}$$

The marginal utility of adding token i to the selected subset is formulated as a score-weighted novelty:

$$G(i \mid \mathcal{S}) = \delta(i \mid \mathcal{S})\, \bar{s}_i^{\,\eta}, \tag{14}$$

where η governs the trade-off between query salience and feature coverage. The selection algorithm modifies classical greedy K-Center: it initializes with the most relevant token $\arg\max_i s_i$ as the seed, and iteratively appends the candidate $\arg\max_{i \notin \mathcal{S}} G(i \mid \mathcal{S})$ until $|\mathcal{S}| = K$. Finally, the selected tokens are restored to their original spatial order to ensure the cloud VLM receives a coherent structural subsequence.

5 Experiments

Experimental setup. We evaluate READ across four representative VLMs: LLaVA-1.5 (Liu et al., 2024a), LLaVA-NeXT (Liu et al., 2024b), Qwen2.5-VL (Bai et al., 2025), and InternVL2.5 (Chen et al., 2025). Baselines include FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024b), VisionZip (Yang et al., 2025), DivPrune (Alvar et al., 2025), VisPruner (Zhang et al., 2025), VisionSelector (Zhu et al., 2025), IDPruner (Tan et al., 2026), and SGP (Zhao et al., 2025). READ is trained on a composite subset of Cambrian-737K (Tong et al., 2024). Evaluation spans ten multimodal benchmarks: VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), POPE (Li et al., 2023), MME (Fu et al., 2023), MMBench and MMBench-CN (Liu et al., 2024c), and MM-Vet (Yu et al., 2023). Following the edge–cloud formulation in Section 3.1, READ selects visual tokens before transmission to the cloud. At a fixed embedding dimension and numerical precision, the retained token budget determines both the visual embedding payload and the visual sequence length processed by the cloud model. We therefore evaluate downstream accuracy across token budgets and separately profile the encoding-and-selection and target-model forward stages. Detailed implementation configurations and evaluation protocols are provided in Appendices A and B.

Table 2: Performance on different VLM architectures. Best Avg. and Rel. in each block are bold.

Method	GQA	SQA	TextVQA	POPE	MMB	MM-Vet	Avg.	Rel.
Qwen2.5-VL-7B-Instruct								
Full	60.9	87.7	75.5	85.5	56.0	65.1	71.8	100.0%
Keep 25% Visual Tokens (↓ 75%)								
FastV	54.6	83.6	76.2	81.3	53.1	51.5	66.7	92.9%
VisPruner	57.9	82.3	66.3	81.3	51.6	46.5	64.3	89.6%
SGP	56.8	82.9	73.2	76.5	53.1	46.9	64.9	90.4%
READ (Ours)	59.1	83.5	74.2	83.9	54.2	57.2	**68.7**	**95.7**%
Keep 10% Visual Tokens (↓ 90%)								
FastV	46.1	78.3	70.2	66.5	40.0	37.7	56.5	78.7%
VisPruner	53.5	78.5	53.4	74.3	47.9	40.4	58.0	80.8%
SGP	51.9	80.0	66.5	65.0	48.9	39.1	58.6	81.6%
READ (Ours)	55.4	81.3	70.7	80.2	49.5	44.6	**63.6**	**88.6**%
InternVL2.5-8B								
Full	63.7	97.9	75.5	89.7	76.0	61.0	77.3	100.0%
Keep 25% Visual Tokens (↓ 75%)								
FastV	59.1	93.3	68.2	85.3	75.1	50.3	71.9	93.0%
VisPruner	60.5	90.1	56.7	89.0	67.2	35.5	66.5	86.0%
SGP	63.3	97.0	73.9	90.1	74.5	50.1	**74.8**	**96.8**%
READ (Ours)	63.2	95.7	73.5	89.9	73.8	48.6	74.1	95.9%
Keep 10% Visual Tokens (↓ 90%)								
FastV	53.3	88.1	62.6	77.0	69.3	35.5	64.3	83.2%
VisPruner	52.3	83.8	47.9	83.0	54.4	26.1	57.9	74.9%
SGP	61.9	93.4	70.3	89.2	71.3	39.4	**70.9**	**91.7**%
READ (Ours)	61.5	91.2	67.0	89.5	69.7	39.8	69.8	90.3%

5.1 Main Results

Table 1 evaluates LLaVA-1.5-7B across ten multimodal benchmarks under budgets of 128, 64, and 32 tokens (22%, 11%, and 5.6% retention). READ consistently maintains top downstream accuracy, expanding its lead under stricter constraints. While query-agnostic baselines remain competitive at 128 tokens by preserving salient foreground patches, they degrade substantially at 64 and 32 tokens on fine-grained reasoning tasks (e.g., TextVQA and GQA). Without query guidance, they fall into the saliency trap by retaining prominent background clutter while discarding subtle task evidence. Meanwhile, in-model pruners (FastV, SparseVLM) rely on uncalibrated early-layer attention and still require full token transmission. In contrast, READ conditions selection directly on prompt embeddings, reliably isolating question-critical cues before uplink transmission.

Query-guided selection is equally critical for mitigating object hallucination under heavy compression. On POPE, heuristic pruners degrade sharply because indiscriminate pruning removes necessary spatial grounding, while learning-based scalar scorers struggle with uncalibrated score distributions. READ preserves POPE accuracy robustly, reaching 86.6% at 32 retained tokens, which closely matches the 85.3% accuracy of the uncompressed model. This stability stems from our complementary training objectives: dynamic prefix ranking maintains reliable evidence ordering across arbitrary retention thresholds, while survival calibration anchors relevance scores for diversity-aware coverage. This synergy ensures that READ consistently isolates task-critical cues, delivering superior accuracy retention and fine-grained visual grounding across diverse compression budgets.

5.2 Evaluation on Diverse VLM Architectures

We evaluate READ on Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and InternVL2.5-8B (Chen et al., 2025), both employing dynamic resolutions and specialized fusion. As shown in Table 2, on Qwen2.5-VL-7B-Instruct, READ consistently achieves the highest overall retention, outperforming SGP by 3.8% at 25% retention and 5.0% at 10% retention on average. This reveals a key drawback of online proxy guidance, where smaller helper models like Qwen2.5-VL-3B exhibit distinct attention distributions from target architectures. On InternVL2.5-8B, READ maintains high accuracy without running an on-device language model. While SGP achieves comparable accuracy on InternVL2.5-8B, it relies on an auxiliary multi-billion-parameter proxy model that incurs substantial edge compute and latency overhead as analyzed in Table 4, whereas READ employs only a compact, lightweight retriever. Heuristics like FastV and VisPruner degrade steeply under tight budgets, confirming that uncalibrated early pruning or query-agnostic selection fails to generalize across dynamic backbones.

Table 3: High-resolution compression on LLaVA-NeXT-7B. Best Avg. and Rel. in each block are bold.

Method	VQAv2	GQA	TextVQA	POPE	MME	Avg.	Rel.
LLaVA-NeXT-7B (2880)	79.6	62.5	60.5	85.7	1518.0	72.8	100.0%
Retain 640 Tokens (↓ 77.8%)							
VisPruner	77.1	61.3	59.8	85.2	1464.5	71.3	97.9%
DivPrune	77.8	62.0	55.8	85.9	1495.3	71.3	97.9%
SparseVLM	78.3	61.8	58.0	83.8	1514.2	71.5	98.2%
READ	78.5	62.6	59.8	86.6	1483.6	**72.3**	**99.3**%
Retain 320 Tokens (↓ 88.9%)							
VisPruner	74.1	59.5	58.8	81.5	1406.6	68.8	94.5%
DivPrune	75.9	60.6	54.5	84.3	1448.3	69.5	95.5%
SparseVLM	73.4	58.0	54.7	77.5	1400.2	66.7	91.6%
READ	77.4	61.4	58.9	86.4	1493.3	**71.8**	**98.6**%
Retain 160 Tokens (↓ 94.4%)							
VisPruner	70.0	56.2	55.8	73.9	1318.0	64.4	88.5%
DivPrune	73.4	58.7	53.6	80.6	1388.8	67.1	92.2%
SparseVLM	66.1	53.0	49.6	68.2	1171.7	59.1	81.2%
READ	76.1	60.7	57.3	85.0	1440.7	**70.2**	**96.4**%

Table 4: Efficiency analysis on InternVL2.5-8B on POPE. Edge covers encoding and selection latency; Transmit indicates uploaded visual token count; Cloud measures target forward latency.

Method	Tokens	FLOPs (T)	Selector Params (M)	GPU Memory (GB)	Edge (ms)	Cloud (ms)
InternVL2.5-8B	1667	30.7	–	17.5	64.2	397.8
VisPruner	166	8.2	–	16.3	81.2	127.0
IDPruner	166	8.3	16.8	16.3	75.8	127.0
SGP	166	18.7	2205.7	21.6	162.6	148.6
READ (Ours)	166	8.3	68.3	16.5	77.4	127.0

5.3 High-Resolution Visual Token Compression

High-resolution architectures like LLaVA-NeXT-7B (Liu et al., 2024b) deploy grid tiling to preserve detail, generating 2,880 tokens per image and introducing extensive spatial redundancy. In Table 3, we compare READ against VisPruner, DivPrune, and SparseVLM under budgets of 640, 320, and 160 tokens. While all methods perform reasonably under moderate compression, baselines degrade sharply under tighter limits: query-agnostic selectors favor prominent foreground patches over fine-grained text, while SparseVLM struggles with diffuse early-layer attention across thousands of tokens. In contrast, READ cross-attends all tiled sub-patches against the prompt, preserving over 96% of uncompressed accuracy at 160 tokens under an aggressive 94.4% reduction. This confirms that READ effectively isolates task-critical evidence from massive spatial redundancy.

5.4 Efficiency Analysis

A practical edge–cloud framework must alleviate uplink transmission and cloud computation without turning the edge device into an execution bottleneck. As shown in Table 4, READ achieves this balanced trade-off by decoupling query-aware scoring from heavy autoregressive reasoning. On the edge, our compact retriever requires only 13.2 ms beyond the initial vision encoding step. This keeps the total on-device latency comparable to query-agnostic selectors like IDPruner and VisPruner, while operating at less than half the latency of proxy-based approaches such as SGP. Prior to offloading, READ filters out 90.0% of visual tokens, directly mitigating uplink transmission bottlenecks over bandwidth-constrained networks. On the cloud, receiving this compact token sequence substantially reduces prefill workload, delivering over $3.1\times$ target forward speedup alongside noticeable reductions in overall FLOPs and memory footprint. In contrast, while SGP also reduces visual sequence length, its multi-billion-parameter auxiliary proxy incurs severe on-device delays. Consequently, READ delivers practical collaborative speedups across both edge and cloud tiers.

6 Conclusion

We presented READ, a query-conditioned evidence retriever for edge–cloud VLM inference. Target-model attention is distilled offline into a compact edge retriever via prefix ranking and survival calibration, bypassing online language models. Relevance-weighted coverage then converts calibrated scores into a compact visual sequence. Across diverse VLM architectures, READ amortizes multimodal reasoning offline to select task-critical tokens, achieving substantial cloud computation and latency speedups with minimal edge overhead.

Reproducibility Statement

All baseline methods are evaluated using publicly released checkpoints, and all benchmark datasets are publicly accessible. Complete architectural specifications, training objectives, and selection algorithms are provided in Section 4 and the Appendix. The complete source code, configuration files, and evaluation scripts are anonymously available at `https://anonymous.4open.science/r/ICLR_READ/` to facilitate exact replication.

AI Declaration

Large language models were used solely for grammar polishing and stylistic improvements. They were not involved in research ideation, method design, or data analysis. The authors retain full responsibility for all findings and contents presented in this manuscript.

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 (CVPR)*, pp. 9392–9401, 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. *arXiv preprint arXiv:2408.10945*, 2024.

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

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*, pp. 19–35. Springer, 2024.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Wen Gao, Siwei Ma, Ling-Yu Duan, Yonghong Tian, Peiyin Xing, Yaowei Wang, Shanshe Wang, Huizhu Jia, and Tiejun Huang. Digital retina: A way to make the city brain more efficient by visual coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4147–4161, 2021. doi: 10.1109/TCSVT.2021.3104305.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2018.

Zixu Hao, Huiqiang Jiang, Shiqi Jiang, Ju Ren, and Ting Cao. Hybrid SLM and LLM for edge–cloud collaborative inference. In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pp. 36–41. Association for Computing Machinery, 2024. doi: 10.1145/3662006.3662067.

Cyril Shih-Huan Hsu, Wig Yuan-Cheng Cheng, and Chrysa Papagianni. Progressive semantic communication for efficient edge-cloud vision-language models. *arXiv preprint arXiv:2604.26508*, 2026.

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 (CVPR)*, pp. 6700–6709, 2019.

Yuanyuan Jia, Shunpu Tang, and Qianqian Yang. CoVSpec: Efficient device-edge co-inference for vision-language models via speculative decoding. *arXiv preprint arXiv:2605.02218*, 2026.

Hongpeng Jin and Yanzhao Wu. CE-CoLLM: Efficient and adaptive large language models through cloud–edge collaboration. *arXiv preprint arXiv:2411.02829*, 2024.

Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. In *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 615–629, 2017.

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 (EMNLP)*, pp. 292–305, 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*, pp. 26296–26306, 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. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, pp. 216–233. Springer, 2024c.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 2507–2521, 2022.

Yoshitomo Matsubara, Marco Levorato, and Francesco Restuccia. Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys*, 55(5):1–30, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ahmed Šabanović, Paul Joe Maliakel, and Ivona Brandić. INAR-VL: Input-aware routing for edge–cloud vision–language inference. *arXiv preprint arXiv:2605.18853*, 2026.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*, 2026.

Jianshu She, Wenhao Zheng, Zhengzhong Liu, Hongyi Wang, Eric P. Xing, Huaxiu Yao, and Qirong Ho. Token level routing inference system for edge devices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 159–166, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-demo.16.

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 (CVPR)*, pp. 8317–8326, 2019.

Soochang Song and Yongjune Kim. Collaborative edge-to-server inference for vision-language models. *arXiv preprint arXiv:2512.16349*, 2025.

Yifan Tan, Yifu Sun, Shirui Huang, Hong Liu, Guanghua Yu, Jianchen Zhu, and Yangdong Deng. IDPruner: Harmonizing importance and diversity in visual token pruning for MLLMs. *arXiv preprint arXiv:2602.13315*, 2026.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Tong Xu, Hailong Shi, and Xingyu Gao. SCoRe: Salience-coverage reduction for vision token pruning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2026.

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*, pp. 19792–19802, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Hanzhang Yuan, Mengxuan Hu, Wenhao Zhang, Tianlong Wang, Zhongliang Zhou, Jiasen Lu, and Sheng Li. Reducing token redundancy in LVLMs: A systematic review of token pruning methods. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7231–7251, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.328.

Liangqi Yuan, Dong-Jun Han, Shiqiang Wang, and Christopher G. Brinton. Device–cloud collaborative LLM inference with multi-modal, multi-task, multi-turn conversations. *arXiv preprint arXiv:2502.11007*, 2025.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. TinyChart: Efficient chart understanding with visual token merging and program-of-thoughts learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16212–16227, 2024a.

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

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024b.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Zhikai Li, Yibing Song, Kai Wang, Zhangyang Wang, and Yang You. A stitch in time saves nine: Small vlm is a precise guidance for accelerating large vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19814–19824, 2025.

Jiaying Zhu, Yurui Zhu, Xin Lu, Wenrui Yan, Dong Li, Kunlin Liu, Xueyang Fu, and Zheng-Jun Zha. VisionSelector: End-to-end learnable visual token compression for efficient multimodal LLMs, 2025.

A Implementation Details

This section provides comprehensive implementation details for the READ framework, covering model architectures, spatial position preservation, training protocols, and optimization configurations.

Target VLMs and vision encoders. We evaluate READ across four representative vision–language model architectures:

- **LLaVA-1.5-7B** (Liu et al., 2024a): Built upon a frozen CLIP ViT-L/14 vision encoder (Radford et al., 2021) operating at 336×336 resolution ($N = 576$ visual tokens with patch size 14×14), a two-layer MLP cross-modal projector, and a Vicuna-v1.5-7B language backbone.
- **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025): Features dynamic-resolution vision encoding with native 2D window attention and 3D Multimodal Rotary Position Embeddings (MRoPE), producing variable visual token lengths depending on input image dimensions.
- **InternVL2.5-8B** (Chen et al., 2025): Utilizes a dynamic pixel-unshuffle vision backbone with tile-based high-resolution image representation, yielding up to several thousand visual tokens per image.
- **LLaVA-NeXT-7B** (Liu et al., 2024b): Employs an AnyRes multi-crop framework that splits high-resolution inputs into a 2×2 grid of local patches alongside a global context thumbnail, generating up to 2,880 visual tokens.

Spatial position indexing under token pruning. For modern VLMs with dynamic resolution and non-trivial spatial encoding mechanisms (such as 3D/2D MRoPE in Qwen2.5-VL and dynamic tile-based embeddings in InternVL2.5), discarding visual tokens requires preserving original spatial coordinates. Rather than naively concatenating retained patches consecutively, READ tracks each selected token's original 2D grid coordinates, tile indices, and native position identifiers. When the cloud VLM receives the sparse token sequence, it binds the retained tokens directly to their original position IDs and spatial embeddings. This preserves the genuine geometric layout, tile boundaries, and relative spatial distances necessary for faithful multimodal reasoning under sparse token inputs.

Retriever architecture and edge deployment. The READ retriever operates entirely on the edge device prior to network transmission. To score token relevance, the retriever consumes projected vision encoder representations, query token embeddings, and the centered log-budget visual prior $\widetilde{\boldsymbol{\pi}}$. Crucially, the retriever executes no autoregressive language-model layers on the edge, maintaining a minimal compute and memory footprint.

Visual prior for backbones without a `[CLS]` token. For vision encoders equipped with a `[CLS]` token (such as CLIP in LLaVA-1.5 and LLaVA-NeXT, or InternViT in InternVL2.5), the prior $\boldsymbol{\pi}$ is derived from `[CLS]`-to-patch attention (with spatial average pooling across pixel-unshuffle tiles for InternVL). For architectures without a dedicated `[CLS]` token (such as Qwen2.5-VL), the visual prior is computed by averaging the self-attention that each visual patch receives from all other visual tokens ($\frac{1}{N}\sum_{j=1}^{N} A_{j,i}$) across all attention heads in the final vision-encoder layer.

Training and optimization protocol. Each compact retriever is trained on a composite $\sim$144K multimodal subset sampled from Cambrian-737K (Tong et al., 2024) for a single epoch, and the training set shares no images or questions with the evaluation benchmarks. Optimization is conducted using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, weight decay of 0.01, a peak learning rate of 2×10^{-4}, and a 5% linear learning rate warmup followed by cosine annealing. We set the loss weights to $\lambda_{\text{rank}} = 1.0$ and $\lambda_{\text{cal}} = 1.0$, and configure the coverage diversity hyperparameter to $\eta = 0.8$. All training experiments are executed on 8 NVIDIA RTX 4090 GPUs.

Offline attention distillation. A key architectural advantage of READ is its fully decoupled offline training scheme. Before retriever optimization, the frozen cloud VLM is executed once in evaluation mode across the training split to extract full-prompt cross-modal attention maps, aggregated across layers as formulated in Section 4. These target attention profiles are serialized and cached to disk. During subsequent retriever training, the multi-billion-parameter target VLM remains completely frozen and is never loaded into GPU memory, entirely eliminating expensive backward passes through

the foundation model. Consequently, retriever training completes in tens of minutes while strictly preserving the generative parameters of the base model.

B Evaluation Protocol and Baselines

This section describes the evaluation benchmarks, summary metrics, and comparative baseline implementations used throughout our experiments.

Evaluation benchmarks. We evaluate visual token compression across ten standard multimodal benchmarks spanning a wide range of perceptual and reasoning capabilities:

- **General Visual Question Answering**: VQAv2 (Goyal et al., 2017) evaluates open-ended visual question answering on natural everyday images; GQA (Hudson & Manning, 2019) evaluates compositionality and multi-step relational reasoning over structured visual scenes.
- **Scientific and Unconstrained Reasoning**: VizWiz (Gurari et al., 2018) consists of real-world queries captured by visually impaired users featuring frequent low image quality, obscure camera angles, and unconstrained answer formats; ScienceQA (Lu et al., 2022) assesses multimodal scientific reasoning across diverse elementary and secondary science topics.
- **Scene Text and Fine-Grained Perception**: TextVQA (Singh et al., 2019) requires reading, localizing, and reasoning over scene text embedded in natural images, serving as a sensitive diagnostic benchmark for whether visual compression prematurely discards fine-grained evidence.
- **Object Hallucination**: POPE (Li et al., 2023) measures object existence verification across random, popular, and adversarial negative splits, providing a targeted metric for probing whether visual token pruning induces spurious hallucinations.
- **Comprehensive Multimodal Benchmarks**: MME (Fu et al., 2023) systematically probes perception and cognition capabilities across 14 fine-grained subtasks; MMBench and MMBench-CN (Liu et al., 2024c) provide bilingual evaluation across hierarchical capability taxonomies using circular evaluation; MM-Vet (Yu et al., 2023) assesses complex, integrated multimodal capabilities across diverse visual-linguistic tasks.

Evaluation metrics. To quantify performance across diverse benchmark suites, we report two summary metrics:

- **Average Score (Avg.)**: The arithmetic mean of the individual benchmark scores reported in each table, rounded to one decimal place. Following standard convention for multi-scale aggregates, MME (ranging up to 2,000) is normalized by dividing by 20 to align its scale to 0–100 prior to computing the mean alongside percentage-based benchmarks.
- **Relative Preservation (Rel.)**: The ratio of the compressed model's overall average score to that of the uncompressed backbone, defined as $\text{Rel.} = (\text{Avg.}_{\text{method}}/\text{Avg.}_{\text{full}}) \times 100\%$, rounded to one decimal place.

Comparative baselines. We benchmark READ against eight representative visual token compression methods, covering all primary compression paradigms across vision-only and query-aware categories:

- **Vision-only edge selectors**: VisPruner (Zhang et al., 2025) first retains visually salient tokens using vision-encoder attention and then removes redundant tokens based on feature similarity; DivPrune (Alvar et al., 2025) selects a diverse token subset rather than clustering by spatial location; VisionZip (Yang et al., 2025) first selects informative visual tokens and merges the remaining ones via feature similarity. These methods execute on the edge prior to transmission but operate without query conditioning.
- **In-model language-guided pruners**: FastV (Chen et al., 2024) computes attention scores within early transformer layers of the cloud language model to drop inactive tokens; SparseVLM (Zhang et al., 2024b) dynamically modulates token sparsity across successive transformer blocks. While query-aware, these methods operate within the cloud LLM after the full sequence has traversed the uplink, failing to alleviate transmission payload or communication latency.
- **Trainable edge selectors**: VisionSelector (Zhu et al., 2025) trains a lightweight visual-token scoring module end-to-end with the downstream VLM objective; IDPruner (Tan et al., 2026) combines visual importance scores with Maximal Marginal Relevance (MMR) heuristics. While

trainable and supporting variable-rate inference, both methods score image tokens purely from visual features without question conditioning, and their scalar supervision lacks explicit ranking calibration across budgets.

- **Auxiliary-guided cross-VLM compression**: SGP (Zhao et al., 2025) executes an auxiliary smaller VLM of the same family on the edge device (e.g., Qwen2.5-VL-3B for Qwen2.5-VL-7B and InternVL2.5-2B for InternVL2.5-8B) to generate query-guided attention signals. However, SGP demands substantial on-device memory and cannot be deployed on target VLMs lacking an architectural counterpart (precluding its application to LLaVA-1.5-7B).

Retention budgets for each evaluation setting are detailed alongside the corresponding experimental results.

C Retriever Architecture

While Section 4.3 outlines the core projection and cross-modal attention design of the retriever, this section provides the detailed layer-wise formulation of its transformer blocks and scoring head. The entire module operates as a lightweight, non-autoregressive feed-forward network without executing multi-billion-parameter language-model layers.

Each of the $L_s = 2$ cross-modal transformer blocks sequentially processes textual context and visual evidence. First, prompt token representations are contextualized via multi-head self-attention (MSA_t) and a feed-forward network (FFN_t):

$$\widehat{\boldsymbol{T}}^{(\ell)} = \boldsymbol{T}^{(\ell-1)} + \mathrm{MSA}_t(\mathrm{LN}(\boldsymbol{T}^{(\ell-1)})), \tag{15}$$

$$\boldsymbol{T}^{(\ell)} = \widehat{\boldsymbol{T}}^{(\ell)} + \mathrm{FFN}_t(\mathrm{LN}(\widehat{\boldsymbol{T}}^{(\ell)})). \tag{16}$$

Next, the visual tokens model mutual spatial relationships through visual self-attention (MSA_v):

$$\widehat{\boldsymbol{V}}^{(\ell)} = \boldsymbol{V}^{(\ell-1)} + \mathrm{MSA}_v(\mathrm{LN}(\boldsymbol{V}^{(\ell-1)})). \tag{17}$$

Finally, each visual patch queries the contextualized prompt representations via multi-head cross-attention ($\mathrm{MCA}_{v\leftarrow t}$), followed by visual feature transformation:

$$\widetilde{\boldsymbol{V}}^{(\ell)} = \widehat{\boldsymbol{V}}^{(\ell)} + \mathrm{MCA}_{v\leftarrow t}\big(\mathrm{LN}(\widehat{\boldsymbol{V}}^{(\ell)}), \mathrm{LN}(\boldsymbol{T}^{(\ell)})\big), \tag{18}$$

$$\boldsymbol{V}^{(\ell)} = \widetilde{\boldsymbol{V}}^{(\ell)} + \mathrm{FFN}_v(\mathrm{LN}(\widetilde{\boldsymbol{V}}^{(\ell)})). \tag{19}$$

The directed image-to-text cross-attention ensures that each visual token independently attends to the full query context in parallel, producing scoreable representations without autoregressive sequence decoding. From the final visual features $\boldsymbol{V}^{(L_s)}$, a lightweight two-layer MLP head computes the residual relevance adjustment in Eq. 9:

$$\Delta_i = \boldsymbol{w}_2^\top \,\mathrm{GELU}\big(W_1\,\mathrm{LN}(\boldsymbol{v}_i^{(L_s)})\big), \tag{20}$$

where $W_1 \in \mathbb{R}^{d\times d}$ and $\boldsymbol{w}_2 \in \mathbb{R}^d$. The default architecture employs hidden dimension $d = 1{,}024$ with 8 attention heads.

D Algorithm and Edge–Cloud Workflow

Algorithm 1 outlines the decoupled workflows of offline attention distillation and online edge inference.

Offline training efficiency. The compact retriever is trained on the composite $\sim$144K multimodal subset sampled from Cambrian-737K (Tong et al., 2024) for a single epoch. While forward evaluation to cache target VLM attention profiles across the dataset requires several hours on cloud hardware, optimizing the lightweight retriever itself takes merely tens of minutes. Because the multi-billion-parameter target VLM remains completely frozen during retriever training and never participates in backpropagation, gradient memory overhead is strictly confined to the lightweight retriever.

Algorithm 1 READ training and edge–cloud collaborative inference

Require: Frozen target VLM F, input image I, query Q, retriever f_θ, target budget K
Phase 1: Offline Attention Distillation
Run frozen target VLM $F(I, Q)$ offline to cache attention weights $\boldsymbol{A}^{(\ell,h)}$
Compute target visual evidence preferences a_i^T and derive teacher ranks $\boldsymbol{r}^T$ (Eq. 3)
Extract input representations $\boldsymbol{V}^{\text{in}}, \boldsymbol{T}^{\text{in}}$ and construct visual prior $\widetilde{\boldsymbol{\pi}}$ (Eq. 5)
Predict relevance logits $\boldsymbol{z} = f_\theta(\boldsymbol{V}^{\text{in}}, \boldsymbol{T}^{\text{in}}, \widetilde{\boldsymbol{\pi}})$
Sample dynamic prefix boundaries and optimize f_θ via Eqs. 10–12
Phase 2: Online Edge–Cloud Inference (Discard teacher attention and target VLM from edge)
On Edge Device:
  Extract visual tokens $\boldsymbol{V}^{\text{in}}$, prompt embeddings $\boldsymbol{T}^{\text{in}}$, and prior $\widetilde{\boldsymbol{\pi}}$ using local front-end
  Predict calibrated patch relevance scores $\boldsymbol{s} = \sigma(f_\theta(\boldsymbol{V}^{\text{in}}, \boldsymbol{T}^{\text{in}}, \widetilde{\boldsymbol{\pi}}))$
  Initialize selected token index set $\mathcal{S} \leftarrow \{\arg\max_i s_i\}$
  while $|\mathcal{S}| < K$ **do**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\arg\max_{i \notin \mathcal{S}} G(i \mid \mathcal{S})\}$ ▷ Greedy coverage selection via Eq. 14
  end while
  Transmit sparse sequence $[\boldsymbol{v}_i]_{i \in \text{sort}(\mathcal{S})}$ and prompt Q over uplink channel
On Cloud Server:
  Bind received tokens with original 2D position embeddings and execute target VLM F
return Generated response text from cloud VLM

Collaborative serving pipeline. In real-world deployment, the edge device runs only the front-end vision encoder and the compact feed-forward retriever. Upon receiving an image and a user query, the edge device scores token relevance, executes relevance-weighted coverage selection, and transmits only the retained K visual tokens over the network. The cloud server receives the compact token sequence, restores original 2D positional embeddings, and performs autoregressive reasoning with the full VLM. This collaborative pipeline simultaneously relieves uplink bandwidth constraints, reduces cloud prefill memory footprint, and eliminates the latency overhead of executing heavy language models on-device.

E Component Ablations

To systematically dissect the contribution of each architectural module, loss objective, and selection mechanism within READ, we conduct comprehensive ablation studies on LLaVA-1.5-7B across five representative benchmarks under three retention budgets $K \in \{128, 64, 32\}$. The experimental findings are compiled in Table 5.

Role of the log-budget visual prior. Incorporating the centered visual prior $\widetilde{\boldsymbol{\pi}}$ provides an inductive anchor that prevents the retriever from overfitting to short prompts or suffering from linguistic drift. When this prior is omitted (*w/o Visual Prior*), downstream retention drops markedly as token budgets tighten (declining to 88.0% at $K = 32$). The prior establishes a robust foreground baseline, guaranteeing that when user queries are concise, generic, or ambiguous, the retriever continues to preserve essential scene topology instead of degenerating into spurious cross-attention noise.

Relevance-weighted coverage vs. naive Top-K selection. Replacing relevance-weighted coverage selection with greedy Top-K ranking based purely on predicted relevance scores (*w/o Coverage*) widens the retention gap as the compression ratio increases. Direct Top-K selection inevitably concentrates scarce token capacity onto spatially adjacent, mutually redundant patches that happen to register high individual relevance, frequently omitting supporting spatial context. By contrast, our coverage utility formulation,

$$G(i \mid \mathcal{S}) = \delta(i \mid \mathcal{S}) \cdot \bar{s}_i^{\eta}, \tag{21}$$

penalizes spatial and feature-space redundancy through the distance penalty $\delta(i \mid \mathcal{S})$. This mechanism achieves a principled equilibrium between isolating query-critical evidence and maintaining broad contextual fidelity.

Table 5: Ablation results on LLaVA-1.5-7B across three retention budgets. Bold numbers indicate the best performance in each budget block.

Method	TextVQA	POPE	MME	MMB	MM-Vet	Avg.	Rel.
LLaVA-1.5-7B (576)	58.2	85.3	1514.8	64.3	29.7	62.6	100.0%
Retain 128 Tokens (↓ 77.8%)							
w/o Visual Prior	57.3	85.6	1430.2	58.8	31.5	60.9	97.3%
w/o Calibration Loss	57.2	85.9	1447.3	58.7	31.1	61.1	97.6%
w/ KL Divergence Loss	56.8	84.2	1412.0	58.2	29.8	59.9	95.7%
w/o Coverage	57.2	85.4	1445.5	58.8	31.4	61.0	97.4%
READ (Ours)	57.2	85.9	1440.5	62.2	32.6	**62.0**	**99.0%**
Retain 64 Tokens (↓ 88.9%)							
w/o Visual Prior	56.1	83.0	1364.1	58.4	27.4	58.6	93.6%
w/o Calibration Loss	56.2	83.8	1374.0	57.6	27.0	58.7	93.8%
w/ KL Divergence Loss	55.3	79.8	1354.2	57.1	26.6	57.3	91.5%
w/o Coverage	56.5	84.8	1420.3	58.1	28.9	59.9	95.7%
READ (Ours)	56.1	86.4	1450.6	61.3	29.3	**61.1**	**97.6%**
Retain 32 Tokens (↓ 94.4%)							
w/o Visual Prior	54.6	77.1	1278.1	55.4	24.5	55.1	88.0%
w/o Calibration Loss	54.3	76.9	1280.0	55.4	23.5	54.8	87.5%
w/ KL Divergence Loss	52.9	69.2	1192.6	53.6	22.5	51.6	82.4%
w/o Coverage	54.8	79.5	1312.2	57.2	26.9	56.8	90.7%
READ (Ours)	55.3	86.6	1348.1	59.7	29.2	**59.6**	**95.2%**

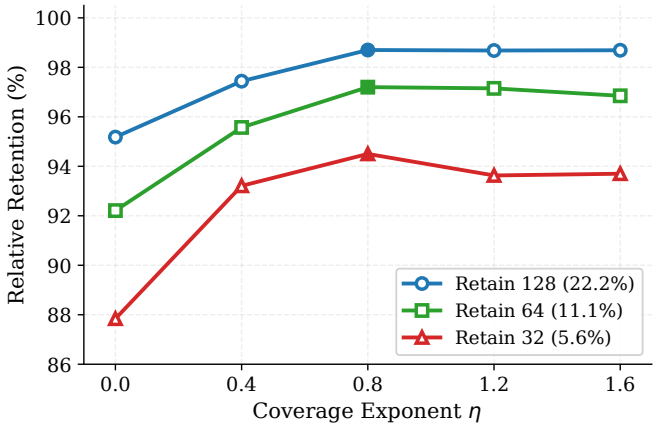


Figure 2: Downstream performance sensitivity across varying coverage exponents $\eta \in [0, 1.6]$ under retention budgets $K \in \{32, 64, 128\}$.

Distillation loss formulation: Rank-survival calibration vs. standard KL divergence. Supervising evidence retrieval via rank-survival calibration and prefix ranking yields distinct advantages over conventional attention distribution distillation:

- **Pathology of standard distribution matching ($\mathcal{L}_{\mathrm{KL}}$)**: Replacing our objectives with standard Kullback–Leibler divergence minimization (*w/ KL Divergence Loss*) causes pronounced degradation across all benchmarks and compression regimes (dropping to 82.4% at $K = 32$). Because teacher cross-attention distributions are inherently heavy-tailed, minimizing KL divergence forces the student to dedicate representational capacity to fitting diffuse probability tails across irrelevant background tokens, rather than establishing sharp ordinal boundaries among critical evidence candidates.
- **Efficacy of rank-survival calibration ($\mathcal{L}_{\mathrm{cal}}$)**: The calibration loss aligns predicted scalar logits with empirical survival probabilities across diverse continuous budgets. Removing this supervision (*w/o Calibration Loss*) noticeably weakens retention under high sparsity. Calibration establishes a reliable absolute scale when relevance scores scale geometric distances ($\bar{s}_i^{\eta}$), ensuring that uncalibrated score distributions do not distort spatial coverage.

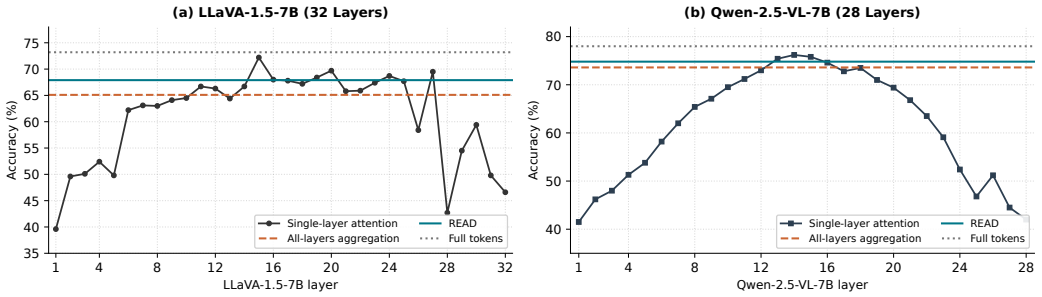


Figure 3: Downstream reasoning accuracy on a 1k VQAv2 subset when selecting the top-32 visual tokens based on prompt-to-vision attention from each individual transformer layer versus all-layer aggregation on (a) LLaVA-1.5-7B (32 layers) and (b) Qwen2.5-VL-7B (28 layers). Single-layer attention exhibits severe depth-wise volatility and divergent peak trajectories across backbones, whereas all-layer aggregation provides stable, architecture-agnostic distillation supervision.

Coverage exponent η and the relevance–diversity trade-off. In the coverage utility function $G(i \mid \mathcal{S}) = \delta(i \mid \mathcal{S}) \cdot \bar{s}_i^{\eta}$, the exponent η directly governs the balance between query-relevance salience $\bar{s}_i$ and spatial–semantic diversity $\delta(i \mid \mathcal{S})$. Figure 2 plots downstream sensitivity across $\eta \in [0, 1.6]$ in terms of the macro-average relative retention across five representative benchmarks (TextVQA, POPE, MME, MMBench, and MM-Vet) under retention budgets $K \in \{32, 64, 128\}$. Across all budget regimes, accuracy follows a consistent inverted-U trajectory.

When $\eta = 0$, relevance modulation is eliminated ($\bar{s}_i^0 = 1$), causing greedy selection after the initial seed token to collapse into unweighted furthest-point sampling on visual features. While this dispersion strategy maintains reasonable retention under generous budgets where capacity is plentiful, performance deteriorates sharply under aggressive pruning ($K = 32$), demonstrating that spatial dispersion alone cannot reliably isolate localized visual evidence. As η increases, query-relevance guidance is restored, with downstream accuracy peaking stably between $\eta = 0.8$ and 1.2. Conversely, setting $\eta > 1.2$ over-emphasizes scalar relevance at the expense of diversity, causing selection to degenerate toward spatially clustered Top-K patches. These empirical trends validate $\eta = 0.8$ as an optimal, budget-agnostic hyperparameter.

F Analysis of Multi-Layer Attention Aggregation

A critical design consideration in attention-guided visual token compression is whether to extract supervisory attention from a single intermediate transformer layer or to aggregate attention across the full network depth of the teacher VLM. While several in-model pruning methods (e.g., Chen et al., 2024) postulate that an early layer provides a sufficient proxy for visual token relevance, we demonstrate that single-layer attention is inherently erratic and exhibits disparate depth dynamics across diverse VLM families. Full-layer aggregation is therefore essential for extracting robust, model-agnostic distillation signals.

Empirical comparison across foundation architectures. In Figure 3, we evaluate downstream answer accuracy after retaining the top-32 visual tokens selected via causal prompt-to-vision attention from each individual transformer layer across two distinct VLM backbones:

1. **LLaVA-1.5-7B** (Figure 3a): 32 transformer layers, standard 1D Rotary Position Embeddings (RoPE), and a two-layer MLP cross-modal projector.
2. **Qwen2.5-VL-7B** (Bai et al., 2025) (Figure 3b): 28 transformer layers, native 3D Multimodal Rotary Position Embeddings (MRoPE), and dynamic-resolution patch tokenization.

For both architectures, we benchmark each individual layer against uniform all-layer aggregation, the full uncompressed sequence, and the trained READ retriever under identical token budgets.

Intra-model volatility across network depth. Across both model architectures, the utility of single-layer attention varies drastically across transformer layers:

- *Shallow layers capture localized low-level primitives*: In early layers, attention remains predominantly localized around low-level textures, edge boundaries, or positional artifacts, providing negligible cross-modal grounding relative to the prompt.
- *Intermediate layers exhibit transient peaks*: Effective prompt-guided cross-modal interaction emerges within intermediate layers. However, performance fluctuates unpredictably between adjacent layers, making post-hoc layer manual tuning highly sensitive to evaluation distributions.
- *Deep layers degrade under delimiter attention sinks*: In the final transformer layers, token selection accuracy degrades precipitously. In these terminal blocks, attention distributions become dominated by attention sinks on special delimiter tokens and specialize toward next-token autoregressive generation, eroding the spatial grounding required for visual evidence retrieval.

Inter-model divergence across VLM families. Crucially, the optimal layer index differs substantially between backbones:

- In LLaVA-1.5-7B, retention rises steadily through mid-to-late representations, maintaining an extended plateau between layers 16 and 26 before collapsing in terminal layers.
- In Qwen2.5-VL-7B, performance peaks much earlier around layer 14, displaying a narrow optimal band followed by a premature decline in subsequent blocks.

Because the most informative depth cannot be determined *a priori* and varies across model families, relying on any single layer introduces an arbitrary architectural dependency that risks catastrophic failure when ported to new models.

All-layer aggregation provides stable, model-agnostic teacher supervision. Aggregating attention matrices across all transformer layers systematically overcomes the vulnerabilities of individual layers:

1. **Noise suppression and stability**: Summing attention across the full network depth dampens low-level noise from shallow layers and sink-induced artifacts from deep layers, establishing a dependable performance floor without requiring manual layer hyperparameter searches.
2. **Multi-depth visual feature integration**: Full-depth aggregation combines spatial structures resolved in early-to-mid layers with abstract semantic associations formulated in deeper layers, capturing a holistic evidence spectrum.
3. **Superior distillation target**: When distilled into our lightweight retriever, this aggregated target provides a smooth, well-conditioned ranking signal. Consequently, the trained READ retriever matches or surpasses the best individual layer trajectories while generalizing seamlessly across downstream tasks.

These cross-model empirical findings confirm that all-layer aggregation provides a principled, foundation-agnostic basis for visual evidence distillation.

G Analysis of Teacher Attention Signals for Distillation

A central design choice in query-conditioned visual token distillation is the selection of the teacher attention supervision signal: specifically, whether to supervise the student retriever using attention extracted solely from the terminal prompt token (*Last Prompt*) or attention aggregated across all prompt tokens (*Full Prompt*). In Table 6, we evaluate this trade-off from two complementary perspectives: (1) utilizing the teacher VLM's native attention maps directly for inference-time Top-K token selection, and (2) employing those attention distributions as distillation supervision targets to optimize the lightweight READ retriever.

VLM direct selection: Last Prompt outperforms Full Prompt. When applying the teacher VLM's online attention directly for token pruning, attention from the terminal prompt token consistently yields higher retention accuracy than full-prompt attention across all budgets, with the

Table 6: Comparison between teacher VLM direct attention-guided selection and distilled READ retrievers under different prompt attention signals on LLaVA-1.5-7B across five multimodal benchmarks.

Budget	Method	GQA	VizWiz	SQA	TextVQA	POPE	Avg.	Rel.
Full (576)	LLaVA-1.5-7B (Uncompressed)	61.9	50.2	69.5	58.2	85.3	**65.0**	**100.0%**
$K = 128$	VLM (Last Prompt)	59.6	52.3	68.6	57.4	86.7	**64.9**	**99.8%**
	VLM (Full Prompt)	60.0	52.0	68.2	57.3	86.5	64.8	99.7%
	READ (Distilled from Last Prompt)	59.2	52.2	68.5	57.2	86.1	64.6	99.4%
	READ (Distilled from Prompt)	59.7	52.3	69.1	57.2	85.9	**64.8**	**99.7%**
$K = 64$	VLM (Last Prompt)	59.0	52.7	68.8	56.8	86.4	**64.7**	**99.5%**
	VLM (Full Prompt)	59.1	51.9	68.6	56.8	86.2	64.5	99.2%
	READ (Distilled from Last Prompt)	57.4	53.0	68.2	56.2	84.8	63.9	98.3%
	READ (Distilled from Prompt)	58.3	53.2	68.7	56.1	86.4	**64.5**	**99.2%**
$K = 32$	VLM (Last Prompt)	58.2	51.9	69.0	55.8	86.3	**64.2**	**98.8%**
	VLM (Full Prompt)	57.1	50.5	68.2	54.8	84.9	63.1	97.1%
	READ (Distilled from Last Prompt)	55.0	53.1	68.5	55.3	81.6	62.7	96.5%
	READ (Distilled from Prompt)	56.7	53.6	68.2	55.3	86.6	**64.1**	**98.6%**

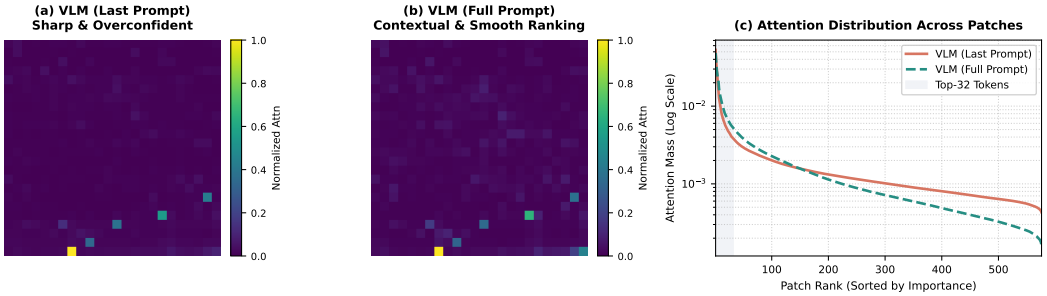


Figure 4: Attention distributions of VLM Last Prompt versus Full Prompt on a representative instance. (a) Last Prompt attention is sparse and sharply peaked, focusing almost exclusively on a localized patch while zeroing out surrounding context. (b) Full Prompt attention distributes probability mass smoothly across relevant object entities and spatial context, creating an informative ranking continuum. (c) Attention distribution across all 576 visual patches in logarithmic scale; whereas Last Prompt decays abruptly into near-zero values, Full Prompt yields well-graded intermediate scores essential for training ranking and calibration objectives.

advantage expanding under aggressive compression ($K = 32$). This empirical behavior reflects the autoregressive causal structure of the base decoder: because the model generates subsequent tokens conditional on the terminal prompt position, attention weights from the final token serve as a concentrated causal bottleneck reflecting immediate visual dependency.

Distillation inversion: Full Prompt outperforms Last Prompt. Remarkably, when these attention profiles are adopted as offline supervision signals to train our compact READ retriever, this relationship inverts: READ trained on Full Prompt supervision consistently surpasses the model trained on Last Prompt across all budgets, widening the performance lead to +1.4% average retention under aggressive compression ($K = 32$). Although Last Prompt functions as an effective direct selector for the teacher model itself, it proves suboptimal as a teacher target for training a student retriever.

Mechanisms underlying the distillation inversion. We identify two primary factors driving this empirical divergence:

1. **Architectural inductive bias and multi-token alignment.** The READ retriever processes prompt queries via cross-attention blocks, where visual tokens attend directly to contextualized

representations of all prompt tokens. Full Prompt supervision aligns naturally with this multi-token cross-attention design, grounding visual patches against the complete linguistic context. Conversely, distilling exclusively from the terminal token introduces an architectural mismatch, forcing a multi-token cross-attention network to collapse its contextual representations toward predicting the isolated focus of a trailing delimiter.

2. **Continuous ranking gradients vs. extreme distribution sparsity.** As illustrated in Figure 4, attention from the final prompt token is sharply peaked on a localized patch, while the vast majority of visual tokens receive near-zero weights. While this sparse focus benefits direct Top-K truncation, it hampers student training:

   - *Vanishing ranking gradients*: In our prefix-ranking objective (Eq. 10), candidate evidence tokens are contrasted against unselected tokens. When teacher attention values across non-focal patches collapse to numerical insignificance, the teacher provides negligible ranking gradients among intermediate and contextual visual patches, depriving the student of meaningful ranking signals.
   - *Overconfidence penalties in calibration*: The extreme dynamic range of single-token attention imposes severe calibration penalties on borderline candidate patches, distorting the student's learned score scale.

   In contrast, Full Prompt attention aggregates relevance across diverse query terms, yielding a smoothly graded importance spectrum across the image grid. This provides informative, stable ranking gradients across diverse candidate patches, enabling the lightweight retriever to generalize robustly across downstream tasks.

H Qualitative Evidence Selection Across Budgets and Queries

To examine how READ isolates question-critical visual evidence under varying compression regimes and textual intents, we provide qualitative token selection visualizations on representative samples from TextVQA (Singh et al., 2019). In scene-text reasoning tasks, critical visual evidence typically occupies localized regions (e.g., text banners, price tags, or brand logos), while prominent foreground objects frequently dominate global visual saliency. We compare READ against three representative baselines: (1) **VisPruner** (Zhang et al., 2025), which combines vision-encoder `[CLS]`-to-patch attention with feature-similarity pruning; (2) **VisionSelector** (Zhu et al., 2025), which scores visual tokens via task-supervised importance projections; and (3) **VisionZip** (Yang et al., 2025), a query-agnostic method that selects informative visual tokens and merges remaining patches based on feature similarity.

H.1 Token Selection Across Compression Budgets

Figure 5 illustrates visual token selection across varying retention budgets ($K \in \{32, 64, 128\}$, corresponding to 5.6%, 11.1%, and 22.2% retention ratios), alongside the uncompressed original image. In this scene, the user query targets a localized brand identifier: *"What is the brand of this television?"* The unpruned full model (576 tokens) correctly reads `Toshiba`.

Under an aggressive retention budget of $K = 32$ (5.6% retention):

- **VisPruner** allocates tokens primarily to salient CRT screen reflections and casing boundaries, missing the bottom brand text and causing the downstream VLM to hallucinate `Lg`.
- **VisionSelector** distributes tokens over the top chassis and tabletop, omitting the logo and erroneously predicting `Samsung`.
- **READ** tightly concentrates its 32 tokens directly on the television's bottom bezel where the brand logo is located, preserving the exact textual evidence and correctly outputting `Toshiba`.

As the budget expands to $K = 64$ and $K = 128$, baselines gradually allocate tokens to peripheral regions, eventually covering the bottom bezel and recovering `Toshiba`. This progression highlights that under tight token allowances where selective precision is paramount, READ successfully prioritizes question-critical visual evidence over general visual saliency.

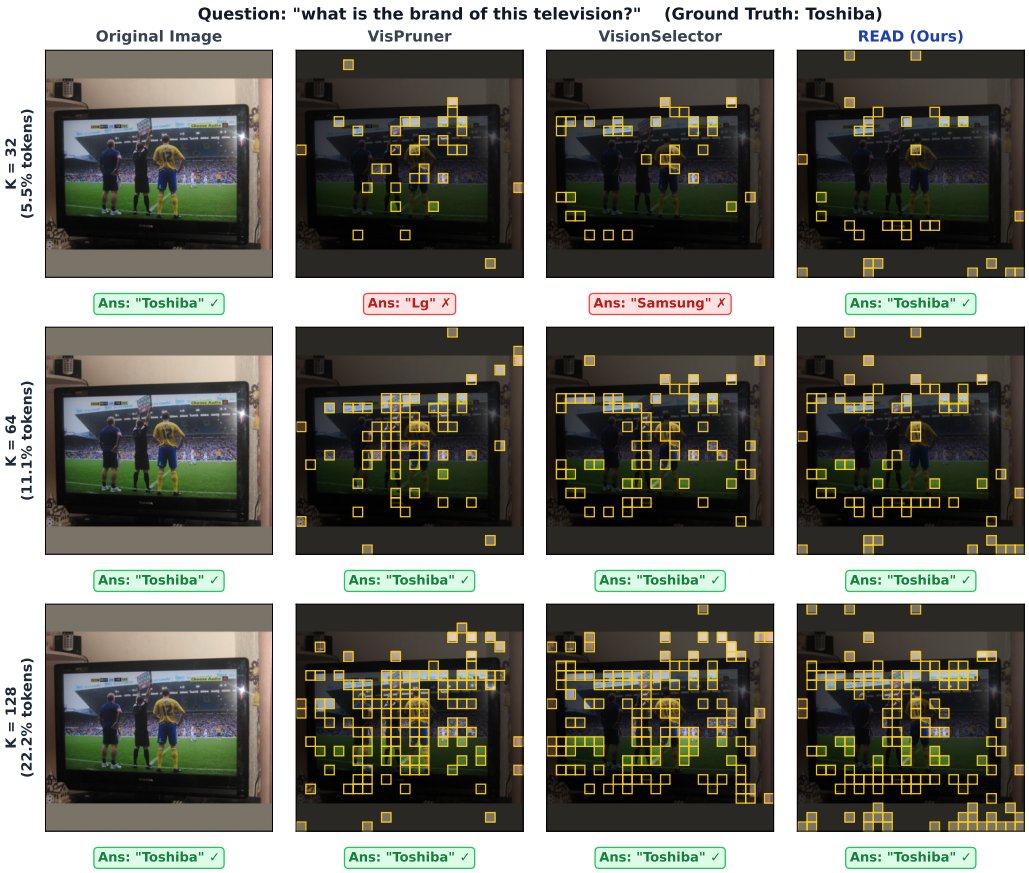


Figure 5: **Visual token selection across varying retention budgets on TextVQA.** The unpruned original image (576 tokens) is displayed on the left column. Retained visual tokens are highlighted with yellow borders, while unselected patches are dimmed. Under aggressive compression ($K = 32$, 5.6% retention), query-agnostic baselines (VisPruner and VisionSelector) scatter scarce token capacity across screen glare and television chassis, missing the brand lettering and hallucinating incorrect electronics brands (`Lg` and `Samsung`). In contrast, READ concentrates retained tokens directly on the brand logo along the bottom bezel, correctly reading `Toshiba` across all compression budgets.

H.2 Query-Dependent Token Selection on the Same Image

A fundamental limitation of query-agnostic selectors is their complete invariance to user intent: when distinct questions are posed regarding the same image, they inevitably transmit the identical token subset regardless of what evidence is required.

Figure 6 illustrates this limitation on a scene containing distinct foreground and background textual entities:

- **Question 1** targets the prominent foreground subject (*"What does his jacket say?"*, ground truth: `Fox`), situated on the person's chest.
- **Question 2** targets the background commercial signage (*"What brand is in white letters with a red background?"*, ground truth: `Coca-Cola`), located in the upper-left wall area.

Because **VisPruner** and **VisionZip** score patches purely from vision-encoder representations without question conditioning, their retained token subsets remain identical across both questions. For Question 1, both baselines allocate tokens to the visually dominant jacket and correctly answer `Fox`. However, when queried with Question 2 about the red background sign, their static token masks retain zero patches on the background signage, completely failing to capture the requisite evidence and hallucinating the foreground text `Fox`.



Figure 6: **Token selection on the identical image under distinct textual queries.** For query-agnostic selectors (VisPruner and VisionZip), retained visual tokens remain 100% identical across distinct prompts, causing severe hallucinations when question evidence lies outside dominant foreground regions. In contrast, READ dynamically redirects token allocation toward the respective evidence regions required by each question.

In contrast, **READ** dynamically adapts token allocation to prompt semantics: under Question 1, it allocates visual tokens to the foreground jacket lettering (yielding `Fox`); under Question 2, it dynamically redirects token capacity to capture the red background sign, enabling the cloud VLM to accurately read and predict `Coca cola`. This comparison concretely validates that READ reconfigures token coverage in response to user queries, overcoming the rigid failure mode of query-agnostic visual pruning.

I Limitations and Future Works

While READ demonstrates strong empirical retention accuracy and computational efficiency across diverse foundation backbones and multimodal benchmarks, several open challenges and promising directions remain for future exploration:

Algorithmic scope and physical communication dynamics. The core objective of this work is to establish a representation-level visual evidence retrieval framework that maximizes information density per transmitted token prior to network offloading. Our evaluation quantifies algorithmic token reduction (slashing transmitted visual tokens by 90.0%) and isolated execution latencies across edge and cloud stages. In practical distributed deployments, end-to-end communication latency is further subject to dynamic physical-layer variables, including wireless channel fading, adaptive modulation and coding (AMC) schemes, network packet overhead, and fluctuating uplink bandwidth. Because READ incurs an edge latency of only 13.2 ms and achieves an order-of-magnitude reduction in data payload (from 1,667 to 166 tokens on InternVL2.5-8B), it provides compact, information-dense inputs that naturally decouple from specific networking stacks. Integrating READ with channel-aware rate adaptation and specialized edge neural processing unit (NPU) pipelining represents a compelling systems-oriented avenue for future research.

Offline caching and cross-architecture generalization. READ requires caching cross-modal attention maps from the target cloud VLM to supervise the compact retriever. Crucially, this offline initialization is highly efficient: unlike methods requiring end-to-end backpropagation through multi-

billion-parameter foundation models, READ extracts attention maps in a single non-autoregressive forward pass and optimizes the retriever in isolation without loading the foundation backbone into GPU memory during gradient descent. Nevertheless, transferring the retriever to an entirely novel VLM architecture still necessitates collecting offline teacher attention profiles. Investigating cross-architecture attention projectors that generalize across heterogeneous VLM backbones without per-model attention caching offers an impactful research direction.

Multi-turn conversation and streaming video reasoning. Our current benchmark evaluation focuses on single-turn visual question answering and image-grounded multimodal reasoning, aligning with established visual token compression conventions. In embodied robotics, autonomous agents, and real-time interactive assistants, edge devices increasingly encounter continuous multi-turn dialogues and high-frame-rate streaming video. In such sequential settings, user intent evolves dynamically across turns, and visual streams exhibit strong spatio-temporal redundancy. Extending READ with cross-turn dialogue memory and temporal evidence tracking represents an exciting next step toward streaming multimodal compression.

Adaptive sample-wise budget allocation. Currently, READ operates under predetermined retention budgets K (e.g., retaining 10% or 25% of visual tokens), conforming to standard benchmarking protocols. In real-world multimodal workloads, however, queries vary substantially in visual difficulty: verifying object existence or recognizing prominent text requires only a handful of localized patches, whereas complex diagram parsing or fine-grained relational reasoning demands broader contextual coverage. Developing adaptive, entropy-guided or confidence-aware stopping rules that dynamically determine the optimal token budget on a per-sample basis could unlock higher average compression ratios while maintaining peak reasoning accuracy.