

When 1+1 Backfires: Rethinking Relevance-Diversity Balancing for Sample-Aware Visual Token Reduction

Yihang Huang
State Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
School of Computer Science and
Technology, University of Chinese
Academy of Sciences
Beijing, China
huangyihang25s@ict.ac.cn

Peisong Wen
School of Computer Science and
Technology, University of Chinese
Academy of Sciences
Beijing, China
wenpeisong@ucas.ac.cn

Chenhao Zhou
State Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
School of Computer Science and
Technology, University of Chinese
Academy of Sciences
Beijing, China
zhouchenhao25@mails.ucas.ac.cn

Difu Feng
State Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
School of Computer Science and
Technology, University of Chinese
Academy of Sciences
Beijing, China
fengdifu24@mails.ucas.ac.cn

Qianqian Xu[✉]
State Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing Academy of Artificial
Intelligence
Beijing, China
xuqianqian@ict.ac.cn

Abstract

Visual token compression has emerged as a critical technique for reducing the inference cost of Vision-Language Models. Existing methods select tokens based on text relevance, visual diversity, or a fixed combination of both—treating the relevance-diversity balance as a universal constant independent of the input. We conduct a systematic analysis to challenge this assumption. At the dataset level, grid-searching the optimal relevance-diversity weight per benchmark reveals substantial cross-task variation. More critically, we further maximize per-sample answer likelihood to recover sample-level oracle weights, finding that their distributions not only differ markedly across benchmarks but also exhibit high within-benchmark variance. Based on these observations, we propose SAVER, a framework for **Sample-Aware** relevance-diversity balanced Visual tokEn Reduction. We first derive a Geometric Intent Prior (GIP) from the spatial concentration of text-relevant tokens to infer a sample-specific relevance-diversity weight, and then apply this weight to jointly select tokens via a Weighted Text Centroid (WTC) for relevance scoring and a Visual Marginal Coverage Gain (VMCG) for diversity. SAVER requires no additional training and achieves sample-aware balancing entirely at inference time. Extensive experiments demonstrate that SAVER consistently outperforms state-of-the-art baselines across diverse benchmarks, retaining 96.5% of full-model performance on LLaVA-1.5-7B with

only 11.1% of the original visual tokens, and further generalizing to LLaVA-NeXT and Mini-Gemini.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Vision-Language Models; Efficient Inference; Visual Token Reduction

ACM Reference Format:

Yihang Huang, Peisong Wen, Chenhao Zhou, Difu Feng, and Qianqian Xu. 2026. When 1+1 Backfires: Rethinking Relevance-Diversity Balancing for Sample-Aware Visual Token Reduction. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3836590>

1 Introduction

Vision-Language Models (VLMs) [2, 5, 7, 14, 19, 22, 23, 25, 26, 30, 37], which jointly process visual and textual inputs, have achieved significant progress across a wide range of multimodal tasks, including visual question answering, image captioning, and document understanding. To faithfully encode image semantics, these models represent visual inputs as a large number of tokens. The resulting long token sequences directly inflate the input length, incurring $O(N^2)$ attention complexity in the Transformer [36] backbone and creating a critical computational bottleneck that limits the practical deployment of LVLMS.



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3836590>

To mitigate this bottleneck, visual token reduction [6, 10, 11, 16, 17, 24, 27, 33, 38, 39, 41–43] has emerged as a promising direction. Existing methods can be broadly grouped into three families. Text-relevance methods [6, 39] prioritize visual tokens that align strongly with the prompt, but often over-focus on a few high-response regions and discard peripheral context. Visual-diversity methods [3, 16] instead maximize coverage of the image content, but remain agnostic to the query and may waste token budget on irrelevant regions. More recent methods [27, 38, 42] attempt to integrate both objectives within a unified pipeline. However, they do not explicitly model the sample-dependent trade-off between relevance preservation and visual coverage. As observed in prior work [21], such input-agnostic balancing does not consistently outperform single-objective baselines, a phenomenon termed “ $1 + 1 < 1$.” These findings suggest that the central challenge lies not merely in combining relevance and diversity, but in estimating an appropriate per-sample allocation between the two objectives.

In this work, we investigate this phenomenon through a systematic two-level analysis (Section 3). Specifically, we formulate the token selection score as a weighted combination of text relevance and visual diversity, controlled by a scalar weight θ , and greedily select tokens in descending order of this score. At the dataset level, we find that each benchmark exhibits a clear optimal relevance-diversity weight that delivers noticeably better performance than alternative allocations, while the locations of these optima vary substantially across tasks. We further investigate the source of this variation at the instance level. Since answer correctness is discrete and does not provide a continuous optimization signal, we use per-sample answer likelihood as a surrogate objective to recover oracle weights for individual inputs. The recovered oracle weights further show that the sample-level preference is consistent with the relevance-diversity tendency observed at the dataset level. **These results suggest that effective token reduction fundamentally requires sample-aware relevance-diversity balancing.**

Closing this gap is non-trivial. A practical sample-aware token reduction method must satisfy two key requirements. 1). The prior must be derived from existing knowledge and the current input, without relying on any task-specific or dataset-level context unavailable at inference time. This is precisely the limitation of MoB [21]: although it provides a principled theoretical account of the relevance-diversity trade-off, its allocation strategy must be configured per dataset. In open-world deployment, however, the dataset identity of an incoming input is simply unknown. 2). The prior must be obtained within the early prefill stage from shallow semantics alone, without auxiliary modules or additional inference beyond the compression pipeline. Relying on an extra predictor, or waiting for more complete semantics from deeper layers, would introduce substantial pre-pruning cost and compromise the efficiency gains of token reduction, thereby undermining its very purpose as an inference acceleration technique.

Crucially, we observe that the spatial distribution of high-relevance visual tokens provides an effective cue for the desired relevance-diversity balance. When high-relevance tokens are spatially concentrated, the textual semantics are strongly aligned with a specific local region of the image, which is likely to contain the information most relevant to answering the query. When high-relevance

tokens are broadly dispersed, the model cannot localize truly relevant visual tokens with high confidence, making relevance scores less reliable as a basis for selection. In this case, a larger diversity weight is preferred to preserve broad visual coverage and maintain prediction accuracy.

Driven by these observations, we propose a Sample-Aware Relevance-Diversity Balanced Token Reduction framework. Its core component is the Geometric Intent Prior (GIP), a training-free prior that quantifies the text-visual coupling of each sample from the spatial covariance of its text-relevant visual tokens. GIP produces a continuous scalar that adaptively modulates the relative weight of each objective, enabling sample-specific objective balancing at negligible cost. Complementing GIP, we introduce two objective-specific improvements: a norm-weighted text centroid for text relevance estimation, free of cross-modal attention and compatible with FlashAttention [9]; and a facility location formulation for visual coverage whose marginal gain jointly accounts for novelty and representativeness. Extensive evaluations demonstrate that our method retains 96.5% of full-model performance using only 11% of the visual tokens, consistently outperforming existing approaches under equivalent compression budgets.

Our main contributions are summarized as follows:

- **Analysis for Relevance-Diversity Balancing.** We conduct a systematic two-level empirical analysis showing that the optimal relevance-diversity balance varies substantially across both tasks and individual samples, establishing that per-sample balancing of the relevance-diversity objectives is necessary for effective visual token compression.
- **Solution to Sample-Aware Objective Balancing.** We propose GIP, a training-free, efficient solution to sample-aware relevance-diversity balancing, which derives a sample-specific weight from the spatial distribution of text-relevant tokens.
- **SAVER Framework.** We propose a Weighted Text Centroid for text relevance scoring and a Visual Marginal Coverage Gain for diversity-aware token selection. By dynamically balancing these two objectives via GIP, our proposed framework SAVER achieves superior performance over competitive baselines across multiple benchmarks and generalizes to diverse model architectures.

2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (LVLMs) [5, 22, 25, 26, 37] have emerged from the foundational success of Large Language Models (LLMs) [1, 4, 8, 13, 35]. The standard LVLM pipeline consists of three core components. A visual encoder, typically a pre-trained vision transformer such as CLIP [32], extracts a sequence of visual tokens from the input image. A text tokenizer independently processes the natural language instruction. Both modalities are then projected into a shared embedding space and fed into a core LLM for joint multimodal reasoning. This architecture has achieved strong performance across a wide range of tasks, including visual question answering, image captioning, and document understanding. Despite these advances, faithfully encoding image semantics requires generating a large number of visual tokens. This directly inflates

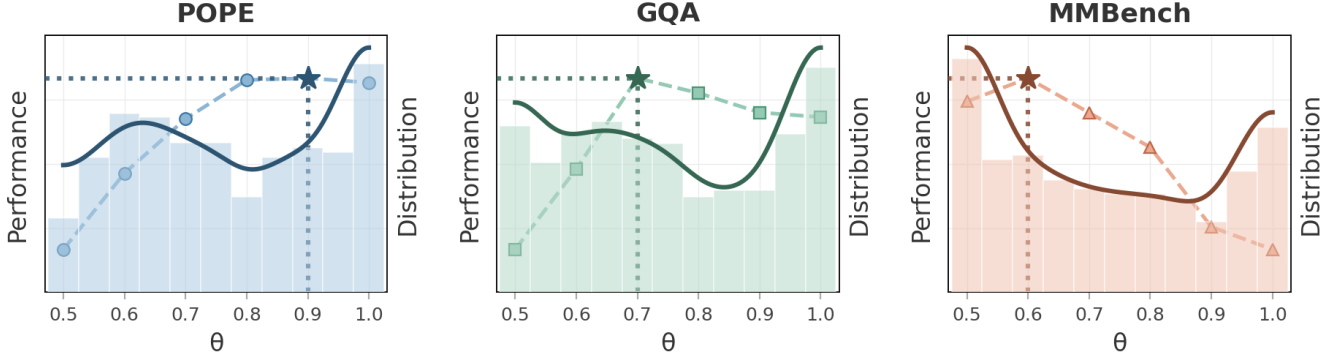


Figure 1: Dataset-level performance vs. θ (dashed line) and instance-level oracle weight distributions θ_i^* (solid curve and histogram) for POPE, GQA, and MMBench. POPE favors high θ (diversity-dominant); MMBench favors low θ (relevance-dominant); GQA peaks at intermediate values. Per-sample distributions closely correspond to dataset-level optima.

the input sequence length, which in turn substantially increases computational overhead and bottlenecks inference efficiency.

2.2 Token Reduction Methods

Existing token reduction methods largely fall into two categories: text-relevance and visual coverage. Text-relevance methods use cross-attention maps to select tokens that align with the text prompt. FastV filters tokens by raw attention scores [6], and PyramidDrop extends this into a progressive, layer-wise pruning scheme [39]. These methods quickly locate core semantic regions. Yet they tend to over-select tokens near these hubs and discard important peripheral context. Visual coverage methods instead maximize the semantic diversity of the selected token set. SAINT prunes via bipartite graph node degrees [16], and DivPrune formulates redundancy reduction as a Max-Min Diversity Problem [3]. These methods recover the global visual context more faithfully. However, they are task-agnostic. They do not consider the text instruction during selection, which can waste token budget on irrelevant regions.

Recent work has attempted to combine both objectives within a unified pipeline [38, 42]. Most such approaches use fixed weighting schemes that do not adapt to the input. In practice, the optimal balance between relevance and diversity varies considerably across different questions. Li et al. formalize this trade-off using ϵ -covering theory [21]. They show that multimodal datasets split into two regimes—weak and strong text-visual coupling—and that each regime requires different allocation parameters for best performance. A limitation of their method is that it requires knowing the target dataset in advance to set these parameters. This limits applicability in open-ended settings where the task intent is unknown at inference time.

3 Motivation

This section presents a systematic empirical study of the relevance-diversity trade-off in visual token compression. We first formulate the dataset-level optimal relevance-diversity weight and identify how it varies across benchmarks. To gain deeper insight into the origins of this cross-dataset variation, we further investigate the distribution of instance-level optimal weights within each benchmark.

Since no continuous metric directly reflects token selection quality at the sample level, we introduce per-sample answer likelihood as a surrogate objective to recover these instance-level weights. Based on both levels of analysis, we report empirical observations that motivate the design of our method.

Dataset-Level Optimal Weight. Given a multimodal input with visual token set $V = \{v_1, \dots, v_N\}$ and text token sequence T , we define the per-token selection score as a weighted combination of text relevance $\text{Rel}(\cdot, T)$ and visual diversity $\text{Div}(\cdot)$:

$$\text{Score}_\theta(v_j, T) = (1 - \theta) \cdot \text{Rel}(v_j, T) + \theta \cdot \text{Div}(v_j), \quad (1)$$

where $\theta \in [0, 1]$ controls the relevance-diversity balance, with larger θ corresponding to greater diversity weight. Under a fixed retention budget K , we select the top- K tokens by score:

$$V_\theta = \text{Top}_K \{ \text{Score}_\theta(v_j, T) \}_{j=1}^N. \quad (2)$$

For each benchmark $\mathcal{D}_k$ with task evaluation metric $\mathcal{M}$, the dataset-level optimal weight is defined as:

$$\theta_k^* = \arg \max_{\theta \in [0, 1]} \frac{1}{|\mathcal{D}_k|} \sum_{(x, Y) \in \mathcal{D}_k} \mathcal{M}(f_\Phi(V_\theta, T), Y), \quad (3)$$

where Y is the ground-truth answer.

Instance-Level Oracle Weight. To probe the distribution of optimal weights at the sample level, we define the *oracle weight* for each sample i as the θ that maximizes the likelihood of the correct answer under the LLM Φ . Formally, let $\mathcal{L}_i(\theta)$ denote the negative log-likelihood (NLL) of the ground-truth answer given V_θ and T :

$$\mathcal{L}_i(\theta) = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log P_{f_\Phi}(y_t | V_\theta, T, y_{<t}). \quad (4)$$

The instance-level oracle weight is then defined as:

$$\theta_i^* = \arg \min_{\theta \in [0, 1]} \mathcal{L}_i(\theta). \quad (5)$$

We recover θ_i^* for sampled instances from each benchmark $\mathcal{D}_k$ and analyze the resulting distributions.

Empirical Observations. As illustrated by the dashed lines in Figure 1, θ_k^* varies substantially across benchmarks: POPE [20] favors high θ (diversity-dominant), MMBench [28] favors low θ

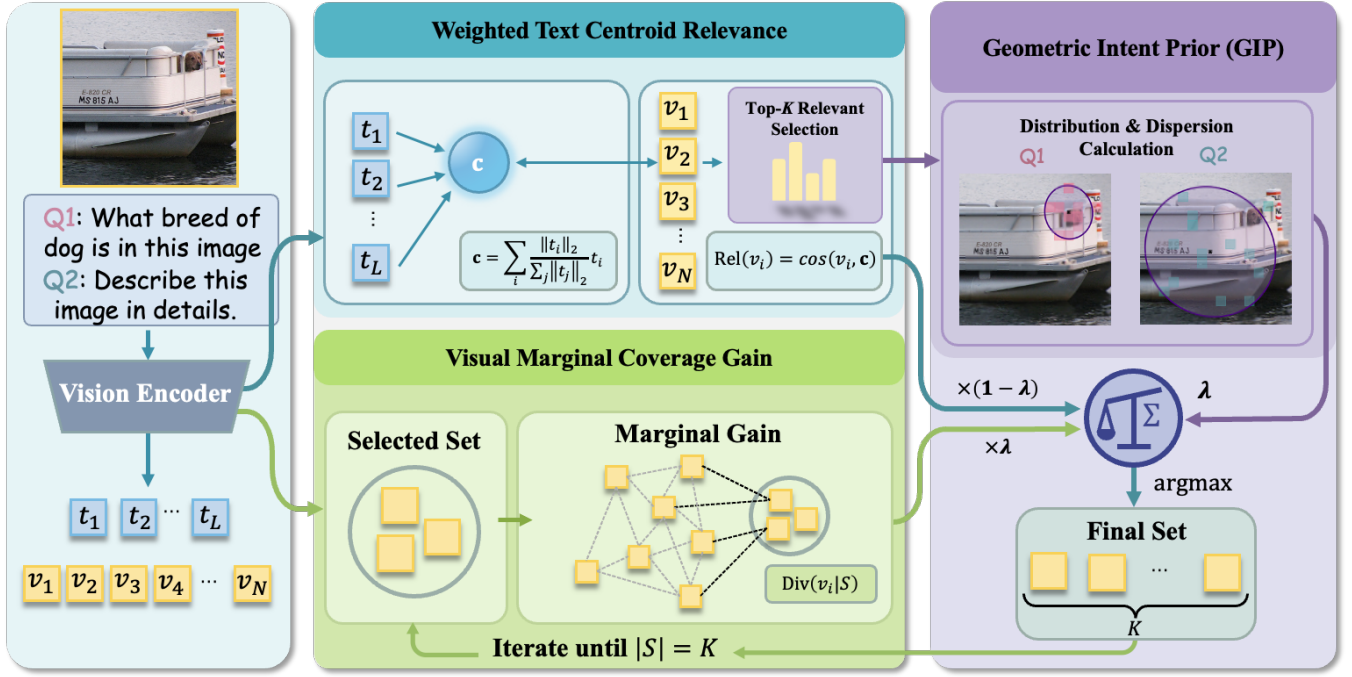


Figure 2: Overview of the SAVER framework. Given a multimodal input, we compute the Geometric Intent Prior (GIP) from the spatial distribution of high-relevance visual tokens, which dynamically weights text relevance (via WTC) and visual diversity (via VMCG) in a unified greedy selection process.

(relevance-dominant), and GQA [15] peaks at an intermediate value. The per-sample distributions $\{\theta_i^*\}$ are well aligned with their respective dataset-level optima, and this pattern is consistent with the intuitive characteristics of each task. POPE primarily probes object hallucination by querying small or absent objects, requiring the model to survey diverse regions across the visual field; MMBench queries fine-grained local details, where focused retrieval of the relevant region is decisive. Critically, both distributions also exhibit substantial within-benchmark variance, indicating that the optimal balance is not a property of the task alone but varies at the level of individual inputs.

4 Methodology

In this section, we present our Sample-Aware Unified Token Reduction framework. We first introduce a Geometric Intent Prior (GIP) in Section 4.2, which infers the sample-specific text-visual coupling from the spatial distribution of text-relevant visual tokens and adaptively balances the two selection objectives. We then describe our reformulated text relevance objective via a Weighted Text Centroid (WTC) in Section 4.3, and our visual diversity objective via a Visual Marginal Coverage Gain (VMCG) in Section 4.4. Finally, we present the unified token selection procedure in Section 4.5. The overall framework is illustrated in Figure 2.

4.1 Preliminaries

Formally, we denote the N visual tokens produced by the visual encoder as $V = \{v_i\}_{i=1}^N$ and the L text tokens from the text tokenizer

Algorithm 1 Sample-Aware Token Reduction

Require: $V = \{v_i\}_{i=1}^N$, $T = \{t_i\}_{i=1}^L$, budget K
Hyperparameters: γ , k , b
Ensure: S with $|S| = K$

- 1: $c \leftarrow \sum_i \frac{\|t_i\|_2}{\sum_j \|t_j\|_2} t_i$; $\text{Rel}(v_i) \leftarrow \cos(v_i, c) \forall i$
- 2: $\mathcal{G} \leftarrow \det(\text{Cov}(\{(x_i, y_i) : v_i \in \text{TopK}(V, \lfloor 0.02N \rfloor)\}))$
- 3: $\hat{\mathcal{G}} \leftarrow \mathcal{N}_{\mathcal{G}}(\mathcal{G})$
- 4: $k_{\text{GIP}} \leftarrow k \mathcal{N}_{\mathcal{D}}\left(1 - \frac{1}{N(N-1)} \sum_{p \neq q} \cos(v_p, v_q)\right)$
- 5: $\theta \leftarrow \text{sigmoid}(k_{\text{GIP}} \cdot \hat{\mathcal{G}} + b)$
- 6: $S \leftarrow \emptyset$
- 7: **while** $|S| < K$ **do**
- 8: $\text{Div}(v_i|S) \leftarrow -\gamma \max_{s \in S} K_{si} + (1-\gamma) \sum_{j \in S, j \neq i} \max(0, K_{ij} - \max_{s \in S} K_{sj})$, $\forall v_i \notin S$
- 9: $v^* \leftarrow \arg \max_{v_i \notin S} (1-\theta) \text{Rel}(v_i) + \theta \text{Div}(v_i|S)$
- 10: $S \leftarrow S \cup \{v^*\}$
- 11: **end while**
- 12: **return** S

as $T = \{t_i\}_{i=1}^L$. The concatenated sequence $[V; T]$ is fed into the LLM decoder f_{ϕ} , which autoregressively generates the response $y = f_{\phi}([V; T])$. Visual token reduction selects a subset $S \subseteq V$ with $|S| = K \ll N$ at an early layer of the LLM, so that the decoder operates on the compressed sequence $[S; T]$ instead, reducing prefilling cost.

Table 1: Comparison with state-of-the-art token reduction methods on LLaVA-1.5-7B across three compression budgets.

Method	GQA	MMBench	MME	POPE	ScienceQA	TextVQA	OCRBench	Avg.
<i>No compression (576 tokens)</i>								
LLaVA-1.5-7B	61.9	64.3	1862	85.9	69.5	58.2	297	100%
<i>192 tokens (66.7% reduction)</i>								
FastV (ECCV'24) [6]	52.7	61.2	1612	64.8	67.3	52.5	291	89.6%
SparseVLM (ICML'25) [42]	57.6	62.5	1721	83.6	69.1	56.1	292	96.3%
PyramidDrop (CVPR'25) [39]	57.1	63.2	1766	82.3	68.8	56.1	290	96.3%
DivPrune (CVPR'25) [3]	59.9	62.4	1752	87.0	68.9	57.0	303	98.3%
DART (EMNLP'25) [38]	60.0	63.6	1856	82.8	69.8	57.4	296	98.7%
VisionZip (CVPR'25) [40]	59.3	63.0	1783	85.3	68.9	57.3	–	97.7%
PruneSID (ICLR'26) [11]	60.1	63.7	1791	86.9	68.5	56.7	294	98.4%
SAVER (Ours)	61.2	63.8	1855	85.7	69.3	57.8	300	99.6%
<i>128 tokens (77.8% reduction)</i>								
FastV (ECCV'24) [6]	49.6	56.1	1490	59.6	60.2	50.6	285	83.8%
SparseVLM (ICML'25) [42]	56.0	60.0	1696	80.5	67.1	54.9	280	93.4%
PyramidDrop (CVPR'25) [39]	56.0	61.1	1644	82.3	68.3	55.1	287	94.2%
DivPrune (CVPR'25) [3]	59.2	62.3	1752	86.6	68.0	54.5	283	96.3%
DART (EMNLP'25) [38]	58.7	63.2	1840	80.1	69.1	56.4	296	97.3%
VisionZip (CVPR'25) [40]	57.6	62.0	1762	83.2	68.9	56.8	–	96.3%
PruneSID (ICLR'26) [11]	58.8	62.1	1749	86.5	68.3	54.7	275	95.9%
SAVER (Ours)	60.9	63.6	1811	85.5	69.2	57.2	287	98.4%
<i>64 tokens (88.9% reduction)</i>								
FastV (ECCV'24) [6]	46.1	48.0	1256	48.0	51.1	47.8	245	72.9%
SparseVLM (ICML'25) [42]	52.7	56.2	1505	75.1	62.2	51.8	180	82.8%
PyramidDrop (CVPR'25) [39]	41.9	33.3	1092	55.9	68.6	45.9	250	72.1%
DivPrune (CVPR'25) [3]	57.6	59.5	1638	85.8	68.6	52.9	275	93.7%
DART (EMNLP'25) [38]	55.9	60.6	1765	73.9	69.8	54.4	270	92.9%
VisionZip (CVPR'25) [40]	55.1	60.1	1690	77.0	69.0	55.5	–	92.9%
PruneSID (ICLR'26) [11]	57.1	58.8	1733	83.8	67.8	54.2	273	93.8%
SAVER (Ours)	60.0	61.8	1758	83.8	68.9	56.2	282	96.5%

4.2 Geometric Intent Prior

As established in Section 3, the optimal relevance-diversity balance is inherently sample-dependent, yet existing methods treat it as a fixed, input-agnostic hyperparameter. Accessing deep cross-modal representations from the LLM backbone could in principle provide rich cues for estimating the sample-specific balance, but doing so requires a full forward pass—the very computation that token reduction is designed to avoid. A practical solution must therefore rely on a proxy derivable from shallow representations in the early prefill stage, before the full depth of the LLM is engaged.

We propose to use the spatial distribution of high-relevance visual tokens as such a proxy. The key observation is that the spatial spread of these tokens reflects the degree to which the input query can be resolved by a localized region of the image. When high-relevance tokens are spatially concentrated, the query is grounded

in a specific local region, and a higher relevance weight is appropriate to retrieve detail-oriented tokens from that region. When they are broadly dispersed, the text tokens lack strong semantic alignment with any specific visual region, making relevance scores unreliable as a selection criterion; in this case, a higher diversity weight is warranted to ensure broad spatial coverage and allow the model to attend to global visual details. This spatial geometry is fully determined from shallow token representations within the LLM, requiring no additional training or full forward pass. We formalize this as the Geometric Intent Prior (GIP).

Formally, let $\mathcal{V} = \{v_i\}_{i=1}^N$ denote the set of visual tokens and let $\text{Rel}(v_i)$ denote the text relevance score of each token (defined in Section 4.3). We select the top- k most text-relevant tokens to form the anchor set:

$$\mathcal{V}_k = \{v_i \in \mathcal{V} \mid \text{Rel}(v_i) \in \text{TopK}(\{\text{Rel}(v_j)\}_{j=1}^N, k)\}. \quad (6)$$

We set $k = \lfloor 0.02N \rfloor$, i.e., the top 2% of visual tokens by relevance score. As validated in our experiments, tokens in this percentile exhibit significantly higher relevance scores than the remainder and carry rich semantic content, making them reliable indicators of the query’s visual grounding.

Each token $v_i \in \mathcal{V}_k$ corresponds to a spatial patch at 2D coordinate $\mathbf{p}_i = (x_i, y_i) \in \mathbb{R}^2$. Let $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^k$ denote the corresponding spatial point set. We compute the 2×2 spatial covariance matrix:

$$\Sigma = \frac{1}{k} \sum_{i=1}^k (\mathbf{p}_i - \bar{\mathbf{p}})(\mathbf{p}_i - \bar{\mathbf{p}})^\top, \quad (7)$$

where $\bar{\mathbf{p}} = \frac{1}{k} \sum_{i=1}^k \mathbf{p}_i$ is the spatial centroid of $\mathcal{P}$. We define the Geometric Intent Prior as its determinant:

$$\mathcal{G} = \det(\Sigma). \quad (8)$$

$\mathcal{G}$ equals the *generalized variance* of $\mathcal{P}$, measuring the area of the covariance ellipsoid spanned by the anchor tokens. A small $\mathcal{G}$ indicates that the high-relevance tokens are spatially concentrated; a large $\mathcal{G}$ indicates broad spatial dispersion. How $\mathcal{G}$ is mapped to the objective weighting scalar θ is described in Section 4.5.

4.3 Text Relevance via Weighted Text Centroid

Previous token reduction methods [6, 38, 42] estimate text relevance from cross-modal attention maps within the LLM, measuring the relevance of each visual token by its accumulated attention weight toward text tokens. However, two limitations hinder this approach in practice. First, modern inference pipelines increasingly adopt FlashAttention [9], which fuses the attention kernel and does not materialize the full attention matrix, rendering attention-based relevance estimation inapplicable. Second, LLM attention maps are susceptible to positional biases [42], where early tokens attract disproportionately high attention irrespective of their semantic content, introducing systematic noise into the relevance scores.

We instead propose to estimate text relevance via the semantic centroid of the text token embeddings, which is both attention-free and compatible with any inference backend. A naive unweighted mean, however, is susceptible to interference from function words and punctuation tokens, which contribute equally despite carrying little semantic content. Empirically, content words exhibit consistently higher L_2 norms in the embedding space than function words and punctuation. We exploit this property to define a norm-weighted centroid, termed Weighted Text Centroid (WTC), which naturally suppresses semantically weak tokens.

Formally, let $\mathcal{T} = \{t_i\}_{i=1}^L$ denote the text token embeddings. The normalized L_2 norm of each token serves as its weight:

$$w_i = \frac{\|t_i\|_2}{\sum_{j=1}^L \|t_j\|_2}. \quad (9)$$

The Weighted Text Centroid is defined as:

$$\mathbf{c} = \sum_{i=1}^L w_i t_i. \quad (10)$$

The text relevance score of each visual token v_i is then its cosine similarity with $\mathbf{c}$:

$$\text{Rel}(v_i) = \frac{v_i \cdot \mathbf{c}}{\|v_i\| \|\mathbf{c}\|}. \quad (11)$$

This formulation requires only a single inner product per visual token and introduces no dependency on attention scores or LLM intermediate states.

4.4 Visual Diversity via Visual Marginal Coverage Gain

Visual diversity in token selection has been approached in two main ways, each with a distinct failure mode. Max-min diversity methods [3] prioritize tokens that are maximally distant from the current selection, which makes them sensitive to outliers: a single token lying in a visually uninformative region can dominate the selection simply by being far from others. Alternatively, coverage-based methods favor tokens that best represent dense groups of similar visual patches, which tends to over-allocate budget to large homogeneous regions (e.g., background) and under-represent compact foreground structures rich in semantic detail.

Inspired by the facility location problem in combinatorial optimization, we propose a selection criterion that avoids both failure modes. In the facility location analogy, each unselected token is a “client” that must be served by the nearest selected token (“facility”). Adding a new token is only beneficial if it meaningfully improves coverage for clients that are currently underserved. This naturally penalizes both redundant selections and outliers.

Formally, let $\mathbf{K} \in \mathbb{R}^{N \times N}$ denote the cosine similarity matrix of the visual tokens, where each entry is defined as:

$$K_{ij} = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}. \quad (12)$$

Given the already-selected set S , the Visual Marginal Coverage Gain (VMCG) of a candidate token v_i is:

$$\begin{aligned} \text{Div}(v_i | S) = & \gamma \cdot \max \left(0, K_{ii} - \max_{s \in S} K_{si} \right) \\ & + (1-\gamma) \cdot \sum_{v_j \notin S, j \neq i} \max \left(0, K_{ij} - \max_{s \in S} K_{sj} \right). \end{aligned} \quad (13)$$

Since $K_{ii} = 1$ for cosine-normalized features, and dropping the additive constant γ , this reduces to:

$$\text{Div}(v_i | S) = \underbrace{-\gamma \max_{s \in S} K_{si}}_{\text{novelty}} + \underbrace{(1-\gamma) \sum_{v_j \notin S, j \neq i} \max \left(0, K_{ij} - \max_{s \in S} K_{sj} \right)}_{\text{representativeness}}, \quad (14)$$

where $\gamma \in (0, 1)$ balances novelty against representativeness. The novelty term measures how dissimilar v_i is from the closest token already in S . In the representativeness term, the $\max(0, \cdot)$ operator encodes a key principle from facility location: an unselected token v_j is only “served” by v_i if v_i is more similar to v_j than any token currently in S —if a better-matched facility already exists, v_j contributes nothing to the score of v_i . Consequently, $\text{Div}(v_i | S)$ is high only when v_i is both novel with respect to the current selection and genuinely representative of a large portion of tokens that remain poorly covered, naturally suppressing both redundant and outlier selections.

Table 2: Generalization results on LLaVA-NeXT.

Method	#Tokens	GQA	MMBench	POPE	MME	ScienceQA	SEEDBench	Avg.
LLaVA-NeXT (Full)	2880	64.2	67.9	86.4	1842	70.2	70.2	100%
VisionZip (CVPR'25) [40]	640	61.3	66.3	86.3	1787	68.1	66.7	97.0%
PruneSID (ICLR'26) [11]	640	61.6	64.2	86.3	1795	68.3	67.3	96.8%
SAVER (Ours)	640	63.6	65.8	86.4	1802	69.8	68.5	98.5%
VisionZip (CVPR'25) [40]	320	59.3	63.1	82.1	1702	67.3	63.4	93.2%
PruneSID (ICLR'26) [11]	320	60.5	63.0	83.1	1754	67.3	65.0	94.5%
SAVER (Ours)	320	62.5	65.0	85.0	1706	69.4	66.2	96.2%

Table 3: Generalization results on Mini-Gemini.

Method	#Tokens	GQA	MMBench	POPE	MME	ScienceQA	SEEDBench	Avg.
Mini-Gemini (Full)	576	62.4	69.3	85.8	1841	70.7	69.7	100%
VisionZip (CVPR'25) [40]	128	58.7	68.1	78.5	1841	70.0	65.6	96.2%
PruneSID (ICLR'26) [11]	128	60.1	66.6	82.4	1821	70.7	66.5	97.1%
SAVER (Ours)	128	60.6	67.2	84.1	1831	71.0	67.2	98.1%
VisionZip (CVPR'25) [40]	64	55.8	65.9	69.6	1737	70.7	61.7	91.4%
PruneSID (ICLR'26) [11]	64	58.3	63.1	76.0	1735	70.6	63.6	93.1%
SAVER (Ours)	64	58.7	64.8	81.0	1742	70.0	64.4	94.7%

4.5 Sample-Aware Token Selection

Given $\text{Rel}(v_i)$ and $\text{Div}(v_i | S)$ defined above, the unified selection score for each candidate token is:

$$\text{Score}(v_i | S) = (1 - \theta) \cdot \text{Rel}(v_i) + \theta \cdot \text{Div}(v_i | S), \quad (15)$$

Let $V = \{v_i\}_{i=1}^N$ be the visual tokens of image I . For a statistic X , define $\mathcal{N}_X(x) = \text{clip}((x - X_{\min}) / (X_{\max} - X_{\min}))$; hence $\tilde{\mathcal{G}} = \mathcal{N}_{\mathcal{G}}(\mathcal{G})$. Raw image diversity is defined as

$$\mathcal{D}(I) = 1 - \frac{1}{N(N-1)} \sum_{i \neq j} \frac{v_i^\top v_j}{\|v_i\|_2 \|v_j\|_2}. \quad (16)$$

Whereas GIP captures prompt-conditioned spatial variation among relevant tokens, $\mathcal{D}(I)$ measures intrinsic visual variation within the image. We normalize it as $\text{DIV}(I) = \mathcal{N}_{\mathcal{D}}(\mathcal{D}(I))$ and use it only to scale the GIP mapping:

$$k_{\text{GIP}} = k \cdot \text{DIV}(I). \quad (17)$$

The sample-specific weight is then

$$\theta = \text{sigmoid}(k_{\text{GIP}} \cdot \tilde{\mathcal{G}} + b). \quad (18)$$

The bounds $\mathcal{G}_{\min}$, $\mathcal{G}_{\max}$, $\mathcal{D}_{\min}$, $\mathcal{D}_{\max}$ used by $\mathcal{N}_{\mathcal{G}}$ and $\mathcal{N}_{\mathcal{D}}$ are estimated from a small held-out calibration subset; clipping prevents out-of-range samples from producing abnormal weights. Here, k and b are global linear coefficients in the sigmoid pre-activation that scale and shift $\tilde{\mathcal{G}}$ into an appropriate range before mapping to $\theta \in (0, 1)$.

At each greedy step, the token with the highest score is appended to S , and the diversity scores are updated incrementally. Selection proceeds until $|S| = K$; the full procedure is given in Algorithm 1.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate on a diverse set of image-based multi-modal benchmarks covering general visual understanding: GQA [15], MMBench [28], MME [12], POPE [20], ScienceQA [31], SEEDBench [18] and TextVQA [34]. We additionally include OCRBench [29] to assess OCR and document understanding. All benchmarks are evaluated under their standard protocols.

Baselines. We compare SAVER against representative token reduction methods spanning single-objective and multi-objective design strategies. Text-relevance-based methods include FastV [6] and PyramidDrop [39]. Visual-diversity-based methods include DivPrune [3], VisionZip [40], and PruneSID [11]. Methods combining both objectives include SparseVLM [42] and DART [38].

Model architectures. We validate SAVER primarily on LLaVA-1.5-7B [26] and further evaluate on LLaVA-NeXT-7B [25] and Mini-Gemini [22] to assess generalizability. All models are evaluated following their original inference configurations, without any additional training.

5.2 Main Results

LLaVA-1.5-7B. LLaVA-1.5-7B is a widely adopted open-source multimodal baseline that resizes input images to a fixed resolution and

Table 4: Efficiency comparison on LLaVA-1.5-7B.

Method	#Tok.	Peak GPU↓	Throughput↑	POPE↑
LLaVA-1.5-7B (Full)	576	14.69	11.21	85.9
SparseVLM [42]	64	18.84	12.09	75.1
DART [38]	64	14.33	17.17	73.9
SAVER (Ours)	64	14.33	17.11	83.8

Table 5: Ablation study on LLaVA-1.5-7B.

Method	POPE	GQA	MMBench	Avg.
WTC relevance	73.7	56.3	61.9	91.0%
VMCG diversity	83.4	58.7	59.9	95.0%
WTC + VMCG(static)	84.0	59.5	60.4	95.9%
SAVER (Ours)	83.8	60.0	61.8	96.9%

encodes them into a fixed set of 576 visual tokens via a CLIP vision encoder. We evaluate at three retention budgets $K \in \{64, 128, 192\}$. As shown in Table 1, SAVER ranks first under all three settings. At $K = 64$, SAVER achieves 96.5%, surpassing PruneSID and VisionZip by 2.7 and 3.6 points. At $K = 128$, SAVER attains 98.4%, maintaining a consistent lead over all baselines. At $K = 192$, SAVER retains 99.6% of full-model performance with negligible degradation.

LLaVA-NeXT. LLaVA-NeXT encodes both the full image and local crops at high resolution, yielding up to 2880 visual tokens and substantially stronger fine-grained perception than single-scale models. We evaluate at $K \in \{320, 640\}$ and report results in Table 2. SAVER achieves 98.5% at 640 tokens and 96.2% at 320 tokens, outperforming PruneSID (96.8%/94.5%) and VisionZip (97.0%/93.2%).

Mini-Gemini. Mini-Gemini employs a dual-encoder architecture that simultaneously extracts high- and low-resolution visual features and fuses them into 576 tokens, enabling richer multi-scale representations. We evaluate at $K \in \{64, 128\}$ and report results in Table 3. SAVER achieves 98.1% at 128 tokens and 94.7% at 64 tokens, consistently ahead of PruneSID (97.1%/93.1%) and VisionZip (96.2%/91.4%).

5.3 Efficiency Analysis

We compare inference efficiency against SparseVLM [42] and DART [38], both of which, like SAVER, apply token reduction during the LLM prefilling stage. We measure peak GPU memory, inference throughput, and POPE accuracy on LLaVA-1.5-7B at $K = 64$ retained tokens. As reported in Table 4, SAVER matches DART in both peak memory and throughput, while achieving a POPE score of 83.8 compared to DART’s 73.9. SparseVLM, despite consuming more GPU memory than the uncompressed model, attains only 75.1 on POPE. These results demonstrate that SAVER’s accuracy gains introduce no additional computational overhead.

5.4 Ablation Study

Component ablation. We ablate the core design choices of SAVER on LLaVA-1.5-7B at $K = 64$ tokens retained, as shown in Table 5. Using relevance alone yields the lowest average (91.0%); without diversity, the selected tokens cluster around query-relevant regions

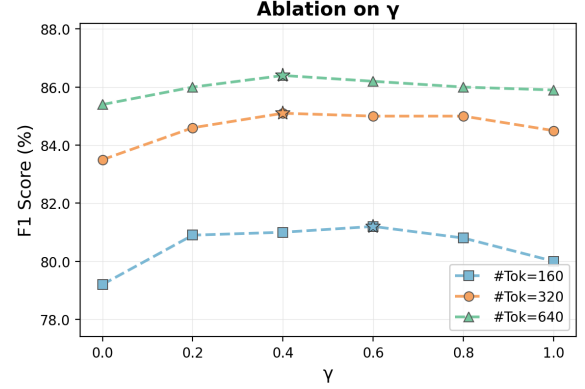


Figure 3: Ablation on γ across different token budgets on LLaVA-NeXT-7B (POPE F1). The effect is amplified under higher compression ratios.

and fail to cover the broader visual context, which is reflected in a notably low POPE score of 73.7. Diversity alone improves the average to 95.0% and raises POPE to 83.4 through wider spatial coverage, but the absence of text guidance leads to a drop in MMBench (59.9), where responses depend on precise localization of query-relevant details. The static combination improves the average to 95.9%, but remains below the sample-aware variant because a fixed relevance-diversity weight cannot accommodate the varying demands of different inputs. SAVER achieves the best average of 96.9% by dynamically inferring a sample-specific balance via GIP, confirming that adaptive, instance-level balancing is the key factor behind its consistent gains.

Discussion on the Novelty–Representativeness Trade-off. The novelty-representativeness trade-off coefficient γ requires more care. As shown in Figure 3, overly small γ values under-emphasize novelty and bias selection toward redundant regions, while overly large γ values over-prioritize outlier tokens at the expense of global coverage. Moderate values yield the best performance in practice. The optimal γ may also vary with the token retention budget.

6 Conclusion

We present a systematic empirical analysis of the relevance–diversity trade-off in visual token reduction, demonstrating that the optimal objective balance varies substantially across both tasks and individual inputs and cannot be captured by any fixed strategy. Motivated by this finding, we propose SAVER, a training-free framework centered on the Geometric Intent Prior (GIP), which infers a sample-specific relevance-diversity weight directly from the spatial distribution of high-relevance tokens. GIP is paired with a Weighted Text Centroid for attention-free relevance estimation and a Visual Marginal Coverage Gain for diversity-aware selection. Extensive experiments show that SAVER consistently outperforms competitive baselines across all compression budgets, while matching their inference efficiency, confirming that instance-level adaptive balancing is a principled approach to visual token reduction.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China: 62525212 and U23B2051, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680201, in part by the China National Postdoctoral Program for Innovative Talents under Grant BX20250377, in part by the China Postdoctoral Science Foundation under Grant 2026M791710, in part by the Beijing Major Science and Technology Project under Contract Z251100008125059, and in part by Beijing Academy of Artificial Intelligence (BAAI).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [3] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. DivPrune: Diversity-based Visual Token Pruning for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9392–9401.
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).
- [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. (2023).
- [6] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2025. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 19–35.
- [7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.
- [8] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) 2, 3 (2023), 6.
- [9] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [10] Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. 2026. MMTok: Multimodal Coverage Maximization for Efficient Inference of VLMs. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=GvPdSWZT31>
- [11] Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. 2026. Prune Redundancy, Preserve Essence: Vision Token Compression in VLMs via Synergistic Importance-Diversity. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=i36E5Ezm0H>
- [12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, et al. 2023. Mm: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394* (2023).
- [13] HIT-SCIR. 2024. Chinese-Mixtral-8x7B: An Open-Source Mixture-of-Experts LLM. <https://github.com/HIT-SCIR/Chinese-Mixtral-8x7B>.
- [14] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. 2024. ReconBoost: Boosting Can Achieve Modality Reconciliation. In *International Conference on Machine Learning*. 19573–19597.
- [15] Drew A Hudson and Christopher D Manning. 2019. 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*. 6700–6709.
- [16] Ahmadreza Jeddi, Negin Baghbanzadeh, Elham Dolatabadi, and Babak Taati. 2025. Similarity-aware token pruning: Your vlm but faster. *arXiv preprint arXiv:2503.11549* (2025).
- [17] Pengfei Jiang, Hanjun Li, Linglan Zhao, Fei Chao, Ke Yan, Shouhong Ding, and Rongrong Ji. 2025. VISA: Group-wise Visual Token Selection and Aggregation via Graph Summarization for Efficient MLLMs Inference. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) (MM '25). Association for Computing Machinery, New York, NY, USA, 11130–11139. doi:10.1145/3746027.3755792
- [18] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* (2023).
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 292–305. doi:10.18653/v1/2023.emnlp-main.20
- [21] Yangfu Li, Hongjian Zhan, Tianyi Chen, Qi Liu, and Yue Lu. 2025. Why 1 + 1 < 1 in Visual Token Pruning: Beyond Naïve Integration via Multi-Objective Balanced Covering. (2025). arXiv:2505.10118 [cs.CV]
- [22] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2026. Mini-Gemini: Mining the Potential of Multi-Modality Vision Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 48, 3 (2026), 3530–3543. doi:10.1109/TPAMI.2025.3637265
- [23] Zixu Li, Yupeng Hu, Zhiwei Chen, Haokun Wen, Xuemeng Song, and Liqiang Nie. 2026. COMBINER: Composed Image Retrieval Guided by Attribute-based Neighbor Relations. *IEEE TIP* (2026).
- [24] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5334–5342.
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914fa369fe6de0-Paper-Conference.pdf
- [27] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjin Yin, and Linfeng Zhang. 2024. Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model. *arXiv preprint arXiv:2411.10803* (2024).
- [28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024. Mmbench: Is your multi-modal model an all-around player?. In *European conference on computer vision*. Springer, 216–233.
- [29] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. OCRBench: on the hidden mystery of OCR in large multimodal models. *Science China Information Sciences* 67, 12 (Dec. 2024). doi:10.1007/s11432-024-4235-6
- [30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525 [cs.AI]
- [31] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [33] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. LLaVA-PrüMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. In *ICCV*.
- [34] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8317–8326.
- [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

- [37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, and Wenbin Ge. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. (2024).
- [38] Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025. 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 9961–9980. doi:10.18653/v1/2025.emnlp-main.505
- [39] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247* (2024).
- [40] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19792–19802.
- [41] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. 2025. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. *arXiv preprint arXiv:2506.10967* (2025).
- [42] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 74840–74857. <https://proceedings.mlr.press/v267/zhang25s.html>
- [43] Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2025. Don't Just Chase "Highlighted Tokens" in MLLMs: Revisiting Visual Holistic Context Retention. *arXiv preprint arXiv:2510.02912* (2025).