

From Inheritance to Saturation: Disentangling the Evolution of Visual Redundancy for Architecture-Aware MLLM Inference Acceleration

Jiaqi Shi^{1*} and Xulong Zhang³ and Yuechan Li^{2†} and Xiaoyang Qu³ and Jianzong Wang^{3‡}

¹University of Science and Technology of China, Hefei, China

²Wuhan University, Wuhan, China

³Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China

Abstract

High-resolution Multimodal Large Language Models (MLLMs) face prohibitive computational costs during inference due to the explosion of visual tokens. Existing acceleration strategies, such as token pruning or layer sparsity, suffer from severe "backbone dependency", performing well on Vicuna or Mistral architectures (e.g., LLaVA) but causing significant performance degradation when transferred to architectures like Qwen. To address this, we leverage truncated matrix entropy to uncover a universal three-stage inference lifecycle, decoupling visual redundancy into universal Intrinsic Visual Redundancy (IVR) and architecture-dependent Secondary Saturation Redundancy (SSR). Guided by this insight, we propose HalfV, a framework that first mitigates IVR via a unified pruning strategy and then adaptively handles SSR based on its specific manifestation. Experiments demonstrate that HalfV achieves superior efficiency-performance trade-offs across diverse backbones. Notably, on Qwen25-VL, it retains 96.8% performance at a $4.1\times$ FLOPs speedup, significantly outperforming state-of-the-art baselines. Our code is available at <https://github.com/civilizwa/HalfV>.

1 Introduction

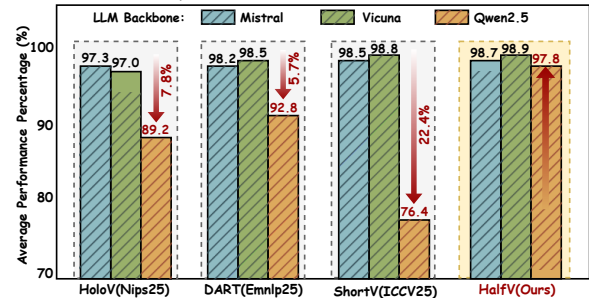
Multimodal Large Language Models (MLLMs) have witnessed remarkable advancements (Liu et al., 2023; Chen et al., 2023; Wang et al., 2024; Liu et al., 2024a; Bai et al., 2025). However, their practical deployment is severely impeded by prohibitive computational costs. Specifically, under high-resolution settings, the input sequence becomes overwhelmingly dominated by visual tokens encoded by the Vision Transformer (ViT) (Dosovitskiy et al., 2021). Given the quadratic

*Core Contributor

†Significant Contributor

‡Corresponding Author

(a) Performance comparison across different LLM Backbones.



(b) A universal three-stage lifecycle across different LLM backbones.

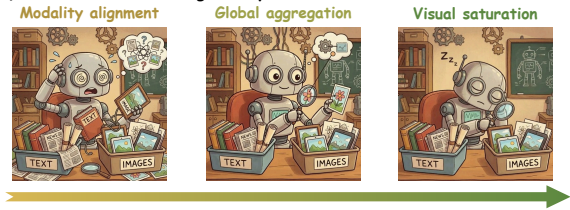


Figure 1: (a): Comparison of token-level methods (HoloV, DART) at a 77.8% pruning ratio and the layer-level method (ShortV) across Vicuna, Mistral and Qwen backbones. Results represent the average relative performance compared to the baseline (%) across POPE, MME, MMBench, and SQA datasets; (b): We find that models with different backbones all exhibit a universal three-stage lifecycle: Modality alignment, global aggregation and visual saturation.

$O(N^2)$ complexity of self-attention, this visual token explosion incurs a massive computational burden primarily during the prefill stage, thereby creating a critical bottleneck for latency-sensitive applications.

Current acceleration efforts, encompassing token-level visual token pruning (Chen et al., 2024a; Shang et al., 2024; Zhang et al., 2024b; Tong et al., 2025; Ye et al., 2025; Xing et al., 2024; Endo et al., 2025) and layer-level layer sparsity (Yuan et al., 2025; Wu et al., 2024), have attempted to mitigate this cost. However, we observe that these approaches exhibit severe *backbone dependency*, often being overfitted to Vicuna or Mistral architectures (e.g., LLaVA series (Liu et al., 2024a,b)) while neglecting the heterogeneity

of underlying LLM backbones. As shown in Figure 1, transferring these strategies to the Qwen2.5 backbone results in performance degradation ranging from 5.7% to 22.4%. Crucially, by utilizing LLaVA-Next (Liu et al., 2024b) which employs dynamic resolution as a control variable, we rule out the possibility that this failure stems from the visual front-end (More details see Appendix B.2). While LLaVA-Next remains robust to existing pruning methods, Qwen2.5-VL exhibits distinct sensitivity. **This firmly establishes that the bottleneck lies in the intrinsic mechanism by which different LLM backbones process visual information.**

To decode this mechanism and break the architecture barrier, we employ *truncated matrix entropy* (Zhang et al., 2024a; Xiong et al., 2025) as a probe to systematically trace the evolution of visual information. Our analysis uncovers a universal three-stage lifecycle across architectures: I. Modality Alignment, II. Global Aggregation and III. Visual Saturation. Based on this evolution, we categorize the observed redundancies into two distinct types:

(1) Intrinsic Visual Redundancy (IVR): Dominating Stage I, this redundancy stems from ViT’s dense tokenization, where highly correlated, spatially adjacent patches are mapped into the LLM space with minimal interaction.

(2) Secondary Saturation Redundancy (SSR): Emerging in Visual Saturation (Stage III), this redundancy is a direct byproduct of Global Aggregation (Stage II). After the LLM aggregates dispersed visual evidence into key semantic regions, the deep layers reach a state of semantic saturation where additional computation yields diminishing information gain. Notably, while IVR is universal, the physical manifestation of SSR is architecture-dependent: it appears as layer-level inactivity in Vicuna/Mistral backbones but as extreme token sparsity in Qwen backbones.

Guided by the distinct mechanisms, we propose HalfV, an architecture-aware acceleration framework that decouples redundancy reduction into two steps. Specifically, HalfV targets the universal IVR with a unified pruning strategy applicable across models, while addressing the architecture-dependent SSR with an adaptive reduction mechanism that tailors the acceleration paradigm to each backbone’s unique saturation manifestation. By disentangling this complex evolution process, HalfV achieves a superior efficiency-performance

frontier. Extensive experiments across diverse architectures validate the effectiveness of our method, providing a principled perspective for designing future universal MLLM acceleration.

Our contributions are summarized as follows:

- **Unveiling Backbone Heterogeneity:** We conduct a systematic evaluation of existing acceleration strategies and identify a critical "backbone dependency". We find that this dependency stems from intrinsic LLM processing mechanisms rather than visual front-end differences.
- **Decoupling Redundancy Mechanisms:** Utilizing an entropy-based probe, we decouple visual redundancy into universal Intrinsic Visual Redundancy (IVR) and architecture-dependent Secondary Saturation Redundancy (SSR).
- **HalfV Framework:** We propose HalfV, an architecture-aware framework that aligns with this redundancy evolution. It employs a unified pruning strategy for the universal IVR and an adaptive reduction mechanism targeting the specific SSR manifestation of each backbone, achieving a superior efficiency-performance frontier across diverse architectures.

2 Related Work

2.1 Multimodal Inference Acceleration

Our work focuses on inference acceleration during the prefill stage of MLLMs. From a redundancy perspective, existing methods fall into two categories: (i) Token-level redundancy, which retains a subset of visual tokens via heuristic scoring rules (Chen et al., 2024a; Zhang et al., 2024b; Endo et al., 2025; Yang et al., 2025b; Xing et al., 2024; Wen et al., 2025; Alvar et al., 2025). (ii) Layer-level redundancy. ShortV (Yuan et al., 2025) shortens the effective inference path by suppressing visual state updates in layers with low contribution to the final output. VTW (Lin et al., 2025) assumes that visual information has been transferred to text in deeper layers, and thus removes all visual tokens in deeper layers for higher speed.

2.2 Internal Evolution of Representations

Understanding the internal evolution of representations has been widely recognized as a prerequisite for model optimization. Prior works analyzed internal signals from three perspectives: (i) intermediate representations (Jiang et al., 2025), using probes or classifiers to evaluate how layers

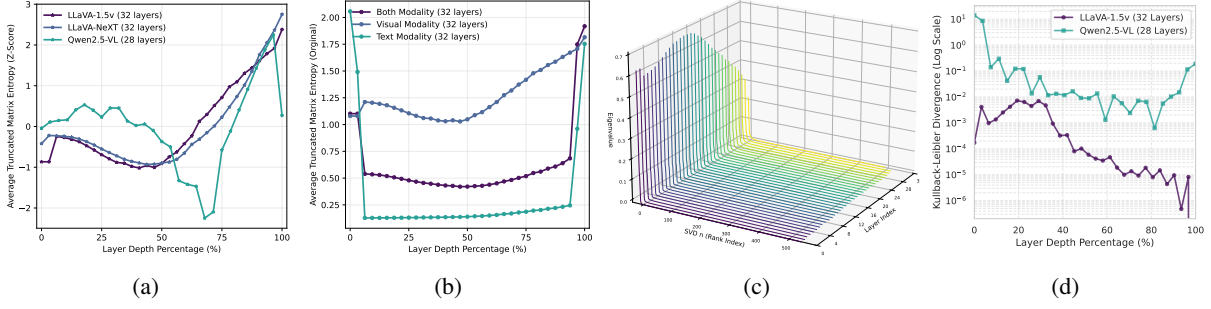


Figure 2: (a) The average truncated matrix entropy for different architectures on GQA dataset. The x-axis is the depth percentage of the layer; (b) The average truncated matrix entropy for different types of tokens on LLaVA-1.5v-7B. Additional results see Appendix B.2; (c) 3D Line Plot of the Average Spectrum of \mathcal{G} across Layers on LLaVA-1.5v-7B. Additional results see Appendix B.1; (d) KL divergence for Vicuna and Qwen2.5 backbones on AI2D dataset.

encode objects, relations, and semantics; (ii) activation patterns and channel sparsity (Chen et al., 2024b), to understand information flow; (iii) attention mechanism, with Wu et al. (2024) summarizing MLLM inference into three stages and using a controller to dynamically discard visual tokens for speedup. However, it relies on heuristic stage boundaries and introduces additional training costs and complexity. In contrast, our work provides a principled foundation for stage-aware acceleration without auxiliary training or controllers.

3 Method

In this section, we first introduce truncated matrix entropy and use it as a unified probe to track redundancy across inference depth. Building on these observations, we present **HalfV**, a two-step acceleration framework that aligns acceleration strategies with redundancy evolution.

3.1 Preliminary: Truncated Matrix Entropy

In each transformer layer, a set of tokens from a given modality is represented by high-dimensional hidden states. Let $\mathbf{h}_i \in \mathbb{R}^D$ denote the hidden states of the i -th token at a certain layer, where D is the hidden dimension. For a group of N tokens, we stack its hidden states into a representation matrix $\mathcal{Z} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times D}$, where N is the number of tokens. The Gram matrix \mathcal{G} is then computed from \mathcal{Z} as:

$$\mathcal{G} = \begin{cases} \mathcal{Z}^\top \mathcal{Z}, & \text{if } N \geq D, \\ \mathcal{Z} \mathcal{Z}^\top, & \text{if } N < D, \end{cases} \quad (1)$$

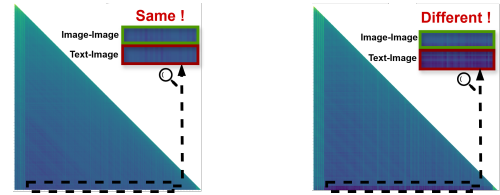
where $\mathcal{G} \in \mathbb{R}^{D \times D}$ or $\mathbb{R}^{N \times N}$ accordingly. The eigenvalue spectrum of \mathcal{G} reflects the effective dimensionality of the representation space. As illus-

trated in Figure 2c, the spectrum typically exhibits an elbow point (Cangelosi and Goriely, 2007), beyond which eigenvalues contribute marginally to the overall variance.

Following prior work (Xiong et al., 2025), we retain the top- k eigenvalues $\{\lambda_1(\mathcal{G}), \dots, \lambda_k(\mathcal{G})\}$ before the elbow point to suppress noise and redundancy. Let \mathcal{G}_k denote the rank- k truncated Gram matrix (keeping the top- k eigenpairs), and define the truncated trace as $\text{tr}_k(\mathcal{G}) := \sum_{j=1}^k \lambda_j(\mathcal{G}) = \text{tr}(\mathcal{G}_k)$. We then define the truncated matrix entropy $\mathcal{H}(\mathcal{Z})$ as:

$$\mathcal{H}(\mathcal{Z}) = - \sum_{i=1}^k p_i \log p_i, \quad p_i := \frac{\lambda_i(\mathcal{G})}{\text{tr}_k(\mathcal{G})} = \frac{\lambda_i(\mathcal{G})}{\text{tr}(\mathcal{G}_k)}. \quad (2)$$

More details are provided in Appendix B and C.



(a) Initial Layer of Stage I (b) Final Layer of Stage I

Figure 3: Comparison of attention heatmaps at the initial and final layers of Stage I on LLaVA-1.5v-7B.

3.2 Unveiling the Redundancy Evolution Lifecycle

By tracking the trajectory of Truncated Matrix Entropy $\mathcal{H}(\mathcal{Z})$ across layers, we uncover a consistent three-stage redundancy lifecycle inherent to MLLM inference. This evolution delineates the distinct origins and manifestations of visual redundancy.

Stage I: Modality Alignment. In early layers, we observe asymmetric entropy dynamics: visual

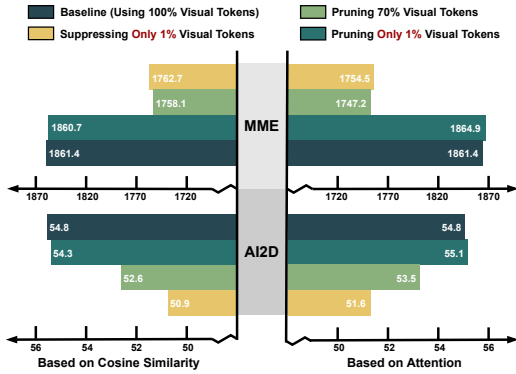


Figure 4: **Suppressing only 1% of visual tokens leads to a performance comparable to pruning 70% of them in Stage II.** This experiment is conducted on the MME (top) and AI2D (bottom) datasets, using cosine similarity (left) and attention scores (right) as criteria. For comparison, we also report the baseline results and the performance when pruning only 1% of visual tokens.

entropy remains consistently high, whereas textual entropy rapidly compresses (see Figure 2b). Concurrently, attention shifts from balanced to text-dominated (see Figure 3), indicating active feature alignment. Crucially, this high-entropy visual state implies that tokens retain their raw, uncompressed structure. Consequently, the sequence is dominated by **Intrinsic Visual Redundancy (IVR)**, which refers to the spatially correlated signals *inherited* from the ViT’s dense tokenization.

Stage II: Global Aggregation. As inference proceeds, the decrease in visual entropy marks the onset of global information aggregation. Our empirical analysis reveals two critical properties of this stage:

(i) *High Sensitivity to Local Perturbation.* We find that suppressing visual state updates for even a tiny fraction of tokens during this stage causes severe performance degradation (see Figure 4). This confirms that the aggregation process is globally coupled, meaning any local interruption disrupts the overall integration pathway.

(ii) *Optimal Timing for One-Shot Pruning.* Given the high sensitivity, we seek the safest intervention point by evaluating the efficiency-performance of pruning. We formalize this via Marginal Utility (MU), defined as the ratio between performance drop $\Delta\mathcal{M}$ and latency gain $\Delta\mathcal{C}$:

$$\text{MU}_{l,r}^{\downarrow} = \frac{-\Delta\mathcal{M}_{l,r}}{\Delta\mathcal{C}_{l,r} + \epsilon}, \quad (3)$$

As detailed in Table 1, minimizing this metric re-

Pruning Layer	Stage	MU \downarrow	Latency Gain \uparrow
Early Layer	I	0.87	46.8%
Stage II Start	II	0.21	41.5%
Stage II Mid	II	0.29	37.8%
Deep Layer	III	0.65	22.4%

Table 1: Marginal utility (MU) of one-shot token pruning at different layers.

Setting	GQA	OCRBench	ChartQA
LLaVA-1.5v-7B	61.9	275	18.2
(a) Withdraw	54.7(-11.6%)	52(-81.1%)	13.6(-25.3%)
(b) Suppress	60.9(-1.6%)	311(+13.1%)	17.0(-6.5%)
Qwen25-VL-7B	60.7	848	84.0
(c) Withdraw	59.7(-1.6%)	698(-17.6%)	75.4(-10.2%)
(d) Suppress	41.2(-32.1%)	117(-86.2%)	67.2(-20.0%)
(e) 5% Tokens	60.6(-0.1%)	827(-2.4%)	83.6(-0.4%)

Table 2: **Comparative experiments on the LLaVA-1.5v-7B and Qwen25-VL-7B models.** We evaluate their performance on the GQA, OCRBench, and ChartQA datasets.

veals that the onset of Stage II offers the optimal trade-off. Thus, Stage II prohibits layer-by-layer intervention but defines the precise window for one-shot IVR elimination at its boundary. More details are provided in Appendix D and E.

Stage III: Visual Saturation. In deep layers, the visual context becomes saturated, inducing **Secondary Saturation Redundancy (SSR)**. Unlike the universal IVR, the physical manifestation of SSR is highly **architecture-dependent**, branching into two distinct patterns:

(i) *Layer-level Inactivity (Vicuna/Mistral).* For architectures like LLaVA, we observe consistently low Kullback-Leibler divergence ($\mathcal{D}_{KL} \approx 0$) in deep layers (Figure 2d), implying minimal information gain. Experimentally, suppressing all visual updates maintains baseline performance (e.g., +13.1% on OCRBench). Here, SSR manifests as entire layers becoming redundant, allowing for static KV reuse. More details are provided in Appendix F.

(ii) *Extreme Token Sparsity (Qwen).* In contrast, Qwen maintains high divergence, and total suppression causes catastrophic failure (e.g., -86.2% on OCRBench). However, we find that the effective information flow collapses onto a minimal subset of dominant tokens \mathcal{S}_{top} . Restricting computation to just this top 5% subset ($\mathcal{S}_{top} \subset \mathcal{V}$) restores near-lossless performance. This indicates that while layers remain active, the redundancy manifests as extreme sparsity within the token sequence.

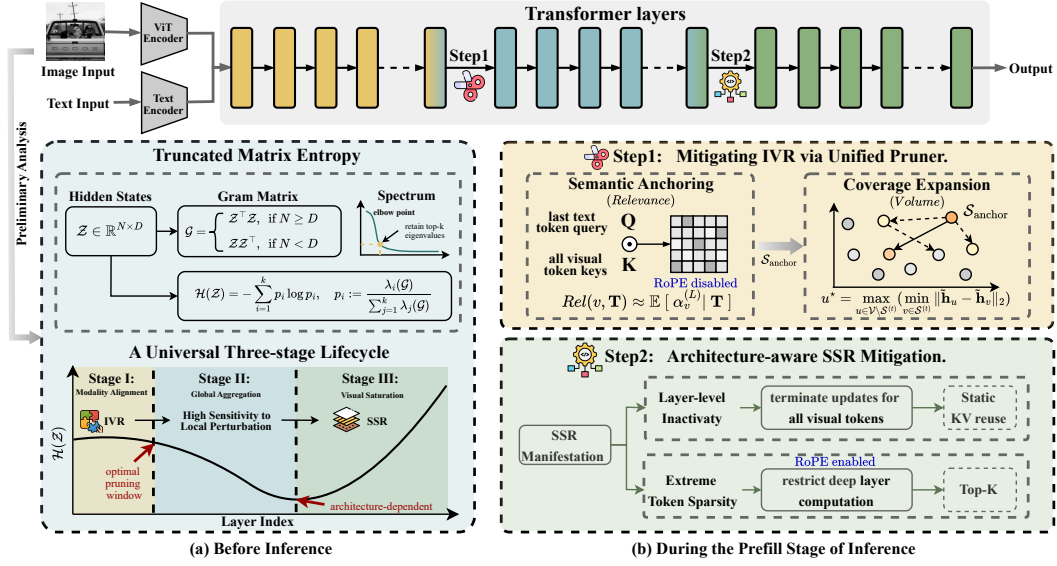


Figure 5: **Overview of HalfV.** (a) Preliminary analysis. We use a small subset of the dataset (100 samples) to identify the onset of the three internal stages in the LLM. (b) Prefill acceleration. Implementation details of our two-step acceleration strategy in the LLM prefill stage.

3.3 HalfV

Based on the observations, we propose HalfV, a two-step inference acceleration method. The overview of HalfV is illustrated in Figure 5.

Step1: Mitigating IVR via Unified Pruner. We perform token pruning at the boundary of Stage I and Stage II. Our goal is to select a subset \mathcal{S} that preserves the cross-modal semantic alignment established in Stage I while mitigating IVR. Crucially, pruning should avoid representational collapse so that the retained tokens provide a stable basis for Stage II global aggregation. Given visual tokens \mathcal{V} and a budget K , we denote the optimal subset by \mathcal{S}^* and formulate:

$$\mathcal{S}^* = \arg \max_{\mathcal{S} \subset \mathcal{V}, |\mathcal{S}|=K} \left(\sum_{v \in \mathcal{S}} \text{Rel}(v, \mathbf{T}) + \lambda \cdot \text{Vol}(\mathcal{S}) \right). \quad (4)$$

The first term $\sum_{v \in \mathcal{S}} \text{Rel}(v, \mathbf{T})$ aggregates token-wise relevance to the textual condition \mathbf{T} , while the second term $\text{Vol}(\mathcal{S})$ promotes geometric spread (coverage) in feature space to prevent degeneration; $\lambda > 0$ balances the two. Since Eq. (4) is NP-hard, we propose **AnchorCover**, a greedy solver that optimizes the two terms sequentially.

(1) **Relevance Maximization (Semantic Anchoring).** We use cross-modal attention $\alpha^{(L)}$ as a utility proxy where L is the pruning layer, i.e., $\text{Rel}(v, \mathbf{T}) \approx \mathbb{E}[\alpha_v^{(L)} | \mathbf{T}]$. For hardware efficiency and conditioned relevance, we compute cross-attention using only the query of the last textual token against the keys of all visual tokens. The

last token summarizes the instruction, and the computation reduces to a single vector product compatible with FlashAttention (Dao et al., 2022). As shown in Tab. 7, this module adds negligible overhead (0.7% TTFT). To reduce positional bias, we disable position embeddings (Su et al., 2024) when computing scores. We then form $\mathcal{S}_{\text{anchor}}$ by selecting the Top- K_S tokens.

(2) **Volume Maximization (Coverage Expansion).** With the relevant anchors fixed, we further aim to maximize the incremental geometric coverage $\Delta \text{Vol}(\mathcal{S})$. It is a known geometric property that the volume of a simplex is maximized when its vertices are mutually distant. Therefore, we employ *Farthest Point Sampling (FPS)* (Gonzalez, 1985) as a greedy surrogate. Let $\tilde{\mathbf{h}}_v$ denote the ℓ_2 -normalized hidden state of token v . At each step, we select the token u^* that is most distant from the current set $\mathcal{S}^{(t)}$:

$$u^* = \arg \max_{u \in \mathcal{V} \setminus \mathcal{S}^{(t)}} \left(\min_{v \in \mathcal{S}^{(t)}} \|\tilde{\mathbf{h}}_u - \tilde{\mathbf{h}}_v\|_2 \right). \quad (5)$$

We iterate u^* into $\mathcal{S}^{(t+1)}$ until $|\mathcal{S}^{(t)}| = K$, obtaining a coverage set $\mathcal{S}_{\text{cover}}$. The final selection

$$\mathcal{S}^* = \mathcal{S}_{\text{anchor}} \cup \mathcal{S}_{\text{cover}} \quad (6)$$

By prioritizing tokens that are nearly orthogonal to the existing anchors, this step effectively mitigates the spatial redundancy while ensuring the final representation forms a robust basis for Stage II.

Step2: Architecture-aware SSR Mitigation. We adopt differentiated strategies based on SSR

Method	TFLOPS	GQA	MME	POPE	SQA	VQA ^{text}	VizWiz	MMB ^{en}	AI2D	Average
<i>LLaVA-1.5v-7B (Liu et al., 2023), Backbone: Vicuna</i>										
Vanilla	8.31	62.0	1859	85.9	70.4	58.2	54.4	64.8	54.8	100.0%
FastV (Chen et al., 2024a)	3.48	57.6	1730	81.0	68.9	52.5	51.3	61.6	49.7	94.6%
VTW (Lin et al., 2025)	4.48	58.7	1792	83.4	<u>69.6</u>	49.7	51.8	63.0	55.4	96.7%
FitPrune (Ye et al., 2025)	3.16	58.5	1776	77.9	68.0	57.4	51.7	62.7	52.4	96.6%
HiRED (Arif et al., 2025)	-	58.7	1737	82.8	68.4	47.4	50.1	62.8	51.7	93.9%
SparseVLM (Zhang et al., 2024b)	3.48	59.5	1787	85.3	68.6	56.1	51.4	60.0	53.5	96.7%
PDrop (Xing et al., 2024)	3.38	57.1	1664	82.3	68.3	56.1	51.0	61.1	50.6	94.0%
VisionZip (Yang et al., 2025c)	3.16	59.3	1782	85.3	68.9	57.3	52.0	63.4	53.4	97.2%
DivPrune (Alvar et al., 2025)	3.16	58.8	1792	85.1	68.4	56.8	51.8	62.1	52.2	97.1%
BTP(Li et al., 2025)	2.62	59.0	1821	<u>85.6</u>	69.1	-	<u>52.8</u>	62.7	53.1	97.8%
DART(Wen et al., 2025)	3.48	<u>60.0</u>	<u>1840</u>	82.8	69.8	57.4	51.2	63.6	53.9	98.2%
HoloV (Zou et al., 2025)	3.16	59.0	1820	85.6	69.8	<u>57.4</u>	50.9	63.9	54.1	97.8%
ShortV [†] (Yuan et al., 2025)	3.77	59.8	1839	84.0	68.7	57.1	<u>52.8</u>	64.8	53.7	98.4%
HalfV (Ours)	3.12	60.5	1862	86.0	70.2	57.6	52.9	<u>63.7</u>	55.6	99.2%
<i>LLaVA-1.5v-13B (Liu et al., 2023), Backbone: Vicuna</i>										
Vanilla	16.21	63.2	1818	85.9	72.9	60.1	56.7	68.7	59.5	100.0%
FastV (Chen et al., 2024a)	6.68	60.0	1752	83.6	72.9	54.7	53.2	67.0	56.2	94.6%
PDrop (Xing et al., 2024)	5.90	60.5	1773	85.1	<u>73.7</u>	57.2	54.2	67.3	56.9	96.0%
DivPrune (Alvar et al., 2025)	6.18	58.8	1741	85.3	72.6	58.4	53.8	65.8	57.7	96.2%
BTP (Li et al., 2025)	5.72	62.2	1819	<u>86.1</u>	72.7	<u>58.3</u>	<u>54.5</u>	68.0	58.3	98.2%
ShortV [†] (Yuan et al., 2025)	7.26	<u>62.0</u>	<u>1802</u>	<u>85.7</u>	73.5	57.9	54.1	68.6	<u>58.9</u>	98.6%
HalfV (Ours)	4.81	61.6	1831	87.2	75.0	58.0	55.1	<u>68.5</u>	59.3	99.4%
<i>LLaVA-NeXT-7B (Liu et al., 2024b), Backbone: Mistral</i>										
Vanilla	40.42	64.8	1824	86.8	78.7	65.8	63.8	68.1	67.3	100.0%
FastV (Chen et al., 2024a)	15.03	60.3	1782	85.5	70.1	59.7	58.4	64.3	66.1	94.1%
DivPrune (Alvar et al., 2025)	13.34	61.4	1703	86.2	71.9	63.7	60.0	65.4	66.5	95.7%
BTP (Li et al., 2025)	11.92	60.6	1802	<u>86.7</u>	74.4	63.2	61.8	66.3	65.7	97.1%
DART (Wen et al., 2025)	15.03	62.4	1789	86.0	75.3	64.5	61.7	<u>67.5</u>	65.9	97.7%
HoloV (Wen et al., 2025)	13.34	62.5	1796	85.9	76.2	64.0	<u>62.2</u>	67.1	66.5	97.8%
ShortV [†] (Yuan et al., 2025)	16.92	63.4	<u>1809</u>	86.5	<u>77.1</u>	63.9	61.4	67.2	<u>66.7</u>	97.9%
HalfV (Ours)	11.22	<u>62.5</u>	1813	87.3	78.8	<u>64.2</u>	62.4	68.0	67.0	98.9%
<i>Qwen2.5-VL-7B (Bai et al., 2025), Backbone: Qwen2.5</i>										
Vanilla	36.63	60.7	2307	86.5	88.7	83.2	70.6	82.5	81.4	100.0%
FastV (Chen et al., 2024a)	10.33	52.7	2036	80.7	78.0	72.5	61.2	74.9	67.1	87.9%
DivPrune (Alvar et al., 2025)	8.30	50.1	2002	83.9	73.0	70.9	62.8	76.9	75.1	88.1%
BTP (Li et al., 2025)	11.16	57.2	2198	84.7	74.1	72.1	62.8	75.2	74.3	91.1%
DART (Wen et al., 2025)	10.33	56.5	2137	82.9	82.1	75.3	67.4	<u>79.3</u>	77.8	93.9%
HoloV (Zou et al., 2025)	8.30	54.3	2043	82.3	79.8	70.3	-	76.5	75.6	90.5%
ShortV [†] (Yuan et al., 2025)	13.57	47.5	1516	68.7	72.8	64.1	59.8	74.1	67.4	80.0%
HalfV (Ours)	8.91	58.4	2242	<u>84.5</u>	87.0	76.7	68.5	81.1	79.7	96.8%

Table 3: Comparison with state-of-the-art inference acceleration methods on LLaVA-1.5-7B/13B, LLaVA-NeXT-7B and Qwen2.5-VL-7B models. The best result in each group is highlighted in bold, and the second-best result is underlined. Results that are higher than the original model are marked in red. † method focuses on layer-level redundancy, while others focus on token-level redundancy.

manifestations.

(1) Handling Layer-level Inactivity. For architectures like LLaVA, we terminate both attention and feed-forward updates for all visual tokens ($\Delta \mathbf{H}_V = \mathbf{0}$) to eliminate redundant computation, while maintaining normal updates for textual tokens:

$$\begin{aligned} \Delta \mathbf{H}_V^{(l)} &= \mathbf{0}, \\ \Delta \mathbf{H}_T^{(l)} &= \mathcal{F}(\mathbf{H}_T^{(l)}, \text{Concat}(\mathbf{H}_V^{(l)}, \mathbf{H}_T^{(l)})). \end{aligned} \quad (7)$$

More details are provided in Appendix G.1.

(2) Handling Extreme Token Sparsity. For architectures like Qwen, we sparsify deep-layer computation by updating only the Top- K_{SSR} visual tokens. Concretely, we rank visual tokens by the cross-modal attention scores induced by the

last textual token with positional embeddings (e.g., RoPE) enabled, and select \mathcal{S}_{top} accordingly. Computation in subsequent layers is then restricted to tokens in \mathcal{S}_{top} , while the remaining visual tokens are pruned. This design differs from our IVR stage, where positional encoding is disabled for scoring. We discuss the rationale and ablations in Appendix G.2.

4 Experiments

4.1 Experimental Settings

We conduct extensive evaluations to verify the robustness of HalfV across varying model architectures. Our benchmarks cover four mainstream MLLMs built upon different linguistic founda-

Method	GQA	MMB	MME	POPE	SQA	VQA ^{Text}	A12D	Avg.
Qwen2-VL-7B	Upper Bound, All Tokens (100%)							
Vanilla	60.7	82.5	2307	86.5	88.7	83.2	81.4	100%
Qwen2-VL-7B	Flops Ratio Reduction (↓ 88.9%)							
+ FastV (ECCV24)	50.1	69.2	1940	78.6	77.4	60.3	68.7	83.6%
+ DART (EMNLP25)	54.3	74.8	2086	76.9	81.7	62.5	72.1	87.9%
+ HoloV (NIPS25)	52.8	72.4	2006	80.7	79.5	61.8	72.6	86.8%
+ HalfV (Ours)	56.8	79.2	2124	82.6	86.5	68.8	74.6	92.7%
Qwen3-VL-4B-FP8	Upper Bound, All Tokens (100%)							
Vanilla	59.0	83.6	-	85.7	90.8	81.7	83.9	100%
Qwen3-VL-4B-FP8	Flops Ratio Reduction (↓ 77.8%)							
+ FastV (ECCV24)	53.7	77.5	-	78.9	84.1	70.3	77.5	91.1%
+ DART (EMNLP25)	55.8	79.2	-	82.5	85.9	72.2	80.3	94.0%
+ HalfV (Ours)	57.3	80.9	-	83.1	87.5	76.4	81.4	96.3%
Qwen3-VL-4B-FP8	Flops Ratio Reduction (↓ 88.9%)							
+ FastV (ECCV24)	49.7	74.9	-	74.6	82.7	66.7	74.2	87.0%
+ DART (EMNLP25)	52.4	76.2	-	78.3	83.9	70.1	76.3	90.1%
+ HalfV (Ours)	54.1	77.4	-	79.2	84.7	75.2	78.1	92.5%

Table 4: More comparative experiments on Qwen2-VL-7B and Qwen3-VL-4B models.

Method	TGIF		MSVD		MSRVT		Avg.	
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
VideoChat-7B	34.4	2.3	56.3	2.8	45.0	2.5	45.1	2.5
LLaMA-Adapter-7B	-	-	54.9	3.1	43.8	2.7	-	-
Video-ChatGPT	51.4	3.0	64.9	3.3	49.3	2.8	55.2	3.0
Video-LLaVA-7B	47.0	3.4	70.2	3.9	57.3	3.5	58.2	3.6
+ FastV (ECCV24)	45.2	3.1	71.0	3.9	55	3.5	57.1	3.5
+ DART (EMNLP25)	46.3	3.4	71.0	4.0	56.7	3.6	58.0	3.7
+ HoloV (NIPS25)	46.1	3.2	71.0	4.0	56.5	3.6	57.8	3.6
+ HalfV (Ours)	46.5	3.5	70.8	4.0	56.9	3.6	58.1	3.7

Table 5: Comparative experiments on video understanding tasks.

tions: the Vicuna-based LLaVA-1.5v-7B/13B (Liu et al., 2023), the Mistral-based LLaVA-1.6v-7B (Liu et al., 2024b), and the Qwen-based Qwen2.5-VL-7B (Bai et al., 2025). For comprehensive implementation details, please refer to Appendix I.

4.2 Main Results

HalfV Maintains High Performance across Diverse MLLM Backbones. As shown in Table 3, our method achieves SOTA or near-SOTA results across various model series and scales. Notably, on the Qwen series, HalfV significantly outperforms previous SOTA methods: (i) HalfV retains **96.8%** performance at $4.1\times$ FLOPs speedup, outperforming DART by **2.9%**; and (ii) at $6.1\times$ FLOPs speedup, this gap widens to **4.8%** (Table 4). Experiments on Qwen3-VL-4B-FP8 (Yang et al., 2025a) further validate our superiority across different acceleration ratios. For video understanding, integrating HalfV into Video-LLaVA (Lin et al., 2024) (Table 5) yields performance comparable to the original model, demonstrating strong reasoning retention even with high-resolution inputs.

HalfV Mitigates Hallucination and Surpasses Baseline. Unlike prior methods that merely aim to preserve accuracy, HalfV surpasses the uncompressed baseline on hallucination benchmarks. As shown in Table 3, HalfV improves POPE on LLaVA-1.5v-7B (86.0% vs. 85.9%) and further

Method	Total Time		TTFT		Prefill Time		Latency		Perf.
	(Min:Sec)	(Millisecond)	(Millisecond)	(Millisecond)	(Millisecond)	(Millisecond)	(Millisecond)		
LLaVA-NeXT-7B (Mistral)									
Vanilla	73:44		391.9		391.0		495.1		86.8
FastV	44:21 (↓ 39.8%)	202.7 (↓ 48.2%)	201.9 (↓ 48.3%)	249.1 (↓ 49.6%)	85.5 (↓ 1.5%)				
DART	46:17 (↓ 37.2%)	205.4 (↓ 47.5%)	205.2 (↓ 47.5%)	277.4 (↓ 43.9%)	86.0 (↓ 0.9%)				
ShortV	59:48 (↓ 18.8%)	284.7 (↓ 27.3%)	284.1 (↓ 27.3%)	382.4 (↓ 22.7%)	86.5 (↓ 0.3%)				
HalfV (Ours)	45:03 (↓ 38.9%)	201.8 (↓ 48.5%)	201.1 (↓ 48.5%)	263.2 (↓ 46.8%)	87.3 (↑ 0.6%)				
Qwen2.5-VL-7B									
Vanilla	49:56		171.6		-		200.5		86.5
FastV	25:03 (↓ 49.8%)	137.2 (↓ 20.0%)	-	165.7 (↓ 17.3%)	80.7 (↓ 6.7%)				
DART	26:44 (↓ 46.5%)	140.5 (↓ 18.1%)	-	169.9 (↓ 15.3%)	82.9 (↓ 4.2%)				
HalfV (Ours)	25:41 (↓ 48.6%)	139.4 (↓ 18.8%)	-	168.4 (↓ 16.0%)	84.5 (↓ 2.3%)				

Table 6: Evaluation of efficiency for different models on POPE. Inference costs of Total Time, Time To First Token (TTFT), Prefill Time, Latency on POPE dataset for LLaVA-NeXT-7B and Qwen2.5-VL-7B models.

#	IVR	SSR	Perf.	TTFT	Speedup	MO	MO/TTFT
LLaVA-NeXT-7B (Mistral)							
(a)	✓		67.1 (↓ 0.3%)	280.9 ms	1.48×	20.6 ms	7.3%
(b)		✓	67.3 (↓ 0.0%)	332.7 ms	1.26×	0.8 ms	0.2%
(c)	✓	✓	67.0 (↓ 0.4%)	220.9 ms	1.89×	21.7 ms	9.8%
Qwen2.5-VL-7B (Qwen2.5)							
(d)	✓		79.5 (↓ 2.3%)	173.0 ms	1.84×	10.2 ms	5.8%
(e)		✓	81.4 (↓ 0.0%)	192.7 ms	1.66×	1.4 ms	0.7%
(f)	✓	✓	79.7 (↓ 2.0%)	166.1 ms	1.92×	10.6 ms	6.3%

Table 7: Ablation study of module efficiency on A12D dataset. **IVR** denotes the module that mitigates *Intrinsic Visual Redundancy*; **SSR** denotes the module that suppresses *Semantic Saturation Redundancy*. **MO** abbreviates *module overhead*, and **MO/TTFT** measures the modules relative time overhead as a fraction of TTFT.

boosts LLaVA-1.5v-13B by **+1.3%** on POPE and **+13** points on MME (reaching 1831). Similar gains are observed on LLaVA-NeXT-7B. We attribute these improvements to the suppression of IVR and SSR, which effectively reduces misleading visual evidence.

4.3 Efficiency Analysis

To assess the practical speedup of HalfV, we benchmark multiple methods on the POPE task using LLaVA-NeXT-7B and Qwen2.5-VL-7B. We report the average total time, time-to-first-token (TTFT), prefill time, end-to-end latency, and performance. Following the same experimental settings as in the main table 3, we evaluate FastV (Chen et al., 2024a), DART (Wen et al., 2025), ShortV (Yuan et al., 2025), and our method. As shown in Table 6, on LLaVA-NeXT-7B, HalfV reduces Total Time by **38.9%**, TTFT by **48.5%**, and latency by **46.8%**, while accuracy not only remains intact but also improves by **0.6%**. Compared with prior methods, HalfV better preserves the original model performance under comparable speedups. On Qwen2.5-VL-7B, HalfV achieves a **48.6%** reduction in Total Time with only a **2.3%** accuracy drop, and improves accuracy by **1.9%** over DART.

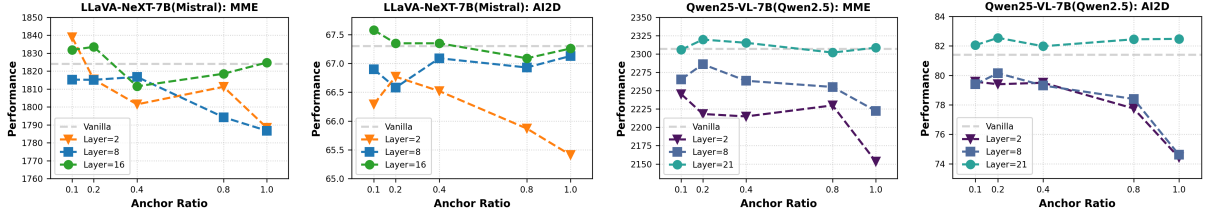


Figure 6: Performance comparison of different models under different pruning layers and anchor retention ratios.

5 Analysis and Discussion

5.1 Ablation Study for Module Efficiency

To quantify each module's contribution, we perform ablations on HalfV's two-step pipeline. The first step mitigates IVR from the ViT encoder, and the second step suppresses SSR caused by global aggregation in the LLM. We evaluate variants on LLaVA-NeXT-7B and Qwen2.5-VL-7B, reporting performance, TTFT, speedup, and module overhead (MO). As shown in Table 7, the IVR module accounts for **7.3%** of TTFT on LLaVA-NeXT-7B and **5.8%** on Qwen2.5-VL-7B. In contrast, the architecture-aware SSR module costs less than **1%** on both models. Notably, on Qwen2.5-VL-7B, enabling both modules yields higher speedup than IVR-only and even improves performance by **0.3%** (Table 7(f) and 7(d)). This indicates that jointly mitigating the two types of redundancy is not a naive concatenation, but a complementary design that yields synergistic benefits from an internal model perspective.

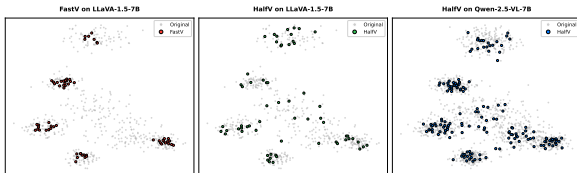


Figure 7: Comparison of feature space coverage

5.2 Sensitivity Analysis of the Anchor Retention Ratio R_S

In practice, we implement the anchor count as a ratio R_S to accommodate dynamic resolutions. We evaluate R_S on LLaVA-NeXT-7B and Qwen2.5-VL-7B, applying the first-step pruner at layers 2, 8, and the Stage III onset (Layer 16/21). As shown in Figure 6, early layers (e.g., Layer 2) benefit from lower R_S , indicating a reliance on token diversity. Conversely, deep layers maintain performance even with high anchor ratios, confirming that Stage II aggregation effectively concentrates key visual evidence, rendering extensive spatial coverage unnecessary in later stages. We

further visualize the selected visual tokens in the IVR stage on LLaVA-1.5v-13B and Qwen2.5-VL-7B. We apply t-SNE (van der Maaten and Hinton, 2008) to project the D -dimensional visual tokens into a 2D space. As shown in Fig. 7, HalfV covers different clusters more uniformly, reducing the risk of spatial collapse.

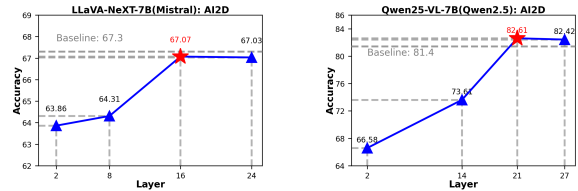


Figure 8: Impact of SSR start layer on different models

5.3 Sensitivity Analysis of the SSR Start Layer

We evaluate the SSR module in isolation on LLaVA-NeXT-7B and Qwen2.5-VL-7B across varying start layers: Layer 2 (IVR), mid-Stage II (8/14), Stage III onset (16/21), and deep Stage III (24/27). As shown in Figure 8, early application significantly degrades performance. Conversely, initiating SSR at the Stage III onset yields near-lossless results, even surpassing the Qwen baseline. Notably, delaying SSR further into Stage III also harms performance. This confirms that the optimal insertion point is not "the later the better," but rather the Stage II-to-III transition (the information aggregation valley).

6 Conclusion

This paper tackles the limited cross-architecture generalization of existing MLLM acceleration methods and proposes an acceleration framework, HalfV. Using truncated matrix entropy as a unified probe, it reveals a three-stage redundancy evolution during inference and decomposes visual redundancy into Intrinsic Visual Redundancy (IVR) and Secondary Saturation Redundancy (SSR). HalfV then applies a two-step acceleration method to improve efficiency across model backbones.

7 Limitations

Although HalfV demonstrates strong and consistent improvements across multiple MLLMs and benchmarks, its applicability is constrained by practical considerations. First, similar to many existing inference acceleration methods, HalfV requires access to token-level hidden states during inference, and thus cannot be directly applied to closed-source black-box models such as the GPT and Gemini families. Second, due to current compute and hardware limitations, we have only deployed and evaluated HalfV on MLLMs in the 4B–13B parameter range, and we have not yet conducted systematic experiments on larger-scale models. In addition, to balance completeness and readability under the page limit, we defer some implementation details of evaluation metrics as well as further analyses and observations about internal model behaviors to the appendix. The main text presents the method, core experiments, and primary findings in full, ensuring that the key contributions and conclusions can be understood without relying on the appendix.

References

- Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. [Divprune: Diversity-based visual token pruning for large multimodal models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9392–9401. Computer Vision Foundation / IEEE.
- Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2025. [Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models](#). In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 1773–1781. AAAI Press.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. [Vizwiz: nearly real-time answers to visual questions](#). In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010*, pages 333–342. ACM.
- Richard Cangelosi and Alain Goriely. 2007. [Component retention in principal component analysis with application to cdna microarray data](#). *Biology direct*, 2:2.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. [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 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pages 19–35. Springer.
- Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024b. [In-context sharpness as alerts: An inner representation perspective for hallucination mitigation](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. [Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). *CoRR*, abs/2312.14238.
- 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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2025. [Feather the throttle: Revisiting visual token pruning for vision-language model acceleration](#). *Preprint*, arXiv:2412.13180.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. [Mme: A comprehensive evaluation benchmark for multimodal large language models](#). *Preprint*, arXiv:2306.13394.
- Teofilo F. Gonzalez. 1985. [Clustering to minimize the maximum intercluster distance](#). *Theoretical Computer Science*, 38:293–306.

- Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. 2025. [Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects?](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6841–6858. Association for Computational Linguistics.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering.](#) In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.
- Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. 2025. [Interpreting and editing vision-language representations to mitigate hallucinations.](#) In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images.](#) In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 235–251. Springer.
- S. Kullback and R. A. Leibler. 1951. [On information and sufficiency.](#) *The Annals of Mathematical Statistics*, 22(1):79–86.
- Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. 2025. [Balanced token pruning: Accelerating vision language models beyond local optimization.](#) *CoRR*, abs/2505.22038.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. [Video-llava: Learning united visual representation by alignment before projection.](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.
- Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. [Boosting multimodal large language models with visual tokens withdrawal for rapid inference.](#) In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 5334–5342. AAAI Press.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning.](#) In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [Llava-next: Improved reasoning, ocr, and world knowledge.](#)
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning.](#) In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. [Mmbench: Is your multi-modal model an all-around player?](#) In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering.](#) In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. [Llava-prumerge: Adaptive token reduction for efficient large multimodal models.](#) *CoRR*, abs/2403.15388.
- 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.
- Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [Roformer: Enhanced transformer with rotary position embedding.](#) *Neurocomputing*, 568:127063.
- Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. 2025. [Flowcut: Rethinking redundancy via information flow for efficient vision-language models.](#) *CoRR*, abs/2505.19536.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne.](#) *Journal of Machine Learning Research*, 9(86):2579–2605.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin

- Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *CoRR*, abs/2409.12191.
- 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](#). *CoRR*, abs/2502.11494.
- Qiong Wu, Wenhao Lin, Weihao Ye, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. 2024. [Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings](#). *CoRR*, abs/2411.19628.
- Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. [Pyramidrop: Accelerating your large vision-language models via pyramid visual redundancy reduction](#). *CoRR*, abs/2410.17247.
- Jing Xiong, Jianghan Shen, Fanghua Ye, Chaofan Tao, Zhongwei Wan, Jianqiao Lu, Xun Wu, Chuanyang Zheng, Zhijiang Guo, Min Yang, Lingpeng Kong, and Ngai Wong. 2025. [UNComp: Can matrix entropy uncover sparsity? — a compressor design from an uncertainty-aware perspective](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4179–4199, Suzhou, China. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponuswamy Sadayappan, Xia Hu, and Bo Yuan. 2025b. [Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19803–19813. Computer Vision Foundation / IEEE.
- Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025c. [Visionzip: Longer is better but not necessary in vision language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 19792–19802. Computer Vision Foundation / IEEE.
- Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025. [Fit and prune: Fast and training-free visual token pruning for multi-modal large language models](#). In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 22128–22136. AAAI Press.
- Qianhao Yuan, Qingyu Zhang, Yanjiang Liu, Jiawei Chen, Yaojie Lu, Hongyu Lin, Jia Zheng, Xianpei Han, and Le Sun. 2025. [Shortv: Efficient multi-modal large language models by freezing visual tokens in ineffective layers](#). *CoRR*, abs/2504.00502.
- Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. 2024a. [Matrix information theory for self-supervised learning](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- 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. 2024b. [Sparse-vm: Visual token sparsification for efficient vision-language model inference](#). *CoRR*, abs/2410.04417.
- 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](#). *CoRR*, abs/2510.02912.

Appendix

A Theoretical Analysis of Truncated Matrix Entropy	12
A.1 From Matrix Entropy to Effective Rank	12
A.2 Why Truncation is Well-Behaved	12
A.3 Connecting Spectral Entropy to Geometry via Curvature	13
A.4 From Curvature Energy to Angle-Based Curvature	13
B Additional Experiments of Truncated Matrix Entropy	14
B.1 Spectrum of Different Models	14
B.2 Top- k Eigenvalues Selection	14
C Theoretical Analysis of Truncated Matrix Entropy	16
C.1 From Matrix Entropy to Effective Rank	16
C.2 Why Truncation is Well-Behaved	18
C.3 Connecting Spectral Entropy to Geometry via Curvature	18

C.4	From Curvature Energy to Angle-Based Curvature	19
D	Additional Cues for Global Aggregation	19
E	Implementation Details of Marginal Utility	20
F	Implementation Details of Redundancy Metric	20
G	Implementation Details of HalfV	21
G.1	Layer-level Inactivity Implementation Details	21
G.2	RoPE: enabled or disabled?	21
G.3	Implementation Details of Qwen3-VL	21
H	Case Study	22
H.1	Llava Model Case	22
H.2	Qwen Model Case	22
I	Detailed Experiment Settings	23
I.1	Datasets	23
I.2	Models	24
I.3	Baselines	24
I.4	Implementation Details	25
I.5	Hyperparameters	25
I.6	Computational Complexity	25
J	Clarifications and Differences from Related Work	26
J.1	Comparison with ShortV	26
J.2	Comparison with BTP	26
J.3	Comparison with PDrop	27
K	Ethics Statement	27
A	Theoretical Analysis of Truncated Matrix Entropy	

Setup. In each layer, we stack the hidden states of N tokens into $\mathcal{Z} = [\mathbf{h}_1, \dots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times D}$. We form the Gram matrix \mathcal{G} as in Eq. 1, and denote its eigenvalues by $\lambda_1(\mathcal{G}) \geq \dots \geq \lambda_r(\mathcal{G}) > 0$ where $r = \text{rank}(\mathcal{G})$. We retain the top- k eigenvalues and define the truncated trace $\text{tr}(\mathcal{G}) = \sum_{i=1}^k \lambda_i(\mathcal{G})$, and the truncated matrix entropy (\mathcal{H}):

$$\mathcal{H}(\mathcal{Z}) = - \sum_{i=1}^k p_i \log p_i, \quad p_i := \frac{\lambda_i(\mathcal{G})}{\sum_{j=1}^k \lambda_j(\mathcal{G})}. \quad (8)$$

Throughout, $p = (p_1, \dots, p_k)$ is a probability vector supported on the principal subspace.

A.1 From Matrix Entropy to Effective Rank

Definition 1 (Entropy-based effective rank). We define the entropy-based effective rank associated with the truncated spectrum as

$$r_{\text{eff}}^{(k)}(\mathcal{Z}) := \exp(\mathcal{H}(\mathcal{Z})). \quad (9)$$

This is the standard effective number of active spectral components implied by Shannon entropy.

Lemma 1 (Basic properties). Let $p_i = \lambda_i(\mathcal{G}) / \sum_{j=1}^k \lambda_j(\mathcal{G})$ for $i = 1, \dots, k$. Then $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$. Defining $r_{\text{eff}}^{(k)}(\mathcal{Z}) := \exp(\mathcal{H}(\mathcal{Z})) = \exp(-\sum_{i=1}^k p_i \log p_i)$, we have

$$1 \leq r_{\text{eff}}^{(k)}(\mathcal{Z}) \leq k.$$

Moreover, $r_{\text{eff}}^{(k)}(\mathcal{Z}) = 1$ iff $p = (1, 0, \dots, 0)$ (spectral mass concentrates on one component), and $r_{\text{eff}}^{(k)}(\mathcal{Z}) = k$ iff $p_i = 1/k$ for all $i = 1, \dots, k$ (uniform truncated spectrum).

Connection to stable rank / participation ratio.

A closely related notion is the participation ratio (a.k.a. stable rank of the truncated spectrum):

$$r_{\text{pr}}^{(k)}(\mathcal{Z}) := \frac{1}{\sum_{i=1}^k p_i^2} = \frac{(\sum_{i=1}^k \lambda_i(\mathcal{G}))^2}{\sum_{i=1}^k \lambda_i(\mathcal{G})^2} = \frac{\text{tr}(\mathcal{G})^2}{\|\mathcal{G}\|_F^2}, \quad (10)$$

where $\|\mathcal{G}\|_F^2 = \sum_{i=1}^k \lambda_i(\mathcal{G})^2$ when only the top- k spectrum is kept.

Lemma 2 (Entropy lower-bounds participation ratio). $r_{\text{eff}}^{(k)}(\mathcal{Z}) \geq r_{\text{pr}}^{(k)}(\mathcal{Z})$.

Proof. Let $H(p) = -\sum_i p_i \log p_i$ be Shannon entropy and $H_2(p) = -\log \sum_i p_i^2$ be Rényi-2 entropy. It is well-known that $H(p) \geq H_2(p)$ for any discrete distribution. Exponentiating both sides gives $\exp(H(p)) \geq \exp(H_2(p)) = 1 / \sum_i p_i^2$. \square

Interpretation for redundancy. For a fixed truncated trace $\text{tr}(\mathcal{G})$ (i.e., fixed total variance inside the principal subspace), a *more redundant / collapsed* representation concentrates spectral mass on fewer eigen-directions, which decreases $H(p)$ and thus decreases both $r_{\text{eff}}^{(k)}$ and $r_{\text{pr}}^{(k)}$. Therefore, $\mathcal{H}(\mathcal{Z})$ is a principled, scale-free proxy for “how many directions are effectively used” by the token representations.

A.2 Why Truncation is Well-Behaved

Truncation controls noise sensitivity. Let \tilde{p} denote the full normalized spectrum over r nonzero

eigenvalues, $\tilde{p}_i = \lambda_i(\mathcal{G}) / \sum_{j=1}^r \lambda_j(\mathcal{G})$. Define the tail mass beyond the elbow as

$$\delta := \sum_{i>k} \tilde{p}_i = \frac{\sum_{i>k} \lambda_i(\mathcal{G})}{\sum_{j=1}^r \lambda_j(\mathcal{G})}. \quad (11)$$

When the spectrum has a clear elbow, δ is small: most variance lies in the principal k components.

Proposition 1 (Entropy perturbation bound (informal)). If the discarded tail mass is $\delta \ll 1$, then the difference between the full matrix entropy and the truncated (renormalized) entropy is controlled:

$$|H(\tilde{p}) - \mathcal{H}(\mathcal{Z})| \leq h_2(\delta) + \delta \log(r-1), \quad (12)$$

where $h_2(\delta) = -\delta \log \delta - (1-\delta) \log(1-\delta)$ is the binary entropy.

Implication. Proposition 1 states that truncation yields a *robust* estimate of the intrinsic spectral diversity: as long as the elbow tail carries little mass (common in practice), \mathcal{H} differs from the full entropy by at most $O(\delta \log r)$. This justifies truncation as a denoising step that suppresses small, unstable eigenvalues caused by stochasticity, finite-sample effects, or irrelevant redundancy.

A.3 Connecting Spectral Entropy to Geometry via Curvature

Discrete curvature surrogate. To relate spectral diversity to geometry, we treat the token representations $\{\mathbf{h}_i\}_{i=1}^N$ as a polyline in \mathbb{R}^D (tokens are naturally ordered, e.g., by spatial rasterization, patch order, or attention order). Define the second-order finite difference $\Delta^2 \mathbf{h}_i := \mathbf{h}_{i+1} - 2\mathbf{h}_i + \mathbf{h}_{i-1}$ for $i = 2, \dots, N-1$, and the curvature energy

$$\mathcal{C}(\mathcal{Z}) := \frac{1}{N-2} \sum_{i=2}^{N-1} \|\Delta^2 \mathbf{h}_i\|_2^2. \quad (13)$$

This quantity is a standard, analytically tractable proxy of geometric ‘‘bending’’: larger $\mathcal{C}(\mathcal{Z})$ indicates more rapid directional change along the token trajectory.

Matrix form (curvature is a Gram-functional).

Let $L \in \mathbb{R}^{(N-2) \times N}$ be the second-difference operator such that $(L\mathcal{Z})_{i,:} = \Delta^2 \mathbf{h}_{i+1}^\top$. Then

$$\mathcal{C}(\mathcal{Z}) = \frac{1}{N-2} \|L\mathcal{Z}\|_F^2 = \frac{1}{N-2} \text{tr}(\mathcal{Z}^\top L^\top L \mathcal{Z}). \quad (14)$$

When $\mathcal{G} = \mathcal{Z}\mathcal{Z}^\top$ (e.g., $N < D$ case), we further have

$$\mathcal{C}(\mathcal{Z}) = \frac{1}{N-2} \text{tr}(L^\top L \mathcal{Z}\mathcal{Z}^\top) = \frac{1}{N-2} \text{tr}(L^\top L \mathcal{G}). \quad (15)$$

Thus, curvature is completely determined by the Gram matrix \mathcal{G} , establishing a direct bridge between a geometric quantity and spectral statistics.

Theorem 1 (Curvature is controlled by spectral concentration). Let \mathcal{G}_k denote the rank- k truncated Gram matrix that keeps the top- k eigenpairs of \mathcal{G} . Define $\mathcal{C}_k(\mathcal{Z}) := \frac{1}{N-2} \text{tr}(L^\top L \mathcal{G}_k)$ as the curvature induced by the principal subspace. Then

$$\begin{aligned} \mathcal{C}_k(\mathcal{Z}) &\leq \frac{\|L^\top L\|_F}{N-2} \|\mathcal{G}_k\|_F = \frac{\|L^\top L\|_F}{N-2} \cdot \frac{\text{tr}(\mathcal{G}_k)}{\sqrt{r_{\text{pr}}^{(k)}(\mathcal{Z})}} \\ &\leq \frac{\|L^\top L\|_F}{N-2} \cdot \frac{\text{tr}(\mathcal{G}_k)}{\sqrt{r_{\text{eff}}^{(k)}(\mathcal{Z})}}. \end{aligned} \quad (16)$$

Proof. The first inequality is Cauchy–Schwarz for the Frobenius inner product: $\text{tr}(AB) \leq \|A\|_F \|B\|_F$ for any matrices A, B , applied to $A = L^\top L$ (PSD) and $B = \mathcal{G}_k$ (PSD). The equality follows from $\|\mathcal{G}_k\|_F = \sqrt{\sum_{i=1}^k \lambda_i(\mathcal{G})^2} = \text{tr}(\mathcal{G}_k) / \sqrt{r_{\text{pr}}^{(k)}(\mathcal{Z})}$ by definition of $r_{\text{pr}}^{(k)}$. The last inequality uses Lemma 4, i.e., $r_{\text{eff}}^{(k)} \geq r_{\text{pr}}^{(k)}$. \square

Interpretation. Eq. (30) provides an explicit link between *geometry* and *information*: for a fixed principal-subspace variance $\text{tr}(\mathcal{G}_k)$ and a fixed token order (hence fixed L), the curvature energy is upper-bounded by a term inversely proportional to $\sqrt{r_{\text{eff}}^{(k)}} = \exp(\mathcal{H}/2)$. Therefore, when representations become spectrally concentrated (lower \mathcal{H} , smaller effective rank), the bound becomes looser, allowing larger geometric bending/instability; conversely, higher \mathcal{H} implies a more spread spectrum, smaller $\|\mathcal{G}_k\|_F$, and thus tighter control on curvature energy. This theoretical coupling supports using \mathcal{H} as a probe of internal redundancy that correlates with geometric complexity.

A.4 From Curvature Energy to Angle-Based Curvature

Our main analysis uses $\mathcal{C}(\mathcal{Z})$ for tractability. If one uses an angle-based curvature (as in prior representation geometry analyses), we can connect them as follows. Let $\mathbf{d}_i = \mathbf{h}_{i+1} - \mathbf{h}_i$ and let θ_i be the turning angle between \mathbf{d}_{i-1} and \mathbf{d}_i . Using the identity $\|\mathbf{d}_i - \mathbf{d}_{i-1}\|_2^2 = \|\Delta^2 \mathbf{h}_i\|_2^2$ and

$1 - \cos \theta_i = \frac{\|\mathbf{d}_i - \mathbf{d}_{i-1}\|_2^2}{2\|\mathbf{d}_i\|_2\|\mathbf{d}_{i-1}\|_2}$, we obtain

$$1 - \cos \theta_i \leq \frac{\|\Delta^2 \mathbf{h}_i\|_2^2}{2s_0^2}, \quad (17)$$

whenever $\|\mathbf{d}_i\|_2, \|\mathbf{d}_{i-1}\|_2 \geq s_0 > 0$.

Averaging over i yields that the mean turning (hence angle-based curvature) is controlled by $\mathcal{C}(\mathcal{Z})$, which in turn is controlled by $\mathcal{H}(\mathcal{Z})$ via Theorem 2. This closes the loop between spectral entropy and geometric curvature.

Takeaway. The above results establish a principled chain:

$$\begin{aligned} \mathcal{H}(\mathcal{Z}) &\iff r_{\text{eff}}^{(k)}(\mathcal{Z}) \implies \|\mathcal{G}_k\|_F \\ &\implies \mathcal{C}_k(\mathcal{Z}) \implies (\text{angle-based curvature}). \end{aligned}$$

Hence, truncated matrix entropy serves as a theoretically grounded and robust spectral probe for redundancy and geometry of internal token representations.

B Additional Experiments of Truncated Matrix Entropy

B.1 Spectrum of Different Models

Extensive experiments verify that the evolution of internal visual redundancy captured by the truncated matrix entropy is primarily an intrinsic property of the model, and is largely insensitive to the dataset type and the number of sampled instances. Specifically, on POPE (He et al., 2025), GQA (Hudson and Manning, 2019), and MME (Fu et al., 2025), we visualize the layer-wise eigenvalue spectra of Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-1.5v-13B (Liu et al., 2024a), and LLaVA-Next-7B (Liu et al., 2024b), computed from random subsets of 10, 100, and 1000 samples as well as from the full datasets. As shown in Figure 9, 10 and 11, all models exhibit a pronounced spectral skewness: *a large fraction of eigenvalues concentrate near zero, while only a few dominant components carry most of the informative variance*. This consistent spectral structure suggests that the tail eigenvalues mainly reflect noise or redundancy, thereby providing direct empirical support for adopting truncated matrix entropy rather than the full matrix entropy as our probe of internal redundancy.

In addition, we observe a pronounced discrepancy in sampling convergence of the eigenvalue

spectrum across model series. For the LLaVA series, the layer-wise eigenvalue spectrum computed from as few as 10 randomly sampled instances is already highly consistent with that obtained from much larger subsets (see Figure 9a and 9d). In contrast, for the Qwen series, the spectrum estimated from 10 samples on MME remains noticeably different from the results under larger sample sizes (see Figure 11e and 11h).

This observation yields two implications. **First, the difference in convergence behavior cannot be simply attributed to whether the ViT backbone uses dynamic resolution.** For example, both LLaVA-Next and Qwen2.5-VL adopt dynamic resolution, yet their spectral stability with respect to sample size differs substantially. **Second, from a statistical estimation perspective, the eigenvalue spectrum is determined by the Gram matrix of token representations, whose estimation error critically depends on the ratio between the representation dimensionality and the number of samples.** When each instance contains substantially more visual tokens, the spectrum becomes more susceptible to finite-sample noise, and therefore requires a larger sample size to obtain a stable, reproducible spectral shape. Concretely, images in MME are typically of higher resolution than those in POPE and GQA, leading the ViT to output 2,000 and even up to 15,000 visual tokens for MME, whereas the other two datasets usually produce fewer than 500 tokens. As a result, for spectral analysis of Qwen models on MME, 10 samples are often insufficient for convergence, making it necessary to increase the sampling budget accordingly.

B.2 Top- k Eigenvalues Selection

In Equation 2, we select the top- k eigenvalues to calculate the truncated matrix entropy based on the location of the inflection point in the model’s eigenvalue spectrum. In this subsection, we investigate the impact of different top- k selections on the entropy curves of various models.

As shown in Figure 9, the inflection point in the eigenvalue spectrum of the LLaVA-1.5v-7B model tends to approach zero across different datasets. This suggests that most of the model’s information is concentrated in a small number of high eigenvalues. Therefore, we select top-2, 4, 8, and 16 as truncation points and compute the truncated matrix entropy for the model’s hidden states across different layers in the GQA and MME datasets. To maintain consistent units, we set the number

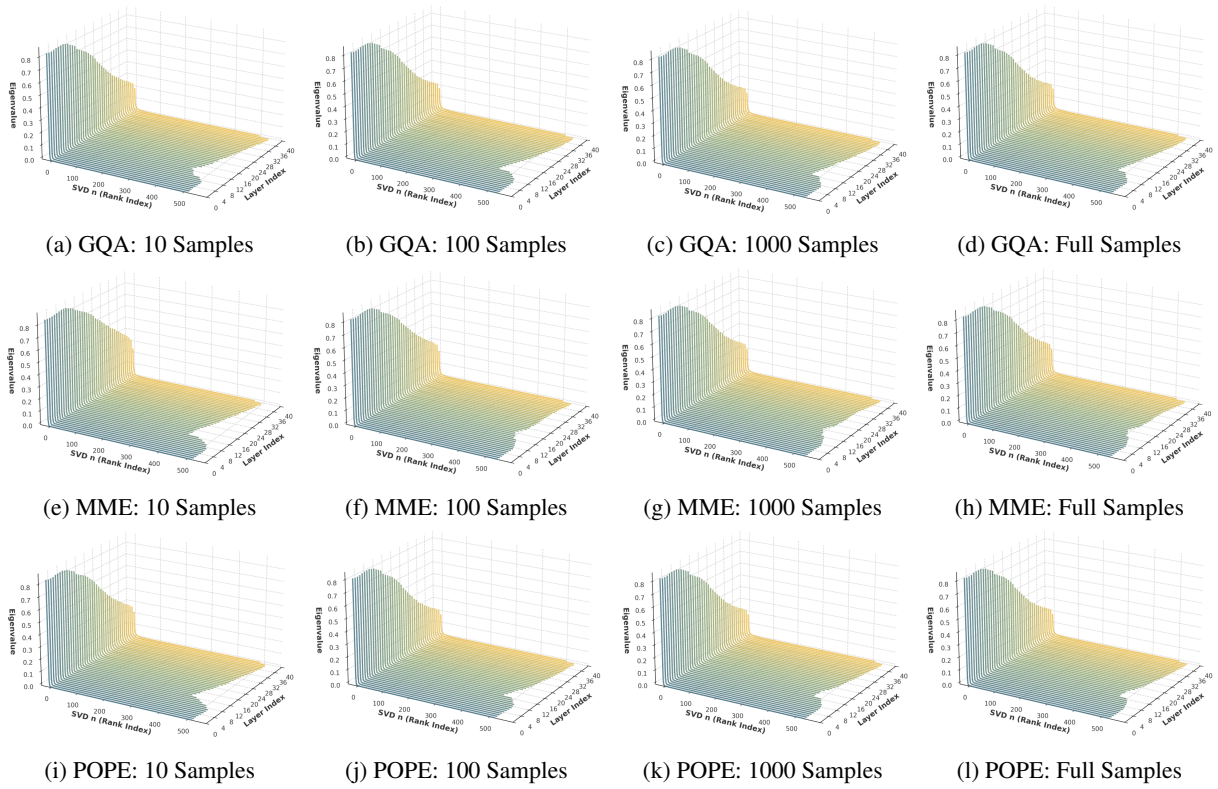


Figure 9: Spectrum of \mathcal{G} across layers for LLaVA-1.5v-13B

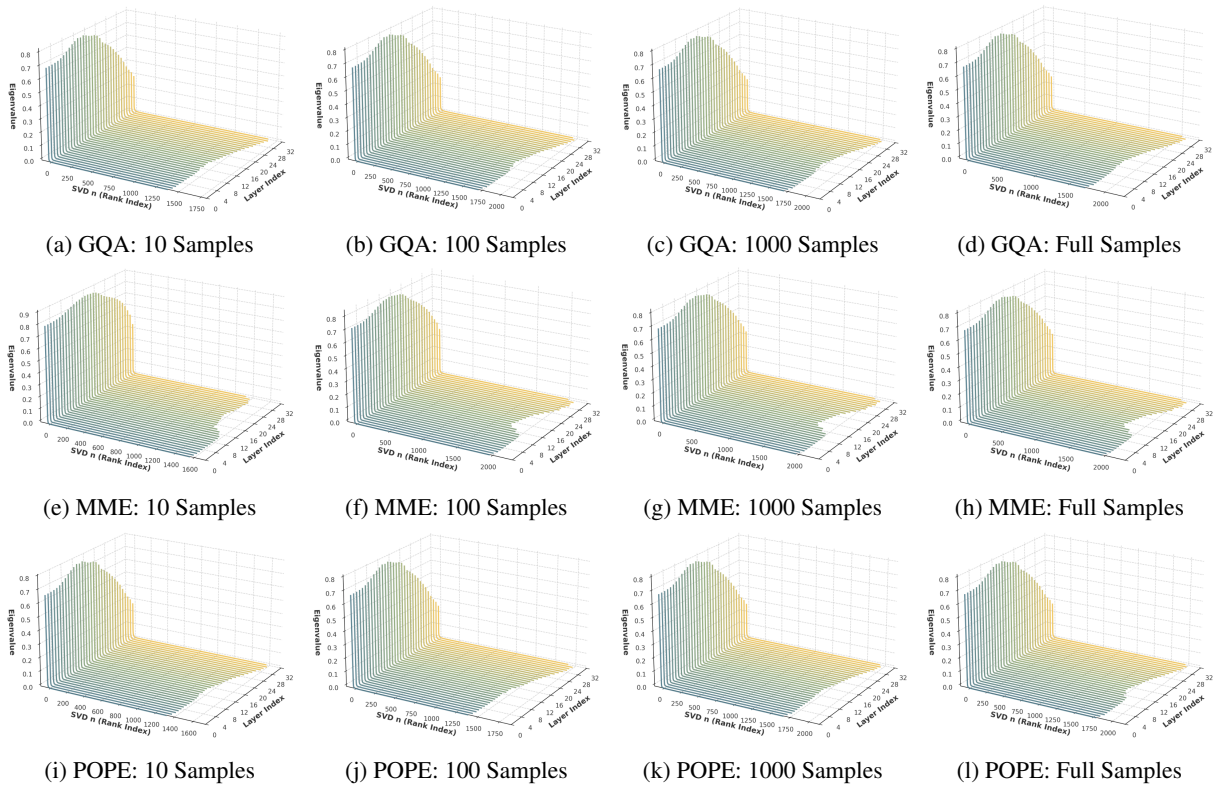


Figure 10: Spectrum of \mathcal{G} across Layers for LLaVA-NeXT-7B (Mistral)

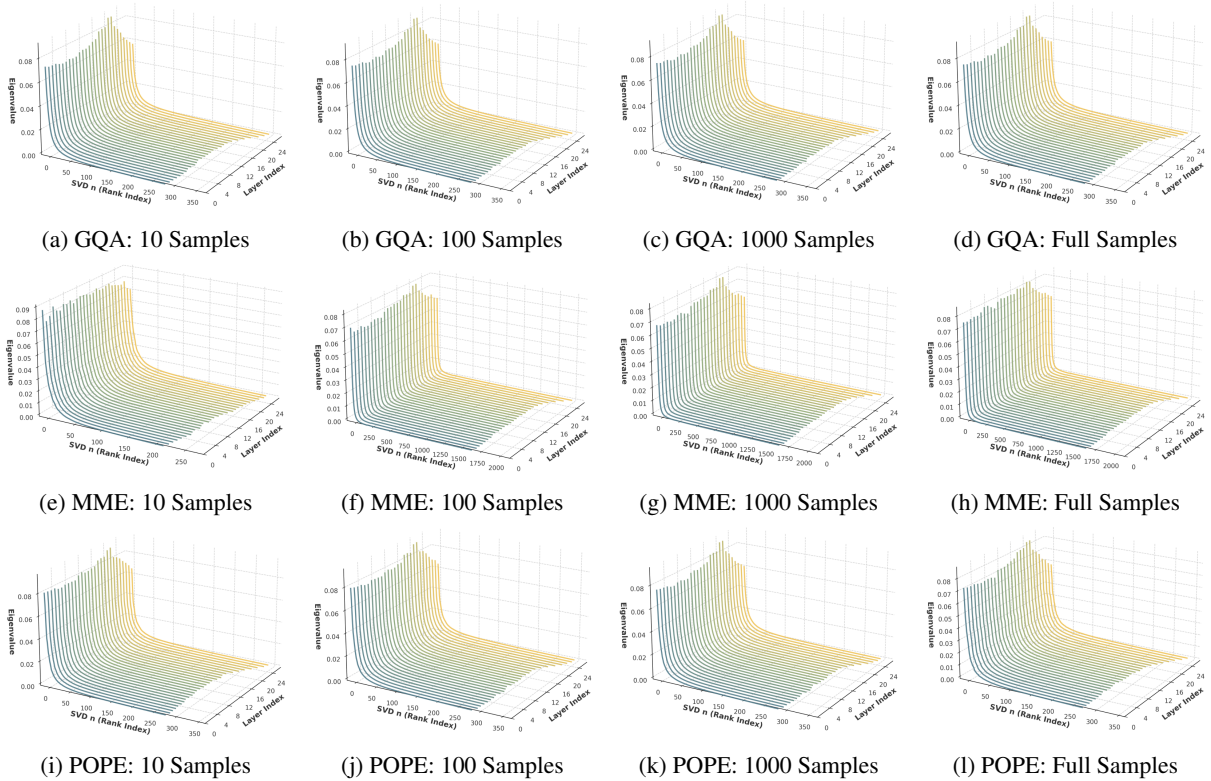


Figure 11: Spectrum of \mathcal{G} across Layers for Qwen25-VL-7B

of samples to 100.

The experimental results are shown in Figure 12, where the **green line** represents the entropy variation of the visual modality’s hidden layer states across layers. We observe that, regardless of the chosen k value, the entropy of the visual modality still shows an aggregation process (a valley), with the lowest point occurring at the 25% position. Additionally, we visualized the entropy curves for the LLaVA-NeXT-7B (Mistral) and Qwen25-VL-7B models. For the Mistral-based model, the lowest point of aggregation appears at the 50% position, while for the Qwen2.5-based model, the lowest point appears at the 75% position. Moreover, we notice that the entropy trend for the text modality of LLaVA models differs significantly from that of Qwen models, which is also an interesting observation. However, as our work currently focuses on the visual modality’s entropy curves, we will not delve into this aspect in detail here.

C Theoretical Analysis of Truncated Matrix Entropy

Setup. In each layer, we stack the hidden states of N tokens into $\mathcal{Z} = [\mathbf{h}_1, \dots, \mathbf{h}_N]^\top \in \mathbb{R}^{N \times D}$.

We form the Gram matrix \mathcal{G} as in Eq. 1, and denote its eigenvalues by $\lambda_1(\mathcal{G}) \geq \dots \geq \lambda_r(\mathcal{G}) > 0$, where $r = \text{rank}(\mathcal{G})$. Throughout we assume $1 \leq k \leq r$.

Let \mathcal{G}_k denote the rank- k truncation of \mathcal{G} (keeping the top- k eigenpairs), and define the truncated trace

$$\text{tr}_k(\mathcal{G}) := \sum_{i=1}^k \lambda_i(\mathcal{G}) = \text{tr}(\mathcal{G}_k).$$

We then define the truncated matrix entropy (\mathcal{H}):

$$\mathcal{H}(\mathcal{Z}) = - \sum_{i=1}^k p_i \log p_i, \quad p_i := \frac{\lambda_i(\mathcal{G})}{\text{tr}_k(\mathcal{G})}. \quad (18)$$

Here $p = (p_1, \dots, p_k)$ is a probability vector supported on the principal subspace. Unless otherwise stated, $\log(\cdot)$ denotes the natural logarithm, so that $\exp(\mathcal{H})$ is consistent.

C.1 From Matrix Entropy to Effective Rank

Definition 2 (Entropy-based effective rank). We define the entropy-based effective rank associated with the truncated spectrum as

$$r_{\text{eff}}^{(k)}(\mathcal{Z}) := \exp(\mathcal{H}(\mathcal{Z})). \quad (19)$$

This is the effective number of active spectral components implied by Shannon entropy.

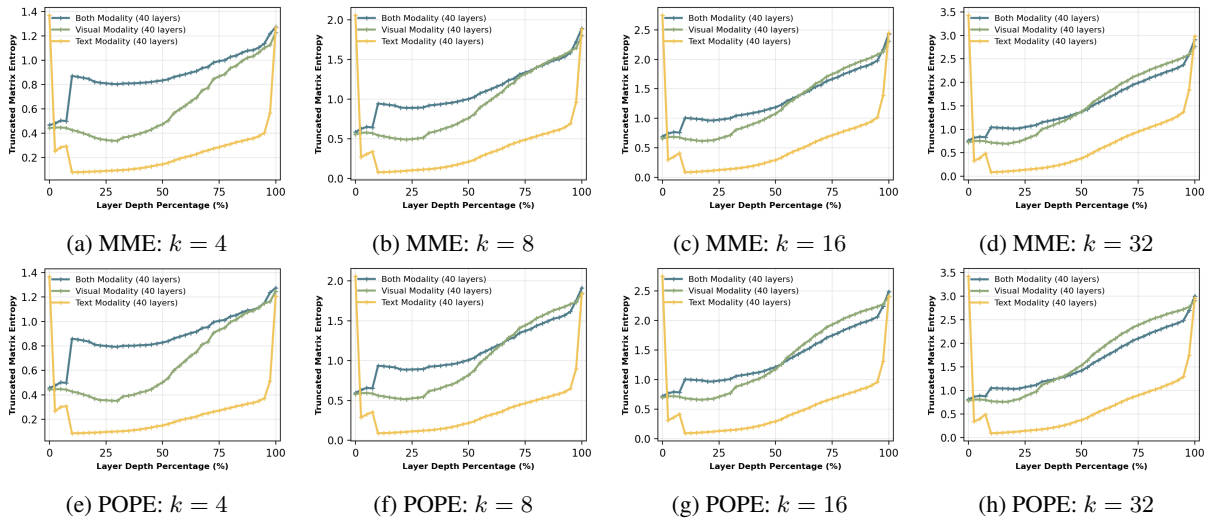


Figure 12: Truncated Matrix Entropy for LLaVA-1.5v-13B

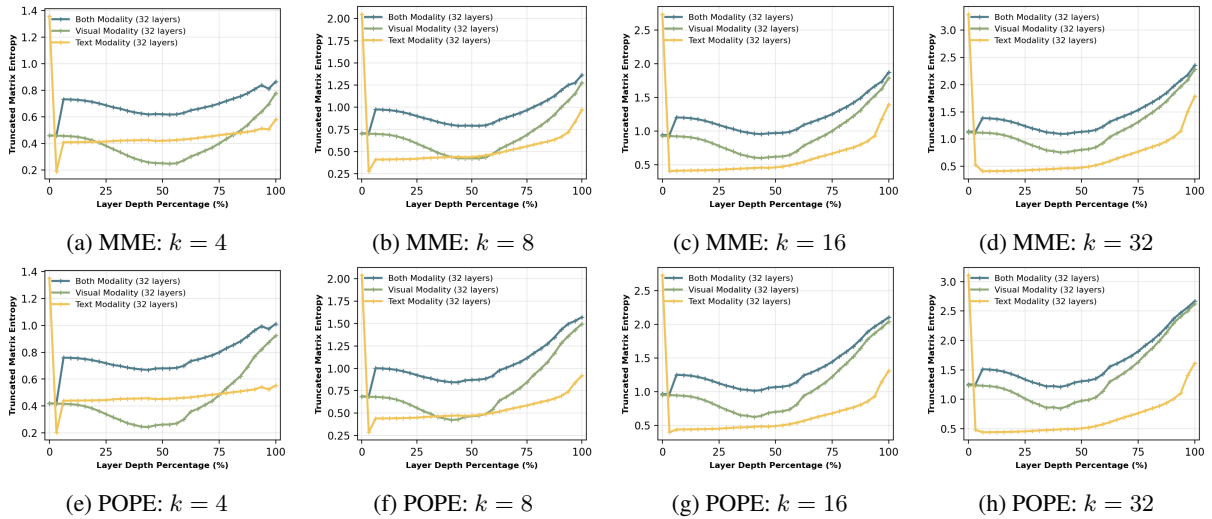


Figure 13: Truncated Matrix Entropy for LLaVA-NeXT-7B (Mistral)

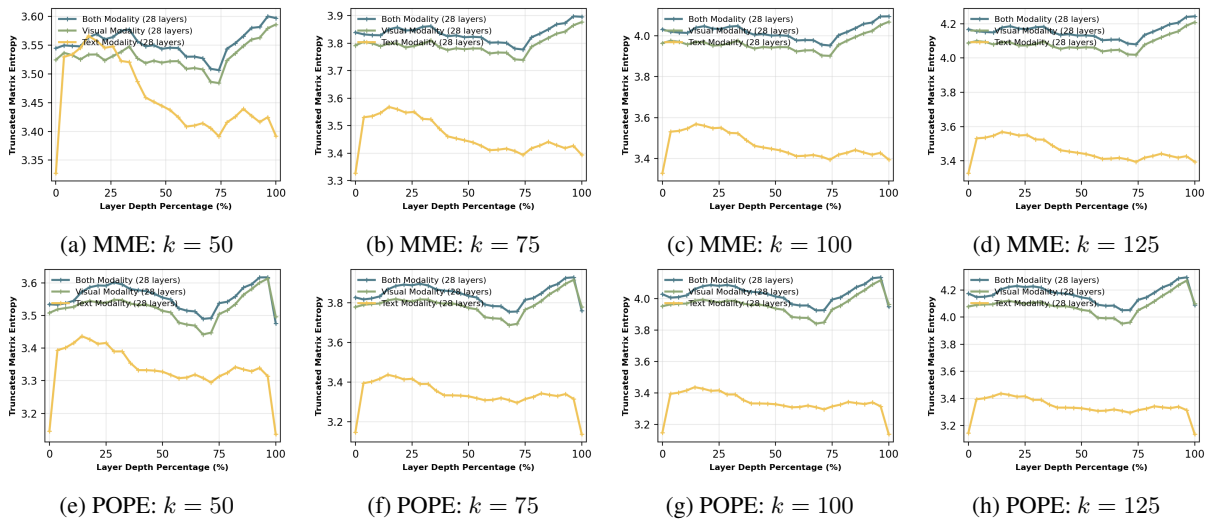


Figure 14: Truncated Matrix Entropy for Qwen25-VL-7B

Lemma 3 (Basic properties). Assume $1 \leq k \leq r = \text{rank}(\mathcal{G})$, and let $p_i = \lambda_i(\mathcal{G})/\text{tr}_k(\mathcal{G})$ for $i = 1, \dots, k$. Then $p_i > 0$ and $\sum_{i=1}^k p_i = 1$. Moreover,

$$1 \leq r_{\text{eff}}^{(k)}(\mathcal{Z}) \leq k.$$

In addition, $r_{\text{eff}}^{(k)}(\mathcal{Z}) = 1$ iff $p = (1, 0, \dots, 0)$ (all mass on one component), and $r_{\text{eff}}^{(k)}(\mathcal{Z}) = k$ iff $p_i = 1/k$ for all $i = 1, \dots, k$ (uniform truncated spectrum).

Connection to participation ratio / Rényi-2 effective rank. A closely related notion is the participation ratio (also known as the Rényi-2 effective rank) of the truncated spectrum:

$$r_{\text{pr}}^{(k)}(\mathcal{Z}) := \frac{1}{\sum_{i=1}^k p_i^2} = \frac{(\sum_{i=1}^k \lambda_i(\mathcal{G}))^2}{\sum_{i=1}^k \lambda_i(\mathcal{G})^2} = \frac{\text{tr}(\mathcal{G}_k)^2}{\|\mathcal{G}_k\|_F^2}, \quad (20)$$

where $\|\mathcal{G}_k\|_F^2 = \sum_{i=1}^k \lambda_i(\mathcal{G})^2$.

Lemma 4 (Shannon effective rank lower-bounds participation ratio). $r_{\text{eff}}^{(k)}(\mathcal{Z}) \geq r_{\text{pr}}^{(k)}(\mathcal{Z})$.

Proof. Let $H(p) = -\sum_i p_i \log p_i$ be Shannon entropy and $H_2(p) = -\log \sum_i p_i^2$ be Rényi-2 entropy. It is standard that $H(p) \geq H_2(p)$ for any discrete distribution. Exponentiating both sides yields $\exp(H(p)) \geq \exp(H_2(p)) = 1/\sum_i p_i^2$, which is exactly $r_{\text{eff}}^{(k)} \geq r_{\text{pr}}^{(k)}$. \square

Interpretation for redundancy. For a fixed principal-subspace variance $\text{tr}(\mathcal{G}_k)$, a more redundant or collapsed representation concentrates spectral mass on fewer eigen-directions, which decreases $\mathcal{H}(\mathcal{Z})$ and therefore decreases both $r_{\text{eff}}^{(k)}$ and $r_{\text{pr}}^{(k)}$. Hence, $\mathcal{H}(\mathcal{Z})$ is a principled, scale-free proxy for how many directions are effectively used by the token representations within the top- k subspace.

C.2 Why Truncation is Well-Behaved

A decomposition of full entropy into head and tail. Let \tilde{p} denote the fully normalized spectrum over the r nonzero eigenvalues:

$$\tilde{p}_i = \frac{\lambda_i(\mathcal{G})}{\sum_{j=1}^r \lambda_j(\mathcal{G})}, \quad i = 1, \dots, r. \quad (21)$$

Define the tail mass beyond the top- k components as

$$\delta := \sum_{i>k} \tilde{p}_i = \frac{\sum_{i>k} \lambda_i(\mathcal{G})}{\sum_{j=1}^r \lambda_j(\mathcal{G})}. \quad (22)$$

When the spectrum has a clear elbow, δ is small. Let p be the renormalized head distribution, i.e.,

$p_i = \tilde{p}_i/(1 - \delta)$ for $i \leq k$. Also define the normalized tail distribution $q_j = \tilde{p}_{k+j}/\delta$ for $j = 1, \dots, r - k$ (when $\delta > 0$). Then the full entropy admits an exact decomposition:

Proposition 2 (Head-tail entropy decomposition). Let $H(\cdot)$ denote Shannon entropy and $h_2(\delta) = -\delta \log \delta - (1 - \delta) \log(1 - \delta)$ be binary entropy. For $\delta \in (0, 1)$, we have

$$H(\tilde{p}) = (1 - \delta) H(p) + h_2(\delta) + \delta H(q). \quad (23)$$

Proof. Write the head probabilities as $\tilde{p}_i = (1 - \delta)p_i$ for $i \leq k$ and the tail probabilities as $\tilde{p}_{k+j} = \delta q_j$ for $j \leq r - k$. Then

$$\begin{aligned} -\sum_{i \leq k} \tilde{p}_i \log \tilde{p}_i &= -(1 - \delta) \sum_{i \leq k} p_i \log((1 - \delta)p_i) \\ &= (1 - \delta) H(p) - (1 - \delta) \log(1 - \delta). \end{aligned} \quad (24)$$

and similarly

$$-\sum_{j \leq r-k} \tilde{p}_{k+j} \log \tilde{p}_{k+j} = \delta H(q) - \delta \log \delta. \quad (25)$$

Summing the two parts gives Eq. (23). \square

Implication. Eq. (23) shows that truncation is well-behaved when δ is small. In particular, since $0 \leq H(q) \leq \log(r - k)$, Proposition 2 implies

$$\frac{H(\tilde{p}) - h_2(\delta) - \delta \log(r - k)}{1 - \delta} \leq H(p) \leq \frac{H(\tilde{p}) - h_2(\delta)}{1 - \delta}. \quad (26)$$

Thus, the truncated (renormalized) entropy $H(p) = \mathcal{H}(\mathcal{Z})$ tracks the spectral diversity of the full distribution up to an $O(\delta \log r)$ correction, supporting truncation as a denoising step that suppresses small, unstable eigenvalues due to stochasticity, finite-sample effects, or irrelevant long-tail variations.

C.3 Connecting Spectral Entropy to Geometry via Curvature

Discrete curvature surrogate. To relate spectral diversity to geometry, we treat the token representations $\{\mathbf{h}_i\}_{i=1}^N$ as a polyline in \mathbb{R}^D (e.g., ordered by spatial rasterization of patches). Define the second-order finite difference $\Delta^2 \mathbf{h}_i := \mathbf{h}_{i+1} - 2\mathbf{h}_i + \mathbf{h}_{i-1}$ for $i = 2, \dots, N - 1$, and the curvature energy

$$\mathcal{C}(\mathcal{Z}) := \frac{1}{N - 2} \sum_{i=2}^{N-1} \|\Delta^2 \mathbf{h}_i\|_2^2. \quad (27)$$

Matrix form (curvature is a Gram-functional).

Let $L \in \mathbb{R}^{(N-2) \times N}$ be the second-difference operator such that $(L\mathcal{Z})_{i,:} = \Delta^2 \mathbf{h}_{i+1}^\top$. Then

$$\begin{aligned} \mathcal{C}(\mathcal{Z}) &= \frac{1}{N-2} \|L\mathcal{Z}\|_F^2 \\ &= \frac{1}{N-2} \text{tr}(\mathcal{Z}^\top L^\top L \mathcal{Z}) \\ &= \frac{1}{N-2} \text{tr}(L^\top L \mathcal{Z} \mathcal{Z}^\top) \\ &= \frac{1}{N-2} \text{tr}(L^\top L \mathcal{G}). \end{aligned} \quad (28)$$

Hence, curvature is completely determined by the Gram matrix \mathcal{G} .

Theorem 2 (Curvature is controlled by spectral concentration). Let \mathcal{G}_k be the rank- k truncated Gram matrix, and define

$$\mathcal{C}_k(\mathcal{Z}) := \frac{1}{N-2} \text{tr}(L^\top L \mathcal{G}_k), \quad (29)$$

i.e., the curvature energy induced by the principal subspace. Then

$$\mathcal{C}_k(\mathcal{Z}) \leq \frac{\|L^\top L\|_F}{N-2} \|\mathcal{G}_k\|_F = \frac{\|L^\top L\|_F}{N-2} \cdot \frac{\text{tr}(\mathcal{G}_k)}{\sqrt{r_{\text{pr}}^{(k)}(\mathcal{Z})}}. \quad (30)$$

Proof. By Cauchy–Schwarz for the Frobenius inner product, $|\text{tr}(A^\top B)| \leq \|A\|_F \|B\|_F$ for any conformable matrices A, B . Since $L^\top L \succeq 0$ and $\mathcal{G}_k \succeq 0$, we have $\text{tr}(L^\top L \mathcal{G}_k) \geq 0$ and hence

$$\mathcal{C}_k(\mathcal{Z}) = \frac{1}{N-2} \text{tr}(L^\top L \mathcal{G}_k) \leq \frac{1}{N-2} \|L^\top L\|_F \|\mathcal{G}_k\|_F. \quad (31)$$

Moreover, $\|\mathcal{G}_k\|_F = \sqrt{\sum_{i=1}^k \lambda_i(\mathcal{G})^2} = \text{tr}(\mathcal{G}_k) \sqrt{\sum_{i=1}^k p_i^2} = \text{tr}(\mathcal{G}_k) / \sqrt{r_{\text{pr}}^{(k)}(\mathcal{Z})}$, which gives Eq. (30). \square

Interpretation. Eq. (30) links *geometry* and *spectral concentration*: for fixed $\text{tr}(\mathcal{G}_k)$ and a fixed token order (hence fixed L), the curvature energy is upper-bounded by a term inversely proportional to $\sqrt{r_{\text{pr}}^{(k)}}$. Thus, a more spread spectrum (larger $r_{\text{pr}}^{(k)}$) yields a smaller upper bound on curvature. Since $r_{\text{eff}}^{(k)} = \exp(\mathcal{H})$ satisfies $r_{\text{eff}}^{(k)} \geq r_{\text{pr}}^{(k)}$ (Lemma 4), \mathcal{H} provides a convenient, scale-free summary of spectral diversity that empirically correlates with $r_{\text{pr}}^{(k)}$ and therefore with geometric stability.

C.4 From Curvature Energy to Angle-Based Curvature

Our main analysis uses $\mathcal{C}(\mathcal{Z})$ for tractability. If one uses an angle-based curvature, we can connect them as follows. Let $\mathbf{d}_i = \mathbf{h}_{i+1} - \mathbf{h}_i$ and let θ_i be the turning angle between \mathbf{d}_{i-1} and \mathbf{d}_i . Using $\|\mathbf{d}_i - \mathbf{d}_{i-1}\|_2^2 = \|\Delta^2 \mathbf{h}_i\|_2^2$ and

$$1 - \cos \theta_i \leq \frac{\|\mathbf{d}_i - \mathbf{d}_{i-1}\|_2^2}{2\|\mathbf{d}_i\|_2 \|\mathbf{d}_{i-1}\|_2}, \quad (32)$$

we obtain

$$1 - \cos \theta_i \leq \frac{\|\Delta^2 \mathbf{h}_i\|_2^2}{2s_0^2},$$

whenever $\|\mathbf{d}_i\|_2, \|\mathbf{d}_{i-1}\|_2 \geq s_0 > 0$. (33)

Averaging over i shows that the mean turning (hence angle-based curvature) is controlled by $\mathcal{C}(\mathcal{Z})$, which in turn is controlled by spectral concentration through Theorem 2.

Takeaway. The above results establish a principled chain:

$$\begin{aligned} \mathcal{H}(\mathcal{Z}) &\iff r_{\text{eff}}^{(k)}(\mathcal{Z}) \gtrsim r_{\text{pr}}^{(k)}(\mathcal{Z}) \implies \|\mathcal{G}_k\|_F \\ &\implies \mathcal{C}_k(\mathcal{Z}) \implies (\text{angle-based curvature}). \end{aligned} \quad (34)$$

Hence, truncated matrix entropy serves as a robust spectral probe for redundancy and (via its connection to spectral concentration) correlates with geometric complexity of internal token representations.

D Additional Cues for Global Aggregation

In this section, we further characterize the global aggregation behavior in Stage II by visualizing information flow in LLaVA-1.5v-7B across the three stages. Given an attention map $\mathbb{A} \in \mathbb{R}^{N \times N}$, we define the information inflow of each token as its primary information source, i.e., $\text{argmax}(\mathbb{A}[i, :])$. Based on this definition, we visualize the information flow in Stage I (Layers 12), Stage II (Layers 35), and Stage III (Layers 18, 19, 20, 21, and 23).

As shown in Figure 15, in the shallow layers of Stage I (Layers 12), visual information predominantly propagates among neighboring visual tokens, exhibiting localized and short-range flows. In Stage II, a clear global aggregation pattern emerges, where a small subset of visual tokens act as hub tokens that collect global information and

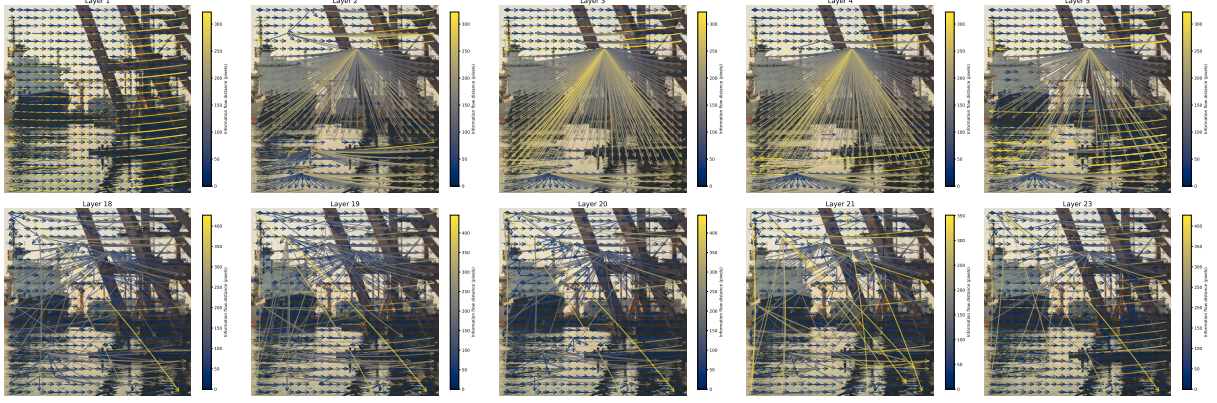


Figure 15: Layer-wise Information Flow in LLaVA-1.5v-7B

redistribute it to other visual tokens. Consequently, global aggregation in this stage can be viewed as **hub-driven global information exchange**. When the model enters Stage III, the influence of hub tokens diminishes and the information flow becomes more scattered and less structured, which is consistent with the observed increase in visual entropy in Stage III.

E Implementation Details of Marginal Utility

Definition of Marginal Utility. To quantify the cost-effectiveness of token pruning at different depths, we define the marginal utility (MU) as the ratio between performance degradation and efficiency gain. Specifically, for one-shot pruning applied at layer l with retention ratio r , we compute

$$\text{MU}_{l,r}^{\downarrow} = \frac{-\Delta\mathcal{M}_{l,r}}{\Delta\mathcal{C}_{l,r} + \epsilon}, \quad (35)$$

where $\Delta\mathcal{M}_{l,r}$ denotes the change in task performance relative to the unpruned baseline, $\Delta\mathcal{C}_{l,r}$ denotes the reduction in inference latency, and ϵ is a small constant for numerical stability. Lower MU indicates a more favorable efficiency-performance trade-off.

Experimental Protocol. For each pruning layer l , we apply one-shot token pruning once and keep the pruned token set fixed for all subsequent layers. All experiments use identical decoding settings and benchmarks. We evaluate multiple retention ratios $r \in \{1/2, 1/4, 1/8\}$ and report both performance and latency.

F Implementation Details of Redundancy Metric

We denote the LLM backbone as F with L Transformer layers. The input to layer ℓ is $\mathbf{Z}^{(\ell)} \in \mathbb{R}^{N \times D}$ and the standard block update is

$$\mathbf{z}^{(\ell+1)} = \mathcal{B}^{(\ell)}(\mathbf{z}^{(\ell)}), \quad (36)$$

where $\mathcal{B}^{(\ell)}$ consists of a self-attention module and a feed-forward network (FFN), with residual connections and normalization omitted for brevity. Let $\mathbf{z}_{\text{last}}^{(L+1)}$ be the hidden state at the last token position after the final layer. The next-token logits are computed by the language-model head $g(\cdot)$:

$$\mathbf{s}(F) = g\left(\mathbf{z}_{\text{last}}^{(L+1)}\right). \quad (37)$$

To quantify the contribution of layer ℓ on a token subset $\Omega \subset \{1, \dots, N\}$, we construct a terminated layer $\tilde{\mathcal{B}}_{\Omega}^{(\ell)}$ that keeps tokens in Ω unchanged at layer ℓ . The resulting model is denoted as $F^{(\ell, \Omega)}$. Concretely, for any input \mathbf{Z} , we define

$$\tilde{\mathcal{B}}_{\Omega}^{(\ell)}(\mathbf{Z})[i] = \begin{cases} \mathbf{Z}[i], & i \in \Omega, \\ \mathcal{B}^{(\ell)}(\mathbf{Z})[i], & i \in \bar{\Omega}, \end{cases} \quad (38)$$

where $\bar{\Omega}$ is the complement of Ω . In implementation, we apply this termination to both submodules: (i) in self-attention, we only compute updates for query positions $i \in \bar{\Omega}$, while positions $i \in \Omega$ bypass the attention update; (ii) in the FFN, we only forward tokens in $\bar{\Omega}$ and skip FFN updates for tokens in Ω . In practice, Ω can be chosen as the set of visual tokens Ω_V or text tokens Ω_T .

We define the layer redundancy degree of layer ℓ with respect to token subset Ω as the KL divergence between the output distributions of

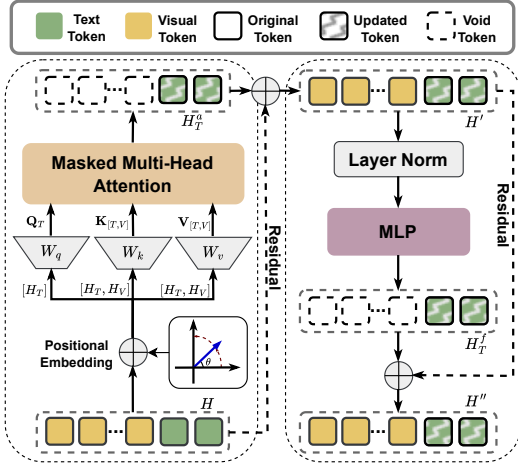


Figure 16: Illustration of Handling Inactivity Layers

the original model and the terminated model: $\text{KL}(\text{softmax}(s(F)) \parallel \text{softmax}(s(F^{(\ell, \Omega)})))$. A smaller value indicates that terminating Ω at layer ℓ barely changes the final prediction, suggesting that the transformations on Ω in this layer contribute little to the model output and may be ineffective or redundant.

G Implementation Details of HalfV

G.1 Layer-level Inactivity Implementation Details

As shown in Figure 16, we consider a decoder layer with input hidden states $\mathbf{H} \in \mathbb{R}^{N \times d}$, where $N = N_V + N_T$ denotes the total number of tokens, consisting of N_V visual tokens and N_T textual tokens. For notational convenience, we write the sequence as $\mathbf{H} = \text{Concat}(\mathbf{H}_V, \mathbf{H}_T)$ and similarly decompose the query matrix $\mathbf{Q} = \text{Concat}(\mathbf{Q}_V, \mathbf{Q}_T)$. Keys and values are computed from all tokens, yielding $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_k}$.

Self-Attention with Visual Update Termination.

We only compute attention outputs for textual queries, while terminating the attention update for visual tokens. Specifically, the attention-induced updates are defined as

$$\Delta \mathbf{H}_T^a = \text{softmax}\left(\frac{\mathbf{Q}_T \mathbf{K}^\top}{\sqrt{d_k}}\right) \mathbf{V}, \quad \Delta \mathbf{H}_V^a = \mathbf{0}, \quad (39)$$

where $\frac{\mathbf{Q}_T \mathbf{K}^\top}{\sqrt{d_k}} \in \mathbb{R}^{N_T \times N}$ and d_k is the key dimension. After the residual connection, the post-attention hidden states are

$$\mathbf{H}' = \mathbf{H} + \text{Concat}(\Delta \mathbf{H}_V^a, \Delta \mathbf{H}_T^a). \quad (40)$$

Note that this construction preserves visual states in the attention sublayer, i.e., $\mathbf{H}'_V = \mathbf{H}_V$.

FFN with Visual Update Termination. In the FFN sublayer, we similarly update only textual tokens and terminate updates for visual tokens:

$$\Delta \mathbf{H}_T^f = \text{FFN}(\mathbf{H}'_T), \quad \Delta \mathbf{H}_V^f = \mathbf{0}. \quad (41)$$

The final output of the layer is then given by

$$\mathbf{H}'' = \mathbf{H}' + \text{Concat}(\Delta \mathbf{H}_V^f, \Delta \mathbf{H}_T^f), \quad (42)$$

so that visual tokens remain unchanged throughout this layer, while textual tokens are updated as usual.

G.2 RoPE: enabled or disabled?

To study the impact of enabling or disabling positional encoding at different stages, we conduct experiments on **POPE** and **ChartQA** with Qwen2.5-VL-7B. Specifically, for both the IVR and SSR modules, we toggle positional encoding at the points where attention scores are used for decision making, and report the results in Table 8. The results show that disabling positional encoding in the IVR module improves performance, suggesting that early layers suffer from positional bias induced by positional encoding, and removing it can partially mitigate this bias. In contrast, enabling positional encoding in the SSR module yields better performance, indicating that after Stage-II global aggregation, the attention mechanism can more reliably capture the truly relevant information and positional bias becomes less pronounced during forward inference. These findings suggest that positional encoding should be configured in a stage-aware manner, enabling or disabling it at different points of inference to improve localization and reasoning.

G.3 Implementation Details of Qwen3-VL

When implementing acceleration on Qwen3-VL-4B-FP, we account for the DeepStack mechanism used in the Qwen3-VL family, which injects visual features from the vision encoder at layers 5/11/17 into the decoder layers 1/2/3 via residual connections. Accordingly, for the IVR module, we shift the pruning location to decoder layer 4, i.e., after the third decoder layer. For a fair comparison, all baseline methods also perform visual token pruning at decoder layer 4.

Setting	IVR	SSR	w/ rope	POPE				ChartQA		
				Accuracy	F1 Score	Precision	Recall	Augmented Split	Human Split	Overall
$R_{IVR} = 0.5, R_S = 0.2$	✓	✗	✓	86.94	85.50	96.14	76.98	92.64±0.74	64.56±1.35	78.60±0.82
$R_{IVR} = 0.5, R_S = 0.2$	✓	✗	✗	87.12	85.71	96.24	77.27	92.16±0.76	65.12±1.35	78.64±0.82
$R_{IVR} = 0.25, R_S = 0.1$	✓	✗	✓	86.14	84.41	96.46	75.04	83.20±1.06	48.16±1.41	65.68±0.95
$R_{IVR} = 0.25, R_S = 0.1$	✓	✗	✗	86.32	84.65	96.45	75.42	82.8±1.07	49.12±1.41	66.00±0.95
$R_{SSR} = 0.15$	✗	✓	✓	87.70	86.45	96.34	78.40	93.60±0.69	72.32±1.27	82.96±0.75
$R_{SSR} = 0.15$	✗	✓	✗	87.63	86.36	96.26	78.31	93.76±0.68	71.76±1.27	82.76±0.76
$R_{SSR} = 0.05$	✗	✓	✓	87.73	86.48	96.32	78.47	92.96±0.72	70.00±1.30	81.48±0.78
$R_{SSR} = 0.05$	✗	✓	✗	87.69	86.42	96.31	78.38	92.80±0.73	69.84±1.30	81.32±0.78

Table 8: Performance comparison of Qwen2.5-VL-7B under different settings when enabling or disabling RoPE in the IVR and SSR stages.

H Case Study

H.1 Llava Model Case

We provide a comprehensive qualitative analysis of LLaVA-1.5-7B across three diverse benchmarks MME, POPE, and ScienceQA as visualized in Figure 17, 18 and 19.

Sensitivity to Stylistic Textures (MME). Figure 17a illustrates the impact of pruning on artwork perception. In the *left example*, when queried if an oil painting is “architecture,” **FastV** incorrectly answers “Yes.” This failure suggests that saliency-based pruning aggressively removes background stylistic tokens (e.g., brushstrokes) while retaining structural objects like the fence, leading the model to misinterpret the scene as a physical building. Similarly, in the *right example* (a sketch), **FastV** hallucinates that the religious scene is a “still-life.” The attention mechanism assigns low scores to the sparse, thin ink lines, causing **FastV** to prune the structural outlines and destroy the semantic meaning of the crowd. In contrast, **HalfV** preserves textural diversity both the brushstrokes and the high-contrast ink lines enabling the model to correctly identify the artwork forms.

Mitigation of Object Hallucination (POPE). We further examine robustness against hallucination in Figure 18a. For the *skiing scene* (left), **FastV** hallucinates a “dining table” due to the loss of environmental context; by pruning the low-saliency snowy background, the model loses the scene constraints needed to reject the unlikely object. In the *urban scene* (right), **FastV** suffers from saliency bias, focusing heavily on the elevated train while pruning the cluttered street level. Lacking fine-grained evidence of the specific vehicles, it hallucinates a “truck” based on street scene priors. **HalfV** effectively mitigates these issues by retaining diverse background and non-salient

tokens, providing the necessary context to verify object absence.

Preservation of Visual Cues for Reasoning (ScienceQA). Finally, Figure 19a investigates multimodal reasoning. In the *state change task* (left), **FastV** misidentifies melting ice as a “chemical change” because the melted water on the dark surface has significantly lower saliency than the ice cubes. Pruning the “water” tokens leaves the model with a static view, leading to hallucinated descriptions of bubbles. In the *map reading task* (right), **FastV** fails to locate the equator, incorrectly selecting “Australia.” This is attributed to the pruning of grid lines, which are thin and occupy minimal pixel area. Without these coordinate references, spatial reasoning collapses. By recognizing the high feature diversity of liquid textures and geometric grid lines, **HalfV** preserves these critical cues, enabling accurate scientific and spatial reasoning.

H.2 Qwen Model Case

We extend our qualitative analysis to the Qwen-2.5-VL-7B backbone, examining its performance across the MME, POPE, and ScienceQA benchmarks.

Perception of Peripheral and Fine-Grained Details (MME). Figure 17b highlights the challenge of preserving information in non-salient regions. In the *movie poster task* (left), the title “THE NEON DEMON” is positioned at the extreme bottom edge. **FastV** incorrectly answers “Yes” to an adversarial title (“Man on Fire”), indicating that saliency-based pruning assigned low scores to the image periphery, effectively discarding the textual anchor. Without the text, the model relied on the high-contrast visual style to hallucinate a match. In the *cluttered movie scene* (right), **FastV** fails to reject a false title due to

background interference. The high-saliency red wallpaper dominates the attention map, causing the model to prune the less salient but identity-defining facial features of the actors (e.g., Robin Williams). **HalfV** overcomes both issues by identifying the high feature diversity of the neon text and the unique facial textures, preserving these critical tokens regardless of their position or background clutter.

Robustness to Color Distraction (POPE). Figure 18b demonstrates a specific failure mode we term *Color Distraction*. When queried about a “giraffe” (left) or a “banana” (right), **FastV** fails to reject their presence, misled by a yellow taxi and yellow pavement lines, respectively. This suggests a failure to disentangle color from object identity: saliency-based pruning retains the dominant “yellow” color tokens but discards the high-frequency tokens representing the taxi’s metallic texture or the pavement lines’ geometric edges. Consequently, the language model, primed by the color cue, hallucinates objects associated with yellow. **HalfV** mitigates this by utilizing diversity-based pruning to retain distinct textural and geometric constraints (e.g., sharp edges), enabling the model to correctly identify the visual inputs as a vehicle and 2D lines.

Visual Evidence for Complex Reasoning (ScienceQA). Finally, Figure 19b investigates the impact of pruning on multi-hop reasoning. In the *map reading task* (left), **FastV** incorrectly predicts “10°C” because it fails to perceive the legend. The legend, containing fine text at the periphery, is treated as background noise and pruned, blinding the model to the reference scale. Similarly, in the *trade reasoning task* (right), **FastV** proposes an incorrect trade involving “almonds” because it cannot find the “broccoli” requested in the text prompt. The broccoli, being a small, dark object in the corner, has low saliency and is pruned. **HalfV** succeeds in both cases by recognizing the high feature diversity of the legend’s text-grid transitions and the broccoli’s texture, preserving the visual evidence necessary for accurate reasoning.

I Detailed Experiment Settings

I.1 Datasets

We performed experiments on eight widely used benchmarks, including GQA (Hudson and Manning, 2019), MME (Fu et al., 2025), POPE (He et al., 2025), SQA (Lu et al., 2022), VizWiz

(Bigham et al., 2010), TextVQA (VQA^{text}) (Singh et al., 2019), MMB (Liu et al., 2024c) and AI2D (Kembhavi et al., 2016).

GQA consists of three key components: scene graphs, questions, and images. The image component includes the images themselves, as well as their spatial features and the attributes of all objects depicted within them. The questions in GQA are carefully designed to evaluate the model’s ability to understand visual scenes and reason about various elements of the images.

MME is intended to rigorously assess a model’s perceptual and cognitive capabilities through 14 distinct subtasks. It uses carefully designed instruction-answer pairs and clear instructions to minimize data leakage and ensure fair evaluation. This setup provides a reliable measure of a model’s performance across a variety of tasks.

POPE is specifically designed to evaluate object hallucination. It consists of a series of binary questions regarding the presence of objects in images, using accuracy, recall, precision, and F1 score as metrics. This method provides a precise assessment of hallucination levels under various sampling strategies.

SQA covers a broad range of domains, including natural sciences, language, and social sciences. The questions are organized hierarchically into 26 topics, 127 categories, and 379 skills, offering a diverse and comprehensive framework for evaluating multimodal understanding, multi-step reasoning, and interoperability.

VizWiz is a visual benchmark created to support visually impaired individuals. It consists of real-world images taken by blind users, each paired with questions they ask about the images. The dataset includes 20,523 training, 4,319 validation, and 8,000 test image-question pairs, with each question having 10 human-annotated answers. VizWiz challenges models to either provide accurate answers to the questions or determine if a question is answerable, emphasizing practical visual understanding and accessibility.

TextVQA focuses on the integration of textual information within images. It tests a model’s ability to read and reason about text embedded in visual content, requiring the model to understand both visual and textual elements in order to answer questions correctly.

MMB provides a hierarchical evaluation framework that categorizes model capabilities into three levels. The first level (L-1) focuses on perception

and reasoning. The second level (L-2) expands on this by introducing six sub-abilities, while the third level (L-3) further refines these into 20 specific dimensions. This structured approach enables a detailed and comprehensive assessment of a models diverse capabilities.

AI2D is a dataset that contains over 5,000 elementary school science diagrams, with more than 150,000 rich annotations, their basic factual syntactic parsing, and over 15,000 corresponding multiple-choice questions.

I.2 Models

We evaluate HalfV using various open-source MLLMs. In particular, we assess our method on four prominent models: LLaVA-v1.5-7B*, LLaVA-v1.5-13B[†] (Liu et al., 2023), LLaVA-NeXT-7B[‡] (Liu et al., 2024b), and Qwen2.5-VL 7B-Instruct[§] (Bai et al., 2025).

LLaVA-v1.5. LLaVA-v1.5 models process images with a 336×336 resolution and treat each image as 576 tokens. Each image is divided into non-overlapping patches, with each patch being treated as a token, enabling the model to process the image in a way that aligns with the transformer architecture used for multimodal tasks.

LLaVA-NeXT. LLaVA-NeXT divides high-resolution images into smaller subimages and encodes both the subimages and downsampled original images independently. This approach enables the model to scale the input to any arbitrary resolution without the need for positional embedding interpolation, which is typically required for Vision Transformers (ViTs) (Dosovitskiy et al., 2021). Compared to LLaVA-1.5, LLaVA-NeXT scales the input image resolution by 4 \times and increases the number of visual tokens by up to 5 \times , resulting in 2880 tokens per image.

Qwen2.5-VL. Qwen2.5-VL’s architecture is designed to handle large-scale vision-language tasks by efficiently processing high-resolution images, making it more capable of understanding and generating accurate descriptions of images. This makes it particularly effective for real-world applications that require both high visual fidelity and

multimodal reasoning. Additionally, Qwen2.5-VL employs advanced attention mechanisms that allow it to efficiently scale to larger datasets while maintaining accuracy across a wide range of benchmarks.

I.3 Baselines

To evaluate the effectiveness of HalfV, we compare it with mainstream methods based on token-level redundancy and layer-level redundancy. Token-level redundancy-based methods mainly focus on how to select and prune visual tokens, which can be further divided into attention-based visual token pruning and diversity-based visual token pruning. In the comparative experiments, we use FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024b), PDrop (Xing et al., 2024), VisionZip (Yang et al., 2025c), HoloV (Zou et al., 2025), and other attention-based methods, as well as DivPrune (Alvar et al., 2025) and other diversity-based visual methods as baselines. Additionally, we include layer-level redundancy-based visual token pruning methods such as BTP (Li et al., 2025) and VTW (Lin et al., 2025) for comparison. As a novel approach based on layer locking, ShortV (Yuan et al., 2025) is also included as one of our baselines for comparison.

FastV concentrates on pruning tokens in the early stages by utilizing attention maps, thereby significantly reducing computational costs in the initial layers.

SparseVLM assesses token importance through cross-modal attention and incorporates adaptive sparsity ratios, along with an innovative token recycling mechanism.

PDrop employs a progressive token-dropping approach throughout the model stages, creating a pyramid-shaped token structure that optimizes both efficiency and performance.

VisionZip determines token importance using attention in the encoder and clusters the remaining tokens based on key similarity.

DivPrune addresses the token pruning challenge in LMMs by formulating it as a Max-Min Diversity Problem (MMDP), selecting a subset of tokens that maximize diversity.

BTP reduces the number of vision tokens by pruning in multiple stages, initially focusing on the global impact on subsequent layers and later emphasizing the preservation of local output consistency.

VTW drops all visual tokens after the K-th layer,

*<https://huggingface.co/liuhaotian/llava-v1.5-7b>

[†]<https://huggingface.co/liuhaotian/llava-v1.5-13b>

[‡]<https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b>

[§]<https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>

Model	L_{IVR}	R_{IVR}	R_S	L_{SSR}
LLaVA-1.5v-7B (32 Layers)	3	50%	0.2	15
LLaVA-1.5v-13B (40 Layers)	3	50%	0.2	15
LLaVA-NeXT-7B (32 Layers)	2	50%	0.1	16
Qwen2.5-VL-7B (28 Layers)	2	[25%,5%]	0.1	21

Table 9: Hyperparameters for HalfV

enabling only text tokens to engage in the subsequent layers.

ShortV uses a novel metric, Layer Contribution (LC), to identify ineffective layers in Multimodal Large Language Models (MLLMs) and freezes visual token updates in these layers, significantly reducing computational costs while maintaining performance.

I.4 Implementation Details

All of our experiments are conducted on NVIDIA GeForce RTX 4080 SUPER (32G) GPU. The implementation was carried out in Python 3.10, utilizing PyTorch 2.1.2, and CUDA 13.0. All baseline settings follow the original paper.

I.5 Hyperparameters

In the real test environment, four hyperparameters need to be defined for the inference process: the number of IVR layer L_{IVR} , the retention ratio R_{IVR} , the anchor retention ratio R_S , and the SSR starting layer L_{SSR} . To facilitate the reproduction of our experimental results, we list the parameter settings used for different models in Table 9. Note that L_{IVR} and L_{SSR} correspond to the starting layers of Stage II and Stage III, respectively. The detailed recognition process for each stage can be found in Section 3.1 and Appendix B.

The experiment also includes other baseline methods, with specific experimental parameters shown in Table 10. For single-stage visual token pruning methods, such as FastV, FitPrune, and SparseVLM, the pruning layer number K and the pruning ratio R . For multi-stage visual token pruning methods, the list of pruning layers and the ratio of retained tokens after each pruning stage must be provided. For the reproduction of the VTW method, we follow the approach in the original paper, removing all visual tokens at the $L/2$ layer, where L is the total number of layers. For the reproduction of the ShortV method, we lock the visual tokens based on the layer number N selected in the original paper.

In the efficiency evaluation and the module-efficiency ablation experiments, we use the same

Model	Method	R	K	N
LLaVA-1.5v-7B	Single-stage	33.3%	2	-
	Multi-stage	[50%,75%,87.5%]	[5,10,20]	-
	ShortV	-	-	19
LLaVA-1.5v-13B	Single-stage	33.3%	2	-
	Multi-stage	[50%,75%,87.5%]	[5,10,20]	-
	ShortV	-	-	24
LLaVA-NeXT-7B	Single-stage	33.3%	2	-
	Multi-stage	[50%,75%,87.5%]	[5,10,20]	-
	ShortV	-	-	24
Qwen2.5-VL-7B	Single-stage	22.2%	2	-
	Multi-stage	[50%,75%,87.5%]	[5,10,20]	-
	ShortV	-	-	19

Table 10: Baseline Settings

settings as in the main experiments for both LLaVA-NeXT-7B and Qwen2.5-VL-7B. For the anchor retention ratio sensitivity study, we set the overall pruning ratio to 50% on LLaVA-NeXT-7B and 77.8% on Qwen2.5-VL-7B. For visualization, we fix the overall pruning ratio to 77.8% for both LLaVA-1.5v-7B and Qwen2.5-VL-7B. In addition, for the sensitivity study on the SSR starting layer, we set the sparse-token retention ratio to 5% on Qwen2.5-VL-7B.

I.6 Computational Complexity

We analyze the computational operations in the self-attention mechanisms and FFNs within the layers of the underlying LLM architecture. Let t denote the count of text tokens, v is the count of visual tokens, h is the hidden state dimension, and m is the intermediate dimension of the FFNs. The FLOPs in the three stages are calculated as:

$$\begin{cases} F_I = 2(t+v)(4h+3m)h + 4(t+v)^2h, \\ F_{II} = 2(t+v')(4h+3m)h + 4(t+v')^2h, \\ F_{III} = \begin{cases} 2t(4h+3m)h + 4t(t+v')h, \\ 2(t+v_{ssr})(4h+3m)h + 4(t+v_{ssr})^2h, \end{cases} \end{cases} \quad (43)$$

where v' is the number of visual tokens after Stage II pruning. For Stage III, F_{III} depends on the architecture-aware strategy: for *Layer-level Inactivity* (e.g., LLaVA), visual updates are terminated ($v_{active} = 0$), so only text tokens consume compute while attending to the full context v' ; for *Extreme Token Sparsity* (e.g., Qwen), computation is restricted to the top active visual tokens v_{ssr} (where $v_{ssr} \ll v'$). The total FLOPs for the entire process is then the weighted sum of the FLOPs for each stage, calculated as:

$$FLOPs = L_I \times F_I + L_{II} \times F_{II} + L_{III} \times F_{III}, \quad (44)$$

where L_I , L_{II} , L_{III} represent the number of layers for each corresponding stage.

Model	t	v	h	m
LLaVA-1.5v-7B	50	576	4096	11008
LLaVA-1.5v-13B	50	576	5120	13824
LLaVA-NeXT-7B	50	2352	4096	14336
Qwen2.5-VL-7B	50	2352	3584	18944

Table 11: Computational Settings

In the computational complexity analysis of Section I.6, the number of text tokens t and visual tokens v are required. Since the number of text tokens varies in real-world testing environments, we assume $t = 50$ for consistency. Additionally, for the LLaVA-1.5v series models, the number of visual tokens generated is fixed at 576. However, for the LLaVA-NeXT and Qwen series models, the number of visual tokens is dynamically changing. Based on the demand for high-resolution images in real-world scenarios, we set $v = 2352$. The other parameter settings are determined by the specific characteristics of each model, with the detailed data presented in Table 11.

J Clarifications and Differences from Related Work

Current approaches for accelerating multimodal large models during the prefill stage can be broadly categorized into two main directions: token-level acceleration and layer-level acceleration. The former studies how to design effective selection strategies that retain informative tokens while removing redundant ones to reduce computation. The latter focuses on identifying inefficient or redundant layers and lowering overall cost through layer skipping, early exiting, or freezing. We systematically analyze the strengths and limitations of both lines of work.

Token-level methods can significantly speed up prefill by retaining only a small number of visual tokens in early layers. However, under aggressive pruning ratios, especially for training-free schemes, model performance often degrades substantially, making the final outputs fall short of practical expectations. In contrast, layer-level methods typically preserve model quality more reliably, but their speedup is constrained by a clear bottleneck. When processing high-resolution images or videos that produce a large number of visual tokens, even keeping only a few dense layers can still incur considerable computation, so the overall acceleration often struggles to exceed 2 \times .

Moreover, we find that such methods are highly model-dependent. For architectures with more active inter-layer representations and few ineffective layers, it is difficult to maintain original performance while sparsifying computation across layers.

Prior work usually designs acceleration strategies from a single perspective. Our observations suggest that a multi-level, model-aware acceleration design grounded in internal inference dynamics can improve efficiency while preserving performance to a much greater extent. In this section, we further clarify and distinguish our method from several related works that share partially similar designs from certain viewpoints.

J.1 Comparison with ShortV

ShortV (Yuan et al., 2025) proposes a layer contribution metric to estimate how much each layer affects the model output on a calibration set, thereby identifying redundant layers. It then reduces computation by freezing visual tokens in these layers, avoiding their matrix operations during the forward pass. We find that this strategy is highly dependent on the models internal redundancy structure. For the LLaVA family, where inter-layer redundancy is relatively high, ShortV can effectively skip visual-token computation in redundant layers and thus improve inference speed.

However, when transferring this approach to the Qwen family, freezing even 5–6 relatively redundant layers leads to a substantial performance drop. We further analyze the layer-wise redundancy of Qwen2.5-VL-7B and observe that its redundancy is 2–3 orders of magnitude lower than that of LLaVA. This indicates that Qwen models exhibit very limited inter-layer redundancy, making layer-redundancy-based acceleration less suitable.

In contrast, we observe that Qwen, despite its low layer redundancy, still exhibits a visual saturation phenomenon similar to LLaVA. Moreover, in Qwen this phenomenon manifests as extreme token sparsity. We explicitly leverage this architecture-dependent property to achieve a better balance between high speedup and strong performance on the Qwen family.

J.2 Comparison with BTP

BTP (Li et al., 2025) applies multi-stage pruning and selects pruning layers via inter-layer co-

sine similarity on a fixed sample set. It thus relies on a static, data-driven layer profile. By contrast, our method is grounded in the functional lifecycle of redundancy. We explicitly align acceleration with how redundancy emerges and evolves.



J.3 Comparison with PDrop

PDrops (Xing et al., 2024) key contribution is to reveal an important phenomenon: LVLMs are much more sensitive to visual-token pruning in shallow layers, while pruning in deeper layers is considerably more robust. However, PDrop does not further investigate the underlying mechanism behind this observation. In contrast, our work provides a principled explanation through Truncated Matrix Entropy (TME). We show that the robustness in deep layers stems from a global aggregation stage inside the model, after which critical information can be consolidated into a small subset of visual tokens.



Moreover, PDrop adopts a uniform stage partition and prunes tokens progressively from shallow to deep layers based on intuition. Different from this heuristic design, our method leverages internal model properties to identify more appropriate processing layers. Empirically, we achieve better performance than PDrop under comparable speedup ratios, which we attribute to our model-aware design rather than an intuitive, uniform pruning schedule.

K Ethics Statement

This work utilizes public datasets and open-source models, and we identify no significant ethical issues regarding data privacy or potential misuse. In accordance with the ACL Policy on AI Writing Assistance, we declare that AI tools were employed solely for the purpose of refining the writing style, correcting grammatical errors, and improving readability. The generation of core ideas, experimental design, data analysis, and the formulation of scientific conclusions were conducted entirely by the human authors.


<p>Does this artwork exist in the form of painting? Please answer yes or no.</p> 	<p>Does this artwork belong to the type of religious? Please answer yes or no.</p> 
<p>Does this artwork exist in the form of architecture? Please answer yes or no.</p>	<p>Does this artwork belong to the type of still-life? Please answer yes or no.</p>
<p>Yes.</p> <p>Yes.</p> <p>Yes.</p>	<p>Yes.</p> <p>Yes.</p> <p>Yes.</p>
<p>No.</p> <p>Yes.</p> <p>No.</p>	<p>No.</p> <p>Yes.</p> <p>No.</p>
<p>Llava-1.5-7B (FastV)</p> <p>Llava-1.5-7B (HalfV)</p>	<p>Llava-1.5-7B (FastV)</p> <p>Llava-1.5-7B (HalfV)</p>

(a) **MME (Artwork):** FastV fails to reject false attributes (e.g., classifying a painting as architecture) due to the loss of stylistic textures.

<p>Is this movie titled the neon demon (2016)? Please answer yes or no.</p> 	<p>Is this movie titled the fisher king (1991)? Please answer yes or no.</p> 
<p>Is this movie titled man on fire (2004)? Please answer yes or no.</p>	<p>Is this movie titled a good day to die hard (2013)? Please answer yes or no.</p>
<p>Yes.</p> <p>Yes.</p> <p>Yes.</p>	<p>Yes.</p> <p>Yes.</p> <p>Yes.</p>
<p>No.</p> <p>Yes.</p> <p>No.</p>	<p>No.</p> <p>Yes.</p> <p>No.</p>
<p>Qwen-2.5-VL-7B (FastV)</p> <p>Qwen-2.5-VL-7B (HalfV)</p>	<p>Qwen-2.5-VL-7B (FastV)</p> <p>Qwen-2.5-VL-7B (HalfV)</p>

(b) **MME (Fine-Grained Perception):** FastV fails adversarial queries by pruning critical peripheral text (Left) or fine-grained facial features (Right).

Figure 17: Qualitative Case Studies on MME



Is there a person in the image?

Yes, there is a person in the image. I can see a skier wearing a blue jacket and orange helmet coming down the slope.

Yes, there is a person.

Yes, there is a person in the image.

Is there a backpack in the image?

Yes, there is a backpack. The skier is carrying a backpack, which appears to be orange or brown.

Yes, there is a backpack.

Yes, there is a backpack visible on the skier's back.


Is there a dining table in the image?

No, there is no dining table in the image.

Yes, there is a dining table.

No, there is no dining table in the image.

Llava-1.5-7B (FastV) Llava-1.5-7B (FastV) Llava-1.5-7B (HalfV)



Is there a train in the image?

Yes, there is a train in the image. It is moving along the elevated railway track above the street.

Yes, there is a train.

Yes, I can clearly see a train on the bridge.

Is there a truck in the image?

No, there is no truck in the image.

Yes, there is a truck on the road.

No, there is no truck in the image.

Is there a dog in the image?

No, there is no dog in the image.

No, there is no dog.

No, I do not see a dog.

Llava-1.5-7B (FastV) Llava-1.5-7B (FastV) Llava-1.5-7B (HalfV)

(a) **POPE (Hallucination):** FastV hallucinates objects (e.g., a dining table or truck) caused by pruning environmental context or non-salient details.



Is there a person in the image?

Yes, there is a person in the image. There are several people walking on the wet street, some holding umbrellas to shelter from the rain.

Yes, there is a person.

Yes, I can see people walking in the street.

Is there a giraffe in the image?

No, there is no giraffe in the image.

Yes, there is a giraffe.

No, there is no giraffe in the image.

Is there a handbag in the image?

Yes, there is a handbag. The woman on the right, holding a red umbrella, is carrying a patterned shoulder bag.

Yes, there is a handbag.

Yes, I can clearly see a handbag carried by the woman on the right.

Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (HalfV)



Is there a truck in the image?

Yes, there is a truck in the image. It appears to be a white commercial van or broadcasting vehicle equipped with a satellite dish.

Yes, there is a truck.

Yes, I can see a white truck or van.

Is there a banana in the image?

No, there is no banana in the image.

Yes, there is a banana.

No, there is no banana in the image.

Is there a car in the image?

Yes, there is a car in the image. The white van can be considered a type of car, and there may be other vehicles in the background.

No, there is no dog.

No, I do not see a dog.


Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (HalfV)

(b) **POPE (Color Distraction):** FastV suffers from color-induced hallucinations, misinterpreting yellow objects (a taxi or pavement markings) as a “giraffe” or “banana” due to the loss of texture and geometry.

Figure 18: Qualitative Case Studies on POPE

Which term matches the picture?
["chemical change", "physical change"]

hint: Read the text. When a substance goes through a physical change, its form or shape changes, but its chemical composition does not. For example, when water freezes and becomes a solid, its form changes, but it is still made up of the same molecules. In a chemical change, by contrast, one or more of the molecules in a substance changes. For example, when vinegar is mixed with baking soda, water and carbon dioxide molecules form.




The term that matches the picture is **physical change**. The image shows ice cubes melting into water on a surface. According to the hint, a physical change occurs when a substance changes its form but not its chemical composition. Since melting changes water from solid to liquid without changing the molecules, it is a physical change.

The term that matches the picture is **chemical change**. The image displays ice cubes. I can see bubbles forming inside the ice, which suggests that a gas is being released or a reaction is taking place. The hint states that in a chemical change, new molecules are formed. Therefore, the formation of bubbles indicates a chemical change.

The term that matches the picture is **physical change**. The image clearly shows a process where ice cubes are turning into liquid water. This is a change in the state of matter, specifically melting. Based on the hint, when water freezes or melts, its form changes but it is still made of the same molecules. Thus, this is a physical change.

Llava-1.5-7B (FastV) Llava-1.5-7B (HalfV)

Which of these continents does the equator intersect?
["North America", "Australia", "Africa"]



The correct answer is **Africa**. Based on the lecture, the equator is the line located at 0° latitude. Looking at the map, I can identify the horizontal line labeled '0°' in the middle. Tracing this line across the map, it clearly passes through the continent labeled **South America** and the continent labeled **Africa**. It lies to the south of North America and to the north of Australia. Since 'Africa' is the only correct option provided, that is the answer.

The correct answer is **Australia**. The equator is the imaginary line in the middle of the earth. In the image, I can see the continents. Australia appears to be located near the central region of the map where the equator would be. North America is too far up. Therefore, the equator intersects Australia.

The correct answer is **Africa**. According to the hint, the equator is at 0° latitude. On the map, I can clearly see the grid lines. The line marked 0° runs horizontally across the center. By following this specific line, I can see it intersects the landmass labeled **Africa**. In contrast, the line is above Australia and below North America. Thus, the equator passes through Africa.

Llava-1.5-7B (FastV) Llava-1.5-7B (HalfV)

(a) **ScienceQA (Reasoning):** FastV fails in multimodal reasoning tasks (state changes and map reading) due to the loss of critical fine-grained visual cues.

Which air temperature was measured within the outlined area shown? ["32°C", "10°C", "5°C"]

hint: The map below shows air temperatures in the lower atmosphere on October 1, 2015. The outlined area shows an air mass that influenced weather in North America on that day. Look at the map. Then, answer the question below. Data source: United States National Oceanic and Atmospheric Administration/Earth System Research Laboratory, Physical Sciences Division.



The correct answer is **32°C**. First, I located the area outlined in white on the map, which is in the southwestern region. The color inside this outline is a deep dark orange. Next, I looked at the legend at the bottom of the map. The legend shows that the dark orange/brown color corresponds to the temperature range between 30°C and 35°C. Among the choices, 32°C is the only value that fits within this range.

The correct answer is **10°C**. The map shows air temperatures across North America. I can see the outlined area, but based on the general color patterns, it seems to represent a moderate temperature. The legend is not clearly visible to define the exact range, but light orange usually suggests around 10°C.

The correct answer is **32°C**. I can clearly see the white oval outline enclosing a dark orange region. Although the legend at the bottom is small, I can read the scale. The color of the outlined area matches the far right side of the legend, which indicates temperatures above 25°C, specifically in the 30-35°C bracket. Therefore, 32°C is the correct temperature.

Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (HalfV)

What can Riley and Ken trade to each get what they want?
["Riley can trade her tomatoes for Ken's broccoli.", "Ken can trade his broccoli for Riley's oranges.", "Riley can trade her tomatoes for Ken's sandwich.", "Ken can trade his almonds for Riley's tomatoes."]

hint: Trade happens when people agree to exchange goods and services. People give up something to get something else. Sometimes people barter, or directly exchange one good or service for another. Riley and Ken open their lunch boxes in the school cafeteria. Both of them could be happier with their lunches. Riley wanted broccoli in her lunch and Ken was hoping for tomatoes. Look at the images of their lunches. Then answer the question below.



The correct answer is **Riley can trade her tomatoes for Ken's broccoli**. According to the hint, Riley wants broccoli and Ken wants tomatoes. Looking at the images: **Riley's lunch** (left) contains cherry tomatoes. **Ken's lunch** (right) contains broccoli florets. Since Riley has what Ken wants (tomatoes) and Ken has what Riley wants (broccoli), they can trade these two items so that both get what they desire.

The correct answer is **Ken can trade his almonds for Riley's tomatoes**. The hint says Ken wants tomatoes, and I can see tomatoes in Riley's lunch. Ken has almonds in his lunch. Trading almonds for tomatoes would give Ken what he wants.

The correct answer is **Riley can trade her tomatoes for Ken's broccoli**. The hint states that Riley specifically wants broccoli and Ken wants tomatoes. In the image, I can identify the specific contents of the lunchboxes. Riley has tomatoes to give. Ken has broccoli in the green compartment of his lunchbox. By trading tomatoes for broccoli, both specific needs mentioned in the hint are met.

Qwen-2.5-VL-7B (FastV) Qwen-2.5-VL-7B (HalfV)

(b) **ScienceQA (Multimodal Reasoning):** FastV fails in complex reasoning tasks due to the loss of visual reference tools (the map legend) or small, low-saliency objects (the broccoli).

Figure 19: Qualitative Case Studies on SQA