

GeoPrune: A Robust and Engine-Friendly Importance Signal Beyond Attention for Vision Token Pruning in VLMs

Kang Lou Shijin Zhang Ming Chen* Junzhe Cao

Northwestern Polytechnical University

{loukang, marscjz}@mail.nwpu.edu.cn

{zhangshijin, chenmingnwpu}@nwpu.edu.cn

Abstract

Vision token pruning can accelerate vision language model inference by removing redundant visual tokens. However, we observe that ViT attention based pruning drops sharply on Qwen3-VL and Qwen3.5. Empirically, we identify the root cause: the last-block MLP of the vision encoder significantly alters the spectral structure of hidden states and induces spectral collapse. Thus, pre-MLP attention rankings no longer reflect downstream token importance. Through further analysis of vision token selection when ViT attention based pruning still succeeds, we find that selected tokens strongly anti-correlate with raw L2 norm in the vision encoder output space S_{out} , while favoring tokens with larger residual energy. Motivated by **shared components mask token-specific importance**, we propose **GeoPrune**, a new *importance-based* pruning method that operates in S_{out} , applies deflation to remove dominant components, and ranks tokens by residual L2 norm. At aggressive pruning rates, GeoPrune is more robust and surpasses ViT attention based pruning by +4.4/+6.3 recovery points on Qwen3-VL and Qwen3.5. We further integrate GeoPrune into the production-grade inference engine vLLM, delivering 5.4%–10.5% mean E2E latency reductions with only slight accuracy drops, advancing vision token pruning toward practical deployment.

1 Introduction

Vision-language models (VLMs) (Liu et al., 2024a; Bai et al., 2025b,a) typically follow a three-stage pipeline: (1) a ViT-based (Dosovitskiy et al., 2020) vision encoder that converts visual inputs into vision tokens; (2) a multimodal merger that aligns different modality spaces; and (3) a large language model (LLM) that fuses text and vision tokens for text generation. However, this pipeline often produces hundreds of vision tokens per image, many

*Corresponding author.

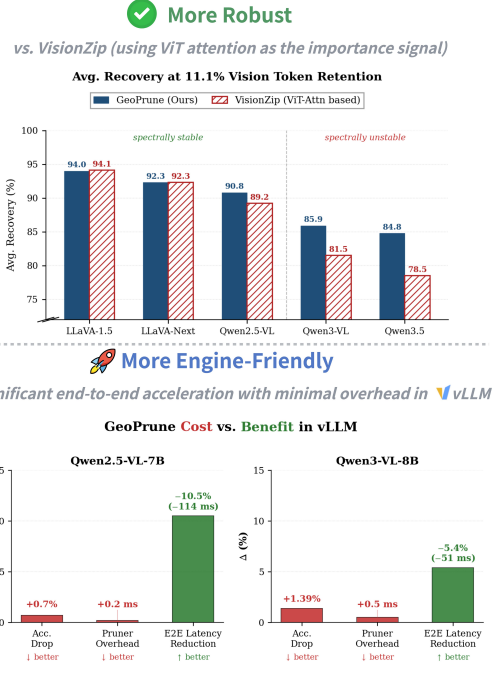


Figure 1: Overview of GeoPrune’s core advantages: 1) More robust; 2) More inference-engine friendly.

of which are redundant. Vision token pruning has thus emerged as a simple and effective training-free acceleration technique (Yang et al., 2025; Zhang et al., 2025b; Alvar et al., 2025; Dong et al., 2025; Chen et al., 2024), substantially reducing inference cost and letting a fixed context window accommodate more images and text.

Vision token pruning mainly divides into two families: *importance-based* methods and *diversity-based* methods. Within the *importance-based* family, two signals dominate in practice:

(1) Attention. Attention-based methods use attention as a proxy for token importance: *vision tokens attended by more tokens are more important*. Some use attention inside the vision encoder (Yang et al., 2025; Zhang et al., 2025b; Deng et al., 2026; Liu et al., 2026), while others use attention inside the LLM (Chen et al., 2024; Zhang et al., 2024; Yu et al., 2026). However, LLM attention based

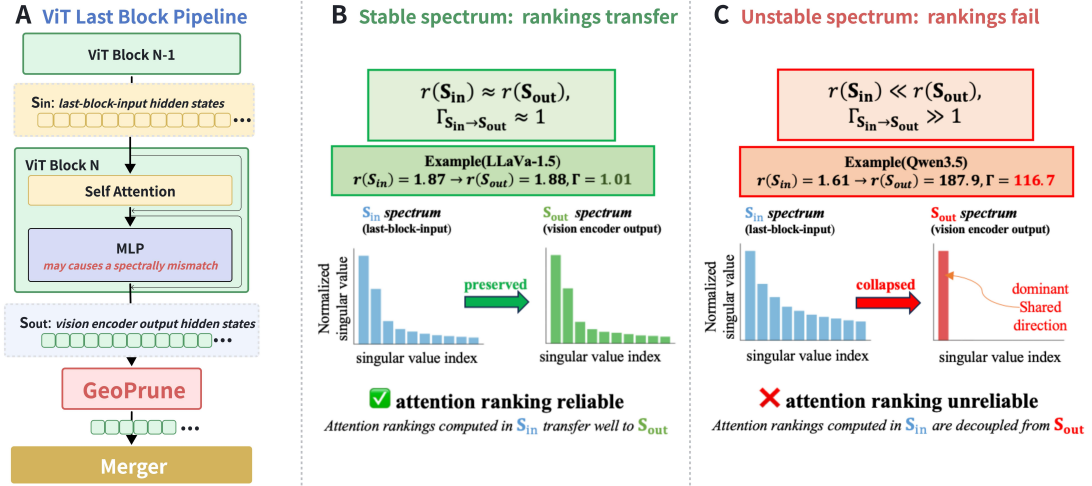


Figure 2: (A) **ViT last-block pipeline**: attention is computed before the last-block MLP and may be obstructed by it, whereas GeoPrune computes importance directly in S_{out} . (B) Attention rankings transfer when $\Gamma_{S_{in} \rightarrow S_{out}} \approx 1$, but (C) attention rankings fail when $\Gamma_{S_{in} \rightarrow S_{out}} \gg 1$, as the last-block MLP collapses S_{out} onto a shared direction.

scores have been shown to suffer from positional bias (Endo et al., 2025; Wen et al., 2025). For ViT attention based pruning, prior work shows that attention redundancy can hurt accuracy (Fang et al., 2026); we further observe severe degradation on recent models such as Qwen3-VL and Qwen3.5, which cannot be explained by redundancy alone.

(2) **Text**. A second line of work scores each vision token by its relevance to the text tokens (Zhang et al., 2024; Yu et al., 2026; Zhang et al., 2026; Dong et al., 2025): *vision tokens more relevant to the prompt are more important*. However, prior work (Yang et al., 2025) points out that fitting token scores to a single prompt is not well suited to multi-turn dialogues over the same image, and computing text–image relevance depends heavily on the quality of the embedding model.

At the same time, existing importance signals are not friendly to efficient inference engines including vLLM and SGLang (Kwon et al., 2023; Zheng et al., 2024): attention-based methods need to read attention maps and thus conflict with widely used high-performance kernels such as FlashAttention (Dao et al., 2022), while text-based methods add an extra text–image relevance computation pathway. There is **a missing piece: a robust and inference engine friendly importance signal**. Meanwhile, recent pruning methods **increasingly require importance and diversity to be jointly combined** (Baek et al., 2026; Deng et al., 2026).

We naturally discover such an importance signal by analyzing when ViT attention based pruning succeeds and fails. When attention fails, the last-

block MLP of the vision encoder causes a sharp spectral mismatch, invalidating attention rankings computed before the MLP. When attention succeeds, its selected tokens are negatively correlated with raw ℓ_2 -norm while aligning with high residual energy in S_{out} . Inspired by these findings, we propose GeoPrune, which removes dominant components by deflation and ranks tokens by residual ℓ_2 -norm in S_{out} .

Across diverse models and benchmarks, GeoPrune shows more robust performance. At aggressive pruning rates that better reflect token selection ability, GeoPrune significantly outperforms ViT attention based pruning by +4.4/+6.3 recovery points on Qwen3-VL and Qwen3.5. Notably, to our knowledge, prior vision token pruning methods have not demonstrated integration with production-grade LLM inference engines, making it difficult to quantify their actual benefits in real-world serving environments. GeoPrune **bridges this gap** by integrating into vLLM, thereby bringing vision token pruning one step closer to practical serving. In summary, our contributions are:

- A systematic analysis of ViT attention based pruning, showing token selection can be sharply interfered with by the common component.
- GeoPrune, an attention-free and text-agnostic *importance-based* method. At 11.1% retention, it outperforms ViT attention based pruning by +4.4/+6.3 points on Qwen3-VL/Qwen3.5;
- An integration of GeoPrune into vLLM, adding only **0.2–0.5 ms** overhead while notably reducing single-request mean E2E latency by **5.4–**

10.5%, with larger gains at higher concurrency.

2 Related Work

2.1 Vision Token Pruning

ViT-attention-based pruning methods. These methods use ViT-internal attention as the saliency signal: VisionZip (Yang et al., 2025) uses CLS as query or key attention, SCOPE (Deng et al., 2026) combines attention saliency with iterative coverage gain, and HiPrune (Liu et al., 2026) exploits hierarchical attention across layers of the vision encoder. These methods require attention maps and therefore cannot use efficient attention kernels such as FlashAttention (Dao et al., 2022). They are also not robust: in our experiments, the spectral mismatch can make ViT-attention-based pruning fail.

Diversity-based pruning methods. These methods operate on the hidden states after visual encoding and before LLM decoding. DivPrune (Alvar et al., 2025) formulates pruning as a max-min diversity problem, while PruneSID (Fang et al., 2026) groups tokens by PCA components and applies intra-group NMS. These methods are engine-friendly, but diversity alone (Baek et al., 2026) can under-sample important regions due to its global coverage objective, and therefore needs to be combined with *importance-based* pruning.

LLM-side or text-image pruning methods. Two other families prune vision tokens inside the LLM or instead leverage text-image relevance for selection. FastV (Chen et al., 2024) prunes vision tokens inside the LLM, SparseVLM (Zhang et al., 2024) uses LLM-side and text-guided scores, and CD-Pruner (Zhang et al., 2026) models pruning as a determinantal point process. MMTok (Dong et al., 2025) requires selected tokens to cover both the global image content and text-relevant semantics. These methods are difficult to combine with optimized inference-engine techniques, substantially reducing the inference gains brought by pruning.

2.2 Shared Components and Spectral Deflation

In NLP, anisotropic embeddings let shared components skew similarity, yet removing dominant components greatly improves similarity computation (Ethayarajh, 2019; Gao et al., 2019; Arora et al., 2017; Mu et al., 2017; Su et al., 2021). We observe a similar phenomenon when pruning vision tokens in VLMs: dominant principal components can interfere with pruning token selection. How-

ever, we find that anisotropy is not the underlying cause of the failure of ViT attention based pruning.

3 Analysis

3.1 Why ViT Attention Fails

We start with a controlled comparison between VisionZip (Yang et al., 2025), a ViT attention based pruning baseline, and DivPrune (Alvar et al., 2025), a representative *diversity-based* method. As shown in Table 1, VisionZip tracks DivPrune closely on LLaVA-1.5, LLaVA-Next, and Qwen2.5-VL, but drops sharply on Qwen3-VL and Qwen3.5.

Model	VisionZip	DivPrune
LLaVA-1.5-7B	94.1	94.5
LLaVA-Next-7B	92.3	93.8
Qwen2.5-VL-7B	89.2	87.9
Qwen3-VL-8B	81.5 -4.7	86.2
Qwen3.5-8B	78.5 -5.5	84.0

Table 1: **Average recovery (%) at 11.1% retention on five VLMs.** VisionZip and DivPrune are close on the first three models; on Qwen3-VL and Qwen3.5, VisionZip drops sharply.

To understand this recovery drop, we examine the last vision-encoder block where VisionZip computes its attention scores. As Figure 2(A) shows, attention is computed before the last-block MLP in the input space $\mathbf{S}_{\text{in}}$, while the LLM consumes the output after it in the vision-encoder output space $\mathbf{S}_{\text{out}}$. Following the singular-value scree-gap criterion (Cattell, 1966; Von Luxburg, 2007), we define the spectral dominance ratio:

Definition 1 (Spectral Dominance Ratio). *Given a token feature matrix $\mathbf{Z}$ with singular values $\sigma_1(\mathbf{Z}) \geq \sigma_2(\mathbf{Z}) \geq \dots$, the spectral dominance ratio is defined as*

$$r(\mathbf{Z}) = \frac{\sigma_1(\mathbf{Z})}{\sigma_2(\mathbf{Z})}. \quad (1)$$

It measures the top-component dominance within the representation space.

Given two spaces, we further define the spectral-mismatch ratio based on their spectral dominance ratios.

Definition 2 (Spectral-Mismatch Ratio). *Given two token feature matrices $\mathbf{A}$ and $\mathbf{B}$, the spectral-mismatch ratio from $\mathbf{A}$ to $\mathbf{B}$ is defined as*

$$\Gamma_{\mathbf{A} \rightarrow \mathbf{B}} = \frac{r(\mathbf{B})}{r(\mathbf{A})}. \quad (2)$$

Stage	Qwen3-VL	Qwen3.5
S_{in} (block input)	1.00	1.00
After self-attention + residual	0.77	0.92
After MLP + residual ($= S_{out}$)	76.8	116.7

Table 2: **Sub-layer decomposition of the last block.** Values are $\Gamma_{S_{in} \rightarrow S_{out}}$ at each stage.

Model	$r(S_{in})$	$r(S_{out})$	$\Gamma_{S_{in} \rightarrow S_{out}}$	Attn
LLaVA-1.5	1.87 ± 0.32	1.88 ± 0.31	1.00 ± 0.01	✓
LLaVA-Next	1.16 ± 0.13	1.16 ± 0.13	1.00 ± 0.01	✓
Qwen2.5-VL	13.3 ± 1.8	20.3 ± 5.0	1.53 ± 0.57	✓
Qwen3-VL	2.03 ± 0.27	155.9 ± 35.2	76.8 ± 16.1	✗
Qwen3.5	1.61 ± 0.21	187.9 ± 47.7	116.7 ± 29.7	✗

Table 3: **Spectral stability across five VLMs.** Values are mean $\pm$ std over 200 images; **Attn** indicates whether ViT attention based pruning remains reliable.

It measures how much top-component dominance changes across spaces. Thus, $\Gamma \approx 1$ indicates spectral stability, whereas $\Gamma \gg 1$ indicates a sharp spectral mismatch.

The mismatch is localized to the last-block MLP.

Table 2 decomposes the last block of the two failing models. The self-attention sub-layer induces a mild shift ($\Gamma_{S_{in} \rightarrow S_{out}} = 0.77\text{--}0.92$), whereas the MLP drives $\Gamma_{S_{in} \rightarrow S_{out}}$ up to 76.8 (Qwen3-VL) and 116.7 (Qwen3.5). Thus, the last-block MLP accounts for nearly all the spectral mismatch.

Spectral mismatch predicts pruning reliability.

As illustrated in Figure 2(B), when $\Gamma_{S_{in} \rightarrow S_{out}} \approx 1$, attention rankings computed in S_{in} transfer well to S_{out} , and ViT-attention pruning remains reliable. In contrast, when $\Gamma_{S_{in} \rightarrow S_{out}} \gg 1$ (Qwen3-VL 76.8, Qwen3.5 116.7; Figure 2(C)), attention rankings become decoupled from S_{out} , and pruning fails.

Mismatch, not raw dominance, causes failure.

Although S_{out} has strong top-component dominance on Qwen2.5-VL ($r(S_{out})=20.3$), S_{in} is similarly dominated ($r(S_{in})=13.3$), so the cross-space mismatch still remains small ($\Gamma_{S_{in} \rightarrow S_{out}}=1.53$). Accordingly, ViT-attention pruning remains reliable. This shows that raw dominance in S_{out} alone does not break ViT-attention pruning (Table 3); failure is caused by a sharp $S_{in} \rightarrow S_{out}$ spectral mismatch.

The spectral patterns are consistent across images: r and Γ values measured on MMBench and OCRBench (a text-heavy benchmark) are nearly identical (Figure 3), confirming that spectral stability is architectural rather than image-specific.

3.2 When ViT Attention Works

The stable models offer the complementary clue: when ViT attention remains empirically reliable, its selections reveal what kind of token signal survives in S_{out} . Given the shared-component interference observed above, we compare ViT-attention selections with three S_{out} -side selectors: the raw ℓ_2 -norm, the residual ℓ_2 -norm after one deflation step (formalized in Sec. 4.1), and the cosine deviation from the mean token (i.e., $1 - \cos(\mathbf{y}_i, \bar{\mathbf{y}})$, where $\bar{\mathbf{y}} = \frac{1}{N} \sum_j \mathbf{y}_j$). We formalize this agreement with coverage.

Definition 3 (Coverage with ViT-attention Selection). *For each retain ratio k/N , let A_k be the ViT-attention top- k set and B_k be the top- k set induced by a given S_{out} -side selector. The coverage is defined as*

$$\text{Coverage}_k = \frac{|A_k \cap B_k|}{k}. \quad (3)$$

A higher Coverage_k indicates stronger agreement with ViT-attention selection.

As shown in Figure 4, residual ℓ_2 -norm consistently shows high overlap with ViT-attention selections across all three stable models, while raw ℓ_2 -norm is negatively correlated. Cosine deviation overlaps strongly on LLaVA-1.5 and LLaVA-Next, but drops sharply on Qwen2.5-VL.

This drop is explained by the geometry of S_{out} : on Qwen2.5-VL and Qwen3-VL, pairwise cosine similarities in the raw space are uniformly close to 1 (Figure 5), indicating token representations are pulled toward a shared direction. Here, cosine deviation loses discriminative power as an importance score. In contrast, deflation suppresses this

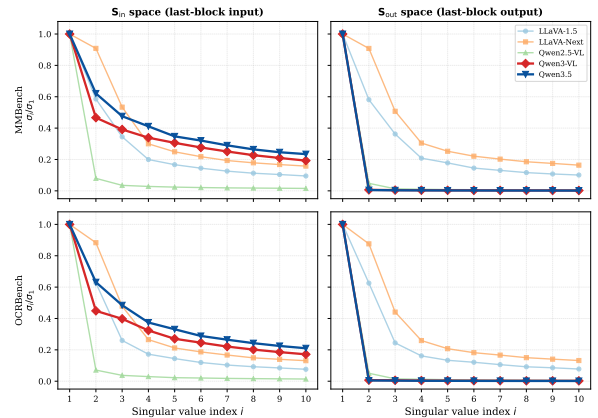


Figure 3: **Spectral structure is stable across image sets.** MMBench (top) and OCRBench (bottom); S_{in} space (left) and S_{out} space (right).

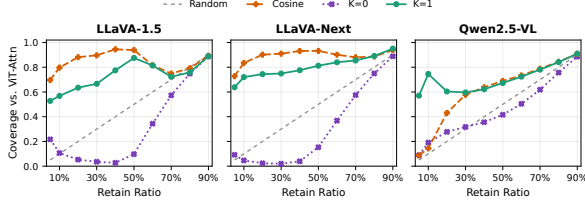


Figure 4: **Coverage with ViT-attention selections in S_{out} when ViT attention works.** Res ℓ_2 -norm strongly overlaps with ViT attention across models, whereas cosine deviation drops on Qwen2.5-VL. Raw ℓ_2 -norm is negatively correlated. Dashed lines(random).

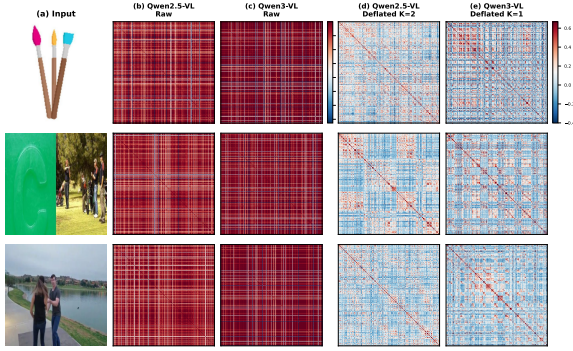


Figure 5: **Pairwise cosine similarity in S_{out} .** Two models show uniformly high cosine similarity in the raw space; deflation restores token separability.

shared component and restores token separability. Combined with the results in Table 1, it is further confirmed that **anisotropy is not the root cause of the failure of ViT attention based pruning.**

The failure mode of raw norm is also visible qualitatively: in Figure 6, raw ℓ_2 -norm selects many background tokens, while residual ℓ_2 -norm concen-

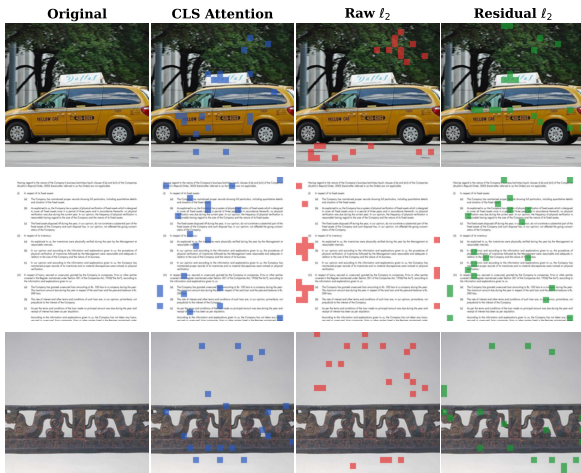


Figure 6: **Top-20 tokens selection on S_{out} (LLaVA-1.5).** Blue: VisionZip; Red: raw ℓ_2 -norm; Green: residual ℓ_2 -norm.

trates on semantically salient foreground tokens.

Summary. The analysis above yields three predictions: (1) Importance computed in S_{out} avoids the $S_{in} \rightarrow S_{out}$ ranking-transfer problem caused by spectral mismatch (§3.1); (2) Deflation recovers discriminative token-specific energy that raw ℓ_2 -norm loses (§3.2); (3) Because both signals live in S_{out} , importance and diversity compose naturally without extra inference pathways.

4 Method: GeoPrune

Based on the analysis above, we propose GeoPrune, a new importance-based pruning method with two design principles: (1) compute importance directly in S_{out} to avoid transferring rankings from S_{in} and efficiently compatible with inference engines, and (2) use residual energy as the importance signal, which successful ViT attention based pruning implicitly tracks. GeoPrune removes dominant spectral components that mask token-specific energy, then scores each token by the residual energy it retains.

4.1 Deflated Residual ℓ_2 -Norm as Importance

Given the raw vision-encoder output hidden states $S_{out} = [y_1, \dots, y_N]^T \in \mathbb{R}^{N \times d}$, GeoPrune computes a per-token importance score as follows.

Step 1: Capped adaptive deflation. Starting from $M^{(1)} = S_{out}$, each power-iteration call estimates the leading singular triplet of the current residual matrix:

$$(\sigma_k, u_k, v_k) = \text{POWERITER}(M^{(k)}, T), \quad (4)$$

where T is the number of power-iteration steps. We then remove this component,

$$M^{(k+1)} = M^{(k)} - \sigma_k u_k v_k^T, \quad M^{(1)} = S_{out}. \quad (5)$$

In practice, echoing the spectral-gap diagnostic in §3.1, we treat the eigengap as a lightweight diagnostic that provides a conservative lower bound on the number of dominant components:

$$\Delta_k = \frac{\sigma_k}{\sigma_{k+1}}. \quad (6)$$

We use this detected k as a starting point and explore up to a small maximum depth K_{max} , keeping deflation shallow while adapting to the observed spectrum.

Step 2: Residual ℓ_2 -norm as importance. After obtaining the top- K right singular directions, the

Model	Method	GQA	MMB	MME	POPE	SQA	TextVQA	Avg.%
Qwen2.5-VL-7B (stable)	Full model (100%)	60.90	83.51	2318	87.80	88.90	82.69	
	VisionZip (CVPR'25)	52.66	76.89	2063	82.18	83.79	66.09	89.2
	DivPrune (CVPR'25)	56.05	74.23	1982	83.39	80.42	62.64	87.9
	MMTok (ICLR'26)	54.77	75.95	2049	81.30	82.05	65.22	88.8
	Raw ℓ_2	51.60	74.48	1970	79.31	80.37	59.66	85.3
	GeoPrune (imp)	53.62	74.31	2176	82.44	82.50	71.88	90.8 +1.6
	GeoPrune (imp+div)	56.81	78.61	2231	85.31	84.48	72.60	93.9 +4.7
Qwen3-VL-8B (unstable)	Full model (100%)	61.63	84.97	2359	88.77	94.35	80.19	
	VisionZip (CVPR'25)	51.24	66.58	1573	81.79	82.20	65.20	81.5
	DivPrune (CVPR'25)	53.67	70.62	1871	84.71	83.69	67.10	86.2
	Raw ℓ_2	49.53	68.56	1627	76.89	80.47	58.01	79.0
	GeoPrune (imp)	52.87	71.22	1868	83.90	82.50	67.74	85.9 +4.4
	GeoPrune (imp+div)	55.30	71.31	1952	86.02	82.94	70.41	88.2 +2.0
Qwen3.5-8B (unstable)	Full model (100%)	60.83	84.71	2369	88.23	93.31	72.72	
	VisionZip (CVPR'25)	48.74	69.67	1609	75.10	82.70	48.55	78.5
	DivPrune (CVPR'25)	50.98	72.77	1902	81.38	83.94	52.27	84.0
	Raw ℓ_2	48.82	69.93	1694	73.83	81.21	50.74	79.1
	GeoPrune (imp)	50.85	73.20	1968	80.86	80.66	56.33	84.8 +6.3
	GeoPrune (imp+div)	53.35	73.80	2008	83.29	83.09	58.19	87.2 +3.2

Table 4: Main benchmark results at 11.1% token retention on Qwen-family models.

residual of token i and its importance score are

$$\mathbf{r}_i = \mathbf{y}_i - \sum_{j=1}^K (\mathbf{y}_i^\top \mathbf{v}_j) \mathbf{v}_j, \quad q_i = \|\mathbf{r}_i\|_2. \quad (7)$$

q_i measures the token-specific energy left after removing shared spectral modes; it essentially quantifies the residual spectral mass unique to each token. Because it is computed directly in $\mathbf{S}_{\text{out}}$, it avoids the $\mathbf{S}_{\text{in}} \rightarrow \mathbf{S}_{\text{out}}$ ranking-transfer problem.

4.2 Composing with Diversity in the Same Space

Because q_i lives in $\mathbf{S}_{\text{out}}$ —the same space used by diversity-based pruners (Alvar et al., 2025; Fang et al., 2026; Zhang et al., 2026)—importance and diversity can thus be naturally composed with no extra inference pathway. Specifically, we directly plug q_i into a greedy diversity selector (Zhang et al., 2026) and omit the text-conditioned term to keep the pipeline prompt-agnostic and engine-friendly:

$$s_i(S) = \text{div}_i(S) q_i^\beta, \quad (8)$$

where $\text{div}_i(S)$ is the marginal diversity gain of adding token i to the current subset S , and $\beta \geq 0$ trades off importance against diversity. Setting $\beta = 0$ recovers a pure diversity-based selection. The end-to-end procedure is summarized in App. A.

5 Experiments

5.1 Setup

We evaluate five representative VLMs: LLaVA-1.5-7B, LLaVA-Next-7B, Qwen2.5-VL-7B, Qwen3-VL-8B, and Qwen3.5-8B across a broad set of

multimodal benchmarks using lmms-eval (Zhang et al., 2025a). Benchmarks include GQA (Hudson and Manning, 2019), MMBench (Liu et al., 2024b), MME (Fu et al., 2026), POPE (Li et al., 2023), SQA (Lu et al., 2022), TextVQA (Singh et al., 2019), MMMU (Yue et al., 2024), and OCR-Bench (Liu et al., 2024c).

Following prior work (Yang et al., 2025; Alvar et al., 2025), the main accuracy setting retains 11.1% vision tokens, which provides an aggressive regime for evaluating token-selection quality. We report *recovery* as an aggregate metric: for each benchmark metric, we compute the ratio of the pruned-model score to the corresponding full-model score, and then average these ratios over all reported metrics. A recovery of 100% therefore corresponds to the performance of the full, unpruned model.

For GeoPrune, we use $T=10$ power iteration steps and a capped spectral-gap stopping heuristic with threshold $\tau=2$ and $K_{\text{max}}=3$. Here, we evaluate two complementary variants of GeoPrune. GeoPrune (imp) is a system-oriented variant that uses the residual ℓ_2 -norm as an engine-friendly importance score (§4.1), aiming to provide robust pruning performance with low overhead. GeoPrune (imp+div) is an accuracy-oriented variant that further combines the importance score with the diversity term in §4.2; we set $\beta=1.5$ for this variant. These two variants reflect different accuracy–efficiency trade-offs: GeoPrune (imp) emphasizes practical system efficiency, whereas GeoPrune (imp+div) leverages complementary importance and diversity signals to improve accuracy.

Model	Method	GQA	MMB	MME	POPE	SQA	TextVQA	MMMU	Avg.%
LLaVA-1.5-7B (stable)	Full model (100%)	61.90	64.70	1862	85.90	69.50	58.20	36.30	
	VisionZip (CVPR'25)	55.15	60.31	1657	80.56	69.01	55.42	36.00	94.1
	DivPrune (CVPR'25)	57.46	58.16	1658	86.57	67.97	55.09	35.11	94.5
	MMTok (ICLR'26)	58.13	61.17	1688	85.77	69.16	55.69	35.11	95.8
	Raw ℓ_2	38.33	19.93	896	49.96	64.25	41.84	31.67	64.4
	GeoPrune (imp)	55.86	59.02	1662	83.13	69.21	53.46	35.89	94.0 -0.1
	GeoPrune (imp+div)	58.12	60.48	1677	86.32	68.17	55.96	35.89	95.9 $+0.1$
LLaVA-Next-7B (stable)	Full model (100%)	64.20	67.90	1842	86.40	70.20	61.30	35.10	
	VisionZip (CVPR'25)	55.42	59.79	1648	79.61	68.42	56.06	35.56	92.3
	DivPrune (CVPR'25)	58.88	61.68	1625	83.33	66.63	54.71	37.00	93.8
	MMTok (ICLR'26)	60.05	62.97	1716	83.87	67.97	54.33	37.11	95.4
	GeoPrune (imp)	56.82	61.43	1578	82.20	67.67	55.21	35.11	92.3 $+0.0$
	GeoPrune (imp+div)	59.47	63.42	1662	85.12	67.13	56.68	36.56	95.3 -0.1

Table 5: **Main benchmark results at 11.1% token retention on LLaVA-family models.**

5.2 Main Results

Tables 4 and 5 report the full benchmark results at 11.1% token retention, grouped by model family. We select three representative baselines covering the major pruning paradigms. For *importance-based* pruning, prior methods use *attention* or *text* to score token importance: VisionZip (Yang et al., 2025) uses ViT attention, while MMTok (Dong et al., 2025) uses text-image relevance and further encourages global visual coverage. DivPrune (Alvar et al., 2025) serves as a representative *diversity-based* pruning method.

The results show that the residual ℓ_2 -norm in the vision-encoder output space $\mathbf{S}_{\text{out}}$ is a more robust importance signal than ViT attention. Moreover, combining this importance signal with diversity in the same $\mathbf{S}_{\text{out}}$ space performs comparably to MMTok on LLaVA models and surpasses it by +5.1 recovery points on Qwen2.5-VL, without adding any extra text-vision relevance computation pathway. Our analysis in §3 makes three predictions about the properties of a robust pruning signal. We verify each below.

(1) Residual-norm is a more robust importance signal than ViT attention. The head-to-head comparison with ViT attention based VisionZip is isolated in Table 6. GeoPrune (imp) matches attention on spectrally stable models and grows increasingly stronger when $\Gamma \gg 1$, improving over VisionZip by +4.4/+6.3 recovery points on Qwen3-VL/Qwen3.5.

(2) Deflation is what unlocks importance in $\mathbf{S}_{\text{out}}$. Table 7 summarizes the average recovery of raw ℓ_2 -norm and deflated residual ℓ_2 -norm at 11.1% retention. Raw ℓ_2 -norm collapses on every model, with text-dense benchmarks suffering the most, whereas deflation removes dominant spectral components and recovers performance across all benchmarks.

Model	VisionZip	GeoPrune (imp)	Δ
LLaVA-1.5	94.1	94.0	−0.1
LLaVA-Next	92.3	92.3	+0.0
Qwen2.5-VL	89.2	90.8	+1.6
Qwen3-VL	81.5	85.9	+4.4
Qwen3.5	78.5	84.8	+6.3

Table 6: **Average recovery (%) at 11.1% retention: residual-norm importance vs. ViT-attention importance.**

Model	Raw ℓ_2	Residual ℓ_2	Δ
LLaVA-1.5-7B	64.4%	94.0%	+29.6
Qwen2.5-VL-7B	85.3%	90.8%	+5.5
Qwen3-VL-8B	79.0%	85.9%	+6.9
Qwen3.5-8B	79.1%	84.8%	+5.7

Table 7: **Average recovery at 11.1% retention: raw vs. deflated residual ℓ_2 -norm in $\mathbf{S}_{\text{out}}$.**

(3) Importance and diversity compose naturally in $\mathbf{S}_{\text{out}}$. Because both signals already live in $\mathbf{S}_{\text{out}}$, the residual ℓ_2 -norm importance q_i can be plugged into a diversity selector at essentially zero extra inference cost, attention-free and text-agnostic (§4.2). The per-benchmark ablation (Table 8) shows that combining residual-norm importance with diversity selection consistently improves over the stronger single component by +3.1/+1.7/+2.4 recovery points on Qwen2.5-VL/Qwen3-VL/Qwen3.5, indicating that the two signals are genuinely complementary rather than redundant.

5.3 Real-World Inference-Engine Speedup with vLLM

Beyond accuracy, we integrate GeoPrune into the production inference engine vLLM (Kwon et al., 2023) on Qwen2.5-VL and Qwen3-VL, patching the model classes with three minimal modifications: (i) the visual encoder also returns the post-ViT pre-

Model	Variant	GQA	MMB	MME	POPE	SQA	TextVQA	Avg.%
Qwen2.5-VL	Div	55.89	75.00	2056	83.43	79.47	64.35	88.8
	GeoPrune (imp)	53.62	74.31	2176	82.44	82.50	71.88	90.8
	GeoPrune (imp+div)	56.81	78.61	2231	85.31	84.48	72.60	93.9+3.1
Qwen3-VL	Div	53.97	71.31	1882	84.38	83.34	67.54	86.5
	GeoPrune (imp)	52.87	71.22	1868	83.90	82.50	67.74	85.9
	GeoPrune (imp+div)	55.30	71.31	1952	86.02	82.94	70.41	88.2+1.7
Qwen3.5	Div	51.12	72.94	1898	81.76	84.18	53.23	84.4
	GeoPrune (imp)	50.85	73.20	1968	80.86	80.66	56.33	84.8
	GeoPrune (imp+div)	53.35	73.80	2008	83.29	83.09	58.19	87.2+2.4

Table 8: **Importance–diversity composition at 11.1% retention (per-benchmark breakdown).** Div: diversity-only selection ($\beta=0$). **GeoPrune (imp)**: residual ℓ_2 -norm top- k selection. **GeoPrune (imp+div)**: combining both signals. Red annotations show the gain over the stronger single component.

merger hidden states; (ii) GeoPrune scores and prunes the visual embeddings via the residual-norm signal; (iii) the image placeholders in the rendered prompt are shrunk so that the LLM KV cache is allocated against the pruned sequence length.

Accuracy at serving retention rates. We select two mild retention rates suitable for practical serving: 50% for Qwen2.5-VL and 70% for Qwen3-VL. At these rates, GeoPrune (imp) preserves **99.3%** of full-model recovery on Qwen2.5-VL (Table 9) and **98.61%** on Qwen3-VL (Table 10).

Method	POPE	MME	OCR	TextVQA	SQA	MMB	Avg.%
Full model	87.80	2318	84.1	82.69	88.90	83.51	100.0
GeoPrune (imp)	87.79	2321	81.8	82.41	88.02	83.51	99.3

Table 9: **Qwen2.5-VL-7B accuracy at 50% retention with GeoPrune (imp).**

Method	GQA	MMB	MME	POPE	SQA	Avg.%
Full model	61.63	84.97	2359	88.77	94.35	100.0
GeoPrune (imp)	60.91	83.32	2318	88.31	92.84	98.61

Table 10: **Qwen3-VL-8B accuracy at 70% retention with GeoPrune (imp).**

Single-request E2E latency. First, following vllm bench mm-processor, we measure the vision encoder forward time and single-request end-to-end latency (preprocessing + prefill + 32-token decode) on 448×448 resolution images using a single H800 (80 GB), bfloat16, 300 text tokens and 50 requests. At these retention rates, the matrix-deflation scoring step is essentially free: the ViT forward time grows by only **0.2–0.5 ms**. Nevertheless, the mean end-to-end latency drops by **10.5%** on Qwen2.5-VL and **5.4%** on Qwen3-VL, because reducing the vision-token count shortens every downstream stage: merger, prompt rendering, KV-cache allocation, prefill, and per-step decode attention.

	Vanilla	Pruned
<i>Qwen2.5-VL (50% retain)</i>		
ViT fwd (ms)	7.7	7.9 (+0.2 ms)
E2E mean (ms)	1087	973 ↓10.5%
E2E p99 (ms)	1732	1489 ↓14.0%
<i>Qwen3-VL (70% retain)</i>		
ViT fwd (ms)	4.1	4.6 (+0.5 ms)
E2E mean (ms)	938	887 ↓5.4%
E2E p99 (ms)	1553	1473 ↓5.1%

Table 11: **Single-request E2E latency on vLLM.**

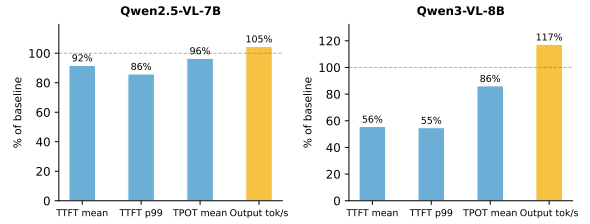


Figure 7: **vLLM serving metrics at $c=16$ (% of no-pruning baseline).**

Batched serving throughput. We further evaluate the serving-level benefit of vision token pruning at high concurrency ($c=16$) on vLLM (Figure 7). With slight accuracy loss, the reduced vision-token count shortens prefill and shrinks the KV cache, improving both TTFT and per-step decode efficiency; the net effect is higher output throughput. Specifically, on Qwen2.5-VL (50% retention), TTFT decreases to 92%/86% (mean/p99) with throughput increasing to 105%; on Qwen3-VL (70% retention), TTFT decreases to 56%/55% with throughput increasing to 117%, further amplified by relief of baseline prefill queueing. Lower-concurrency results ($c=1, 4, 8$) are deferred to App. J.

6 Conclusion

Through a spectral-geometric analysis of ViT attention based pruning, we demonstrate that the dom-

inant spectral component of S_{out} interferes with token selection: when the last-block MLP of the vision encoder injects a large new dominant direction, it causes a spectral mismatch; when the spectrum is preserved, ViT attention implicitly prefers off-common-direction tokens. Consequently, we propose GeoPrune, an attention-free, text-agnostic importance signal that lives directly in S_{out} : deflation removes the shared spectral modes and the residual ℓ_2 -norm serves as a plug-and-play score, naturally composing with diversity selection in the same space. For accuracy, GeoPrune exhibits more robust performance than ViT attention based pruning, especially on spectrally unstable models, while remaining competitive on stable models. For efficiency, its integration into vLLM trades minimal overhead and slight accuracy loss for substantial end-to-end latency reduction, bringing vision token pruning closer to practical deployment.

Limitations

Our deflation depth uses the eigengap as a conservative lower bound capped by K_{max} ; a more principled rule for choosing K remains future work. We observe experimentally that the optimization objectives of certain diversity-based methods are not compatible with our importance score; importance signals that are compatible with that class of diversity algorithms still need to be explored. Our vLLM serving results indicate that a qualitative jump in throughput can occur between adjacent pruning rates, suggesting that algorithms capable of maintaining higher accuracy at slightly more aggressive rates warrant further investigation. From a training perspective, why the last-block ViT MLP develops dominant outlier channels that directly collapse the spectrum of the final layer remains without a training-level explanation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 52335001), the China Postdoctoral Science Foundation (Grant Nos. GZC20252728 and 2026M794817), and the Jiangsu Province Outstanding Postdoctoral Fellowship Program (Grant No. 2025ZB735).

Ethical considerations

To the best of our knowledge, we do not foresee any ethical concerns associated with this work. This

work focuses on accelerating inference for vision-language models through training-free token pruning, and uses only publicly released model checkpoints and standard academic benchmarks. Additionally, no new data is collected and no human subjects are involved. Finally, we used AI assistants for writing polishing and grammar checking.

References

- Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401.
- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.
- Changwoo Baek, Jouwon Song, Sohyeon Kim, and Kyeongbo Kong. 2026. Agilepruner: An empirical study of attention and diversity for adaptive visual token pruning in large vision-language models. *arXiv preprint arXiv:2603.01236*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Raymond B Cattell. 1966. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.
- Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.
- Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. 2026. Scope: Saliency-coverage oriented token pruning for efficient multimodal llms. *Advances in*

- Neural Information Processing Systems*, 38:161527–161552.
- Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. 2025. Mmtok: Multimodal coverage maximization for efficient inference of vlms. *arXiv preprint arXiv:2508.18264*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2025. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22826–22835.
- Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.
- Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. 2026. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity. *arXiv preprint arXiv:2603.09480*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.
- Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Jizhihui Liu, Guangdao Zhu, and Feiyi Du. 2026. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 41275–41277.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocr-bench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.
- Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.
- Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.
- Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Stop looking for “important tokens” in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9972–9991.

Zhenkai Wu, Xiaowen Ma, Zhenliang Ni, Dengming Zhang, Han Shu, Xin Jiang, and Xinghao Chen. 2026. Vlm-pruner: Buffering for spatial sparsity in an efficient vlm centrifugal token pruning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31952–31961.

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

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. 2026. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

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

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. 2026. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.

A Full Pruning Algorithm

Algorithm 1 summarizes the full GeoPrune procedure, combining power-iteration deflation with the eigengap-based diagnostic (§4.1) and greedy importance-weighted diversity selection.

Algorithm 1 GeoPrune

Require: $\mathbf{S}_{\text{out}} \in \mathbb{R}^{N \times d}$; target count k ; gap threshold τ ; exponent β ; power-iteration steps T ; max deflation depth K_{max}

Ensure: Selected token indices S , $|S| = k$

```

1:  $\mathbf{M} \leftarrow \mathbf{S}_{\text{out}}$ ;  $K \leftarrow 0$ 
2: repeat
3:    $(\sigma, \mathbf{u}, \mathbf{v}) \leftarrow \text{POWERITERATION}(\mathbf{M}, T)$ 
4:    $K \leftarrow K + 1$ 
5:    $(\sigma_K, \mathbf{v}_K) \leftarrow (\sigma, \mathbf{v})$ 
6:    $\mathbf{M} \leftarrow \mathbf{M} - \sigma \mathbf{u} \mathbf{v}^\top$ 
7: until  $K \geq 1$  and  $\sigma_K / \sigma < \tau$  or  $K = K_{\text{max}}$ 
8: for  $i = 1, \dots, N$  do
9:    $\mathbf{r}_i \leftarrow \mathbf{y}_i - \sum_{j=1}^K (\mathbf{y}_i^\top \mathbf{v}_j) \mathbf{v}_j$ 
10:   $q_i \leftarrow \|\mathbf{r}_i\|_2$ 
11: end for
12:  $S \leftarrow \{\arg \max_i q_i\}$ 
13: while  $|S| < k$  do
14:    $S \leftarrow S \cup \{\arg \max_{i \notin S} \text{div}_i(S) \cdot q_i^\beta\}$ 
15: end while
16: return  $S$ 

```

B Metric Choice for Spectral Mismatch

In §3.1, we use σ_1 / σ_2 as the building block of the spectral-mismatch ratio $\Gamma_{\mathbf{S}_{\text{in}} \rightarrow \mathbf{S}_{\text{out}}} = r(\mathbf{S}_{\text{out}}) / r(\mathbf{S}_{\text{in}})$. A natural alternative is the top-1 energy fraction $\sigma_1^2 / \sum_i \sigma_i^2$. As shown in Table 12, energy-based Γ is far less discriminative than r -based Γ : even on collapsed models the energy-based ratio changes only to 1.71/1.75, whereas r -based Γ clearly separates the stable Qwen2.5-VL (1.53) from the unstable Qwen3-VL/Qwen3.5 (76.8/116.7).

Model	Γ_{sdr}	Γ_{energy}
LLaVA-1.5	1.00	0.99
LLaVA-Next	1.00	1.00
Qwen2.5-VL	1.53	1.02
Qwen3-VL	76.8	1.71
Qwen3.5	116.7	1.75

Table 12: **Cross-space spectral mismatch: Γ based on r vs. top-1 energy.** σ_1 / σ_2 provides a non-saturating dynamic range that clearly separates stable from unstable models, whereas energy-based Γ compresses the difference.

Model	$K=0$	$K=1$	$K=2$	$K=3$
LLaVA-1.5	0.34	0.04	0.03	0.006
LLaVA-Next	0.34	0.14	0.03	0.002
Qwen2.5-VL	0.95	0.19	0.08	0.013
Qwen3-VL	0.97	0.06	0.002	0.001
Qwen3.5	0.98	0.07	0.002	0.000

Table 13: **Mean pairwise cosine similarity in S_{out} vs. deflation depth K** (200 MMBench images).

C Additional Analysis

C.1 Cosine Similarity Decay with Deflation

Table 13 reports mean pairwise cosine similarity in S_{out} as the deflation depth increases. High raw anisotropy ($K=0$) alone does not explain attention failure: Qwen2.5-VL is as anisotropic as Qwen3-VL/Qwen3.5, but its Γ is small and attention remains reliable.

D Additional Ablation Studies

D.1 Diversity in the Residual Space

Table 14 reports a diagnostic variant that applies diversity-only selection in the deflated residual space and compares it with diversity in the original space. Residual-space diversity can improve over original-space diversity on Qwen2.5-VL, but degrades on Qwen3-VL/Qwen3.5. We hypothesize that deflation exposes more token-specific residual variation, which is useful when used as an importance prior, but a pure diversity selector may oversample noisy residual outliers without the residual-norm weighting used by GeoPrune.

D.2 Deflation Depth K

Table 15 validates that a small capped deflation depth is sufficient. We vary K on representative settings; $K=0$ corresponds to raw ℓ_2 -norm without deflation. Performance recovers sharply once the leading components are removed and then plateaus, supporting the use of the eigengap as a conservative lower bound with a small K_{max} cap rather than heavy tuning.

E Additional Model Results

E.1 LLaVA-1.5-7B and LLaVA-Next-7B

Tables 16 and 17 report two additional spectrally stable models, LLaVA-1.5-7B and LLaVA-Next-7B, at three retention rates each. Both models preserve last-block spectral structure (Table 3), so attention already provides a reasonable importance proxy and all methods perform comparably.

F Pairwise Cosine Similarity Heatmaps

Figure 8 visualizes the full pairwise cosine similarity matrices among vision tokens in S_{out} across three Qwen-family models. The left group (b–d) shows *raw* cosine similarity: all three models exhibit nearly uniform matrices with mean similarity close to 1, consistent with the anisotropy statistics in Table 13. This uniform high similarity causes cosine deviation from the mean token to lose discriminative power as an importance score—when all tokens lie near the same direction, their deviations from the mean are nearly identical and cannot differentiate token importance.

The right group (e–g) shows the same matrices after deflation (Qwen2.5-VL $K=2$; Qwen3-VL and Qwen3.5 $K=1$). Deflation removes the dominant shared component, drastically reducing mean cosine and revealing rich per-token structure that was previously masked. This restored token separability is precisely what enables residual ℓ_2 -norm to serve as a reliable energy-based importance score, as analyzed in §3.2.

G Extended VideoLLM Evaluation

We extend our evaluation to VideoLLMs to examine whether the observations from image-based benchmarks generalize to long-sequence video inputs. We evaluate on MVBench and VideoMME using Qwen3-VL-8B-Instruct and Qwen3.5-9B. Unless otherwise specified, all GeoPrune hyperparameters are kept identical to those used in the image experiments; only the video-specific number of input frames is adjusted.

We compare GeoPrune with VisionZip, DivPrune, and MMTok. In addition to the two main variants, we report GeoPrune (only_div), which uses only the diversity component ($K=0$), to isolate the contribution of diversity-based selection. Tables 19 and 20 present the detailed results at 5% and 50% vision-token retention.

H Comparison with Decoder-Stage Pruning

Our main experiments focus on *pre-decoder vision-token pruning*, where redundant visual tokens are removed before entering the language decoder. Accordingly, the primary baselines in the main paper, including VisionZip, DivPrune, PruneSID, and MMTok, operate in this pre-decoder setting.

For completeness, we additionally compare GeoPrune with VLM-Pruner (Wu et al., 2026),

Model	Variant	GQA	MMB	MME	POPE	SQA	TextVQA	Avg.%
Qwen2.5-VL-7B	Div (original)	55.89	75.00	2056	83.43	79.47	64.35	88.8
	Div (residual)	56.34	77.32	2128	85.04	82.85	67.95	91.5
Qwen3-VL-8B	Div (original)	53.97	71.31	1882	84.38	83.34	67.54	86.5
	Div (residual)	52.84	69.07	1817	82.93	80.12	64.19	83.7
Qwen3.5-8B	Div (original)	51.12	72.94	1898	81.76	84.18	53.23	84.4
	Div (residual)	49.88	69.33	1725	78.11	81.31	51.36	80.5

Table 14: **Original-space vs. residual-space diversity** at 11.1% retention. Pure diversity in the residual space may over-sample noisy residual outliers on Qwen3-VL/Qwen3.5.

which performs token pruning inside the decoder. We reproduce VLM-Pruner using its Qwen2-VL-7B-Instruct configuration, with `min_pixels = max_pixels = 1280 × 28 × 28`. For GeoPrune, we use $K = 1$ and $\beta = 1.5$. We also report GeoPrune (only_div), corresponding to $K = 0$, to isolate the effect of diversity-based selection without our importance signal.

Table 18 reports results at 11.1% vision-token retention. This comparison serves as a reference across different pruning stages rather than a fully controlled system-level comparison, since pre-decoder and decoder-stage methods differ in pruning locations and runtime characteristics. GeoPrune (imp+div) achieves 93.86% recovery, compared with 92.27% for VLM-Pruner under this configuration. The results further suggest that importance and diversity provide complementary signals: GeoPrune (imp) performs relatively well on TextVQA and OCRBench, while combining the two signals yields the highest overall recovery among the compared methods.

I Extended Token Selection Visualization

Figure 9 extends Figure 6 with additional examples on LLaVA-1.5. Consistent with the coverage analysis in §3.2, VisionZip (ViT-attention based pruning) and residual ℓ_2 -norm show high overlap in their selected tokens, while both differ sharply from raw ℓ_2 -norm. Across all inputs, raw ℓ_2 -norm top- k tokens (red) cluster on featureless background patches, whereas residual ℓ_2 -norm top- k tokens (green) concentrate on semantically salient foreground objects, closely tracking vit attention (blue).

J Full Inference-Engine Measurements

J.1 Preprocessing Pipeline Breakdown

Table 21 reports the per-stage mean latency from vllm bench mm-processor (448×448 synthetic images, 1 image/request, 50 requests, 300 text tokens). Across all retention rates, the overhead in-

troduced by GeoPrune—i.e., computing residual-norm scores and selecting the top- k tokens—is less than 1 ms (reflected in the `encoder_forward` row), since GeoPrune merely snapshots the last hidden states *after* the ViT has finished. Despite this negligible overhead, end-to-end latency improves monotonically with stronger pruning, reaching up to -18% at 11.1% retention. Since the number of decode tokens is fixed at 32, the throughput gain primarily comes from prefill, which is consistent with expectation: vision token pruning significantly accelerates prefill inference by reducing the large number of vision tokens.

J.2 End-to-End Serving Metrics

Table 22 reports the full vLLM bench serve output for both models across two retention settings (baseline / mild retention with minimal accuracy loss) on ShareGPT4V chat (100 prompts), unlike the offline preprocessing-pipeline profiler which fixes the number of decode tokens at 32. We consider it unsound to benchmark at extremely low retention rates where model accuracy is substantially degraded (e.g., 11.1%), because the number of decoded tokens can be affected by model accuracy. We observe that **end-to-end acceleration becomes more pronounced as concurrency increases**, highlighting the significant advantage of vision token pruning in high-concurrency inference serving. Notably, for Qwen3-VL-8B at concurrency $c=8$, the baseline suffered from severe per-batch prefill queueing (TTFT p99 6302.7 ms); the TTFT reduction in this case should be interpreted as relief of a saturation outlier rather than a typical compute speedup. We also observe a temporal trade-off between TTFT and TPOT, which is tied to the scheduling behavior of the production-grade inference engine. Nevertheless, the output tokens per second consistently increase, demonstrating that vision token pruning brings substantial throughput gains.

Model	Bench.	$K=0$	$K=1$	$K=2$	$K=3$	K^*
LLaVA-1.5-7B 11.1% retention	GQA	38.33	55.48	55.86	55.83	
	MMBench	19.93	57.47	59.02	57.99	
	MME	896	1675	1662	1602	
	MMMU	31.67	35.44	35.89	35.81	2
	POPE	49.96	82.01	83.13	82.29	
	SQA	64.25	68.77	69.21	68.17	
	TextVQA	41.84	51.14	53.46	51.34	
Qwen3-VL-4B 30% retention	OCR	37.1	56.5	52.7	53.3	
	MME	1824	2138	2078	1980	
	SQA	85.66	89.08	88.73	89.01	1
	POPE	81.30	88.08	88.48	88.23	

Table 15: **Effect of deflation depth** K (residual ℓ_2 -norm top- k pruning). $K=0$ (raw ℓ_2 -norm) is consistently harmful; performance recovers at small K and then plateaus. K^* denotes the chosen deflation depth for each model.

Method	Signal	GQA	MMB	MME	POPE	SQA	TextVQA	MMMU	Avg. %
<i>LLaVA-1.5-7B — Total 576 tokens (100%)</i>									
Full model	–	61.90	64.70	1862	85.90	69.50	58.20	36.30	100%
<i>Retain 192 tokens ($\downarrow$ 67%)</i>									
VisionZip	Attn	59.24	63.57	1783	86.64	68.52	57.32	36.33	98.3%
DivPrune	Div	59.82	62.71	1749	87.89	68.27	56.96	35.33	97.6%
MMTok	Txt	60.07	63.40	1774	86.42	68.76	57.47	35.89	98.2%
Raw ℓ_2 ($K=0$)	Hid	48.99	34.36	1210	67.03	63.66	45.79	33.56	76.9%
GeoPrune (imp)	Hid	60.03	63.32	1793	87.22	68.42	55.88	35.78	98.0%
GeoPrune (imp+div)	Hid+Div	60.18	63.32	1793	87.66	68.47	57.46	36.33	98.7%
<i>Retain 128 tokens ($\downarrow$ 78%)</i>									
VisionZip	Attn	57.51	62.03	1761	84.71	68.82	56.98	37.00	97.3%
DivPrune	Div	59.21	61.17	1683	87.41	68.42	56.16	35.67	96.5%
MMTok	Txt	59.29	62.29	1753	86.25	68.82	56.91	34.78	97.0%
Raw ℓ_2 ($K=0$)	Hid	39.84	22.08	960	51.36	63.71	42.64	32.67	66.4%
GeoPrune (imp)	Hid	58.75	61.86	1750	86.11	68.47	53.87	36.56	96.7%
GeoPrune (imp+div)	Hid+Div	59.45	62.03	1769	87.49	68.91	56.47	36.11	97.8%
<i>Retain 64 tokens ($\downarrow$ 89%)</i>									
VisionZip	Attn	55.15	60.31	1657	80.56	69.01	55.42	36.00	94.1%
DivPrune	Div	57.46	58.16	1658	86.57	67.97	55.09	35.11	94.5%
MMTok	Txt	58.13	61.17	1688	85.77	69.16	55.69	35.11	95.8%
Raw ℓ_2 ($K=0$)	Hid	38.33	19.93	896	49.96	64.25	41.84	31.67	64.4%
GeoPrune (imp)	Hid	55.86	59.02	1662	83.13	69.21	53.46	35.89	94.0%
GeoPrune (imp+div)	Hid+Div	58.12	60.48	1677	86.32	68.17	55.96	35.89	95.9%

Table 16: **LLaVA-1.5-7B results** across three retention rates (33.3%, 22.2%, 11.1%), fixed $K=2$

Method	Signal	GQA	MMB	MME	MMMU	POPE	SQA	TextVQA	Avg. %
<i>LLaVA-Next-7B — Baseline (100%)</i>									
Full model	—	64.20	67.90	1842	35.10	86.40	70.20	61.30	100%
<i>Retain 128 tokens</i>									
VisionZip	Attn	61.15	65.55	1787	36.00	87.48	68.12	60.18	98.3%
DivPrune	Div	61.78	64.69	1798	37.44	87.82	67.92	56.44	98.0%
MMTok	Txt	62.14	64.95	1829	36.44	86.74	68.32	58.82	98.5%
GeoPrune (imp)	Hid	61.20	64.78	1768	36.22	87.59	67.28	59.67	97.8%
GeoPrune (imp+div)	Hid+Div	62.21	65.12	1827	37.11	88.26	67.82	58.96	99.0%
<i>Retain 64 tokens</i>									
VisionZip	Attn	59.01	63.92	1713	35.89	84.51	67.58	59.15	96.0%
DivPrune	Div	60.27	63.23	1754	37.89	86.22	67.92	55.83	96.8%
MMTok	Txt	60.96	64.00	1799	36.89	85.76	67.43	56.89	97.2%
GeoPrune (imp)	Hid	58.92	63.86	1686	36.33	85.12	67.72	58.10	95.8%
GeoPrune (imp+div)	Hid+Div	60.75	64.52	1794	36.44	87.24	67.58	57.86	97.5%
<i>Retain 32 tokens</i>									
VisionZip	Attn	55.42	59.79	1648	35.56	79.61	68.42	56.06	92.3%
DivPrune	Div	58.88	61.68	1625	37.00	83.33	66.63	54.71	93.8%
MMTok	Txt	60.05	62.97	1716	37.11	83.87	67.97	54.33	95.4%
GeoPrune (imp)	Hid	56.82	61.43	1578	35.11	82.20	67.67	55.21	92.3%
GeoPrune (imp+div)	Hid+Div	59.47	63.42	1662	36.56	85.12	67.13	56.68	95.3%

Table 17: **LLaVA-Next-7B results** across three retention rates (33.3%, 22.2%, 11.1%), fixed $K=3$.

Setting	Method	GQA	MMB	MME	POPE	SQA	TextVQA	OCRBench	OK-VQA	Recovery
Full	<i>Baseline, no pruning (100%)</i>	61.97	80.33	2302.77	88.88	85.87	82.37	80.3	50.63	100.00%
11.1% retention	<i>Pre-decoder pruning</i>									
	DivPrune (CVPR’25)	57.92	65.21	1871.36	87.71	79.03	69.63	47.8	35.52	82.61%
	GeoPrune (only_div)	57.90	66.24	1952.95	87.78	79.62	70.96	50.9	36.74	84.29%
	GeoPrune (imp)	53.08	67.44	1972.50	84.79	80.07	76.58	61.7	35.64	85.51%
	GeoPrune (imp+div)	60.22	76.29	2233.59	88.77	81.21	77.78	64.1	47.09	93.86%
	<i>Decoder-stage pruning</i>									
	VLM-Pruner (CVPR’26)	60.18	76.55	2142.61	87.09	80.57	76.15	59.5	47.77	92.27%

Table 18: **Comparison between pre-decoder and decoder-stage pruning on Qwen2-VL-7B-Instruct.** GeoPrune and the other methods in the upper block perform pre-decoder vision-token pruning, whereas VLM-Pruner (Wu et al., 2026) prunes tokens within the language decoder. Results are reported at 11.1% vision-token retention under the Qwen2-VL configuration used by VLM-Pruner. Recovery is measured relative to the full, unpruned model.

Setting	Method	MVBench	VideoMME-All	Short	Medium	Long	Recovery
Full	<i>Baseline, no pruning (100%)</i>	68.62	64.3	75.8	60.1	56.9	100.0%
5% retention	VisionZip (CVPR’25)	57.08	55.5	63.2	52.3	50.9	84.7%
	DivPrune (CVPR’25)	62.33	57.7	68.9	53.3	50.8	90.3%
	MMTok (ICLR’26)	56.10	58.2	67.2	54.6	52.7	86.1%
	GeoPrune (only_div)	57.17	56.5	65.6	52.8	51.1	85.6%
	GeoPrune (imp)	57.50	58.1	67.0	54.9	52.6	87.0%
	GeoPrune (imp+div)	62.17	58.3	67.8	55.7	51.6	90.6%
50% retention	GeoPrune (imp)	65.90	63.4	74.3	59.8	56.1	97.3%
	GeoPrune (imp+div)	66.35	63.3	74.7	60.0	55.2	97.5%

Table 19: **Video benchmark results on Qwen3-VL-8B-Instruct.** Results are evaluated on MVBench and VideoMME at different vision-token retention ratios. Recovery denotes the relative average performance retained with respect to the full model.

Setting	Method	MVBench	VideoMME-All	Short	Medium	Long	Recovery
Full	<i>Baseline, no pruning (100%)</i>	<i>69.60</i>	<i>68.1</i>	<i>79.6</i>	<i>67.2</i>	<i>57.4</i>	<i>100.0%</i>
5% retention	VisionZip (CVPR'25)	56.20	59.2	68.2	57.8	51.7	83.8%
	DivPrune (CVPR'25)	60.62	59.3	68.6	58.6	50.8	87.1%
	MMTok (ICLR'26)	53.00	57.5	65.0	56.0	51.6	80.3%
	GeoPrune (only_div)	54.73	59.8	68.8	57.7	52.9	83.2%
	GeoPrune (imp)	58.75	61.0	71.6	59.3	52.2	87.0%
	GeoPrune (imp+div)	61.40	61.9	73.3	59.0	53.4	89.5%
50% retention	GeoPrune (imp)	68.97	67.7	79.1	66.8	57.2	99.3%
	GeoPrune (imp+div)	68.90	67.1	79.2	65.8	56.3	98.8%

Table 20: **Video benchmark results on Qwen3.5-9B**. Results are evaluated on MVBench and VideoMME at different vision-token retention ratios. Recovery denotes the relative average performance retained with respect to the full model.

Stage (mean, ms)	<i>Qwen2.5-VL-7B</i>				<i>Qwen3-VL-8B</i>			
	Vanilla	50%	30%	11.1%	Vanilla	70%	30%	11.1%
apply_hf_processor	6.13	4.97	5.54	5.61	5.01	4.89	5.11	4.92
apply_prompt_updates	0.54	0.30	0.36	0.23	0.43	0.28	0.25	0.22
preprocessor_total	6.84	5.42	6.05	5.99	5.59	5.31	5.51	5.29
encoder_forward	7.73	7.94	7.82	8.16	4.07	4.62	4.75	4.47
E2E (mean, ms)	1087	973	933	887	938	887	850	775
Δ_{E2E}	—	−10%	−14%	−18%	—	−5%	−9%	−17%

Table 21: **Offline mm-processor breakdown** (vllm bench mm-processor, 50 requests, 448×448 resolution images, 32 decode tokens, mean per-request latency).

<i>Qwen2.5-VL-7B</i>								
Metric	<i>c=1</i>		<i>c=4</i>		<i>c=8</i>		<i>c=16</i>	
	Vanilla	50%	Vanilla	50%	Vanilla	50%	Vanilla	50%
TTFT mean (ms)	84.3	84.5	106.2	107.2	89.0	84.3	139.8	128.2
TTFT p99 (ms)	133.9	128.5	304.7	374.2	221.9	205.3	393.9	338.3
TPOT mean (ms)	6.04	6.15	7.08	7.02	8.00	7.96	9.99	9.64
Output tok/s	151.4	148.7	505.1	508.5	891.2	897.0	1355.9	1417.5
<i>Qwen3-VL-8B</i>								
Metric	<i>c=1</i>		<i>c=4</i>		<i>c=8</i>		<i>c=16</i>	
	Vanilla	70%	Vanilla	70%	Vanilla	70%	Vanilla	70%
TTFT mean (ms)	67.1	64.3	77.8	81.7	555.5	38.0	94.3	52.5
TTFT p99 (ms)	203.8	98.6	289.3	352.5	6302.7	91.1	286.5	157.0
TPOT mean (ms)	6.99	6.87	7.76	7.73	7.88	7.23	8.95	7.71
Output tok/s	137.0	139.4	485.2	485.4	730.0	1042.5	1574.8	1847.2

Table 22: **End-to-end serving on vLLM** across four concurrencies.

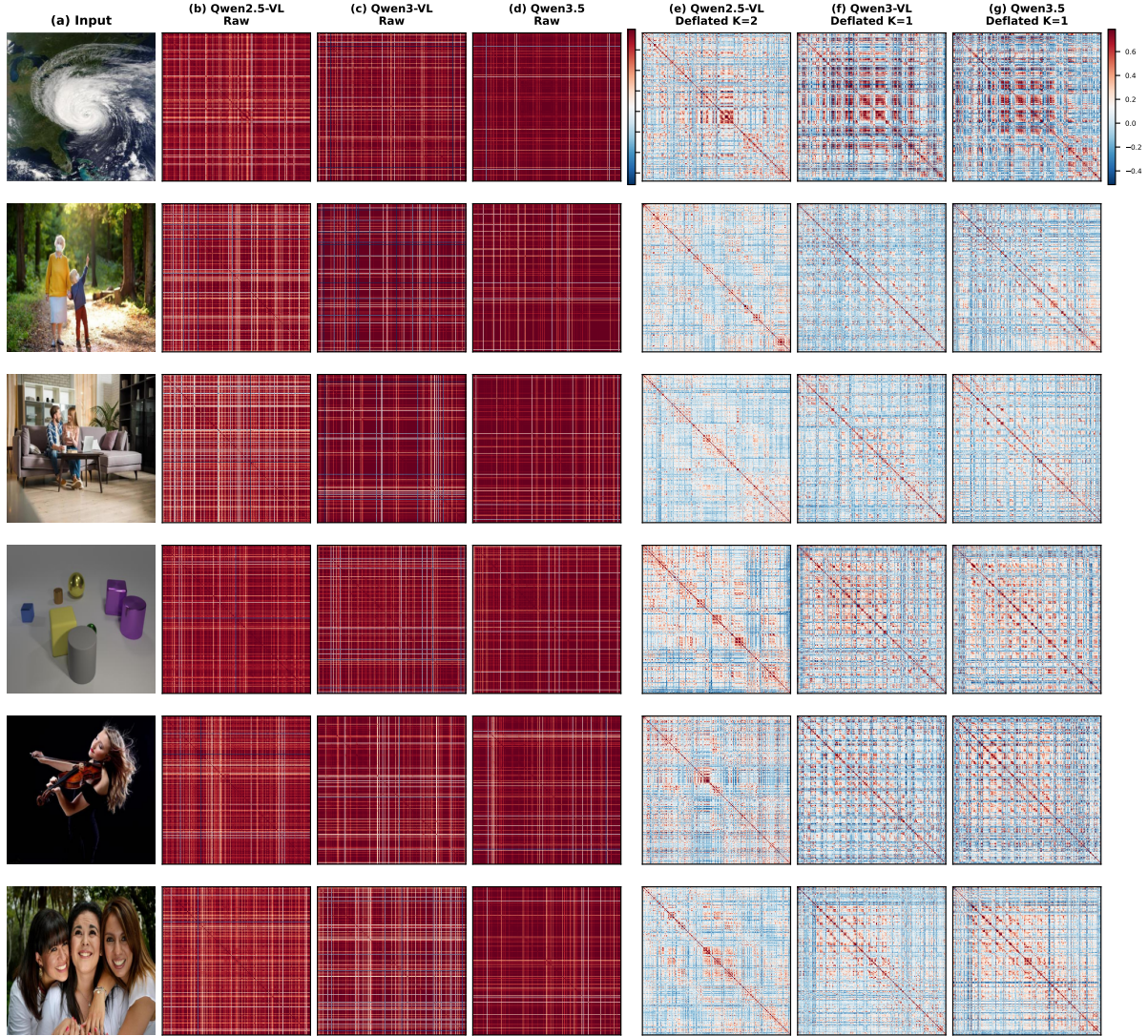


Figure 8: **Pairwise cosine similarity in S_{out} : raw vs. deflated** across three Qwen-family models. (b–d) Raw similarity is nearly uniform, causing cosine deviation from the mean to lose discriminative power as an importance score. (e–g) After deflation, rich token-level structure emerges, restoring token separability and enabling residual ℓ_2 -norm to serve as a reliable importance signal.

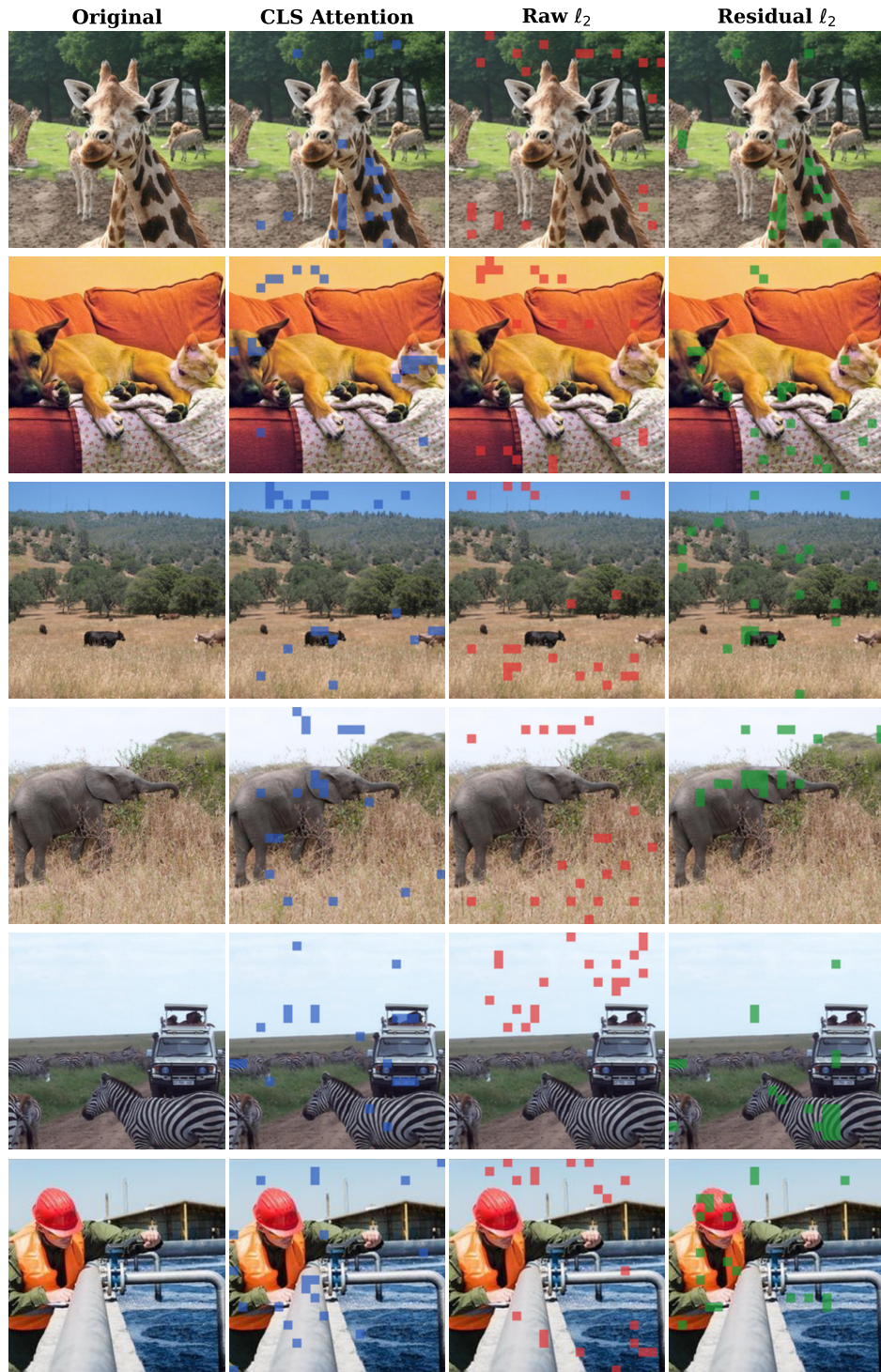


Figure 9: **Extended raw vs. residual ℓ_2 -norm token selection on LLaVA-1.5.** Columns: input image, VisionZip top-20 (blue), raw ℓ_2 -norm top-20 (red), and residual ℓ_2 -norm top-20 (green).