BEYOND TOKEN-WISE IMPORTANCE: VISION TOKEN PRUNING WITH LOCAL RELATIONAL REDUNDANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) process hundreds or even thousands of visual tokens, resulting in substantial computational and memory overhead. Since these tokens contribute unequally to response generation, *vision token pruning* aims to reduce this overhead by retaining only the most informative tokens. Existing methods typically estimate token importance from individual attention scores, visual features, or token-level redundancy. In this paper, we propose worst-case local dissimilarity as a simple prior for token selection, where high similarity even to a token's K-th most similar visual token indicates strong redundancy. Based on this prior, we introduce R^{2}PRUNER, a training-free vision token pruning method that uses local representation structure to guide progressive token selection. R^{2}PRUNER applies the prior at both the token and neighborhood levels, using local dissimilarity as a multiplicative weight on token importance. We find that this prior is effective both on its own and in combination with existing token-importance signals. When combined with an existing spectral evolution score, R^{2}PRUNER retains 96.4% of the unpruned model's performance on average across eight multimodal benchmarks at an average visual-token budget of only 64 out of the original 576 tokens. This exceeds the previous state-of-the-art by 1.7 percentage points at the same token budget.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities across a wide range of vision-language tasks, including visual question answering, visual reasoning, and document understanding (Alayrac et al., 2022; Liu et al., 2023; Li et al., 2023b; Dai et al., 2023; Bai et al., 2023). Despite their strong performance, MLLMs typically represent an input image with hundreds or even thousands of visual tokens, all of which are processed by the language model (Wang et al., 2024; Chen et al., 2024b). The resulting long visual sequence incurs substantial computational and memory overhead, particularly during the prefill stage (Tang et al., 2025). At the same time, visual tokens often exhibit substantial redundancy, as multiple tokens may encode highly similar visual information (Yang et al., 2025; Alvar et al., 2025). Recent studies have therefore explored vision token pruning as an effective approach to improve MLLM inference efficiency without additional training (Xing et al., 2024; Zhang et al., 2025; Chen et al., 2026c).

A central challenge in vision token pruning is determining which visual tokens should be preserved. Existing methods predominantly address this problem through token-importance estimation using signals such as text-to-vision attention or layer-wise changes of visual representation (Chen et al., 2024a; 2026c;a). While effective for identifying tokens that are likely to contribute useful information, these criteria do not explicitly account for token's redundancy. In particular, a token can receive a high importance score even when its visual information is highly redundant with other similar tokens. Consequently, retaining such a token provides little additional information. This suggests that token selection should consider not only how important a token is, but also how difficult its information is to replace: highly importance score yet redundant tokens can be deprioritized, whereas moderately important but distinctive information tokens may still be worth preserving.

We introduce R^{2}PRUNER, a training-free vision token pruning method that complements token-importance estimation with a worst-case local dissimilarity prior that explicitly show how replaceable a token is with respect to its local representation neighborhood. Our key intuition is simple:

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

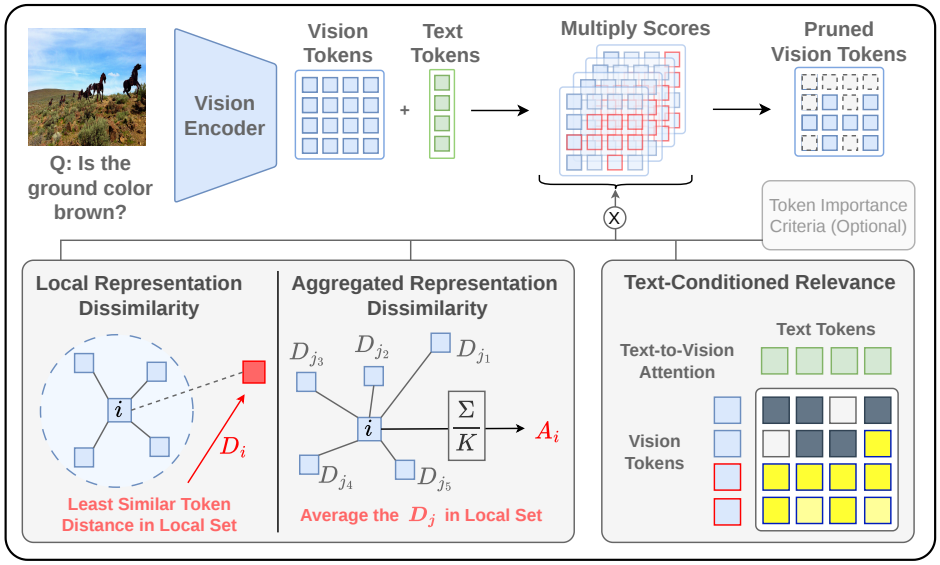


Figure 1: **Overview of R^{2}PRUNER.** R^{2}PRUNER complements a conventional token-importance criterion with two representation-level signals that explicitly capture local redundancy: local and aggregated representation dissimilarity for token- and neighborhood-level distinctiveness, respectively. At the initial pruning stage, these redundancy-aware signals are combined with token importance and question-conditioned relevance to identify informative yet locally non-redundant visual tokens. Subsequent pruning stages progressively reduce the remaining visual tokens using question-conditioned relevance, allowing the tokens to focus on regions most relevant to the current instruction.

if a token remains highly similar even to its K-th most similar visual representation, its information is likely to be well covered by nearby tokens and is therefore more locally redundant. We instantiate this prior at two complementary levels: (*i*) local representation dissimilarity measures the distinctiveness of the token itself relative to its neighboring representations, while (*ii*) aggregated representation dissimilarity captures the overall distinctiveness of the token's local neighborhood. Together, these signals provide whether the information carried by a token is already represented elsewhere in its local neighborhood. The proposed prior can guide token selection on its own or serve as a multiplicative weight on existing token-importance scores; see Fig. 1 for an overview.

Across eight multimodal benchmarks, R^{2}PRUNER consistently maintains the strongest downstream performance across multiple average visual-token budgets, where the budget denotes the number of visual tokens processed across decoder layers after following the first pruning stage. Under the most aggressive setting on LLaVA-1.5-7B, which reduces the original visual-token count by 88.9%, R^{2}PRUNER retains 96.4% of the vanilla model performance and outperforms the strongest competing method by 1.7 percentage points. Considering task performance, inference throughput, and KV-cache memory jointly, R^{2}PRUNER establishes a superior Pareto frontier upon the state-of-the-art approaches, achieving the highest downstream performance with comparable throughput while reducing KV-cache memory by 8.6% on average compared with the strongest baseline, across all the evaluated token-reduction settings.

2 Related Work

Multimodal Large Language Models. Many MLLMs connect a pretrained vision encoder to a language model through a modality connector, allowing visual representations to interact with text tokens (Li et al., 2023b; Dai et al., 2023; Bai et al., 2023). Vision Transformers encode images as sequences of patch-level tokens (Dosovitskiy et al., 2020), while vision-language encoders such as CLIP, SigLIP, and SigLIP2 learn visual representations through large-scale image-text pretraining (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025). For example, LLaVA projects patch-level features from a CLIP vision encoder into the embedding space of language model and

processes them alongside text tokens (Liu et al., 2023; 2024a). Although the visual encoders and connectors vary across architectures, a common design is to convert visual inputs into token representations that the language model can process jointly with text.

Recent MLLMs extend this interface to higher-resolution images and longer multimodal contexts. Qwen2-VL (Wang et al., 2024) introduces dynamic-resolution processing, representing images with variable numbers of visual tokens. InternVL3 (Zhu et al., 2025) explores native multimodal pre-training to jointly develop visual and linguistic capabilities, while Qwen3-VL (Bai et al., 2025) integrates multi-level visual features and supports long interleaved image-text-video contexts. Other architectures explore specialized visual encoders and native-resolution processing for unified image and video perception (An et al., 2026). In particular, support for higher-resolution images and longer visual contexts increases the number of visual tokens processed by the language model. This growing computational and memory cost motivates visual token reduction methods that preserve useful information while improving inference efficiency.

Vision Token Reduction in MLLMs. Visual token reduction methods operate either before visual tokens enter the language model or within its decoder layers. Among methods in the former category, LLaVA-PruMerge (Shang et al., 2025) and VisionZip (Yang et al., 2025) use attention to identify informative tokens and feature similarity to merge redundant information. DivPrune (Alvar et al., 2025) formulates token selection as a max-min diversity problem, while VisPruner (Zhang et al., 2025) combines visual saliency with feature similarity to preserve informative and diverse tokens. These approaches primarily reduce redundancy before language-model processing through similarity-, diversity-, or merging-based criteria.

Within the language model, attention-based methods exploit multimodal interactions to guide token selection. FastV (Chen et al., 2024a) prunes visual tokens using attention from early decoder layers, while SparseVLM (Zhang et al., 2024) uses selected text tokens to estimate visual-token relevance and progressively sparsifies the visual sequence. PriorTR (Chen et al., 2026c) further separates instruction-dependent relevance from an instruction-agnostic attention prior. Beyond attention-based selection, other methods use changes in visual representations across decoder layers: V^{2}Drop (Chen et al., 2026b) measures token-wise representation variation, while CLSE (Chen et al., 2026a) measures spectral evolution to identify tokens undergoing representational transformation. Despite their different formulations, these inner-LLM methods primarily determine which tokens should be retained based on attention, task relevance, or representational change.

R^{2}PRUNER belongs to the training-free, inner-LLM vision token pruning paradigm, but differs in how local representation structure informs token selection. Unlike existing importance criteria, R^{2}PRUNER explicitly characterizes both a token's own local distinctiveness and the redundancy of its representation-space neighborhood. The resulting signals serve as multiplicative weights on token importance, complementing existing pruning without further training.

Local Structure of Neural Representations. Neural representations can be characterized not only through individual feature vectors, but also through the relationships they induce among data points. Global representation similarity measures such as centered kernel alignment (CKA) compare representations through their pairwise similarity structure, providing a coordinate-independent view of representational alignment across layers or models (Kornblith et al., 2019).

Beyond such global comparisons, recent work has shown that local structure in representation space also provides a meaningful description of learned representations. The Platonic Representation Hypothesis (PRH), for example, compares representation spaces across models and modalities using mutual K-nearest-neighbor agreement (Huh et al., 2024). Subsequent work further examines representation alignment under calibrated similarity measures and highlights local neighborhood agreement as a robust indicator of representational similarity (Gröger et al., 2026). These studies motivate the broader view that local neighborhood statistics can provide complementary information beyond that captured by individual features vectors alone.

Building on this perspective, R^{2}PRUNER applies local neighborhood information to a different problem: characterizing visual-token redundancy. Unlike representation comparison methods that use neighborhood relations to assess alignment across representation spaces, R^{2}PRUNER uses them as an auxiliary signal for token selection. This connection motivates our use of neighborhood-level representation information alongside conventional token-importance criteria.

3 METHOD

R^{2}PRUNER is a training-free vision token pruning method that uses local representation structure to guide token selection. It builds on a simple worst-case local dissimilarity prior: a visual token is likely to be locally redundant if it remains highly similar even to its K-th most similar visual representation. We apply this prior at two levels: (*i*) *local representation dissimilarity* D_i measures a token's own local distinctiveness, while (*ii*) *aggregated representation dissimilarity* A_i summarizes the local distinctiveness of its representation-space neighbors for structural characteristics. Together, these scores provide local redundancy information that can be used to weight token importance.

3.1 TOKEN SCORING COMPONENTS

Local Representation Dissimilarity. At the initial pruning stage, let $\mathcal{V}$ denote the visual-token index set and let $\mathbf{v}_i \in \mathbb{R}^d$ denote the hidden representation of visual token $i \in \mathcal{V}$ immediately before pruning. We use raw dot-product similarity without normalizing their feature magnitudes. Unlike cosine similarity, which removes magnitude information through feature normalization, the raw dot product preserves both directional alignment and feature magnitude, providing a richer and more differentiated signal in the original representation space.

Let $\arg\,\mathrm{TopK}_K(\{a_j\}_{j\in\mathcal{J}})$ denote the indices corresponding to the K largest scores over an index set $\mathcal{J}$. For each visual token i, we define its local representation set as

$$\mathcal{L}_i = \arg\,\mathrm{TopK}_K\left(\left\{\mathbf{v}_i^\top \mathbf{v}_j\right\}_{j\in\mathcal{V}\setminus\{i\}}\right), \text{ where } |\mathcal{L}_i| = K. \tag{1}$$

Restricting the comparison to the top-K representations preserves the locality of the redundancy estimate. If all visual tokens were considered, the minimum similarity could be dominated by an unrelated representation elsewhere in the image, causing a locally redundant token to appear distinctive simply because it is dissimilar to a globally unrelated token. The visualization of global vs. local dissimilarity is provided in Appendix E. We then define the worst-case local similarity of token i as

$$M_i = \min_{j\in\mathcal{L}_i} \mathbf{v}_i^\top \mathbf{v}_j. \tag{2}$$

Since $\mathcal{L}_i$ contains the K visual representations with the largest dot-product similarities to token i, M_i corresponds to its similarity with the K-th most similar representation. A high M_i therefore indicates that token i remains strongly similar even to the least similar representation within its local top-K neighborhood, suggesting that it is supported by multiple locally related representations.

Throughout this section, we use sample-wise min-max normalization. For an index set $\mathcal{S}$, we define

$$\mathrm{MM}_{\mathcal{S}}(x_i) = \frac{x_i - \min_{j\in\mathcal{S}} x_j}{\max_{j\in\mathcal{S}} x_j - \min_{j\in\mathcal{S}} x_j + \epsilon}, \tag{3}$$

where ϵ is a small constant for numerical stability. With this normalization, local representation dissimilarity is then defined as

$$D_i = 1 - \mathrm{MM}_{\mathcal{V}}(M_i). \tag{4}$$

Aggregated Representation Dissimilarity. Local representation dissimilarity characterizes each visual token from a token-level perspective. To additionally characterize the local redundancy from a neighborhood-level perspective, we aggregate the dissimilarity scores of the representations in $\mathcal{L}_i$:

$$A_i = \frac{1}{K}\sum_{j\in\mathcal{L}_i} D_j. \tag{5}$$

Importantly, A_i does not aggregate pairwise dissimilarities between the center token i and the representations in $\mathcal{L}_i$. Instead, each D_j measures the local distinctiveness of representation j with respect to its own local representation set. Consequently, A_i summarizes the redundancy characteristics of the representations forming the local neighborhood of token i, providing a neighborhood-level signal complementary to the token-level D_i.

Question-Conditioned Relevance. The representation-level signals described above characterize the local redundancy among visual tokens, but they are independent of the current question. To incorporate task-specific information, we additionally use a standard question-conditioned attention signal that measures the relevance of each visual token to the input question.

Let $\mathcal{Q}$ denote the set of user-question token indices and $\mathcal{H}$ the set of attention heads. For pruning stage $m \in \{1, 2, 3\}$, we define $\mathcal{I}^{(0)} = \mathcal{V}$, where $\mathcal{I}^{(m-1)}$ denotes the set of visual tokens remaining immediately before pruning stage m. For each visual token $i \in \mathcal{I}^{(m-1)}$, we first compute its raw question relevance as

$$\widetilde{R}_i^{(m)} = \frac{1}{|\mathcal{H}||\mathcal{Q}|} \sum_{h \in \mathcal{H}} \sum_{q \in \mathcal{Q}} \alpha_{q,i}^{(h,m)}, \tag{6}$$

where $\alpha_{q,i}^{(h,m)}$ denotes the scalar attention weight from question-token index q to visual token i under attention head h, measured at the decoder layer immediately preceding pruning stage m. The question-conditioned relevance score is then obtained by normalizing the raw relevance scores over the currently remaining visual tokens:

$$R_i^{(m)} = \mathrm{MM}_{\mathcal{I}^{(m-1)}} \left(\widetilde{R}_i^{(m)} \right), \qquad i \in \mathcal{I}^{(m-1)}. \tag{7}$$

Thus, the question-conditioned relevance is recomputed at each pruning stage using the attention weights from the corresponding decoder layer. A larger $R_i^{(m)}$ indicates stronger relative relevance to the input question among the visual tokens available at that stage.

3.2 Progressive Token Selection

R^{2}Pruner performs three pruning stages, indexed by $m \in \{1, 2, 3\}$. At the initial pruning stage ($m = 1$), R^{2}Pruner uses the proposed local dissimilarity signals to weight token importance. Let E_i denote the score assigned to visual token i by a chosen token-importance criterion. In our instantiation, E_i is standardized across visual tokens and mapped to $(0, 1)$ using a temperature-scaled sigmoid before multiplication. We combine this score with local representation dissimilarity D_i, aggregated representation dissimilarity A_i, and question-conditioned relevance $R_i^{(1)}$ as

$$S_i^{(1)} = E_i \cdot D_i \cdot A_i \cdot R_i^{(1)}. \tag{8}$$

The multiplicative formulation prioritizes tokens that simultaneously exhibit high underlying importance, low local redundancy at both the token and neighborhood levels, and strong question-conditioned relevance. Given a target retention count N_1, we retain the tokens with the largest selection scores:

$$\mathcal{I}^{(1)} = \arg \mathrm{TopK}_{N_1} \left(\left\{ S_i^{(1)} \right\}_{i \in \mathcal{V}} \right). \tag{9}$$

Here, $\mathcal{I}^{(1)} \subseteq \mathcal{V}$ denotes the retained indices with respect to the original visual-token sequence. Throughout all pruning stages, $\mathcal{I}^{(m)}$ preserves these original token indices rather than positions in a newly compacted sequence.

The proposed representation-level signals D_i and A_i are used only at the initial pruning stage. At subsequent stages $m \in \{2, 3\}$, the remaining visual tokens are progressively selected using question-conditioned relevance. As described in Sec. 3.1, $R_i^{(m)}$ is recomputed at the decoder layer corresponding to pruning stage m, so that the selection reflects the updated interaction between the remaining visual representations and the input question. Specifically,

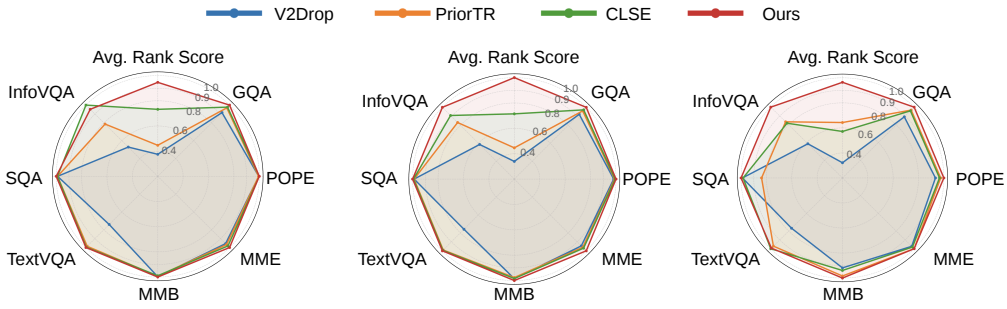
$$\mathcal{I}^{(m)} = \arg \mathrm{TopK}_{N_m} \left(\left\{ R_i^{(m)} \right\}_{i \in \mathcal{I}^{(m-1)}} \right), \qquad m \in \{2, 3\}, \tag{10}$$

where N_m denotes the target retention count at pruning stage m. Thus, $\mathcal{I}^{(3)} \subseteq \mathcal{I}^{(2)} \subseteq \mathcal{I}^{(1)} \subseteq \mathcal{V}$, and each stage further reduces the visual-token set retained by the preceding stage.

In our implementation, we use the spectral evolution score from CLSE (Chen et al., 2026a) as E_i; the definitions of D_i and A_i do not depend on this choice. The subsequent question-conditioned pruning stages follow the progressive pruning design used in PyramidDrop (Xing et al., 2024) and CLSE. The complete pruning procedure of R^{2}Pruner is specified in Appendix A.

Table 1: Comparison with state-of-the-art token pruning methods on LLaVA-1.5-7B, with different numbers of vision tokens preserved; out of 576 in the Vanilla. "Avg." reports the average of normalized results on all datasets. Best results are in **bold** and second-best results are in underlined.

Method	GQA	POPE	MME	MMB	Text VQA	SEEDB	SQA	Info VQA	Avg. (%)
Upper Bound, 576 Tokens (**100%**)									
Vanilla	61.9	85.8	1868	64.0	58.2	66.1	70.4	20.4	100
Average Retain 192 Tokens (↓ 66.7%)									
FastV (Chen et al., 2024a)	52.8	63.8	1612	60.8	52.6	63.4	68.7	19.4	90.0
SparseVLM (Zhang et al., 2024)	59.4	85.3	1786	62.5	56.1	64.2	68.8	20.0	97.2
PruMerge (Shang et al., 2025)	54.3	71.3	1623	59.3	54.4	57.3	68.6	19.6	90.5
VisPruner (Fan et al., 2025)	59.6	**86.1**	1791	63.2	57.5	63.6	68.6	19.5	97.4
V2Drop (Chen et al., 2026b)	59.1	83.8	1811	62.2	55.6	63.7	69.2	20.2	97.1
PriorTR (Chen et al., 2026c)	60.3	83.9	1831	63.3	57.6	64.7	69.9	19.6	98.0
CLSE (Chen et al., 2026a)	61.4	85.2	1836	**63.8**	57.5	65.1	**70.2**	20.1	99.0
R^{2}PRUNER **(Ours)**	**61.6**	85.7	**1880**	63.7	**58.0**	**66.0**	**70.2**	**20.6**	**100.0**
Average Retain 128 Tokens (↓ 77.8%)									
FastV (Chen et al., 2024a)	49.5	59.6	1490	56.7	51.2	58.1	68.9	19.1	85.6
SparseVLM (Zhang et al., 2024)	58.4	85.1	1765	60.0	55.5	63.5	68.6	19.4	95.7
PruMerge (Shang et al., 2025)	53.2	67.0	1530	58.1	54.2	54.9	68.5	19.2	88.0
VisPruner (Fan et al., 2025)	58.3	83.3	1754	61.8	57.0	61.6	69.0	19.8	96.0
V2Drop (Chen et al., 2026b)	58.1	81.6	1790	62.2	52.8	62.3	68.6	19.8	95.2
PriorTR (Chen et al., 2026c)	59.1	81.5	1794	62.2	56.7	63.8	69.9	19.4	96.5
CLSE (Chen et al., 2026a)	60.7	84.3	1807	62.5	56.8	64.6	70.0	20.1	98.0
R^{2}PRUNER **(Ours)**	**61.2**	**85.6**	**1871**	**63.2**	**57.6**	**65.8**	**70.1**	**20.4**	**99.4**
Average Retain 64 Tokens (↓ 88.9%)									
FastV (Chen et al., 2024a)	46.0	53.2	1256	48.0	47.2	49.6	69.8	18.5	78.1
SparseVLM (Zhang et al., 2024)	52.7	77.7	1593	56.0	51.8	56.8	69.4	19.1	89.5
PruMerge (Shang et al., 2025)	51.8	65.4	1567	52.3	54.2	53.3	68.8	19.1	86.3
VisPruner (Fan et al., 2025)	55.4	80.1	1671	59.7	**55.5**	58.2	69.1	19.3	92.7
V2Drop (Chen et al., 2026b)	54.6	71.5	1617	56.2	52.0	58.0	68.6	19.1	89.3
PriorTR (Chen et al., 2026c)	55.9	76.5	1742	61.2	54.7	60.0	**69.9**	19.3	93.4
CLSE (Chen et al., 2026a)	58.2	78.1	1787	61.3	55.2	61.5	69.6	19.3	94.7
R^{2}PRUNER **(Ours)**	**59.0**	**82.2**	**1823**	**61.7**	54.8	**64.1**	69.8	**19.5**	**96.4**



(a) 66.7% Token Reduction **(b) 77.8% Token Reduction** **(c) 88.9% Token Reduction**

Figure 2: **Comparison on Qwen2-VL under visual-token reduction ratios, of 66.7%, 77.8%, and 88.9%, across seven multimodal benchmarks.** Normalized benchmark scores and Avg. Rank Score are shown for each reduction setting. R^{2}PRUNER exhibits consistently strong normalized performance and the highest Avg. Rank Score across all three token-reduction settings. Detailed results and metric definitions are provided in Appendix B.

4 EXPERIMENTS

We evaluate R^{2}PRUNER on LLaVA-1.5-7B across eight multimodal benchmarks and compare it with seven representative training-free vision token reduction methods under the same backbone and evaluation pipeline. We report results under post-pruning average visual-token budgets of 192, 128, and 64 tokens. Detailed descriptions of the datasets, evaluation protocol, and implementation settings, additional experimental results of Qwen2-VL model are provided in Appendix B, qualitative analysis of question-conditioned relevance is provided in Appendix D.

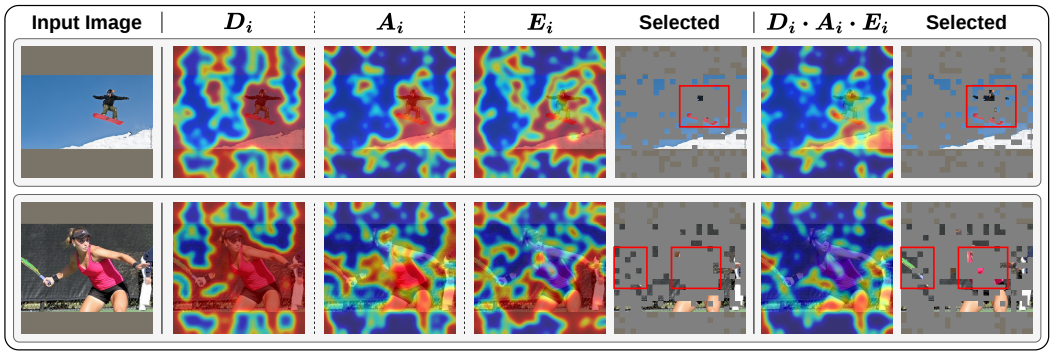


Figure 3: **Qualitative effect of local redundancy-aware token selection.** Comparing with token importance E_i alone, combining local and aggregated representation dissimilarity suppresses locally redundant tokens and preserves visually distinctive object regions. This demonstrates that the proposed representation-level signals complement token importance by favoring informative yet locally less replaceable visual tokens.

4.1 Main Results

Table 1 compares R^{2}Pruner with representative training-free vision token reduction methods under three depth-averaged visual-token budgets. R^{2}Pruner consistently outperforms competing methods across individual benchmarks and achieves the strongest average performance across all evaluated token budgets. R^{2}Pruner performs particularly well on MME (Fu et al., 2023) and SEED-Bench (Li et al., 2023a), both of which evaluate diverse multimodal capabilities rather than a single task category. This suggests that redundancy-aware selection effectively removes replaceable visual representations while preserving distinctive information required across a broad range of perceptual and reasoning tasks.

Notably, the advantage of R^{2}Pruner becomes more pronounced as the token budget decreases. This trend is particularly clear on POPE (Li et al., 2023c). At the 192 token settings, VisPruner (Zhang et al., 2025) slightly outperforms R^{2}Pruner on POPE, it achieves the best performance at both 128 and 64 tokens, with a larger margin under the most aggressive setting. Since POPE directly evaluates object-presence judgments, this trend suggests that redundancy-aware selection becomes increasingly beneficial when visual-token budget is highly constrained. However, an exception to this trend is TextVQA Singh et al. (2019). Under the most aggressive 64-token setting, R^{2}Pruner becomes less competitive, as accurate OCR requires preserving fine-grained visual details. This sensitivity is consistent with the redundancy-aware nature of R^{2}Pruner: visually similar text patches may appear replaceable in representation space while still encoding distinct character-level information, making aggressive pruning more likely to discard useful OCR evidence.

We further evaluate R^{2}Pruner on Qwen2-VL under 66.7%, 77.8%, and 88.9% visual token reduction ratios. As shown in Fig. 2, R^{2}Pruner maintains the strongest overall performance across the three settings. Interestingly, the same task-specific sensitivity reappears under the most aggressive reduction setting, where R^{2}Pruner is relatively weaker on TextVQA. This consistent pattern across two MLLM backbones suggests that the limitation is associated with aggressively pruning fine-grained, locally similar textual information. Detailed results are provided in Appendix B.

4.2 Analyses

Qualitative analysis. Figure 3 qualitatively examines how the proposed representation-level signals affect token selection. We compare the underlying token-importance score E_i with the redundancy-aware score $D_i \cdot A_i \cdot E_i$. While E_i identifies visually informative regions, high importance alone does not indicate whether the represented information is locally replaceable. Consequently, importance-only selection can allocate tokens to locally redundant regions while discarding distinctive parts of foreground objects.

This behavior is consistently observed across the examples. In the first image, E_i preserves only a small portion of the snowboarder while selecting many background regions, whereas $D_i \cdot A_i \cdot E_i$

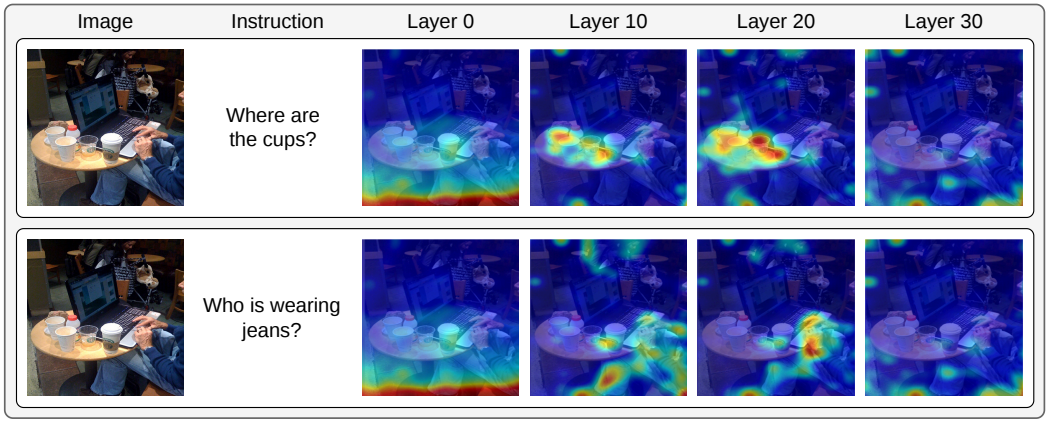
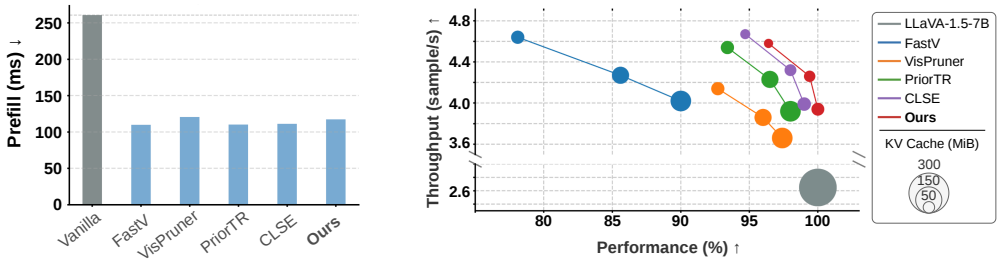


Figure 4: **Visualization of layer-wise text-conditioned relevance in LLaVA-1.5-7B.** LLaVA-1.5-7B consists of 32 LLM layers. We visualize the text-conditioned relevance maps at Layer 0, 10, 20, and 30 for the same image under two different instructions. As representations propagate through the decoder, the relevance maps become increasingly conditioned on the given instruction, highlighting different image regions depending on the queried content.



(a) Prefill latency

(b) Performance *vs.* Throughput *vs.* KV-cache trade-off

Figure 5: **Efficiency comparison across vision token pruning methods on LLaVA-1.5-7B.** (a) Complete multimodal prefill latency. (b) Joint trade-off among performance, inference throughput, and KV-cache memory across different visual-token budgets. Each trajectory corresponds to increasingly aggressive token reduction, and marker size represents the KV-cache footprint, with smaller markers indicating lower memory usage.

retains substantially more tokens on the person. In the second example, selection based only on E_i removes most tokens corresponding to the player's torso and racket, where incorporating D_i and A_i preserves both regions. These results illustrate that the proposed local representation signals complement token importance by suppressing locally replaceable representations and preserving informative regions that are less redundant within their representation-space neighborhoods.

Evolution of text-condition across layers. To examine how text-conditioned visual relevance evolves across decoder layers, we visualize relevance maps from LLaVA-1.5-7B, which consists of 32 decoder layers, in Fig. 4. At the earliest layer, the maps exhibit a strong positional bias and remain highly similar across different instructions. In contrast, Layers 10 and 20 show much clearer instruction-dependent responses, highlighting different image regions according to the query. Interestingly, this distinction becomes less pronounced again near the final layer, suggesting that instruction-conditioned relevance is most discriminative at intermediate depths. Additional qualitative results are in Appendix D.

This observation motivates our progressive pruning design. Aggressive pruning at an early layer may remove tokens before instruction-specific relevance is sufficiently developed, whereas intermediate layers provide more reliable signals for refining the retained token set. We therefore perform pruning progressively, using complementary visual cues at the initial stage and increasingly relying on instruction-conditioned relevance in subsequent stages.

Table 2: Ablation study of the proposed pruning signals under average visual-token budgets of 64 tokens. E_i denotes the existing token-importance criterion from Chen et al. (2026a). D_i, A_i, and R_i denote local representation dissimilarity, aggregated representation dissimilarity, and question-conditioned relevance, respectively.

	$\mathbf{E}_i$	$\mathbf{D}_i$	$\mathbf{A}_i$	$\mathbf{R}_i$	GQA	POPE	MME	SEEDB
R^2PRUNER	✓	✓	✓	✓	**59.0**	**82.2**	**1823**	**64.1**
w/o E_i	✗	✓	✓	✓	57.7	80.8	1792	63.2
w/o E_i, D_i	✗	✗	✓	✓	57.7	80.7	1791	63.2
w/o E_i, A_i	✗	✓	✗	✓	57.3	79.2	1801	62.5
w/o E_i, R_i	✗	✓	✓	✗	57.0	74.8	1700	62.0
w/o D_i, A_i	✓	✗	✗	✓	58.2	81.5	1800	63.3
w/o E_i, D_i, A_i	✗	✗	✗	✓	55.7	75.5	1764	61.5

Efficiency analysis. We measure complete multimodal prefill latency, including the vision encoder, multimodal projector, method-specific pruning operations, multimodal input construction, and LLM prefill, while excluding host-to-device transfer. We use identical multimodal prompt, "Describe this image. Answer with a single word or phrase.", for all methods during efficiency evaluation. We use FP16 inference and the eager attention implementation for all methods. KV-cache memory is reported in MiB, and throughput is reported in samples per second. All runtime measurements are averaged over 100 runs after 10 warm-up iterations.

Fig. 5 evaluates the practical efficiency of different vision token pruning methods in terms of prefill latency, inference throughput, and KV-cache memory. As shown in Fig. 5a, all pruning methods substantially reduce multimodal prefill latency compared with the unpruned LLaVA-1.5-7B baseline. Although R^2PRUNER performs additional redundancy-aware computations on top of underlying token importance criterion, it maintains competitive prefill latency with existing pruning methods.

Fig. 5b jointly visualizes the trade-off among task performance, inference throughput, and KV-cache memory. Although R^2PRUNER incurs slightly higher prefill latency due to progressive redundancy-aware pruning, it produces a smaller final visual-token set and consequently the smallest KV-cache footprint. The reduced decoding-side memory and computation contribute to higher throughput than several competing methods, yielding a favorable performance-efficiency trade-off.

Ablation study. Table 2 examines the contribution of each scoring component under an average visual-token budget of 64 tokens. Removing either E_i or the redundancy-aware signals D_i and A_i consistently degrades performance, confirming their complementary roles in token selection. Among the redundancy-aware signals, removing D_i causes only a modest degradation, likely because A_i aggregates neighboring D_j scores and therefore partially retains local-dissimilarity information. In contrast, removing A_i results in a larger performance drop, highlighting the value of neighborhood-level redundancy information. Meanwhile R_i alone yields the weakest results, suggesting that question-conditioned relevance is insufficient as a standalone pruning criterion. This may be partly due to positional bias in early-layer attention, which can make text-conditioned scores less reliable. Overall, the full combination of four signals achieves the strongest performance across all benchmarks.

5 CONCLUSION

We presented R^2PRUNER, a training-free vision token pruning method that complements common token-importance estimation signals with local redundancy information. Our key idea is that a visually important token is not necessarily worth preserving if nearby tokens already encode similar information. Conversely, tokens that are both informative and locally distinctive are more difficult to replace and therefore more valuable to retain. R^2PRUNER captures this intuition by characterizing token-level and neighborhood-level local distinctiveness with common token importance signals. Extensive experiments across multiple benchmarks and token budgets demonstrate that R^2PRUNER consistently preserves strong task performance under aggressive token reduction. Overall, our results highlight local redundancy as a complementary signal for selecting visual tokens whose information is not readily represented elsewhere.

AI USE STATEMENT

In this work, we used generative AI tools to create and edit software code, support language translation, create or modify scientific figures, draft parts of the manuscript, summarize existing literature, and search for relevant information. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401. IEEE, 2025.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Bin Chen, Yuxiang Cai, Yadan Luo, Yi Zhang, Jianwei Yin, and Zhi Chen. Spectral evolution-guided token pruning in multimodal large language models. *arXiv preprint arXiv:2606.24165*, 2026a.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3489–3499, 2026b.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zengjie Chen, Yuxiang Cai, Jingcai Guo, Taotao Cai, Jianwei Yin, and Zhi Chen. Accelerating multimodal large language models with prior-corrected token reduction. *arXiv preprint arXiv:2606.24156*, 2026c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. VisiPruner: Decoding discontinuous cross-modal dynamics for efficient multimodal LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18896–18913, 2025.

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

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6693–6702. IEEE, 2019.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296. IEEE, 2024a.

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

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

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2582–2591. IEEE, 2022.

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

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867. IEEE, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8309–8318. IEEE, 2019.

Zicong Tang, Ziyang Ma, Suqing Wang, Zuchao Li, Lefei Zhang, Hai Zhao, Yun Li, and Qianren Wang. CoViPAL: Layer-wise contextualized visual token pruning for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20701–20714, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952. IEEE, 2023.

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

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.