

Hi-Lo Prune: Look at What You’ll Lose before Pruning with Hierarchical Token Selection

Zixun Sun¹, Yubo Dong², Hehe Fan^{2,3†}, Yi Yang¹

[†]Corresponding author

¹The State Key Lab of Brain-Machine Intelligence, Zhejiang University

²CCAI, Zhejiang University ³State Key Laboratory of CAD&CG, Zhejiang University

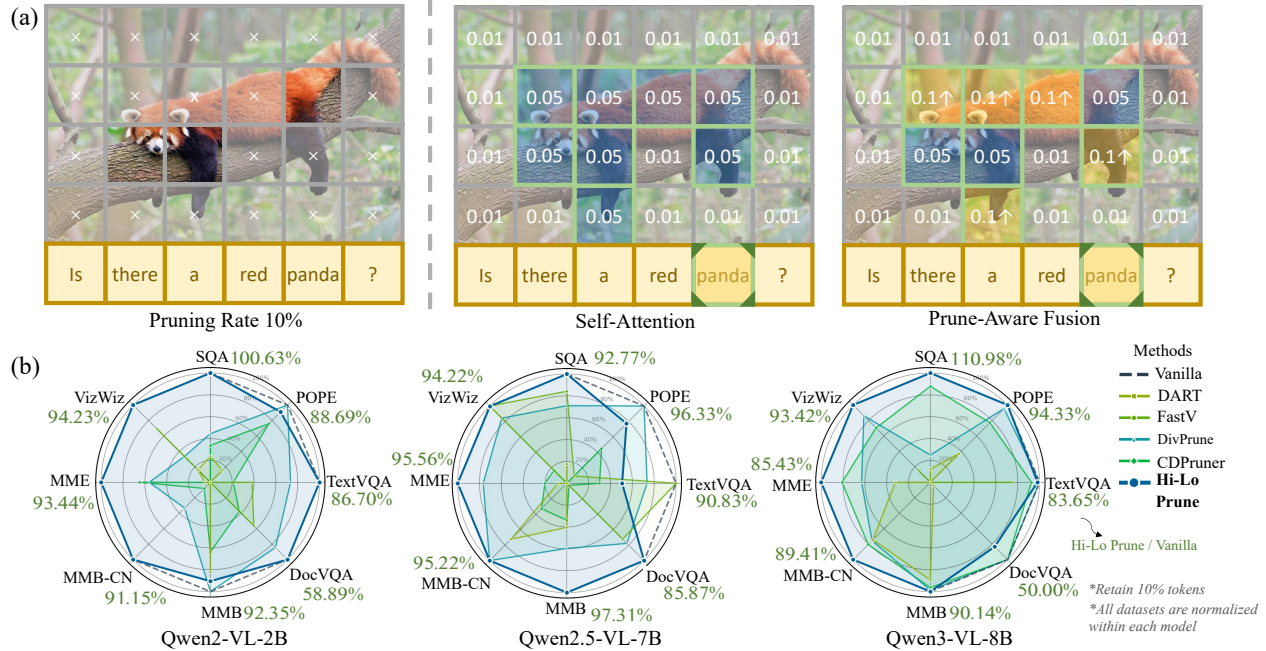


Figure 1. (a) Difference between self-attention and prune-aware fusion in solving the relevant elements (the blue patches), prune-aware fusion gained the relevant elements from tokens to drop (the yellow patches), for the query image token or text token (the patches with four corners green). Our Prune-Aware Fusion transfers information from pruning candidates to retained tokens before removal, enabling the model to “look at what you’ll lose.” (b) Performance comparison across eight datasets on Qwen series models. The radial axis shows min-max normalized scores across all methods, with the dashed line marking baseline (100%).

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in vision-language understanding, yet processing long visual token sequences remains computationally expensive. Existing approaches mitigate this cost by reducing image tokens, either by discarding them after the visual encoder or by pruning them in the early Transformer layers of the LLM. While these strategies improve efficiency, they inevitably discard informative visual content and risk degrading downstream reasoning performance. To address this challenge, we introduce *Hi-Lo Prune*, a training-free pruning strategy that is

built on a simple principle: **look at what you will lose before you remove it**. Instead of directly dropping tokens, *Hi-Lo Prune* first identifies which tokens are safe to prune through a coarse-to-fine selection process, and then encourages the model to absorb their information before pruning occurs. Specifically, the framework consists of three stages: (1) **Hierarchical Pruning Token Selection**. After visual encoding, we apply a coarse-to-fine process that identifies tokens to retain and selects a critical set of pruning candidates from redundant ones. (2) **Attention-Guided Candidate Token Merge**. Before removing selected tokens, an attention mechanism is applied to the early LLM layers, which explicitly transfers information from these

candidates to the retained tokens. (3) **Low-informative Candidate Token Removal.** At a designated Transformer layer, the pruned tokens are removed, reducing computation for all subsequent layers. This design enables aggressive early-layer pruning while preserving critical visual cues. Experiments on Qwen2-VL, Qwen2.5-VL, and Qwen3-VL demonstrate that Hi-Lo Prune consistently outperforms existing pruning methods across multiple benchmarks, achieving strong performance even under high pruning ratios without any fine-tuning. The code is available at https://github.com/sealost/Hi-Lo_Prune.

1. Introduction

Multimodal Large Language Models (MLLMs) have recently demonstrated impressive capabilities across diverse vision-language tasks [5, 12, 57], such as image captioning, visual question answering, and embodied reasoning [1, 22, 31, 45, 46]. Foundation models such as Qwen [5, 6, 48] and LLaVA [30] are increasingly adopted in real-world scenarios ranging from robotics to multimodal interaction systems [8, 47], driven by their improving complex reasoning capabilities [14, 39, 41]. However, their scalability is fundamentally constrained by the computational cost of processing dense visual inputs [21, 47, 48]. High-resolution images and long videos can generate tens of thousands of visual tokens [9, 19], leading to quadratic complexity in transformer-based models [18, 34, 49] and severely limiting deployment in resource-constrained environments [17].

Training-free visual token pruning methods have shown that reducing tokens by 50% to 90% can significantly accelerate inference [3, 10, 59] without additional fine-tuning, enabling seamless generalization across datasets and architectures. However, these approaches face a critical limitation: they perform token pruning in shallow layers where the model has not yet fully processed the visual content. Consequently, the model cannot adequately absorb information from dropped tokens, resulting in inevitable performance degradation, especially at high pruning ratios. While progressive multilayer pruning [4, 55] attempts to address this, it introduces computational overhead and still cannot fully prevent early-stage information loss.

In this work, we explore a simple yet overlooked principle: *look at what you will lose*. Instead of directly dropping tokens, we argue that the model should first examine and transfer information from the tokens to be pruned, ensuring that their essential content is preserved within the remaining token representations. To realize this, we propose Hi-Lo Prune, a training-free token pruning method that combines hierarchical token selection with a novel Prune-Aware Fusion (PA-Fusion) mechanism that transfers visual context from to-be-pruned tokens to retained ones. Specifically, Hi-Lo Prune operates in three stages: (1) Hierarchical

Pruning Token Selection: A coarse-to-fine process identifies tokens to retain and selects a significant set of pruning candidates from redundant ones; (2) Attention-Guided Candidate Token Merge: Before actual pruning, early layers transfer key visual information from these candidates to the retained tokens; (3) Low-informative Candidate Token Removal: The remaining low-informative tokens are then removed to accelerate subsequent inference. This design enables aggressive pruning in the early layers while minimizing performance degradation. Our main contributions are summarized as follows:

- We propose Hi-Lo Prune, a training-free pruning method guided by the principle of "look at what you will lose", which enables early-layer pruning by transferring visual context from pruned tokens to kept ones.
- We introduce Prune-Aware Fusion (PA-Fusion), a lightweight mechanism that allows retained tokens to absorb information from pruning candidates before removal, ensuring that critical visual content is preserved.
- We perform a coarse-to-fine token selection: coarse selection on pixel-level features identifies redundant regions, followed by fine-grained selection on visual token embeddings to determine the final set of tokens to retain, while maintaining computational efficiency.
- As shown in Figure 1, extensive experiments on Qwen2-VL, Qwen2.5-VL, and Qwen3-VL across 8 benchmarks demonstrate that Hi-Lo Prune outperforms existing methods, achieving strong performance even at high pruning ratios (up to 90%) without requiring fine-tuning.

2. Related works

Multimodal Large Language Models. Recent advances in MLLMs have shown the effectiveness of integrating large language models with visual encoders for unified multimodal understanding. Early examples such as Flamingo [2] and PaLI [11] demonstrated the potential of large-scale vision-language pretraining. Representative models such as Qwen2-VL, Qwen2.5-VL, and the latest Qwen3-VL [5, 6, 48] further bridge the gap between visual perception and language generation, supporting diverse tasks including visual question answering, image-conditioned dialogue, and scene reasoning. Meanwhile, lightweight architectures such as the Llava family [20, 27, 30] and Mini-GPT4 [62] show that simplicity can still yield competitive performance. However, a key limitation of current MLLMs is the large number of visual tokens generated from high-resolution images or video inputs. This leads to substantial computational and memory overhead during both training and inference [19, 21]. Thus, efficient visual token compression strategies are increasingly essential.

Visual Token Pruning. To reduce the computational overhead of visual token processing in MLLMs, recent studies explore visual token pruning strategies, which can

be training-based or training-free. Training-based methods [7, 23, 42, 44] fine-tune models or add modules to learn token importance. While effective, they require re-training with task-specific data, limiting their applicability in resource-constrained settings. Training-free methods [3, 10, 15, 26, 36, 40, 50, 51, 59, 60] directly operate on pre-trained MLLMs without fine-tuning and follow two main strategies: importance-based and diversity-based. Importance-based approaches, assessing the importance of visual tokens with text-visual attention such as FastV [10], SparseVLM [60], and PyramidDrop [51]. However, recent work demonstrates that attention-shift poses a fundamental challenge to such approaches, limiting their effectiveness [58]. Diversity-based methods, including DART [50], DivPrune [3], and CDPruner [59], retain complementary tokens using clustering or semantic distance. Despite their efficiency, existing methods make pruning decisions purely based on token features. Unlike SparseVLM [60], which fuses pruned tokens after removal, we let to-be-pruned tokens contribute effectively before removal, rather than fixing the loss afterward.

3. Method

In this section, we present the Hi-Lo Prune framework in a clear and structured manner. We first review existing visual token pruning methods in Section 3.1, then provide an overview of Hi-Lo Prune in Section 3.2. Section 3.3 details our hierarchical token selection from coarse pixel-level features to fine-grained embeddings, and Section 3.4 offers a detailed exploration of the Prune-Aware Fusion mechanism.

3.1. Visual Token Pruning in MLLMs

Modern MLLMs generally consist of a vision encoder f_v , a projector f_p , and a large language model f_ℓ . Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a text query, the vision encoder extracts visual features from the image. The projector then aligns these features with the language model’s embedding space, producing visual tokens $\mathbf{V} = f_p(f_v(\mathbf{I})) \in \mathbb{R}^{N \times D}$, where N is the number of visual tokens and D is the hidden dimension. Simultaneously, the text query is encoded into text tokens $\mathbf{T} \in \mathbb{R}^{L \times D}$, where L is the length of the text sequence. These visual and text tokens are concatenated as $[\mathbf{V}; \mathbf{T}]$ and fed into the LLM for further processing. The key challenge is that $N \gg L$ in most cases, especially for high-resolution images, leading to substantial computational costs. Visual token pruning addresses this by selecting a subset of important tokens $\hat{\mathbf{V}} \subset \mathbf{V}$ with $|\hat{\mathbf{V}}| = K$ (where $K < N$) while maintaining model performance. The pruning objective is formulated as follows:

$$\hat{\mathbf{V}}^* = \arg \min_{\hat{\mathbf{V}} \subseteq \mathbf{V}, |\hat{\mathbf{V}}|=K} \mathcal{D} \left(f_\ell([\hat{\mathbf{V}}; \mathbf{T}]), f_\ell([\mathbf{V}; \mathbf{T}]) \right), \quad (1)$$

where $\mathcal{D}$ measures the difference in model outputs before and after pruning. The pruning ratio is defined as $r = 1 - K/N$. Existing methods typically prune tokens in early layers based on attention scores [10, 51] or diversity [3]. Some approaches merge pruned tokens after pruning [60], but this often results in a larger number of retained tokens, which introduces additional computational costs due to the token merging process.

3.2. Hi-Lo Prune

We propose Hi-Lo Prune, a training-free visual token pruning framework designed to achieve efficient inference in MLLMs while maintaining performance. Hi-Lo Prune combines hierarchical token selection with prune-aware fusion, allowing for significant reductions in computational overhead while minimally impacting model accuracy. The complete pipeline is illustrated in Figure 2. Before the LLM decoding phase, given an input image $\mathbf{I}$ and a target pruning ratio r , we first apply the hierarchical selection (§3.3) to identify the final retained token set $\mathcal{K}$ of size $K = (1 - r) \cdot N$. In parallel, we generate an intermediate set $\mathcal{F}$, which contains approximately 30% of the total visual tokens and represents the most important tokens among those to be dropped. Next, we apply the prune-aware fusion mechanism (§3.4) in the layers before the target prune layer to transfer information from $\mathcal{F}$ to $\mathcal{K}$. At the prune layer, we remove all other tokens, retaining only $\mathcal{K}$ for subsequent layers. By combining these techniques, Hi-Lo Prune effectively reduces the number of visual tokens fed into the LLM, enabling faster and more efficient inference than recent methods.

3.3. Hierarchical Token Selection

Hierarchical feature representations have proven essential in modern vision systems [29, 33], where shallow layers capture fine-grained texture details while deep layers encode abstract semantic concepts [56]. Inspired by the functional multi-level features, we propose a hierarchical token selection strategy that combines texture information from pixel-level features with semantic-level features aligned to the LLM’s embedding space.

As shown in Algorithm 1, given an input image $\mathbf{I}$ of size $C \times H \times W$, we first partition it into non-overlapping patches of size $p \times p$, resulting in $N = \frac{H \times W}{p^2}$ patches. Each patch is flattened into a vector, forming pixel-level features $\mathbf{P} \in \mathbb{R}^{N \times D_p}$, where $D_p = Cp^2$. Subsequently, we extract semantic-level features $\mathbf{V} = f_p(f_v(\mathbf{I}))$ from the vision encoder and the projector. By using these features, we perform the selection process in two stages: (1) **Coarse selection** uses pixel-level features $\mathbf{P}$ to select αK candidate tokens (where relaxation factor $\alpha > 1$), filtering out globally redundant patches; (2) **Fine selection** refines the selected set $\mathcal{S}$ using semantic-level features $\mathbf{V}$ to select the

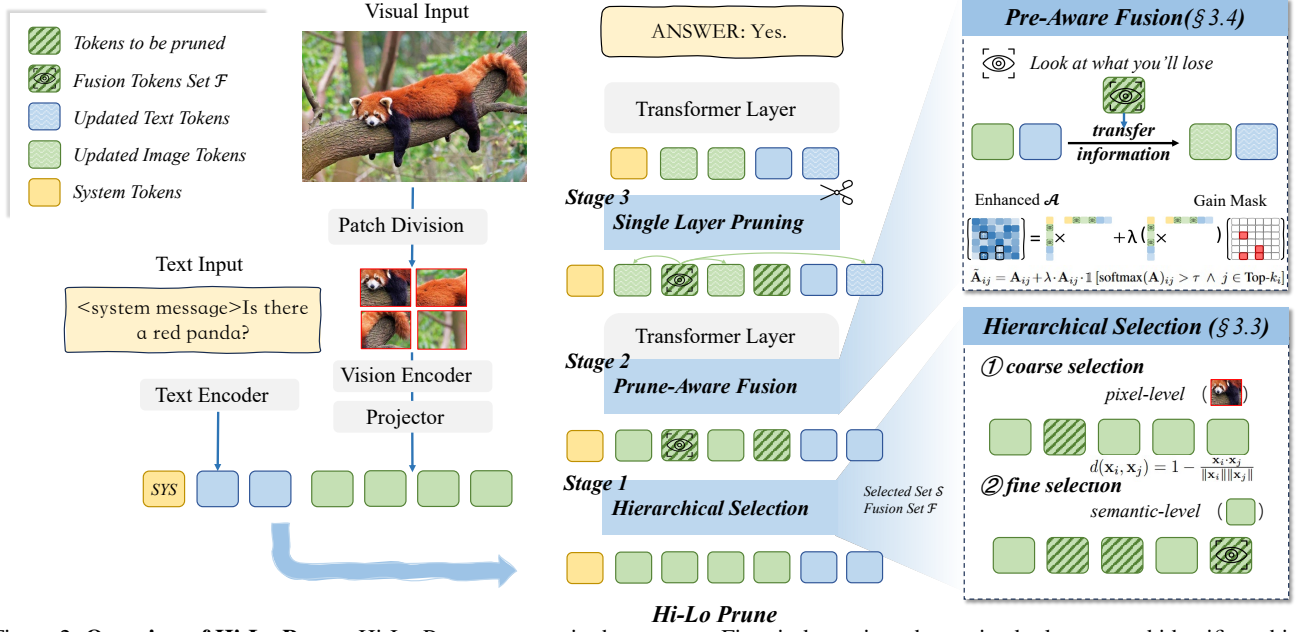


Figure 2. **Overview of Hi-Lo Prune.** Hi-Lo Prune operates in three stages. First, it determines the retained token set and identifies which tokens should be enhanced through coarse selection using texture features and fine selection using semantic features before LLM prefilling. Second, prune-aware fusion is applied to boost the attention scores of selected to-be-pruned tokens across related retained tokens before the pruning layer, thereby smoothly reducing visual information loss. Third, single-layer pruning is performed to retain only the selected tokens. This design enables efficient inference while preserving model performance.

final K tokens. After the selection process, we obtain both the final set of retained tokens and a subset of pruned tokens $\mathcal{F}$. Specifically, $\mathcal{F}$ consists of tokens that are selected in the coarse stage but discarded in the fine stage, accounting for approximately 30% of the original tokens. These pruned tokens will be leveraged for information transfer in the subsequent fusion stage. Both selection steps use the same greedy algorithm, which maximizes diversity by iteratively selecting tokens that are farthest apart from already selected tokens. We use cosine similarity $d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$ to measure the distance between tokens.

3.4. Prune-Aware Fusion

To mitigate the information loss from early-layer pruning, we modify the attention mechanism in the first few layers to enable information transfer from to-be-pruned tokens to retained tokens before discarding.

Adaptive Fusion. Given the retained token set $\mathcal{K}$ and prune-aware fusion set $\mathcal{F}$ identified in §3.3, we selectively boost attention between tokens in $\mathcal{K}$ and those in $\mathcal{F}$, where the boosting strength is proportional to the original attention affinity. As illustrated in Figure 2, let $\mathbf{A} = \mathbf{QK}^T \in \mathbb{R}^{q \times k}$ denote the attention logits in layer ℓ , where q is the number of queries and k is the number of keys. For each position (i, j) where $i \notin \mathcal{K}$ (retained visual token and text token) and $j \in \mathcal{F}$ (to-be-pruned token), we compute:

$$\tilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij} + \lambda \cdot \mathbf{A}_{ij} \cdot \mathbb{1}[\text{softmax}(\mathbf{A})_{ij} > \tau \wedge j \in \text{Top-}k_i], \quad (2)$$

where λ is the fusion strength, τ is an adaptive threshold, $\text{Top-}k_i$ denotes the top- k keys for query i , and $\mathbb{1}[\cdot]$ is the indicator function that equals 1 when the condition is true and 0 otherwise. The softmax operation is subsequently applied to $\tilde{\mathbf{A}}$ within the early fusion layers. This design ensures that to-be-pruned tokens with high attention affinity to retained tokens have their connections significantly strengthened, while irrelevant token pairs remain unaffected to preserve the original pattern.

4. Experiment

In this section, we provide an overview of the experiments evaluating Hi-Lo Prune. §4.1 details the datasets, models, and metrics. §4.2 presents key insights, followed by a comparison with existing methods in §4.3. The efficiency of our approach is analyzed in §4.4, and §4.5 examines the contribution of the main components.

4.1. Experimental Settings

Baselines and Models We compare Hi-Lo Prune with two sets of baselines. The first includes single-layer pruning methods—FastV [10], DART [50], CDPruner [59], and DivPrune [3]—which, like Hi-Lo Prune, are plug-and-play and require no fine-tuning. For multi-layer pruning, we compare with PyramidDrop [51] and SparseVLM [60], using fixed pruning layers and ratios for fairness. Additionally, Our primary comparison is based on the Qwen series, including Qwen2-VL-2B [48], Qwen2.5-VL-7B [6], and

Algorithm 1: Hierarchical Token Selection

Input: Image $\mathbf{I}$, target token count K , relaxation factor α

Output: Selected indices $\mathcal{S}$, fusion indices $\mathcal{F}$

```

1 Extract pixel-level features  $\mathbf{P} \in \mathbb{R}^{N \times D_p}$ ;
2 Extract semantic-level features  $\mathbf{V} \in \mathbb{R}^{N \times D}$ ;
3 Initialize  $\mathcal{S} = \{1, \dots, N\}$ ; // All token indices
4 for  $(\mathbf{F}, M) \in [(\mathbf{P}, \alpha K), (\mathbf{V}_S, K)]$  do
5   // Select  $M$  tokens from features  $\mathbf{F}$ 
6   if  $M = \alpha K$  then
7      $\mathcal{S}_{\text{coarse}} \leftarrow \mathcal{S}$ 
8   ; // Store coarse selection
9   Initialize  $\mathcal{T} = \emptyset$  and candidate pool  $\mathcal{R} = \mathcal{S}$ ;
10  Compute pairwise distances  $d(\mathbf{F}_i, \mathbf{F}_j)$  for all
11   $i, j \in \mathcal{R}$ ;
12  // Select first token
13  for  $i \in \mathcal{R}$  do
14     $d_{\min}^{(i)} = \min_{j \in \mathcal{R}, j \neq i} d(\mathbf{F}_i, \mathbf{F}_j)$ ;
15     $k = \arg \max_{i \in \mathcal{R}} d_{\min}^{(i)}$ ;
16    Move  $k$  from  $\mathcal{R}$  to  $\mathcal{T}$ ;
17  // Iteratively select subsequent tokens
18  while  $|\mathcal{T}| < M$  do
19    for  $i \in \mathcal{R}$  do
20       $d_{\min}^{(i)} = \min_{j \in \mathcal{T}, j \neq i} d(\mathbf{F}_i, \mathbf{F}_j)$ ;
21       $k = \arg \max_{i \in \mathcal{R}} d_{\min}^{(i)}$ ;
22      Move  $k$  from  $\mathcal{R}$  to  $\mathcal{T}$ ;
23   $\mathcal{S} \leftarrow \mathcal{T}$ ; // Update selected set
24   $\mathcal{F} \leftarrow \mathcal{S}_{\text{coarse}} \setminus \mathcal{S}$ ; // Pruned tokens for fusion
25 return  $\mathcal{S}, \mathcal{F}$ ;

```

Qwen3-VL-8B [54], as they represent state-of-the-art multimodal models widely used in recent research. Additionally, we also conduct experiments on the Llava 1.5-7B [30] model, demonstrating the generalizability and robustness of our approach across different architectures.

Datasets and Metrics We evaluate Hi-Lo Prune on nine benchmark datasets covering diverse visual reasoning tasks. These include VQA benchmarks (SQA [35], ChartQA [37], VizWiz [16]), high-resolution visual inputs (DocVQA [38], TextVQA [43]), visual illusions (POPE [25]), and higher-level reasoning tasks (MME [61], MMB [32], MMB-CN [32]). Additionally, we measure total and pre-filling times on the same device to assess the practical feasibility of our approach. Following the methodology of FastV [10], we also compute inference FLOPs to compare the computational costs of different methods. This dual evaluation of accuracy and efficiency provides a thorough analysis of Hi-Lo Prune’s effectiveness and its practical applicability.

Implementation Details In our experiments, pruning is performed at fixed layers. For single-layer pruning, we prune at the second layer, while for multi-layer pruning, we prune at the 4th, 8th, and 16th layers. To evaluate pruning performance at higher pruning ratios, we set pruning rates to 10% and 20%. Additionally, we apply Prune-Aware Fu-

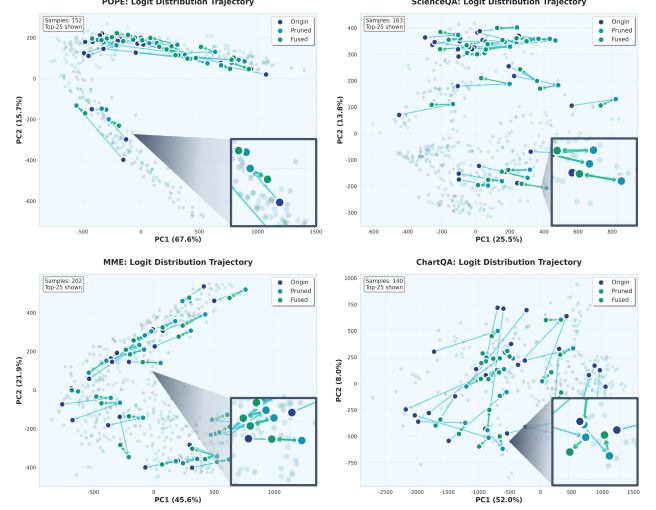


Figure 3. Logit distribution trajectories in PCA-projected 2D space. Cyan arrows show deviation caused by pruning (Origin $\rightarrow$ Pruned), while green arrows show recovery through prune-aware fusion (Pruned $\rightarrow$ Fused). Top-25 samples with largest improvements are shown for clarity.

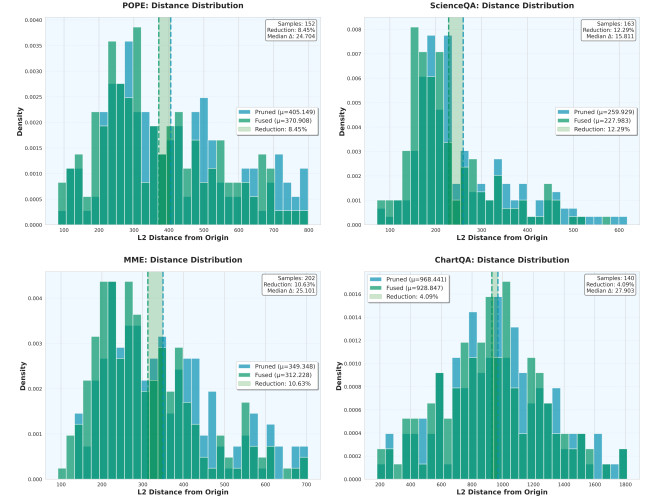


Figure 4. Distribution of L2 distances from origin logits. Prune-aware fusion (green) consistently reduces the distance compared to pruning alone (cyan), indicating better preservation of original prediction patterns.

sion (PA-Fusion) for information transfer in the first two pruned layers. Notably, our method supports FlashAttention [13] integration. In early layers, the fusion process, which operates on the fusion token subset $\mathcal{F}$, can be computed in parallel with the standard self-attention accelerated via FlashAttention, further improving overall efficiency.

4.2. Insights

To understand how prune-aware fusion affects model predictions, we analyze the logit distributions across four

Setting	Methods	SQA	POPE	TextVQA	DocVQA	MMB ^{Val}	MMB-CN ^{Val}	MME	VizWiz ^{Val}	Avg.
Retain 20%	Vanilla Qwen2-VL-2B	65.34%	88.47%	74.80%	90.00%	78.52%	78.61%	1873.55	39.01%	100.00%
	FastV(eccv24)	62.62%	80.41%	65.39%	57.00%	73.71%	71.05%	1687.60	37.75%	88.57%
	DART(emnlp25)	63.11%	80.06%	62.49%	36.00%	72.94%	70.70%	1693.65	31.62%	83.11%
	DivPrune(cvpr25)	63.46%	84.63%	68.09%	67.00%	74.40%	72.77%	1741.55	36.11%	91.39%
	CDPruner(nips25)	62.72%	82.60%	61.33%	48.00%	72.59%	71.13%	1705.93	33.64%	85.61%
	Hi-Lo Prune	66.19%	83.70%	69.68%	69.00%	76.46%	73.80%	1878.72	36.85%	93.97%
Retain 10%	FastV(eccv24)	59.30%	70.21%	56.63%	39.00%	64.35%	66.38%	1504.51	34.87%	78.16%
	DART(emnlp25)	60.83%	71.61%	51.53%	21.00%	71.74%	66.24%	1517.31	31.62%	75.49%
	DivPrune(cvpr25)	62.17%	79.27%	61.32%	48.00%	73.37%	68.04%	1642.32	32.69%	83.94%
	CDPruner(nips25)	61.48%	77.08%	54.33%	33.00%	70.10%	66.67%	1664.26	30.64%	79.00%
	Hi-Lo Prune	65.75%	78.46%	64.85%	53.00%	72.51%	71.65%	1750.61	36.76%	88.26%
Retain 20%	Vanilla Qwen2.5-VL-7B	84.38%	88.38%	68.92%	92.00%	86.25%	86.25%	2245.20	40.80%	100.00%
	FastV(eccv24)	77.74%	82.58%	67.02%	69.00%	78.61%	77.78%	1863.23	38.46%	89.55%
	DART(emnlp25)	75.46%	82.22%	58.07%	43.00%	80.76%	80.93%	1886.26	33.04%	83.24%
	DivPrune(cvpr25)	77.29%	86.00%	64.65%	71.00%	81.78%	82.13%	2079.52	37.61%	91.84%
	CDPruner(nips25)	74.86%	83.92%	60.75%	44.00%	80.41%	79.21%	1919.88	33.94%	84.18%
	Hi-Lo Prune	78.28%	85.14%	62.60%	79.00%	83.93%	82.13%	2145.62	38.44%	95.04%
Retain 10%	FastV(eccv24)	72.73%	74.46%	62.47%	44.00%	64.26%	62.68%	1517.59	35.26%	82.61%
	DART(emnlp25)	72.88%	74.82%	49.10%	23.00%	76.46%	75.32%	1625.71	27.62%	79.76%
	DivPrune(cvpr25)	75.36%	83.22%	58.20%	53.00%	79.98%	78.26%	1906.50	35.14%	92.42%
	CDPruner(nips25)	72.38%	79.52%	53.99%	34.00%	76.20%	74.89%	1681.66	30.52%	83.87%
	Hi-Lo Prune	77.40%	83.72%	60.65%	56.00%	81.62%	79.30%	1948.78	37.08%	93.51%
Retain 20%	Vanilla Qwen3-VL-8B	68.22%	88.53%	76.85%	96.00%	89.78%	89.26%	2390.83	48.01%	100.00%
	FastV(eccv24)	62.27%	81.92%	63.90%	54.00%	81.53%	80.83%	1880.90	44.04%	84.37%
	DART(emnlp25)	65.49%	83.21%	62.85%	46.00%	84.19%	83.25%	2017.75	41.62%	84.73%
	DivPrune(cvpr25)	65.29%	86.80%	70.08%	72.00%	84.45%	82.82%	2087.74	45.92%	91.22%
	CDPruner(nips25)	74.76%	85.13%	67.86%	70.00%	84.71%	82.63%	2121.00	45.08%	92.06%
	Hi-Lo Prune	71.29%	86.74%	70.60%	68.00%	85.82%	84.28%	2156.45	45.98%	92.64%
Retain 10%	FastV(eccv24)	61.82%	74.02%	52.69%	28.00%	73.54%	72.77%	1649.08	38.27%	73.01%
	DART(emnlp25)	63.51%	77.53%	13.70%	29.00%	80.15%	78.01%	1775.15	36.99%	69.58%
	DivPrune(cvpr25)	65.29%	83.08%	65.21%	52.00%	80.84%	78.09%	1896.92	43.76%	84.57%
	CDPruner(nips25)	74.02%	81.34%	61.41%	52.00%	80.66%	78.44%	1967.56	42.52%	85.38%
	Hi-Lo Prune	75.71%	83.51%	64.29%	48.00%	80.93%	79.81%	2042.58	44.85%	87.17%

Table 1. **Comparison of Hi-Lo Prune with baselines across different Qwen-VL models.** Results are reported on eight benchmarks under 80% and 90% token pruning. **Hi-Lo Prune** consistently achieves the best average performance across all models, showing strong robustness under aggressive pruning.

benchmarks (POPE, SQA, MME, ChartQA). It aims to examine how prune-aware fusion mitigates the distribution shift caused by aggressive pruning, helping the model’s predictions stay closer to the original distribution seen with full token sets, thereby preserving prediction consistency.

Benchmark	L2 Distance ↓		Cosine Sim. ↑		L2 Reduction (%)
	Pruned	Fused	Pruned	Fused	
POPE	405.15	370.91	0.9812	0.9830	8.45
SQA	259.93	227.98	0.9818	0.9858	12.29
MME	349.35	312.23	0.9826	0.9849	10.63
ChartQA	968.44	928.85	0.8121	0.8259	4.09

Table 2. **Logit distribution statistics across benchmarks.** All metrics compare against the original (full tokens).

For each sample, we collect the final-layer logits from three settings: (1) **Origin**: full visual tokens without pruning; (2) **Pruned**: the result of hierarchical token selection with a 90% pruning ratio, where no prune-aware fusion is applied; (3) **Fused**: the result of hierarchical token selection with prune-aware fusion. Since logits are high-dimensional vectors (vocabulary size), we apply PCA to project them into 2D space for visualization, using the same transformation for all three settings to ensure comparable projections.

Figure 3 visualizes the logit distribution trajectories for selected samples. Each arrow represents one sample’s movement: cyan arrows show *Origin* → *Pruned*, and green arrows show *Pruned* → *Fused*. Our observations are as follows: (1) Aggressive pruning (90%) pushes the logit distri-

butions away from the origin, causing prediction deviation; (2) prune-aware fusion consistently recovers lost information, with green arrows pointing back toward the origin; (3) This recovery pattern appears across all benchmarks with different visual characteristics, suggesting that prune-aware fusion effectively mitigates the negative effects of pruning.

Figure 4 quantifies the deviation using L2 distance between logits. Table 2 shows that prune-aware fusion reduces L2 distance by **4.09%–12.29%** and improves cosine similarity by **0.18%–1.38%** across benchmarks. ChartQA shows smaller improvements, likely due to its complex visual reasoning that requires more visual content. These results show that prune-aware fusion effectively mitigates information loss from aggressive pruning, bringing model predictions closer to those from the full-token baseline.

4.3. Main Results

We mainly conduct a comprehensive evaluation on the Qwen series models. As presented in Table 1, we conduct a thorough evaluation of the single-layer pruning task on Qwen2-VL-2B, Qwen2.5-VL-7B, and Qwen3-VL-8B with 80% and 90% visual token pruning. The results show that our method consistently outperforms all baselines across the Qwen models. Notably, with 90% pruning, our approach significantly exceeds the performance of all baselines, demonstrating its ability to preserve critical information even under aggressive pruning. In summary, these results highlight the effectiveness and scalability of our approach in single-layer pruning across various pruning scenarios.

Methods	SQA	ChartQA	MME	MMB-CN ^{Val}	Avg.
Vanilla	65.34%	76.24%	1873.55	78.61%	100.00%
PDrop	64.75%	54.52%	1829.19	76.29%	91.32%
20% SparseVLM	65.49%	55.88%	1857.11	75.95%	92.32%
Hi-Lo Prune	65.69%	54.97%	1879.76	76.33%	92.52%
PDrop	63.76%	30.80%	1779.73	75.00%	82.10%
10% SparseVLM	64.15%	34.84%	1740.44	75.17%	83.41%
Hi-Lo Prune	65.64%	35.96%	1819.42	74.94%	85.02%

Table 3. **Evaluation of Hi-Lo Prune against multi-layer pruning baselines on Qwen2-VL-2B.** Across four benchmarks and two pruning ratios, **Hi-Lo Prune** delivers the best average performance, highlighting its advantage in preserving task performance under strong compression.

Fair Comparison with Multi-layer Pruning Methods.

To further validate our approach, we extend Hi-Lo Prune to multi-layer pruning and compare it with previous multi-layer pruning methods, PyramidDrop [51] and SparseVLM [60]. As shown in Table 3, with pruning ratios of 80% and 90%, our method consistently outperforms both baselines. Additionally, with 90% pruning, our method achieves a score of 1819.42 on the MME dataset, demonstrating

its strong advantage in cognitive and perceptual reasoning tasks. Unlike SparseVLM, our hierarchical pruning approach does not require token recovery but still effectively collects and utilizes information from pruned tokens, highlighting the flexibility and efficiency of our method.

Methods	SQA	ChartQA	MME	MMB-CN ^{Val}	Avg.
Vanilla	62.37%	18.20%	1781.94	67.53%	100.00%
FastV	62.62%	15.68%	1637.89	63.75%	93.22%
20% DART	62.52%	15.56%	1681.74	63.40%	93.50%
DivPrune	62.92%	16.08%	1687.09	63.11%	94.34%
CDPruner	62.62%	16.56%	1678.61	63.20%	94.80%
Hi-Lo Prune	62.92%	16.72%	1733.04	63.97%	96.18%
FastV	62.62%	14.16%	1479.73	60.14%	87.57%
DART	62.12%	14.88%	1634.80	60.57%	90.70%
10% DivPrune	62.07%	15.84%	1684.19	59.24%	92.20%
CDPruner	62.22%	15.76%	1599.94	60.36%	91.38%
Hi-Lo Prune	63.01%	16.00%	1677.47	62.42%	93.88%

Table 4. **Pruning results of Hi-Lo Prune and baselines on Llava 1.5-7B.** Performance is evaluated on four benchmarks with 80% and 90% token pruning. **Hi-Lo Prune** maintains strong accuracy and achieves the highest overall average under both pruning levels.

Results on Llava architecture. To further assess the effectiveness of our method, we conduct additional evaluations on the LLaVA 1.5-7B model. As shown in Table 4, the results on LLaVA mirror those on the Qwen models, with our approach demonstrating a clear advantage under both 80% and 90% pruning ratios. Specifically, with 90% pruning, our method achieves an average score of 93.88%, surpassing the second-best method by 2.5%. This further demonstrates the robustness and flexibility of our approach.

Overall, these results demonstrate the significant advantage of our method on the different models, highlighting its superior performance across various tasks and its adaptability to advanced models.

4.4. Efficiency Analysis

To evaluate efficiency, we compare Hi-Lo Prune with state-of-the-art pruning methods on Qwen3-VL-8B using the MME benchmark under consistent output lengths. The experiment is performed on a single NVIDIA L40S GPU. As shown in Table 5, by reducing 90% visual tokens, Hi-Lo Prune achieves a reduction of 85.35% FLOPs (from 2.576 TFLOPs to 0.377 TFLOPs) and reduces total inference time by 43.75% with a reduction in prefill time of 49.79%. Our method, which integrates prune-aware fusion mechanisms with optimized Flash Attention support, significantly outperforms FastV and DART in both efficiency and accuracy. Compared to DivPrune and CDPruner, Hi-Lo Prune achieves comparable efficiency while maintaining superior performance with an MME score of 2042.58, demonstrating the trade-off between accuracy and efficiency with only

Q: Is this movie originated from the country or region of west germany? Please answer yes or no.

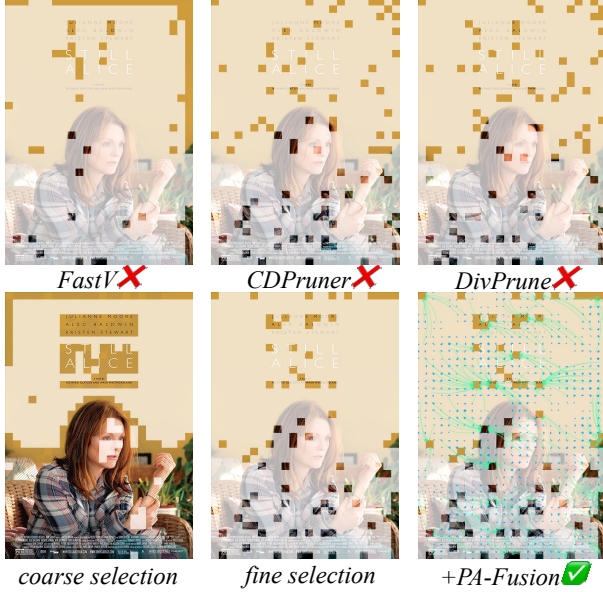


Figure 5. **Qualitative comparison of pruning with and without PA-Fusion.** PA-Fusion preserves critical regions that baseline methods may discard, leading to more accurate and reliable answers under aggressive pruning.

14.6% degradation from the vanilla model.

Methods	Reduced Total Time	Reduced Prefilling Time	Infer Flops (T)	MME ↑ Score
Vanilla Qwen3-VL-8B	0%	0%	2.576	2390.83
+ FastV	9.64%	7.14%	0.377	1649.08
+ DART	11.30%	3.27%	0.377	1775.15
+ DivPrune	45.34%	53.13%	0.377	1896.92
+ CDPPruner	41.52%	48.82%	0.377	1967.56
+ Hi-Lo Prune	43.75%	49.79%	0.377	2042.58

Table 5. **Efficiency comparison on the MME benchmark.** All reductions are measured relative to the vanilla Qwen3-VL-8B model. **Hi-Lo Prune** achieves a substantial reduction in total and prefilling time while maintaining the highest MME score among pruning methods.

4.5. Ablation Study

Effectiveness of Prune-Aware Fusion. To demonstrate the effectiveness of our PA-Fusion, we conduct extensive experiments by integrating it into existing training-free, text-agnostic pruning methods, including CDPPruner and DivPrune. We evaluated these methods with and without PA-Fusion across the Qwen models on multiple benchmarks.

As shown in Table 6, PA-Fusion consistently improves performance across nearly all these methods. Notably, with 90% of visual tokens pruned, CDPPruner achieves a 0.45% improvement on ScienceQA, DivPrune gains 67.96 points

on MME, and our Hi-Lo Prune improves by 1.28% on ChartQA. These consistent improvements across the Qwen series models validate the effectiveness and generalizability of our Prune-Aware Fusion mechanism.

Qualitative Analysis. To further illustrate how PA-Fusion preserves critical information before pruning, Figure 5 presents qualitative comparisons between the baseline methods and our approach with and without PA-Fusion. During stage 1 of Hi-Lo pruning, the retained boxes represent candidate tokens for pixel-level pruning. Boxes marked with green corner triangles indicate tokens that transfer their semantic content to retained tokens through the attention mechanism, and the green arrows visualize selected visual token transfer trajectories. As shown in the first row, baseline methods sometimes discard critical visual context relevant to the questions, leading to incorrect answers and hallucinations. In contrast, our PA-Fusion mechanism, guided by the principle of *look at what you’ll lose*, successfully preserves these essential regions and produces correct answers. These visualizations demonstrate that PA-Fusion enables more informed pruning decisions by considering what information would be lost, thereby maintaining semantic integrity while achieving aggressive compression.

	Methods	SQA	ChartQA	MME	MMB-CN ^{Val}
20%	CDPruner	74.76%	54.28%	2121.00	82.63%
	+PA-Fusion	74.86%	54.64%	2145.10	83.42%
	DivPrune	65.29%	55.20%	2087.74	82.82%
	+PA-Fusion	69.06%	55.72%	2130.75	83.08%
	Hi-Lo Prune	70.00%	52.80%	2122.28	84.19%
	+PA-Fusion	71.29%	52.88%	2156.45	84.28%
10%	CDPruner	74.02%	37.52%	1967.56	78.44%
	+PA-Fusion	74.47%	37.68%	1959.34	78.69%
	DivPrune	65.29%	35.76%	1896.92	78.09%
	+PA-Fusion	73.13%	43.40%	1964.84	78.52%
	Hi-Lo Prune	73.87%	42.64%	2042.10	79.04%
	+PA-Fusion	75.71%	43.92%	2042.58	79.81%

Table 6. **Impact of PA-Fusion on pruning methods for Qwen3-VL-8B.** Adding PA-Fusion consistently improves performance for all pruning methods, highlighting its effectiveness in mitigating performance loss under aggressive token reduction.

5. Conclusion

We introduce Hi-Lo Prune, a training-free visual token pruning method that preserves critical information before removal. By hierarchically selecting tokens and explicitly transferring their information to retained ones, Hi-Lo Prune enables aggressive pruning without significant performance loss. Experiments on Qwen2.5-VL-7B model demonstrate that our method retains over 93.51% of baseline accuracy when keeping only 10% of visual tokens, outperforming existing pruning methods across multiple benchmarks.

Acknowledgments

This work was supported by the National Key R&D Program of China (2023YFC3305600) and the National Natural Science Foundation of China (92570101).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [2] 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. 2
- [3] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025. 2, 3, 4, 1
- [4] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1773–1781, 2025. 2
- [5] 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. 2
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4
- [7] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024. 3
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2
- [9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024. 2
- [10] 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*, pages 19–35. Springer, 2024. 2, 3, 4, 5, 1
- [11] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2
- [12] 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. 2
- [13] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. 5
- [14] Yubo Dong and Hehe Fan. Enhancing large language models through structured reasoning. *arXiv preprint arXiv:2506.20241*, 2025. 2
- [15] Yichen Guo, Hanze Li, Zonghao Zhang, Jinhao You, Kai Tang, and Xiande Huang. Star: Stage-wise attention-guided token reduction for efficient large vision-language models inference. *arXiv preprint arXiv:2505.12359*, 2025. 3
- [16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 5, 1
- [17] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 2
- [18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2
- [19] Bo Li, Peiyuan Zhang, Jingkan Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multimodality model. *arXiv preprint arXiv:2311.04219*, 2023. 2
- [20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2
- [21] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 2
- [22] 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*, pages 19730–19742. PMLR, 2023. 2
- [23] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, pages 1–19, 2025. 3
- [24] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description.

- In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 1
- [25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5, 1
- [26] Yangfu Li, Hongjian Zhan, Tianyi Chen, Qi Liu, and Yue Lu. Why 1+ 1_i 1 in visual token pruning: Beyond naive integration via multi-objective balanced covering. *arXiv preprint arXiv:2505.10118*, 2025. 3
- [27] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2
- [28] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024. 1
- [29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 5
- [31] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29139–29149, 2025. 2
- [32] 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*, pages 216–233. Springer, 2024. 5, 1
- [33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3
- [34] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2
- [35] 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. 5, 1
- [36] Bozhi Luan, Wengang Zhou, Hao Feng, Zhe Wang, Xiaosong Li, and Houqiang Li. Multi-cue adaptive visual token pruning for large vision-language models. *arXiv preprint arXiv:2503.08019*, 2025. 3
- [37] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 5, 1
- [38] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 5, 1
- [39] Chunyang Meng, Shijie Song, Haogang Tong, Maolin Pan, and Yang Yu. Deepscaler: Holistic autoscaling for microservices based on spatiotemporal gnn with adaptive graph learning. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 53–65. IEEE, 2023. 2
- [40] Yu Meng, Kaiyuan Li, Chenran Huang, Chen Gao, Xinlei Chen, Yong Li, and Xiaoping Zhang. Plphp: Per-layer per-head vision token pruning for efficient large vision-language models. *arXiv preprint arXiv:2502.14504*, 2025. 3
- [41] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2
- [42] Zhenwei Shao, Mingyang Wang, Zhou Yu, Wenwen Pan, Yan Yang, Tao Wei, Hongyuan Zhang, Ning Mao, Wei Chen, and Jun Yu. Growing a twig to accelerate large vision-language models. *arXiv preprint arXiv:2503.14075*, 2025. 3
- [43] 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*, pages 8317–8326, 2019. 5, 1
- [44] Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*, 2025. 3
- [45] Songbai Tan, Xuerui Qiu, Yao Shu, Gang Xu, Linrui Xu, Xiangyu Xu, Huiping Zhuang, Ming Li, and Fei Yu. Wmarkgpt: Watermarked image understanding via multi-modal large language models. In *Forty-second International Conference on Machine Learning*, 2025. 2
- [46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2
- [47] InternLM Team. Internlm: A multilingual language model with progressively enhanced capabilities, 2023. 2
- [48] 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. 2, 4
- [49] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 2

- [50] Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. [3](#), [4](#), [1](#)
- [51] 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. [3](#), [4](#), [7](#)
- [52] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. [1](#)
- [53] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. [1](#)
- [54] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. [5](#)
- [55] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. [2](#)
- [56] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. [3](#)
- [57] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. [2](#)
- [58] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pages arXiv–2412, 2024. [3](#)
- [59] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025. [2](#), [3](#), [4](#), [1](#)
- [60] 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. [3](#), [4](#), [7](#)
- [61] Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 18, 2021. [5](#), [1](#)
- [62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)