Accelerating Multimodal Large Language Models with Prior-Corrected Token Reduction

Zengjie Chen[1,2], Yuxiang Cai*[1,2], Jingcai Guo[3], Taotao Cai[4], Jianwei Yin[1,2], and Zhi Chen*[4]

[1] School of Software Technology, Zhejiang University, Ningbo, China
[2] Zhejiang Key Laboratory of Digital-Intelligence Service Technology, China
[3] Hong Kong Polytechnic University, Hong Kong SAR
[4] The University of Southern Queensland, Toowoomba, QLD, Australia
uqzhichen@gmail.com, caiyuxiang@zju.edu.cn
https://github.com/CodeChildCZJ/PriorTR

Abstract. Visual token reduction has emerged as an effective strategy for accelerating Multimodal Large Language Models (MLLMs). Many existing methods prune tokens by ranking text–visual attention scores. However, we show that attention is often dominated by a model-induced prior: even without textual instruction, MLLMs tend to focus on certain task-agnostic regions. Consequently, the attention scores of instruction-conditioned tokens are suppressed, increasing the risk that these tokens are discarded during pruning. To address this issue, we propose Prior-Corrected Token Reduction (PriorTR), a training-free token reduction method that explicitly separates task-conditioned attention from the model-induced prior. PriorTR estimates the attention map of the prior, and contrasts it with the task-conditioned attention distribution to measure the additional usable information contributed by each visual token. Importantly, PriorTR computes both the model-induced prior and the task-conditioned posterior within a single forward pass by introducing a null token that serves as an instruction-agnostic probe in the attention block. This design avoids duplicated propagation. Extensive experiments across multiple multimodal benchmarks and MLLMs demonstrate that PriorTR consistently improves the trade-off between accuracy and efficiency over strong training-free baselines, particularly under aggressive token budgets.

Keywords: MLLMs · Token Reduction · Prior Correction

1 Introduction

Multimodal Large Language Models (MLLMs) [4,7,27,29] extend large language models (LLMs) [5,35] with visual perception, enabling instruction following over images for tasks such as visual question answering, reasoning, and interactive assistants. A common design converts an image into a long sequence of visual

* Corresponding Authors: Zhi Chen, Yuxiang Cai

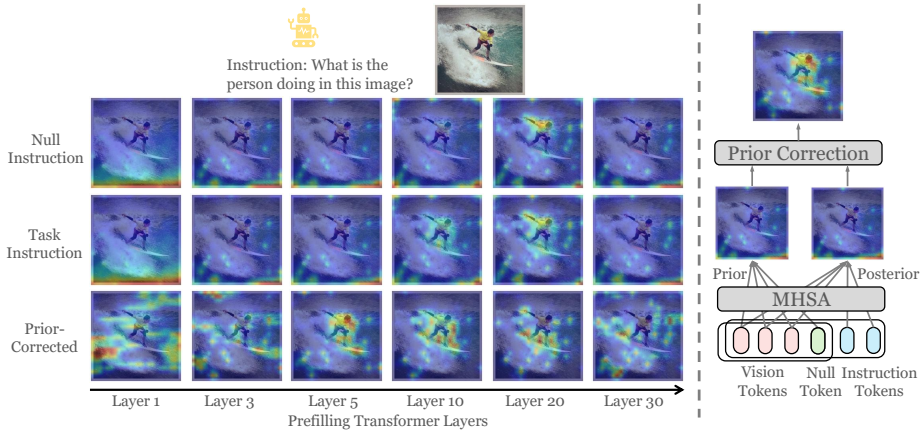

Fig. 1: Left: Visualization of text-to-visual attention maps across LLM decoding layers for a sample instruction ("*What is the person doing in this image?*"). (a) **Null instruction**: even without textual input, the model consistently attends to certain instruction-agnostic regions, revealing a strong model-induced prior. (b) **Task instruction**: attention distribution is partially influenced by the model-induced prior. (c) **Prior-corrected attention**: by correcting the prior, the resulting attention better highlights instruction-conditioned regions. **Right**: schematic illustration of our idea. A null token estimates the prior attention while the instruction produces the posterior attention within the same attention block; their contrast yields a prior-corrected signal used for visual token selection.

tokens that are fused with text tokens and processed by an LLM. While effective, this token-heavy interface makes inference expensive. The cost becomes especially pronounced for high-resolution images [23] and videos [26]. Consequently, reducing the number of visual tokens during inference has become a practical solution to accelerate MLLMs [47].

Recent methods [6, 36, 51, 66] observe that text–visual attention scores can serve as a training-free signal for visual token pruning. At an early layer, visual tokens are ranked according to an attention-derived importance score, and only the top-K tokens are retained for subsequent layers. This plug-and-play paradigm is attractive because it requires no additional training and can be readily applied across different model backbones. However, these methods implicitly rely on a key assumption: the magnitude of attention weights is a reliable proxy for instruction-conditioned semantic relevance.

In this paper, we show that this assumption may fail due to a *model-induced prior* in attention, as illustrated in Fig. 1. Even without any instruction, MLLMs tend to allocate high attention to certain task-agnostic regions. When pruning relies on absolute posterior attention scores, this prior can dominate the ranking. As a result, task-agnostic regions may be preserved, while instruction-conditioned evidence is suppressed. The problem becomes particularly severe under strict token budgets, where a small number of incorrect selections can

irreversibly remove the visual evidence required for correct reasoning. This observation motivates a central question: *how can we rank visual tokens according to instruction-conditioned semantics rather than model-induced prior?*

To address this challenge, we propose *PriorTR*, a training-free visual token reduction method that ranks tokens using token-level $\mathcal{V}$-usable information. As illustrated in Fig. 1, the key idea is to explicitly disentangle two factors that are entangled in raw attention: (i) a model-induced prior that reflects what the model attends to even without instruction, and (ii) the additional semantic evidence introduced by the instruction. By contrasting these two components, PriorTR scores tokens according to how much *usable* instruction-conditioned information they contribute, rather than their raw attention magnitude. This relative ranking criterion directly mitigates model-induced prior and provides a more reliable token selection strategy, particularly under aggressive token budgets.

PriorTR is implemented efficiently within a single forward pass. Instead of executing a separate prior stream, we exploit the causal attention structure of autoregressive decoders: the null token (e.g., the separator `\n`) placed immediately after the image tokens cannot attend to any instruction tokens under the causal mask, making its attention distribution over visual tokens a natural instruction-agnostic prior. At the designated pruning layer, we simultaneously obtain the task-conditioned posterior from the attention of instruction tokens to the visual sequence within the same attention matrix. We then apply our prior-corrected ranking rule and physically prune the visual sequence by retaining only the selected hidden states and KV-cache entries, allowing subsequent layers to operate on a truly shortened sequence. In this way, PriorTR achieves the effect of two forward passes at the cost of one, remains fully plug-and-play, requires no additional training, and provides real reductions in computation and memory during inference. Notably, our method is orthogonal to existing attention-based pruning approaches and can be seamlessly integrated with them. Extensive experiments on multiple MLLMs and multimodal benchmarks demonstrate that PriorTR consistently improves the trade-off between accuracy and efficiency compared with strong training-free baselines.

Our main contributions are:

- We identify and formalize a *model-induced prior* in attention-based token pruning for MLLMs, showing that absolute attention ranking can be dominated by instruction-agnostic attention patterns, particularly under aggressive token budgets.
- We propose PriorTR, a training-free token reduction method that ranks visual tokens using $\mathcal{V}$-usable information by explicitly separating instruction-induced semantics from the model-induced prior.
- We design an efficient single-pass implementation that estimates the prior using a null token and performs physical pruning of hidden states and KV cache, enabling real compute and memory savings during inference.
- Extensive experiments across multiple MLLMs and multimodal benchmarks demonstrate that PriorTR consistently improves the accuracy–efficiency trade-off and can be seamlessly integrated with existing token pruning methods.

2 Related Work

Multimodal Large Language Models (MLLMs) [4, 7, 27, 29] commonly convert images into hundreds or thousands of visual tokens that are then processed jointly with text, leading to substantial inference cost. To improve efficiency, existing efforts explore architectural changes (e.g., more compact visual encoders [49, 67] or fewer visual tokens during training [2, 43]), model compression [22, 44], parameter-efficient fine-tuning [37, 56, 59], and inference-time acceleration [6, 19, 66]. Reducing visual token redundancy has emerged as an effective strategy to accelerate multimodal large language models (MLLMs) without retraining. Training-free token reduction methods typically identify informative visual tokens during inference and discard redundant ones before they are processed by later Transformer layers. Existing approaches can be broadly categorized into four groups based on the criteria used to estimate token importance.

Similarity-based methods remove tokens that are highly similar to others in the feature space, assuming that redundant tokens contribute little additional information. Representative methods include DART [48], which prunes tokens based on similarity to surrounding patches, and subsequent variants that refine similarity-based selection using feature clustering or local redundancy estimation [3, 14, 18, 39, 55, 60]. While similarity-based pruning effectively reduces spatial redundancy, these methods rely on instantaneous feature similarity and often ignore how token representations evolve across layers.

Attention-based methods use attention scores as an importance signal. FastV [6] is a representative plug-and-play approach that ranks image tokens using attention weights at a selected Transformer layer and prunes tokens with the lowest scores. Subsequent work further explores attention-based token selection in multimodal Transformers [11, 12, 25, 30, 41, 42, 45, 50, 58, 64, 68]. These approaches are simple to implement and often achieve significant speedup. However, attention magnitude may not always reflect true semantic contribution, and it can be biased toward structurally dominant tokens.

Diversity-based methods aim to preserve tokens that maximize coverage of the feature space. Methods such as EntropyPrune [46] and related approaches [63, 65] select tokens that maintain a diverse set of visual representations. Although diversity criteria can mitigate redundancy more effectively than simple similarity measures, they still rely on single-layer statistics and may overlook the dynamic evolution of token representations across network depth.

Hybrid methods. combine multiple importance signals to improve robustness. For example, VisPruner [65] integrates attention-based importance with diversity constraints, while ToDre [21] jointly considers attention and redundancy to select representative tokens. Hybrid strategies often improve pruning quality but introduce additional heuristics and still primarily rely on instantaneous token statistics.

Token merging methods. Beyond pruning, **token merging methods** as another line of work reduce token count through token merging, where simi-

lar tokens are aggregated instead of discarded. Representative approaches include methods such as LightVLM [13], CrossGet [38], VisionZip [57], and SparseVLM [66]. Token merging can preserve more information than pruning, but it introduces additional merging operations and may alter the semantic structure of token representations. In contrast to these approaches, our method revisits the fundamental assumption underlying attention-based token ranking. We observe that attention scores are often influenced by saliency prior rather than instruction-specific relevance. PriorTR explicitly corrects this saliency prior and ranks tokens based on their additional usable information for the given instruction.

3 Method

3.1 Preliminaries

We consider a Multimodal Large Language Model (MLLM), in which the input comprises a visual token sequence $\mathbf{V} = \{v_1, \ldots, v_N\}$ and an instruction token sequence $\mathbf{X}$. During the LLM autoregressive decoding process, the full visual token sequence $\mathbf{V}$ is involved in the computation at every token generation step. We formulate visual token pruning as a constrained feature selection problem. Given a strict budget constraint $|\hat{\mathbf{V}}| \leq K$, the objective is to identify an optimal subset $\hat{\mathbf{V}}^*$ that maximizes the conditional likelihood of the target response $\mathbf{Y}$:

$$\hat{\mathbf{V}}^* = \underset{\hat{\mathbf{V}} \subset \mathbf{V}}{\arg\max}\, P(\mathbf{Y} \mid \hat{\mathbf{V}}, \mathbf{X}). \tag{1}$$

To approximate the objective above, existing works typically adopt a scoring and selection paradigm that uses the attention weights from the causal self-attention from the transformer as indicators of importance. Consequently, existing works interpret this score as the model's approximate posterior belief conditioned on the instruction:

$$\mathcal{S}_{\text{base}}(v) \approx P_\theta(v \mid \mathbf{X}). \tag{2}$$

We abuse notation and write $P_\theta(v_i \mid \mathbf{X})$ to denote the normalized attention mass assigned to token index i. This practice implicitly treats the attention distribution as a proxy for semantic relevance.

3.2 Motivation

Model-Induced Prior. Empirical observations, as shown in Fig. 1, indicate that MLLMs exhibit a strong model-induced prior over visual tokens: even without any instruction, the model tends to concentrate attention on certain instruction-agnostic regions that are visually prominent but may not be relevant to the given task. This implies that the observed posterior distribution $P_\theta(v \mid \mathbf{X})$ is not a pure task signal, but rather a mixture confounded by prior noise and

task semantics. To decouple this prior, we first define the model-induced prior to represent the model's attention response to the *null* instruction input:

$$Q_\theta(v) \triangleq P_\theta(v \mid \emptyset), \tag{3}$$

where $\emptyset$ means a task-agnostic null token, such as a separator `\n`. It captures the model's intrinsic saliency preferences. Subsequently, we decompose the posterior attention $P_\theta(v \mid \mathbf{X})$ into the product of a prior and a relative gain:

$$P_\theta(v \mid \mathbf{X}) \equiv Q_\theta(v) \cdot \underbrace{\left(\frac{P_\theta(v \mid \mathbf{X})}{Q_\theta(v)}\right)}_{\text{Semantic Lift } \mathcal{L}(v,\mathbf{X})}, \tag{4}$$

where the term $\mathcal{L}(v, \mathbf{X})$ is defined as the semantic lift, which quantifies the multiplicative increase in the importance of token v relative to its prior, induced by the intervention of instruction $\mathbf{X}$.

Prior-Dominated Masking. We refer to this failure mode as prior-dominated masking: an important token receives stronger instruction-conditioned evidence but is ranked lower under absolute posterior magnitude due to the model-induced prior. The outcome of posterior ranking depends on both the prior term Q and the lift term $\mathcal{L}$. When a background token v_{bg} possesses extremely high prior such that $Q_\theta(v_{bg}) \gg Q_\theta(v_{tgt})$, even if a target token v_{tgt} exhibits a high semantic lift $\mathcal{L}_{tgt} \gg \mathcal{L}_{bg}$, its final posterior probability may still be suppressed. Specifically, the sufficient condition for ranking failure can be characterized as:

$$P_\theta(v_{tgt} \mid \mathbf{X}) < P_\theta(v_{bg} \mid \mathbf{X}) \iff \frac{\mathcal{L}(v_{tgt}, \mathbf{X})}{\mathcal{L}(v_{bg}, \mathbf{X})} < \frac{Q_\theta(v_{bg})}{Q_\theta(v_{tgt})}. \tag{5}$$

Under attention-based pruning, this condition is sufficient for the background token to be retained while the task-conditioned token is discarded. This indicates that as long as the prior discrepancy between background and target is greater than the semantic lift discrepancy induced by the instruction in magnitude, the baseline method $\mathcal{S}_{\text{base}}$ will inevitably preserve background noise while discarding semantic signals. Therefore, the ideal metric should be the decoupled $\mathcal{L}$ or its logarithmic form, rather than the raw P_θ.

3.3 $\mathcal{V}$-Usable Information for Budgeted Token Selection

Figure 2 provides an overview of PriorTR. Our goal is to select a compact set of K visual tokens that preserves the information *induced by the instruction* $\mathbf{X}$, rather than the model-induced prior. To this end, we contrast the task-conditioned posterior token distribution $P_\theta(\cdot \mid \mathbf{X})$ with a model-induced prior baseline $Q_\theta(\cdot)$ computed from a null instruction in §3.2. Following the usable-information perspective under computational constraints [54], we quantify Pointwise $\mathcal{V}$-Information (PVI) with instruction-induced evidence at the token level via a log-likelihood ratio:

$$\text{PVI}(v) \triangleq \log \frac{P_\theta(v \mid \mathbf{X})}{Q_\theta(v)}, \tag{6}$$

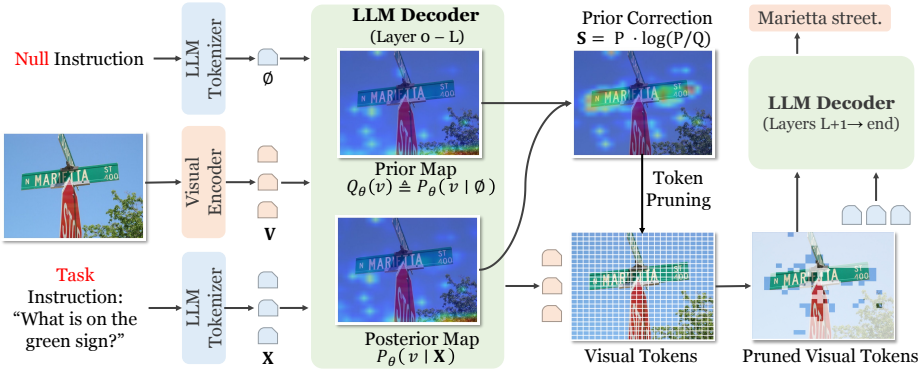

Fig. 2: Overview of PriorTR. Given an input image and a task instruction, PriorTR performs a single forward pass up to layer L of the shared LLM decoder. At layer L, attention from the null token to visual tokens estimates the model-induced saliency prior, while attention from instruction tokens to visual tokens estimates the task-conditioned posterior. Visual tokens are then scored by their $\mathcal{V}$-usable information contribution, and the top-K tokens are physically retained. The LLM decoder continues from layer $L{+}1$ using the pruned visual tokens to generate the final response, achieving efficient inference while preserving instruction-relevant visual evidence.

which measures the *semantic lift* contributed by $\mathbf{X}$ beyond the prior. Importantly, for prior-driven regions where $P_\theta(v \mid \mathbf{X}) \approx Q_\theta(v)$, $\mathrm{PVI}(v) \approx 0$, naturally suppressing prior-only tokens.

Masked Total Usable Information. To model a strict token budget, we introduce a binary selection mask $m \in \{0,1\}^N$ and $\sum_{i=1}^{N} m_i = K$. We define the usable information retained after masking as the posterior expectation of pointwise lift restricted to selected tokens:

$$I_{\mathrm{mask}}(m; \mathbf{X} \to \mathbf{V}) \triangleq \sum_{i=1}^{N} m_i \, P_\theta(v_i \mid \mathbf{X}) \, \log \frac{P_\theta(v_i \mid \mathbf{X})}{Q_\theta(v_i)}, \tag{7}$$

which is an additive masked surrogate of the full divergence $D_{\mathrm{KL}}(P_\theta(\cdot \mid \mathbf{X}) \,\|\, Q_\theta(\cdot))$ and explicitly decomposes token utility into two factors: (i) *relevance mass* $P_\theta(v_i \mid \mathbf{X})$, which avoids selecting negligible-probability outliers, and (ii) *debiasing density* $\log \frac{P_\theta(v_i \mid \mathbf{X})}{Q_\theta(v_i)}$, which discounts prior-dominated tokens. This objective optimizes a masked additive surrogate of the full divergence under a strict token budget, rather than the original likelihood objective in Eq. (1).

Closed-Form Optimal Selection. The optimal subset is obtained by selecting the K tokens with the largest per-token contributions. This yields the PriorTR importance score:

$$S_{\mathrm{PriorTR}}(v_i) \triangleq P_\theta(v_i \mid \mathbf{X}) \, \log \frac{P_\theta(v_i \mid \mathbf{X})}{Q_\theta(v_i)}. \tag{8}$$

Algorithm 1 PriorTR: Prior-Corrected Visual Token Reduction

Input: Model weights θ, Image tokens $\mathbf{V}$, Instruction tokens $\mathbf{X}$, Pruning layer L, Budget K.
Output: Response $\mathbf{Y}$.
// Phase 1: Single-Pass Prior and Posterior Estimation
Insert null token $\emptyset$ between the visual and instruction tokens:
  $\mathbf{Z} \leftarrow [\,\mathbf{V}, \emptyset, \mathbf{X}\,]$
Forward pass up to layer L:
  $(\mathbf{H}, \mathbf{KV}, \mathbf{A}) \leftarrow \textsc{Forward}(\mathbf{Z}; \theta, \text{depth}{=}L)$
Extract saliency prior from null-token attention:
  $\mathbf{Q} \leftarrow \textsc{Agg}(\mathbf{A}[\emptyset \rightarrow \mathbf{V}])$ ▷ $Q_i \approx P_\theta(v_i \mid \emptyset)$
Extract task-conditioned posterior from instruction attention:
  $\mathbf{P} \leftarrow \textsc{Agg}(\mathbf{A}[\mathbf{X} \rightarrow \mathbf{V}])$ ▷ $P_i \approx P_\theta(v_i \mid \mathbf{X})$
// Phase 2: Prior-Corrected Scoring
Compute PriorTR scores:
  $\mathbf{S} \leftarrow \mathbf{P} \odot \log\left((\mathbf{P} + \epsilon)/(\mathbf{Q} + \epsilon)\right)$
Select top-K token indices:
  $\mathcal{I}_{\text{top}} \leftarrow \textsc{TopK}(\mathbf{S}, K)$
// Phase 3: Physical Pruning and Decoding
Gather selected visual hidden states and KV cache:
  $\hat{\mathbf{H}} \leftarrow \mathbf{H}[\mathcal{I}_{\text{top}}]$; $\widehat{\mathbf{KV}} \leftarrow \textsc{Gather}(\mathbf{KV}, \mathcal{I}_{\text{top}})$
Continue decoding from layer $L{+}1$:
  $\mathbf{Y} \leftarrow \textsc{Generate}(\theta; \hat{\mathbf{H}}, \widehat{\mathbf{KV}}, \text{depth}{>}L)$

We rank tokens by S_{PriorTR} in descending order and keep the top-K tokens for subsequent layers. In §3.4, we describe an efficient single forward pass to simultaneously estimate $P_\theta(\cdot \mid \mathbf{X})$ and $Q_\theta(\cdot)$ with negligible overhead.

3.4 Prior-Corrected Token Reduction Procedure

Single-Pass Prior Estimation. Algorithm 1 shows the token reduction procedure. Instead of performing two separate forward passes, PriorTR estimates both the prior and the posterior within a single forward pass. Specifically, we utilize the null token (implemented as the text token `\n` in LLaVA models) between the visual tokens and the instruction tokens. This token acts as an instruction-agnostic probe that attends to the visual sequence without being influenced by the instruction. Because the decoder is causal, the null token precedes the instruction tokens and therefore cannot attend to them, ensuring that its attention over visual tokens remains instruction-independent.

Attention Aggregation. At the pruning layer L, we obtain two distributions over visual token indices: (i) attention from the null token to visual tokens, which serves as the saliency prior, and (ii) attention from instruction tokens to visual tokens, which captures task-conditioned relevance. We aggregate attention across heads and normalize each to form comparable categorical distributions; the explicit aggregation is defined in Sec. A of the Supplementary Material.

Token Scoring and Selection. Using these two distributions, we compute per-token importance scores according to the prior-corrected formulation in §3.3. The computation is performed element-wise in a fully vectorized manner with a small numerical stabilizer ϵ. We then select the top-K token indices via a standard Top-K operation, without requiring sorting across layers or iterative optimization.

Physical Pruning and Continued Decoding. PriorTR performs *physical* pruning rather than masking. We gather only the hidden states and KV-cache entries corresponding to the selected visual tokens. All other visual tokens are removed from memory. The decoder then resumes autoregressive generation from layer $L{+}1$ using the compacted hidden states and KV cache. Consequently, both attention and feed-forward computation in deeper layers scale with K instead of the original visual token count.

4 Experiments

4.1 Image Understanding Tasks

Experimental Settings. We evaluate PriorTR on twelve multimodal benchmarks spanning visual reasoning (GQA [15], SQA [32], VizWiz [10], OKVQA [33]), hallucination detection (POPE [24]), OCR (TextVQA [40]), captioning (Flickr [34], NoCaps [1]), and holistic understanding (MME [9], MMB [31], SEED [20], MMVet [61]). We compare against five state-of-the-art training-free methods: FastV [6], PDrop [51], SparseVLM [66], VisPruner [65], and PruMerge [36]. To assess cross-model generalizability, we evaluate on LLaVA-1.5-7B [29], LLaVA-1.5-13B [29], and Qwen3-VL-8B [4].

Implementation Details. We apply PriorTR to each model's standard inference pipeline without modifying model weights or requiring additional training. Following FastV [6], we prune at layer $L{=}2$; early pruning maximizes downstream savings, and Sec. B.4 of the Supplementary Material confirms PriorTR is robust. As detailed in §3.4, both distributions are extracted from a single forward pass: the prior Q is derived from the post-image separator token, and the task-conditioned posterior P is aggregated from the instruction tokens. The prior-corrected score is computed per visual token, and the top-K tokens are physically retained in hidden states and KV cache. All experiments are conducted on NVIDIA RTX PRO 6000 GPUs.

Results on LLaVA-1.5-7B. We evaluate all methods under three token budgets ($K \in \{192, 128, 64\}$). As shown in Table 1, PriorTR achieves the highest average accuracy across all compression levels. The margin over FastV [6] is largest at the tightest budget ($K{=}64$, 89% token reduction), precisely the regime in which prior ranking most severely misidentifies instruction-relevant tokens. Moreover, PriorTR maintains the highest absolute accuracy at every evaluated budget, confirming that prior-corrected scoring consistently preserves instruction-relevant tokens under increasingly aggressive compression.

Table 1: Performance comparison with state-of-the-art methods with different vision tokens preserved in LLaVA-1.5-7B. There are 576 visual tokens in the vanilla setting. **Avg.** reports the average of normalized results on all datasets. Best results in **Bold**.

Method	GQA	POPE	MME	MMB	TextVQA	SEED	VizWiz	SQA	Flickr	NoCaps	OKVQA	MMVet	Avg. (%)
Upper Bound, 576 Tokens **(100%)**													
Vanilla	61.9	85.9	1864	64.6	58.3	66.2	50.0	69.5	74.9	1.05	53.4	30.9	100
Retain 192 Tokens (↓ 66.7%)													
FastV (ECCV24)	52.6	61.0	1605	61.0	52.5	63.9	50.8	52.5	72.8	1.03	51.3	26.7	89.8
PDrop (CVPR25)	57.1	82.3	1766	63.2	56.1	54.7	**51.1**	**68.8**	74.2	1.02	51.8	30.5	96.0
SparseVLM (ICML25)	57.6	83.6	1721	62.5	56.1	64.1	50.5	68.7	72.0	1.01	51.9	**33.1**	97.4
PruMerge (ICCV25)	54.3	71.3	1626	58.9	54.3	57.4	50.1	67.9	58.4	0.89	46.1	29.0	89.1
VisPruner (ICCV25)	59.6	**86.2**	1796	63.3	**57.7**	63.7	50.3	68.2	72.8	1.02	**52.1**	32.4	98.5
PriorTR (Ours)	**60.4**	83.8	**1845**	**63.6**	57.5	**64.8**	50.9	68.6	**75.6**	**1.06**	**52.1**	32.5	**99.5**
Retain 128 Tokens (↓ 77.8%)													
FastV (ECCV24)	49.6	53.4	1490	56.1	50.5	48.1	51.3	60.2	69.1	0.99	49.1	26.3	85.1
PDrop (CVPR25)	56.0	82.3	1644	61.1	55.1	53.3	51.0	68.3	65.6	0.92	49.8	30.8	92.7
SparseVLM (ICML25)	56.0	80.5	1696	60.0	54.9	58.2	**51.4**	67.1	58.2	0.82	51.0	29.0	91.2
PruMerge (ICCV25)	53.3	67.1	1544	58.1	54.3	55.0	50.3	67.1	54.9	0.83	43.7	24.4	85.3
VisPruner (ICCV25)	58.1	**84.3**	1761	62.2	**57.0**	61.6	51.2	**68.7**	70.4	0.99	50.3	**31.6**	96.6
PriorTR (Ours)	**59.1**	81.7	**1820**	**62.4**	56.9	**63.8**	51.3	68.6	**74.8**	**1.05**	**51.4**	31.4	**98.2**
Retain 64 Tokens (↓ 88.9%)													
FastV (ECCV24)	46.1	38.2	1255	47.2	47.8	43.7	50.8	51.1	45.1	0.71	40.0	19.6	70.7
PDrop (CVPR25)	41.9	55.9	1092	33.3	45.9	40.0	49.5	68.6	51.1	0.70	42.2	**30.7**	74.4
SparseVLM (ICML25)	52.7	75.1	1505	56.2	51.8	52.2	50.1	62.2	42.0	0.59	46.1	24.9	81.4
PruMerge (ICCV25)	51.9	65.3	1549	55.1	54.0	53.7	50.1	68.1	52.0	0.77	43.4	22.2	83.0
VisPruner (ICCV25)	55.4	**80.4**	1687	59.6	**55.3**	57.8	**52.2**	68.6	63.4	0.89	47.2	28.3	91.7
PriorTR (Ours)	**56.6**	76.3	**1745**	**61.2**	54.9	**60.1**	51.5	**68.8**	**71.2**	**1.01**	**49.2**	29.4	**94.5**

Generalization to other MLLMs. Table 2 extends our evaluation to Qwen3-VL-8B under dynamic resolution, comparing PriorTR against the posterior baseline FastV [6] to isolate the benefits of prior correction relative to raw attention ranking. PriorTR consistently outperforms FastV and remains advantaged compared to more recent sophisticated methods PDrop and VisPruner. This accelerating advantage confirms that correcting for intrinsic saliency bias becomes increasingly critical as visual token scarcity intensifies.

Table 2: Performance comparison on Qwen3-VL-8B under different token ratios. Avg. reports the mean of per-dataset scores normalized by the baseline.

Method	GQA	MME	MMB	POPE	TextVQA	Avg. (%)
Baseline (Full Tokens)						
Qwen3-VL	61.5	2346 (100.0)	84.7	88.0	80.2	100.0
Ratio 33.3%						
FastV	55.5	2190 (93.4)	80.4	85.4	75.4	93.9
PDrop	57.4	2090 (89.1)	78.4	85.4	72.3	92.4
VisPruner	**58.7**	2224 (94.8)	**81.8**	**86.7**	**76.0**	96.0
PriorTR (Ours)	58.3	**2333 (99.5)**	81.1	86.0	**76.0**	**96.5**
Ratio 22.2%						
FastV	52.1	1961 (83.6)	75.2	81.3	72.5	88.0
PDrop	54.5	2008 (85.6)	75.9	81.2	69.3	88.5
VisPruner	56.5	2026 (86.4)	78.1	**84.6**	72.0	91.3
PriorTR (Ours)	**56.7**	**2208 (94.1)**	**79.3**	84.4	**74.0**	**93.6**
Ratio 11.1%						
FastV	46.0	1619 (69.0)	64.1	68.9	64.0	75.5
PDrop	41.7	1542 (65.7)	55.8	58.2	50.5	65.7
VisPruner	50.3	1713 (73.0)	67.3	74.7	60.7	79.0
PriorTR (Ours)	**53.5**	**2047 (87.3)**	**75.7**	**78.8**	**68.8**	**87.8**

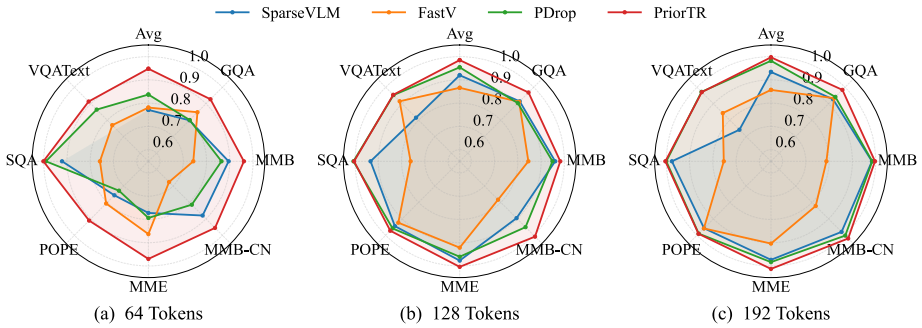

Fig. 3: Performance comparison under different token budgets on LLaVA-1.5-13B. These radar plots illustrate normalized performance across six representative benchmarks for four pruning methods under three token budgets. PriorTR consistently achieves the largest coverage area, indicating stronger overall robustness across tasks, particularly under aggressive compression (64 tokens).

Scalability. In Fig. 3, we further evaluate on LLaVA-1.5-13B to assess cross-scale generalizability. PriorTR achieves the highest accuracy at all three token pruning ratios. Fig. 3 visualizes the accuracy–budget trade-off across token budgets and full per-benchmark results are reported in Sec. B of the Supplementary Material against six representative benchmarks. These results confirm that decoupling instruction-induced semantics from the model-induced prior yields a consistently superior accuracy–efficiency trade-off over all evaluated training-free baselines when generalizing to a larger model.

4.2 Video Understanding Tasks

Experimental Settings. We apply PriorTR to Video-LLaVA [26] and evaluate on four video question-answering benchmarks: MSVD-QA [52], MSRVTT-QA [53], TGIF-QA [17], and ActivityNet-QA [62]. The 2048 video tokens are compressed to 194 (over 90% reduction) at layer L=2. We compare against two attention-based baselines that share the same raw-attention scoring but differ in downstream strategy. FastV applies hard pruning, discarding low-scored tokens entirely. SparseVLM [66] augments pruning with token merging by clustering and reintegrating discarded tokens as compact cluster representatives to preserve spatial-temporal context. We also test our method by combining it with token merging as another variant, as the merging strategy is particularly useful in video tasks with many duplicate visual tokens. We report Accuracy (%) and Score (0–5) under both GPT-4o-mini [16] and Gemini-2.5-Flash [8] evaluation.

Results. As shown in Table 3, among pruning-only methods, PriorTR outperforms FastV by +4.6 average accuracy points (GPT), demonstrating that prior-corrected scores identify more instruction-relevant tokens. SparseVLM involves token merging that effectively preserves spatial-temporal context, bringing it to

Table 3: The results of Video-LLaVA with different pruning strategies on video question answering tasks. The original number of video tokens is 2048, while our experiment collectively prunes it down to 194 tokens. We report Accuracy (%) and Score (0-5) evaluated by GPT-4o-mini (GPT) and Gemini-2.5-Flash (Gem).

Method	Strategy	Eval	MSVD		MSRVTT		TGIF		ActivityNet		Avg.	
			Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
Video-LLaVA	Vanilla	GPT	58.9	3.31	42.7	2.69	11.3	1.19	42.0	2.20	38.7	2.35
		Gem	57.3	2.93	47.5	2.32	14.2	0.77	42.5	2.12	40.4	2.04
FastV	Pruning	GPT	35.9	2.36	31.4	2.19	2.8	1.16	34.2	1.83	26.1	1.89
		Gem	35.5	1.88	34.3	1.69	2.3	0.14	34.2	1.74	26.6	1.36
PriorTR (Ours)		GPT	43.6	2.65	35.1	2.35	4.1	**1.20**	40.0	2.04	30.7	2.06
		Gem	45.3	2.33	39.4	1.95	4.8	0.27	40.1	2.00	32.4	1.64
SparseVLM	Merging	GPT	56.0	3.20	41.9	2.63	10.7	1.03	43.6	2.21	38.1	2.27
		Gem	58.2	3.05	48.0	**2.38**	13.1	0.71	44.0	2.20	40.8	2.09
PriorTR(Ours)		GPT	**56.6**	**3.26**	**42.9**	**2.66**	**10.9**	1.05	**44.3**	**2.25**	**38.7**	**2.31**
+Merging		Gem	**59.2**	**3.08**	**48.3**	2.36	**14.2**	**0.74**	**44.8**	**2.24**	**41.6**	**2.11**

Table 4: Efficiency comparison on LLaVA-1.5-7B. All measurements averaged over 100 runs (10-run warmup). Prefill latency excludes ViT/Proj. FLOPs as fraction of vanilla. Performance from Table 1.

Method	Prefill ↓ (ms)	KV Cache ↓ (MB)	FLOPs ↓	Performance ↑	Throughput ↑ (samples/s)	Speedup ↑
Vanilla LLaVA-1.5-7B	40.7	312	100.0%	100%	21.2	1.00×
Retain 192 Tokens (↓ 66.7%)						
+ FastV (ECCV24)	**20.6**	**120**	49.5%	89.8%	34.2	**1.98**×
+ PDrop (CVPR25)	33.7	205	70.0%	96.0%	24.0	1.21×
+ SparseVLM (ICML25)	34.0	128	**45.7%**	97.4%	25.2	1.20×
+ **PriorTR** (Ours)	20.7	**120**	49.5%	**99.5%**	**34.5**	1.97×
Retain 128 Tokens (↓ 77.8%)						
+ FastV (ECCV24)	**19.9**	**88**	41.1%	85.1%	**35.5**	**2.05**×
+ PDrop (CVPR25)	34.9	182	63.3%	92.7%	23.0	1.17×
+ SparseVLM (ICML25)	29.5	97	**36.8%**	91.2%	24.2	1.38×
+ **PriorTR** (Ours)	20.1	**88**	41.1%	**98.2%**	35.2	2.02×
Retain 64 Tokens (↓ 88.9%)						
+ FastV (ECCV24)	**18.0**	**56**	32.8%	70.7%	**38.0**	**2.26**×
+ PDrop (CVPR25)	29.9	156	55.4%	74.4%	26.1	1.36×
+ SparseVLM (ICML25)	31.0	66	**28.1%**	81.4%	24.8	1.31×
+ **PriorTR** (Ours)	18.4	**56**	32.8%	**94.5%**	37.9	2.21×

near-baseline accuracy. However, because it still relies on biased raw-attention scores to dictate which tokens are pruned and merged, the quality of its aggregated representations remains limited by prior noise. PriorTR+Merging applies a complementary token merging strategy guided by prior-corrected scores instead, so it matches the uncompressed baseline under GPT (38.7%) and exceeds it under Gemini (41.6% vs. 40.4%), surpassing SparseVLM in both settings.

4.3 Analysis

Efficiency Analysis. Table 4 presents a comprehensive efficiency evaluation on LLaVA-1.5-7B under different token budgets. We report prefilling latency, KV cache size, FLOPs (relative to vanilla) and throughput. All measurements are averaged over 100 runs (with 10-run warmup). PriorTR consistently achieves the fastest or near-fastest prefilling latency across all budgets. At 64 retained tokens, prefilling is reduced to 18.4 ms, corresponding to a 2.21× speedup over the vanilla model. Importantly, PriorTR matches the smallest KV cache footprint (56 MB), demonstrating that physical token pruning effectively reduces memory usage. PriorTR achieves substantial computational savings. Across budgets, FLOPs scale proportionally with retained tokens, confirming that pruning directly reduces attention and feed-forward computation rather than merely masking tokens. Despite aggressive token reduction, PriorTR consistently achieves the highest throughput in all settings. Notably, unlike SparseVLM, which achieves the lowest FLOPs but fails to translate this into throughput gains, PriorTR consistently translates theoretical compute reduction into practical runtime gains. Across all budgets, PriorTR achieves the strongest balance between efficiency and accuracy. While other methods may achieve similar FLOPs reduction, they incur larger performance drops.

Ablation on Prior Correction Functions. Table 5 compares PriorTR's scoring rule against several alternative criteria, including saliency prior (Q), posterior attention magnitude (P), entropy-based measures ($P \log P - Q \log Q$), log-ratio scoring ($\log(P/Q)$), and simple difference ($P - Q$). The prior baseline performs worse than posterior-based methods, confirming that saliency alone is insufficient for token selection. Posterior attention (P) improves performance but remains limited under strict compression, indicating that absolute attention magnitude still conflates saliency and instruction relevance. Entropy-based scoring and pure log-ratio ranking underperform compared to difference-based metrics, suggesting that overly aggressive normalization or scale-invariant ranking can discard useful relevance magnitude. The difference rule ($P - Q$) consistently improves over posterior-only scoring, validating the importance of subtracting the saliency baseline. However, PriorTR's $\mathcal{V}$-information formulation $P \cdot \log(P/Q)$ achieves the best overall performance across most benchmarks and token budgets. Unlike the simple difference rule, PriorTR jointly accounts for both relevance magnitude (P) and relative lift over the prior ($\log(P/Q)$), yielding more stable and discriminative token ranking. The advantage of PriorTR is most pronounced at smaller budgets, where incorrect token ranking has a larger impact.

Debiasing Qualitative Analysis. Figure 4 provides qualitative comparisons between the intrinsic saliency prior, task-conditioned posterior attention, and the prior-corrected importance maps produced by PriorTR. Across diverse question types—including object recognition, attribute reasoning, action understanding, and scene description—we observe a consistent pattern. The saliency prior, estimated via the null token, strongly activates on high-contrast regions and textured backgrounds, independent of the task instruction. While posterior at-

Table 5: Ablation study of token scoring functions on LLaVA-1.5-7B. We compare our PriorTR against other statistical metrics under different token budgets (K). **Bold** denotes the best results.

K	Method	Formula	MME	MMB	GQA	POPE	TextVQA	Avg.(%)
576	Vanilla		1864	64.6	61.9	85.9	58.3	100
128	Prior	Q	1722.3	61.6	55.2	72.8	54.1	90.9
	Posterior	P	1774.3	**62.4**	56.7	76.1	56.2	93.7
	Entropy	$P\log P - Q\log Q$	1603.8	55.2	54.9	66.4	44.1	82.6
	Log Ratio	$\log(P/Q)$	1700.8	57.7	56.7	75.6	53.7	90.5
	Diff	$P - Q$	1757.0	60.6	57.0	78.8	55.7	93.5
	PriorTR	$P \cdot \log(P/Q)$	**1804.0**	**62.4**	**58.5**	**80.5**	**56.9**	**95.8**
64	Prior	Q	1461.0	54.6	50.5	57.3	49.8	79.3
	Posterior	P	1641.6	59.6	52.8	65.2	53.7	86.7
	Entropy	$P\log P - Q\log Q$	1401.8	43.0	50.8	54.3	43.5	72.3
	Log Ratio	$\log(P/Q)$	1501.3	52.3	52.6	64.5	50.7	81.7
	Diff	$P - Q$	1635.2	57.7	53.3	71.5	53.2	87.6
	PriorTR	$P \cdot \log(P/Q)$	**1713.0**	**60.7**	**55.7**	**75.0**	**54.9**	**91.5**
32	Prior	Q	1183.8	36.8	43.8	26.3	45.2	59.9
	Posterior	P	1366.5	51.7	47.8	47.4	50.5	74.5
	Entropy	$P\log P - Q\log Q$	1146.9	29.6	44.3	30.5	42.9	57.6
	Log Ratio	$\log(P/Q)$	1304.5	46.2	48.4	52.9	47.8	72.7
	Diff	$P - Q$	1448.6	51.7	47.6	57.8	48.2	76.9
	PriorTR	$P \cdot \log(P/Q)$	**1627.0**	**56.8**	**51.6**	**63.4**	**52.4**	**84.5**

tention incorporates instruction signals, it often remains partially biased toward these visually dominant regions. This effect is especially evident in scenes with complex backgrounds or strong lighting contrasts. After prior correction, the resulting token importance maps exhibit sharper focus on instruction-relevant visual evidence. For example, in the sports image, the corrected map emphasizes the team logo rather than surrounding crowd or background textures. In reasoning-heavy examples (*e.g.,* identifying animals or describing actions), the corrected maps concentrate on semantically meaningful entities rather than peripheral salient regions. These qualitative results support our hypothesis that absolute attention magnitude conflates intrinsic saliency with task-driven relevance, and demonstrate that prior correction yields more semantically aligned token selection.

Additional Experiments in the Supplementary Material. We provide complementary analyses in the Supplementary Material to further validate PriorTR. These include experimental setup details (*e.g.,* baseline methods, implementation, and datasets), the impact of different null tokens beyond the default separator token `\n`, details of a PriorTR variant with token merging, ablation on pruning at different layers, experiments on additional MLLMs (LLaVA-Next [28]

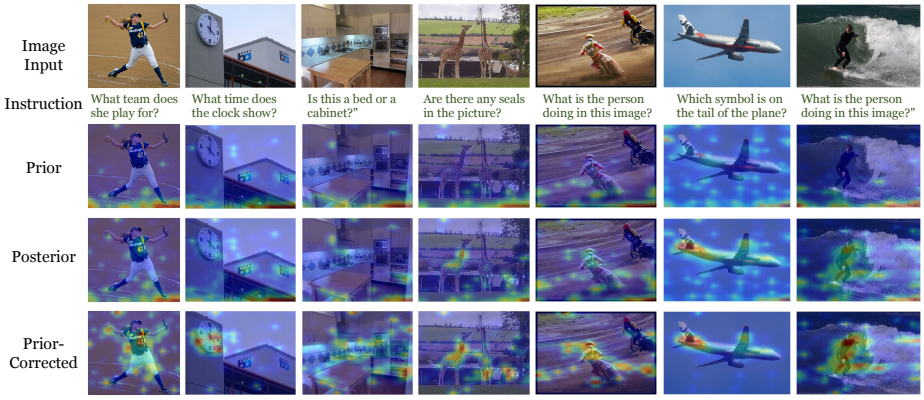

Fig. 4: Qualitative comparison of prior, posterior, and prior-corrected attention maps. For each example, we visualize (row 1) the input image and task instruction, (row 2) attention induced by a null token representing the intrinsic saliency prior, (row 3) task-conditioned posterior attention, and (row 4) prior-corrected token importance computed by PriorTR. The prior maps consistently highlight visually salient regions (e.g., high-contrast edges, textured backgrounds), while posterior attention remains partially influenced by this bias. In contrast, prior-corrected maps suppress saliency-dominated regions and emphasize instruction-relevant evidence, such as the team logo on the jersey, cabinet structures, animals in the scene, or the action performed by a person. These visualizations illustrate how prior correction disentangles intrinsic visual saliency from task-driven relevance.

and InternVL [7]), complete numeric results for LLaVA-1.5-13B, and a detailed computational complexity analysis.

5 Conclusion

We introduced PriorTR, a training-free visual token reduction method for accelerating multimodal large language models. We identify that text–visual attention is often dominated by a model-induced prior, which can suppress instruction-relevant tokens when ranking purely by attention magnitude. PriorTR corrects this bias by selecting tokens based on their additional usable information beyond the prior baseline. Across multiple benchmarks and token budgets, PriorTR consistently improves the accuracy–efficiency trade-off while delivering real reductions in latency, memory usage, and FLOPs through physical pruning. Our results suggest that separating intrinsic saliency from task-driven relevance provides a principled and effective direction for efficient multimodal inference.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62502429, and the Zhejiang Key Laboratory Project (2024E10001)

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9392–9401 (2025)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: 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 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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 (2024)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
10. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
11. Han, Y., Liu, X., Ding, P., Wang, D., Chen, H., Yan, Q., Huang, S.: Rethinking token reduction in mllms: Towards a unified paradigm for training-free acceleration. arXiv preprint arXiv:2411.17686 **2**(3) (2024)
12. He, Y., Chen, F., Liu, J., Shao, W., Zhou, H., Zhang, K., Zhuang, B.: Zipvl: Efficient large vision-language models with dynamic token sparsification. arXiv preprint arXiv:2410.08584 (2024)
13. Hu, L., Shang, F., Feng, W., Wan, L.: Lightvlm: Acceleraing large multimodal models with pyramid token merging and kv cache compression. arXiv preprint arXiv:2509.00419 (2025)
14. Huang, K., Zou, H., Xi, Y., Wang, B., Xie, Z., Yu, L.: Ivtp: Instruction-guided visual token pruning for large vision-language models. In: European conference on computer vision. pp. 214–230. Springer (2024)

15. Hudson, D.A., Manning, C.D.: 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. 6700–6709 (2019)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
18. Jeddi, A., Baghbanzadeh, N., Dolatabadi, E., Taati, B.: Similarity-aware token pruning: Your vlm but faster. arXiv preprint arXiv:2503.11549 (2025)
19. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th symposium on operating systems principles. pp. 611–626 (2023)
20. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
21. Li, D., Yang, Z., Lu, S.: Todre: Visual token pruning via diversity and task awareness for efficient large vision-language models. arXiv e-prints pp. arXiv–2505 (2025)
22. Li, S., Hu, Y., Ning, X., Liu, X., Hong, K., Jia, X., Li, X., Yan, Y., Ran, P., Dai, G., et al.: Mbq: Modality-balanced quantization for large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4167–4177 (2025)
23. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
24. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
25. Liang, X., Guan, C., Lu, J., Chen, H., Wang, H., Hu, H.: Dynamic token reduction during generation for vision language models. arXiv preprint arXiv:2501.14204 (2025)
26. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
27. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
28. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/, accessed: 26 Jun 2026
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
30. Liu, Y., Wu, F., Li, R., Tang, Z., Li, K.: Par: Prompt-aware token reduction method for efficient large multimodal models. arXiv preprint arXiv:2410.07278 (2024)
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)

32. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
33. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
35. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
36. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22857–22867 (2025)
37. Shao, Z., Wang, M., Yu, Z., Pan, W., Yang, Y., Wei, T., Zhang, H., Mao, N., Chen, W., Yu, J.: Growing a twig to accelerate large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20064–20074 (2025)
38. Shi, D., Tao, C., Rao, A., Yang, Z., Yuan, C., Wang, J.: Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. arXiv preprint arXiv:2305.17455 (2023)
39. Si, G., Yin, H., Li, X., Liao, W., He, T., Peng, P., Zhu, W., et al.: Infoprune: Revisiting visual token pruning from an information-theoretic perspective (2025)
40. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
41. Song, D., Wang, W., Chen, S., Wang, X., Guan, M.X., Wang, B.: Less is more: A simple yet effective token reduction method for efficient multi-modal llms. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 7614–7623 (2025)
42. Sun, Z., Ma, Y., Liu, G., Chen, Y., Tang, X., Hu, Y., Xu, Y.: Ivc-prune: Revealing the implicit visual coordinates in lvlms for vision token pruning. arXiv preprint arXiv:2602.03060 (2026)
43. Vasu, P.K.A., Faghri, F., Li, C.L., Koc, C., True, N., Antony, A., Santhanam, G., Gabriel, J., Grasch, P., Tuzel, O., et al.: Fastvlm: Efficient vision encoding for vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19769–19780 (2025)
44. Wang, C., Wang, Z., Xu, X., Tang, Y., Zhou, J., Lu, J.: Q-vlm: Post-training quantization for large vision-language models. Advances in Neural Information Processing Systems **37**, 114553–114573 (2024)
45. Wang, Q., Ye, H., Chung, M.Y., Liu, Y., Lin, Y., Kuo, M., Ma, M., Zhang, J., Chen, Y.: Corematching: A co-adaptive sparse inference framework with token and neuron pruning for comprehensive acceleration of vision-language models. arXiv preprint arXiv:2505.19235 (2025)
46. Wang, Y., Wu, J., Ni, Z., Yang, C., Liu, Y., Yang, L., Zhou, Y., Wen, Y., He, L.: Entropyprune: Matrix entropy guided visual token pruning for multimodal large language models. arXiv preprint arXiv:2602.17196 (2026)

47. Wen, Z., Gao, Y., Li, W., He, C., Zhang, L.: Token pruning in multimodal large language models: Are we solving the right problem? In: Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025. pp. 15537–15549. Association for Computational Linguistics (2025)
48. Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., Zhang, L.: Stop looking for "important tokens" in multimodal language models: Duplication matters more. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9972–9991 (2025)
49. Wu, Q., Xu, W., Liu, W., Tan, T., Liujianfeng, L., Li, A., Luan, J., Wang, B., Shang, S.: Mobilevlm: A vision-language model for better intra-and inter-ui understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 10231–10251 (2024)
50. Xing, L., Wang, A.J., Yan, R., Shu, X., Tang, J.: Vision-centric token compression in large language model. arXiv preprint arXiv:2502.00791 (2025)
51. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2025)
52. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
53. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
54. Xu, Y., Zhao, S., Song, J., Stewart, R., Ermon, S.: A theory of usable information under computational constraints. In: International Conference on Learning Representations (2020)
55. Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Li, C., Yan, J., Bai, Y., Sadayappan, P., Hu, X., et al.: Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19803–19813 (2025)
56. Yang, C., Dong, X., Zhu, X., Su, W., Wang, J., Tian, H., Chen, Z., Wang, W., Lu, L., Dai, J.: Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24939–24949 (2025)
57. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: 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 (2025)
58. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22128–22136 (2025)
59. Ye, X., Gan, Y., Ge, Y., Zhang, X.P., Tang, Y.: Atp-llava: Adaptive token pruning for large vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24972–24982 (2025)
60. Yu, H., Li, W., Qu, X., Wang, S., Chen, J., Zhu, J.: Visiontrim: Unified vision token compression for training-free mllm acceleration. arXiv preprint arXiv:2601.22674 (2026)

61. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: Forty-first International Conference on Machine Learning (2024)
62. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
63. Zamini, M., Shukla, D.: Delta-llava: Base-then-specialize alignment for token-efficient vision-language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3648–3657 (2026)
64. Zeng, Q.S., Li, Y., Wang, Q., Jiang, P.T., Wu, Z., Cheng, M.M., Hou, Q.: A glimpse to compress: Dynamic visual token pruning for large vision-language models. arXiv preprint arXiv:2508.01548 (2025)
65. Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: 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 (2025)
66. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D.A., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. In: Forty-second International Conference on Machine Learning (2025)
67. Zhou, B., Hu, Y., Weng, X., Jia, J., Luo, J., Liu, X., Wu, J., Huang, L.: Tinyllava: A framework of small-scale large multimodal models. arXiv preprint arXiv:2402.14289 (2024)
68. Zhuang, X., Zhu, Z., Xie, Y., Liang, L., Zou, Y.: Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4189–4199 (June 2025)