ToDRE: Effective Visual Token Pruning via Token Diversity and Task Relevance

Duo Li[1*], Zuhao Yang[1*], Xiaoqin Zhang[2], Ling Shao[3], and Shijian Lu[1†]

[1] College of Computing and Data Science, Nanyang Technological University, Singapore
{duo001,yang0756}@e.ntu.edu.sg
[2] College of Computer Science and Technology, Zhejiang University of Technology, China
[3] Terminus AI Lab, University of Chinese Academy of Sciences, China

Abstract. Visual token pruning, which aims to compress and prune redundant visual tokens, plays a critical role in efficient inference with large vision-language models (LVLMs). However, existing methods fail to disentangle intra-modal visual redundancy from cross-modal redundancy between vision and language. We show that visual token diversity and task-specific token relevance are two crucial yet orthogonal factors that complement each other in conveying useful information and should therefore be treated separately for more effective visual token pruning. Building upon this insight, we design **ToDRE**, a two-stage and training-free framework that incorporates **To**ken **D**iversity and task **RE**levance for effective token compression and efficient LVLM inference. Instead of pruning redundant tokens, we introduce a greedy max-sum diversification algorithm that selects and retains a subset of diverse and representative visual tokens after the vision encoder. On top of that, ToDRE leverages an "information migration" mechanism to eliminate task-irrelevant visual tokens within certain decoder layers of the large language model (LLM), further improving token pruning and LVLM inference. Extensive experiments show that ToDRE prunes 90% of visual tokens after the vision encoder as well as all visual tokens in certain LLM decoder layers, leading to a 2.6× speed-up in total inference time while maintaining 95.0% model performance plus excellent model compatibility.

Keywords: Efficient Large Vision-Language Models · Visual Token Pruning · Token Diversity · Task Relevance

1 Introduction

Leveraging the superior reasoning capability of large language models (LLMs) [1, 3, 45, 48, 49], large vision-language models (LVLMs) [5, 20, 47, 52, 60] have

* Equal contribution.
** Corresponding author.

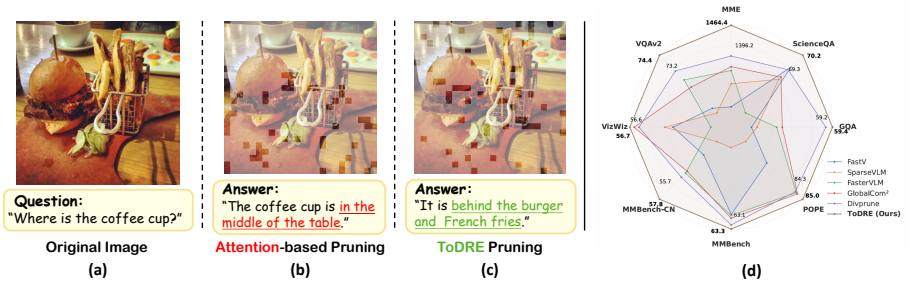

Fig. 1: (a–b) Existing methods fail to disentangle intra-modal visual redundancy from cross-modal vision-language redundancy, leading to erroneous pruning [10, 56]; as shown in (b), attention-guided pruning discards the informative white coffee-cup region. **(c)** ToDRE disentangles redundancy into two orthogonal factors, visual token diversity and task-specific token relevance, and prunes them in dedicated stages, preserving critical visual cues. **(d)** Across eight image-language benchmarks, ToDRE consistently outperforms all baselines.

achieved impressive performance in various multimodal understanding tasks such as visual question answering [16, 17, 19, 37, 44] and video understanding [15, 24, 38, 51, 59]. LVLMs convert visual inputs into visual tokens and align the converted visual tokens with text tokens for various multimodal understanding tasks. However, LVLM inference often incurs prohibitive computational and memory costs due to the massive number of visual tokens involved, significantly restricting the applicability of LVLMs to various downstream tasks.

Two representative approaches have recently been explored for improving the LVLM inference efficiency. The first approach is *model-centric*. It speeds up the inference via knowledge distillation [9], parameter quantization [53], or transformer replacement [40]. However, this approach requires model retraining, which incurs significant computational costs. The second approach is *data-centric*. It works by token pruning [10, 31, 34, 42, 56] or block skipping [43], and has attracted increasing attention due to its training-free and architecture-agnostic nature. Moreover, the *data-centric* approach strikes a strong balance between inference efficiency and model performance, offering a complementary solution to the *model-centric* approach.

Most existing token pruning methods estimate visual redundancy without explicitly differentiating between intra-modal and cross-modal signals. While such metrics—whether based on attention scores [10, 42, 56, 57], feature similarity [6, 21, 58], or output divergence [31, 55]—capture certain aspects of redundancy, they inherently conflate two distinct sources: intra-modal redundancy (visually duplicated content) and cross-modal redundancy (tokens irrelevant to the textual query). This coupling is analyzed in detail in Section 3; here we summarize its practical implications: attention-based methods suffer from positional bias [50], similarity-based merging degrades performance compared to direct pruning [18], and divergence-based approaches require model-specific calibration [31].

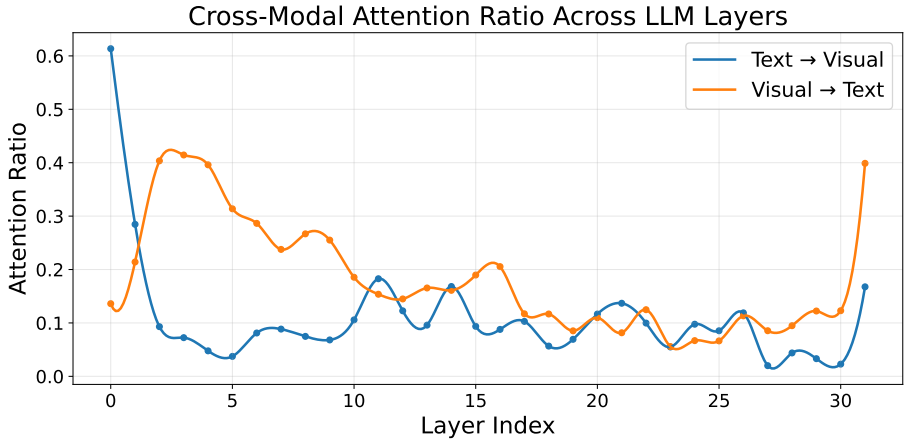

Fig. 2: Text-to-visual attention (blue) and visual-to-text attention (orange) in each LLM decoder layer. We observe a clear pattern of *"information migration"*: cross-modal attention (both visual-to-text and text-to-visual) is high in early layers, reflecting active information exchange, but gradually diminishes in deeper layers as the model shifts toward unimodal text reasoning.

Beyond these issues, we further observe an "*information migration*" phenomenon (Figure 2): cross-modal attention (both visual-to-text and text-to-visual) is concentrated in the early decoder layers and progressively diminishes in deeper layers. Given that output tokens exhibit near-zero attention to visual tokens during decoding (see the supplementary material), this suggests that once visual information has been absorbed into text representations, the remaining visual tokens become task-irrelevant and can be safely removed. However, most existing work [10, 42, 56, 57] passes all remaining visual tokens from the prefilling stage into decoding, thereby incurring unnecessary computations.

We design **ToDRE**, a simple yet effective token pruning technique that incorporates both *visual token diversity* and *task-specific token relevance* for effective token pruning and efficient LVLM inference. ToDRE performs token pruning in the embedding space prior to LLM input and during the LLM prefilling stage. First, we introduce a greedy max-sum diversification algorithm that iteratively identifies and preserves visual tokens that have minimal cumulative similarity to the selected tokens. Such token selection in the LLM embedding space explicitly reduces intra-modal redundancy, thereby preserving a broad spectrum of visual information and enhancing token representativeness under high pruning ratios. In addition, ToDRE leverages the "*information migration*" mechanism by adaptively selecting one layer in the latter half of the LLM decoder (where cross-modal interaction has significantly diminished) and dropping all visual tokens within that layer. This modality-level pruning removes visual tokens irrelevant to the given question and thus further eliminates redundant computation during inference. As a result, this relevance-guided pruning enables

continuous inference-time efficiency gains as the decoding length increases. As shown in Figure 1 (c–d), ToDRE's two-stage design enables effective visual token compression while preserving unique visual information and maintaining strong accuracy.

In summary, the major contributions of this work are threefold:

- **Revisit redundancy indicators.** First, we re-examine the principles of existing indicators on token redundancy and identify their constraints via systematic and comprehensive analysis. On top of that, we prove that inter-token diversity and token-task relevance are two orthogonal factors, and treating them separately enables more effective token pruning.
- **Propose a training-free and plug-and-play framework.** Second, we design a two-stage plug-and-play token pruning technique that is fully compatible with efficient attention operators [13] without requiring any additional training.
- **Conduct extensive empirical validation.** Third, extensive experiments over four widely adopted LVLMs and twelve multimodal benchmarks show that ToDRE achieves superior token pruning consistently.

2 Related Work

2.1 Large Vision-Language Models

Large vision-language models (LVLMs) [5, 47, 60] have demonstrated remarkable advancements by extending the reasoning capabilities of pretrained LLMs [3,48,49] to image and video comprehension tasks. Typically, LVLMs employ a vision encoder to extract visual features, which are subsequently projected into the LLM's embedding space via a visual projector (e.g., Q-Former [23] or MLP [27,33]). To process real-world high-resolution images, previous LVLMs [4, 32] resize input images to a fixed resolution, which introduces geometric distortion and degrades fine-grained local details. To tackle this, subsequent studies adopt dynamic tiling [11, 22, 33], which partitions images into regions and encodes each region independently using a shared vision encoder. However, dynamic tiling can yield thousands of visual tokens, significantly increasing computational overhead. This issue becomes even more pressing in video-based LVLMs [5, 30], since processing multiple video frames demands significantly more visual tokens. These challenges highlight the urgent need for accelerating LVLM inference in resource-constrained real-world environments.

2.2 Token Compression for LVLMs

Given that *spatially redundant* visual tokens outnumber *information-dense* text tokens by tens to hundreds of times [39], one natural solution to optimize LVLM inference is visual token compression. Several early attempts [8, 25, 26, 54] modify model components and introduce additional training costs. More recently, training-free token compression methods have been widely adopted due to their efficiency

and effectiveness. These methods can be categorized into two main groups: (1) Token compression in the vision encoder [6,29,42], the LLM decoder [10,31,57], or both [18]: For example, ToMe [6] reduces tokens in the vision encoding phase by merging redundant tokens via a binary soft-matching algorithm. Other approaches prune tokens during the LLM decoding stage by evaluating token redundancy through criteria such as attention scores with text tokens [10, 57] or observed divergence with LLM outputs [31, 55]. Subsequent studies [18, 34, 61] perform token compression during both stages to further enhance inference efficiency. (2) Token compression in the LLM embedding space [2,56]: A representative example is FasterVLM [56], which measures token redundancy more accurately through the cross-attention between the `[CLS]` token and visual tokens. Unlike previous methods, our proposed ToDRE simultaneously reduces tokens in both the LLM embedding space and the LLM decoder. Our two-stage approach effectively captures both visual token diversity and token-task relevance—two orthogonal yet critical aspects previously overlooked—achieving superior inference efficiency while maintaining competitive performance.

3 Preliminary Analysis

Recently, numerous visual token compression techniques have emerged. Most approaches [2,10,31,50,56] reduce computational redundancy only within specific stages of the LVLM inference process, lacking a systematic analysis and holistic consideration. To bridge this gap, we provide a deeper analysis organized as follows. In Section 3.1, we review the processing flow of existing LVLMs, identifying where redundant computation arises. In Section 3.2, we present a theoretical analysis that justifies disentangling redundancy into intra-modal and cross-modal components, which in turn motivates our two-stage token pruning method. Together, these analyses clarify *where* pruning should be applied and *how* it should be performed.

3.1 Computational Overhead in LVLM Processing Pipeline

Computational Cost Analysis. Prior studies [18,34] have shown that the dominant contributors to inference cost in LVLMs are the vision-encoding stage, the LLM prefilling stage, and the LLM decoding stage, each of which incurs substantial self-attention and feed-forward network (FFN) computations. Following previous studies [10,50], we formulate the calculation of floating-point operations (FLOPs) as follows:

$$\text{FLOPs}_{\text{encoding}} = \text{FLOPs}_{\text{prefilling}} = T \times \left(4nd^2 + 2n^2d + 2ndm\right), \tag{1}$$

$$\begin{aligned} \text{FLOPs}_{\text{decoding}} &= T \sum_{t=1}^{L} \left(4d^2 + 2d(n+t-1) + 2dm\right) \\ &= T \left(4Ld^2 + 2Ldm + dL(2n+L-1)\right), \end{aligned} \tag{2}$$

where T is the number of transformer layers; n and L respectively denote the lengths of the input and output sequences; d is the size of the hidden state; and m is the intermediate dimension of the FFN. We take LLaVA-NeXT-7B [33], which employs a CLIP-ViT-Large-Patch14 [41] vision encoder and a Vicuna-7B-v1.5 [12] LLM decoder, as an example. The relative ratio of FLOPs (with n=3000 and L=20) is approximately encoding:prefilling:decoding $\approx$ 1:**63.6**:0.4. When scaled to LLaVA-NeXT-13B, the relative ratio shifts to 1:**121.1**:0.8, indicating that the LLM's prefilling and decoding stages roughly double their share of the total computational cost. This underscores the importance of pruning visual tokens as early as possible—ideally *prior to* or *during* the LLM prefilling stage—to mitigate the exploding computational burden.

3.2 Intra- and Cross-Modal Redundancy

To justify the disentanglement of redundancy into intra-modal and cross-modal components, we develop the following theoretical analysis.

Notation. Let the visual token embeddings produced by the vision encoder-projector be $V = \{v_i\}_{i=1}^{N} \subset \mathbb{R}^d$ and the text tokens be $T = \{t_j\}_{j=1}^{M} \subset \mathbb{R}^d$. We map the two modalities onto mutually orthogonal subspaces of a shared embedding space:

$$\mathcal{V} = \mathrm{Span}(W_V), \qquad \mathcal{T} = \mathrm{Span}(W_T).$$

Visual data (e.g. images) encode spatial-texture patterns in pixel grids, whereas textual data (e.g. language) convey semantic information through symbol sequences. To preserve this intrinsic heterogeneity inside a multimodal model, we apply an embedding scheme based on orthogonal subspace decomposition. The resulting orthogonality constraint is:

$$W_V^\top W_T = 0 \qquad (\mathcal{V} \perp \mathcal{T}).$$

Intra-modal redundancy.

$$D_\kappa(V) = \frac{1}{N^2} \sum_{i \neq j} \kappa\big(W_V^\top v_i,\, W_V^\top v_j\big), \tag{3}$$

where $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is any non-negative kernel measuring pairwise similarity.

Cross-modal redundancy.

$$R_\rho(V, T) = \frac{1}{N} \sum_{i=1}^{N} \rho\big(W_T^\top v_i,\, T\big), \tag{4}$$

for a redundancy function $\rho : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}_{\geq 0}$. Equations (3)–(4) provide the two quantitative metrics that underpin our compression strategy.

Remark 1 (Limitations of entangled metrics). Existing pruning metrics attempt to capture redundancy with a mixed scoring metric $S(v_i)$, implicitly entangling intra-modal and cross-modal signals. This is inherently sub-optimal: a visual token can be *visually unique yet task-irrelevant*, or *highly relevant yet spatially duplicated*. These two cases are Pareto-incomparable, meaning any single scalar must impose an implicit and often suboptimal trade-off between diversity and relevance, which motivates a principled two-stage design that handles each source of redundancy with a dedicated, optimized criterion.

Lemma 1 (Subspace decorrelation). *Assume $\mathcal{V} \perp \mathcal{T}$ and that the projected components $W_V^\top v$ and $W_T^\top v$ are uncorrelated in expectation.*[4] *Then for any $v_i, v_j, v_k \in \mathcal{V}$,*

$$\mathbb{E}\big[\kappa(v_i, v_j)\,\rho(v_k, T)\big] = \mathbb{E}\big[\kappa(v_i, v_j)\big]\,\mathbb{E}\big[\rho(v_k, T)\big]. \tag{5}$$

Conclusion. Under the orthogonality constraint $W_V^\top W_T = 0$ and the mild decorrelation assumption in Lemma 1,

$$\mathrm{Cov}\big(D_\kappa(V),\, R_\rho(V, T)\big) = 0, \tag{6}$$

which means the two redundancy measures vary along orthogonal statistical directions.

Proof of (6)**.** Set $X_{ij} = \kappa(v_i, v_j)$ and $Y_k = \rho(v_k, T)$. By Eqs. (3) and (4),

$$D_\kappa(V) = \frac{1}{N^2}\sum_{i \neq j} X_{ij}, \qquad R_\rho(V, T) = \frac{1}{N}\sum_{k=1}^{N} Y_k.$$

Expanding the covariance,

$$\mathrm{Cov}(D_\kappa, R_\rho) = \mathbb{E}[D_\kappa R_\rho] - \mathbb{E}[D_\kappa]\mathbb{E}[R_\rho].$$

Step 1: substitute definitions.

$$\mathbb{E}[D_\kappa R_\rho] = \mathbb{E}\Big[\Big(\tfrac{1}{N^2}\sum_{i \neq j} \kappa(W_V^\top v_i, W_V^\top v_j)\Big)\Big(\tfrac{1}{N}\sum_{k=1}^{N} \rho(W_T^\top v_k, T)\Big)\Big].$$

Step 2: apply Lemma 1. Lemma 1 gives $\mathbb{E}[\kappa\,\rho] = \mathbb{E}[\kappa]\,\mathbb{E}[\rho]$, so

$$\mathbb{E}[D_\kappa R_\rho] = \frac{1}{N^3}\sum_{i \neq j}\sum_{k=1}^{N} \mathbb{E}[\kappa]\;\mathbb{E}[\rho] = \mathbb{E}[D_\kappa]\,\mathbb{E}[R_\rho].$$

Step 3: plug back into the covariance.

$$\mathrm{Cov}(D_\kappa, R_\rho) = \mathbb{E}[D_\kappa]\,\mathbb{E}[R_\rho] - \mathbb{E}[D_\kappa]\,\mathbb{E}[R_\rho] = 0. \;\square$$

[4] This is a modeling assumption that holds exactly under a generative model $v = W_V a + W_T b$ where $a \perp b$, and approximately when the vision-language projector preserves the heterogeneity of the two modalities. We verify this approximation empirically and provide a perturbation bound in Remark 2.

Remark 2 (Relaxation to approximate orthogonality). The strict constraint $W_V^\top W_T = 0$ is a simplifying assumption. In practice, the vision-language projector induces a small but non-zero coupling $W_V^\top W_T = E$ with $\|E\|_{\text{op}} \leq \epsilon$. Under this relaxation, the covariance bound becomes $|\text{Cov}(D_\kappa, R_\rho)| = \mathcal{O}(\epsilon)$, indicating that the two redundancy measures remain approximately decorrelated as long as the cross-subspace leakage is small. This is empirically supported by the fact that CLIP-style contrastive pre-training aligns global semantics (via `[CLS]` tokens) rather than patch-level features, leaving the patch-level visual subspace largely orthogonal to the textual subspace.

Thus, intra-modal redundancy and cross-modal redundancy are *uncorrelated* under the orthogonal subspace model, supporting the effectiveness of the two-stage compression paradigm.

4 Visual Token Pruning with Token Diversity and Task Relevance

Building on the preliminary analysis, we introduce ToDRE, a two-stage, training-free, and plug-and-play visual token compression framework (see Figure 3). ToDRE utilizes a greedy search in the LLM embedding space to select a maximally diverse subset of visual tokens, followed by an adaptive task-relevance-based pruning mechanism within the LLM decoder. Next, we elaborate on each stage in detail.

4.1 Diversity-Driven Token Selection

To obtain a maximally diverse subset of visual tokens, we adopt a greedy max-sum diversification algorithm [7] consisting of two steps: (1) initializing a retention set by selecting an initial pivot token, and (2) iteratively adding the token that minimizes its cumulative similarity to the current set. Full pseudocode of our proposed token retention algorithm is provided in the supplementary material.

Pivot Token Selection. To determine the initial pivot, we leverage the `[CLS]` attention from the last layer of the vision encoder [41] as an importance indicator. The attention from the `[CLS]` token $\boldsymbol{z}_{[\text{CLS}]} \in \mathbb{R}^d$ to other visual tokens $\boldsymbol{Z}_v \in \mathbb{R}^{n \times d}$ is calculated as:

$$\begin{aligned} \boldsymbol{q}_{[\text{CLS}]} &= \boldsymbol{z}_{[\text{CLS}]} \boldsymbol{W}_Q, \quad \boldsymbol{K}_v = \boldsymbol{Z}_v \boldsymbol{W}_K, \\ \boldsymbol{a}_{[\text{CLS}]} &= \text{Softmax}\left(\frac{\boldsymbol{q}_{[\text{CLS}]} \boldsymbol{K}_v^\top}{\sqrt{d}} \right), \end{aligned} \tag{7}$$

where n is the length of the visual token sequence; d is the hidden state size of the vision encoder; $\boldsymbol{W}_Q \in \mathbb{R}^{d \times d}$ and $\boldsymbol{W}_K \in \mathbb{R}^{d \times d}$ represent the weight matrices for queries and keys, respectively.

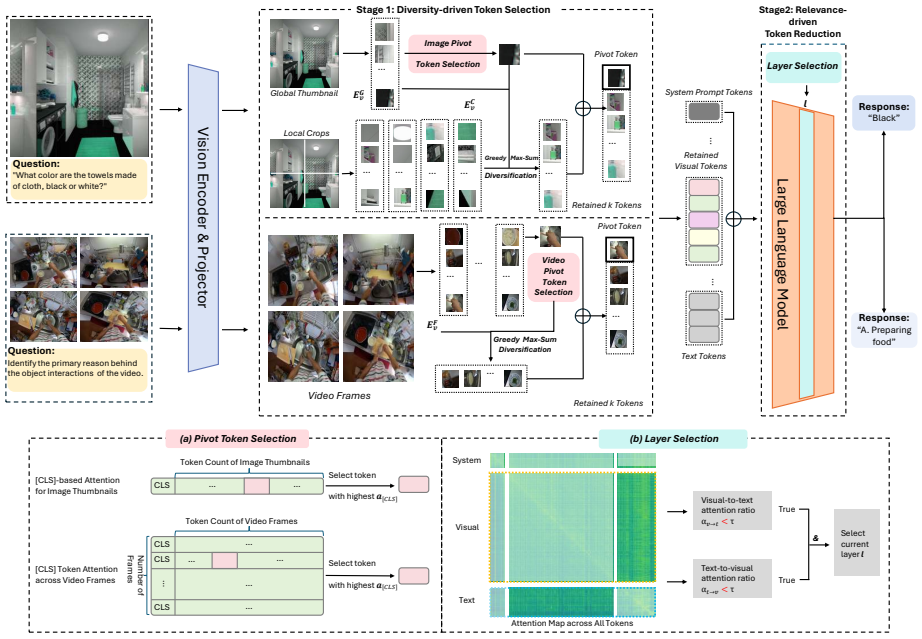

Fig. 3: Overall framework of ToDRE. Given the visual and textual inputs, the proposed *Diversity-driven Token Selection* first selects a pivot token from global thumbnail or video frames with `[CLS]`-based attention and then performs max-sum diversification to retain a diverse set of k visual tokens. The proposed *Relevance-driven Token Reduction* then dynamically identifies a pivot decoder layer and prunes all its visual tokens—the layer is identified if its visual-to-text and text-to-visual attention ratios both fall below a threshold τ. $\boldsymbol{E_v^G}$, $\boldsymbol{E_v^C}$, and $\boldsymbol{E_v^F}$ denote the embeddings of thumbnail, local crops, and video frames, respectively.

As shown in Figure 3-(a), pivot token selection proceeds as follows: (1) *Image Inputs with AnyRes [32] Support*: In this case, LVLM yields one global thumbnail G along with several local crops C. We compute the `[CLS]` attention score for each token in the global thumbnail and choose the token with the highest score as the pivot, since it captures the most comprehensive global information. (2) *Image Inputs without AnyRes Support*: The pivot token is selected from all visual tokens of the original image, using the same `[CLS]`-based criterion. (3) *Video Inputs*: We first identify, for each frame, the visual token with the highest `[CLS]` attention. The final pivot token is then selected as the one with the highest score among these frame-wise candidates.

For MLLMs without a `[CLS]` token in their encoders, a random selection strategy is also acceptable, as it yields performance that is nearly comparable to the original approach. We provide a detailed comparison of different pivot token selection strategies in the supplementary material.

Greedy Max-Sum Diversification. The expansion starts from the designated pivot. At iteration t, we pick a new token index $c^{(t)}$ by minimizing its *cumulative* similarity to the already selected set:

$$c^{(t)} \;=\; \underset{v \in V \setminus \mathcal{C}^{(t-1)}}{\arg\min} \left[\sum_{c \in \mathcal{C}^{(t-1)}} s(\mathbf{x}_v, \mathbf{x}_c) \right], \tag{8}$$

where $\mathbf{x}_v$ and $\mathbf{x}_c$ denote visual token features with indices v and c, and $\mathcal{C}^{(t-1)}$ is the selected set from the previous iteration. The similarity between two tokens is measured with cosine similarity

$$s(\mathbf{x}_v, \mathbf{x}_c) \;=\; \frac{\mathbf{x}_v^\top \mathbf{x}_c}{\|\mathbf{x}_v\| \, \|\mathbf{x}_c\|}. \tag{9}$$

Equivalently, (8) maximizes the *sum of distances* if $d(\cdot,\cdot) = 1 - s(\cdot,\cdot)$. After selecting $c^{(t)}$, we update the cumulative similarities by adding its contribution:

$$\forall v \in V \setminus \mathcal{C}^{(t)} : \quad S_v^{(t)} \;=\; S_v^{(t-1)} \;+\; s(\mathbf{x}_v, \mathbf{x}_{c^{(t)}}), \tag{10}$$

and mask the chosen index. This greedy procedure repeats until k diverse tokens (e.g., k=288, about 10% of visual tokens) are retained, yielding

$$\mathcal{C} \;=\; \{c^{(1)}, c^{(2)}, \ldots, c^{(k)}\}. \tag{11}$$

Finally, all remaining visual tokens are discarded; the retained visual tokens together with all text tokens are fed to the LLM decoder for inference.

4.2 Relevance-Driven Token Compression

While strategies involving partial or multi-stage pruning could be further applied, we argue that such strategies are unnecessary, since the majority of visual tokens have already been removed in the previous stage. We propose a forward-pass metric based on global cross-modal interaction to select the most appropriate decoder layer for token removal. As illustrated in Figure 3-(b), once this layer is identified, we remove all remaining visual tokens immediately after it.

Specifically, let L be the number of decoder layers in the LLM. Based on our observation that cross-modal interaction has largely concluded in the latter half of the decoder (Figure 2), we compute the global cross-modal attention ratios only at a few selected layers in the deeper portion of the LLM to determine whether information migration is complete. Since these attention ratios tend to remain stable across consecutive deeper layers, computing them at every layer would introduce unnecessary overhead. A more detailed ablation of layer selection can be found in the supplementary material. At each selected layer ℓ, we compute two cross-modal attention ratios based on average attention probabilities across all attention heads and tokens:

$$\alpha_{t\to v}^{(\ell)} = \frac{\sum_{i\in T}\sum_{j\in V} A_{ij}^{(\ell)}}{\sum_{i\in T}\sum_{j\in S\cup V\cup T} A_{ij}^{(\ell)}}, \quad \alpha_{v\to t}^{(\ell)} = \frac{\sum_{i\in V}\sum_{j\in T} A_{ij}^{(\ell)}}{\sum_{i\in V}\sum_{j\in S\cup V\cup T} A_{ij}^{(\ell)}}, \tag{12}$$

where $A_{ij}^{(\ell)}$ denotes the softmax-normalized attention weight from query token i to key token j at layer ℓ; S, V, and T represent the system prompt, visual, and textual tokens, respectively. If both ratios indicate that cross-modal interaction has diminished sufficiently, we remove all visual tokens from that layer onward, since the model has already absorbed the necessary visual information into text representations. This further eliminates redundant visual computation in the subsequent prefilling and decoding stages, yielding slight improvements in both efficiency and performance.

5 Experiments

Experimental Setting. We evaluate ToDRE over multiple prevalent LVLMs (including LLaVA-NeXT-7B/13B [33], Qwen2.5-VL-7B-Instruct [5], and InternVL2-8B [46]) and twelve widely adopted benchmarks (including eight on image understanding tasks and four on video understanding tasks). More details on the benchmarks, network backbones, and comparison methods can be found in the supplementary material.

5.1 Benchmarking

Image Understanding Tasks. In Table 1, we report ToDRE's performance on a range of image-understanding benchmarks at different token-retention ratios. First, under the same setup where 75% of visual tokens are pruned in Stage 1—matching competing methods—ToDRE further removes all remaining visual tokens in Stage 2 and achieves a 98.2% average score, outperforming the second-best method by 1.6%. Second, under more extreme compression (only 10% of visual tokens are retained), ToDRE surpasses the second-best approach by 1.5%. Third, ToDRE also achieves top performance on larger models, reaching an average score of 93.6% on the 13B variant—demonstrating strong adaptability across model scales. Note that FastV [10] and SparseVLM [57] are excluded from the 13B comparison, as their 7B-tuned pruning strategies degrade substantially when directly transferred to 13B. This further highlights the robustness and transferability of ToDRE.

Video Understanding Tasks. To further assess ToDRE's generalization ability, we evaluate it on both short- and long-form video understanding benchmarks. As shown in Table 2, ToDRE outperforms the baseline by 3.1% and 0.9% under the same token retention ratios used for images, and surpasses the second-best

Method	MME	ScienceQA	GQA	POPE	MMBench-EN	MMBench-CN	VizWiz	VQAv2	Average
Upper Bound, 2880 Tokens									
LLaVA-NeXT-7B [33]	1519.6	72.0	64.2	87.7	68.5	59.0	60.6	80.1	100.0%
Ratio=25%, Retain up to 720 Tokens									
FastV [10]	1477.3	69.8	60.4	83.1	65.6	55.4	57.2	77.2	95.4%
SparseVLM [57]	1446.1	67.5	60.9	71.0	63.8	55.4	58.6	77.2	93.1%
FasterVLM [56]	1454.6	67.1	61.3	87.2	66.0	56.8	58.4	76.4	96.0%
GlobalCom2 [35]	1468.7	68.1	61.4	**87.6**	64.0	54.4	58.7	76.6	95.6%
DivPrune [2]	1486.5	70.0	61.8	87.4	64.6	56.4	58.5	76.4	96.6%
ToDRE (Ours)	**1504.3**	**70.7**	**63.3**	87.5	**66.6**	**58.0**	**59.5**	**77.5**	**98.2%**
Ratio=10%, Retain up to 288 Tokens									
FastV [10]	1282.9	69.3	55.9	71.7	61.6	53.5	56.1	70.2	88.8%
SparseVLM [57]	1332.2	68.6	56.1	63.2	54.5	52.3	56.2	69.9	86.3%
FasterVLM [56]	1359.2	66.5	56.9	83.6	61.6	55.1	55.6	72.3	91.4%
GlobalCom2 [35]	1365.5	68.7	57.1	83.8	61.8	55.0	56.6	71.8	92.2%
DivPrune [2]	1396.2	69.3	59.2	84.3	63.1	55.7	56.6	73.2	93.5%
ToDRE (Ours)	**1464.4**	**70.2**	**59.4**	**85.0**	**63.3**	**57.8**	**56.7**	**74.4**	**95.0%**
LLaVA-NeXT-13B [33]	1575.2	71.2	65.4	87.5	70.1	66.0	63.6	81.9	100.0%
Ratio=25%, Retain up to 720 Tokens									
FasterVLM [56]	1516.1	71.1	62.3	86.1	67.6	62.1	58.1	76.1	95.6%
GlobalCom2 [35]	1531.2	71.4	62.7	86.5	67.9	61.3	58.2	77.2	96.0%
DivPrune [2]	1530.2	71.4	62.9	87.0	67.7	61.4	**60.6**	77.4	96.6%
ToDRE (Ours)	**1557.0**	**72.8**	**63.8**	**87.5**	**69.1**	**63.9**	57.6	**78.5**	**97.3%**
Ratio=10%, Retain up to 288 Tokens									
FasterVLM [56]	1386.2	70.5	58.1	81.6	61.7	53.5	55.9	77.1	89.1%
GlobalCom2 [35]	1399.5	71.0	58.3	82.4	65.0	56.6	55.6	72.8	90.8%
DivPrune [2]	1463.3	70.7	**60.1**	**86.5**	64.3	53.0	**59.1**	75.4	92.5%
ToDRE (Ours)	**1490.2**	**71.4**	59.9	83.7	**65.3**	**60.5**	56.9	**75.9**	**93.6%**

Table 1: Performance of training-free token compression methods across eight image-language benchmarks. "Average" denotes the mean performance ratio between each token compression method and the vanilla MLLM model. We evaluate all methods at retention ratios of 25% and 10%, with the best results highlighted in bold.

Method	Retain Ratio	# Token	VideoMME	Egoschema	MLVU	LongVideoBench	Average
LLaVA-NeXT-7B [33]	0%	2880	33.3	35.7	20.1	42.5	100.0%
FastV [10]	25%	720	32.3	31.2	16.5	40.3	90.4%
FasterVLM [56]			**33.8**	41.0	**19.6**	40.5	102.3%
DivPrune [2]			33.6	41.8	19.5	40.4	102.5%
ToDRE (Ours)			33.3	**42.4**	**19.6**	**41.0**	**103.1%**
FastV [10]	10%	288	30.4	30.4	11.2	38.7	80.8%
FasterVLM [56]			**34.3**	36.1	**19.1**	36.9	96.4%
DivPrune [2]			33.3	40.9	18.9	40.3	100.7%
ToDRE (Ours)			33.2	**41.9**	18.2	**40.7**	**100.9%**

Table 2: Performance of training-free token compression methods across four video-language benchmarks.

method by 0.6% and 0.2%, respectively. Interestingly, ToDRE even surpasses the baseline model in some cases. We attribute this to the reduced interference from redundant visual tokens, which may otherwise suppress task-relevant information during inference. Similarly, SparseVLM is excluded due to limited transferability, and GlobalCom2 [35] is omitted as it targets image-only inputs. In contrast, ToDRE generalizes well across modalities and model scales.

Benchmark	Qwen2.5-VL-7B-Instruct [5]			InternVL2-8B [46]		
	Original	**Ret. 25%**	**Ret. 10%**	**Original**	**Ret. 25%**	**Ret. 10%**
MME [14]	1687.7	1680.6	1573.9	1628.8	1566.4	1424.4
ScienceQA [37]	88.5	86.4	85.4	96.5	95.3	92.5
GQA [19]	60.9	57.3	52.9	62.8	56.9	52.1
POPE [28]	87.7	85.4	80.8	87.8	83.7	76.9
MMBench-EN [36]	82.9	79.9	72.8	81.2	77.2	71.4
MMBench-CN [36]	81.7	78.0	70.4	80.0	74.2	67.6
VizWiz [17]	70.6	68.7	66.7	60.6	58.6	56.3
VQAv2 [16]	82.9	78.3	72.8	79.0	72.8	68.6
VideoMME [15]	61.5	59.4	57.3	55.0	54.2	52.6
Egoschema [38]	58.3	57.3	55.8	55.9	55.0	52.5
MLVU [59]	59.3	58.7	57.2	47.3	52.8	54.0
LongVideoBench [51]	58.4	57.7	54.8	55.0	52.3	48.8
Average	100.0%	97.1%	92.0%	100.0%	96.8%	91.5%

Table 3: Performance of ToDRE on Qwen2.5-VL-7B-Instruct and InternVL2-8B. Benchmarks as rows. "Ret." = Retention Ratio. Averages are normalized to each model's Original (=100%).

Method	**FLOPs ↓ (T)**	**Memory ↓ (GB)**	**Throughput ↑ (samples/s)**	**Performance ↑**
Upper Bound, 2880 Tokens				
LLaVA-NeXT-7B [33]	31.4	15.9	1.5	100%
Ratio=10%, Retain up to 288 Tokens				
FastV [10]	8.2 (↓73.9%)	14.1 (↓11.3%)	2.1 (1.4×)	88.8%
SparseVLM [57]	6.9 (↓78.0%)	14.1 (↓11.3%)	2.5 (1.7×)	86.3%
FasterVLM [56]	6.1 (↓80.6%)	**13.6** (↓14.5%)	2.7 (1.8×)	91.4%
GlobalCom^2 [35]	6.1 (↓80.6%)	13.9 (↓12.6%)	2.7 (1.8×)	92.2%
DivPrune [2]	**6.0** (↓80.9%)	**13.6** (↓14.5%)	2.8 (1.9×)	93.5%
ToDRE (Ours)	**6.0** (↓80.9%)	**13.6** (↓14.5%)	**2.9** (1.9×)	**95.0%**

Table 4: Inference efficiency comparisons. All experiments were conducted on a single NVIDIA RTX 3090 GPU. "Memory": peak GPU memory usage; "Throughput": number of POPE samples processed per second; "Performance": average score across 8 image understanding benchmarks.

Cross-Model Evaluation. As shown in Table 3, we further evaluate ToDRE on Qwen and InternVL backbones. Specifically, ToDRE retains 97.1% and 96.8% of the original performance on Qwen2.5-VL-7B-Instruct and InternVL2-8B at a 25% retention ratio, respectively, and still maintains more than 90% of the original performance even when only 10% of visual tokens are preserved, demonstrating strong robustness across different model architectures.

Method	Total Time ↓ (Min:Sec)	MME	ScienceQA	GQA	POPE	Average
Upper Bound, 2880 Tokens						
LLaVA-NeXT-7B [33]	77:04	1519.6	72.0	64.2	87.7	100.0%
Stage 2 only	70:15	1522.7	71.9	64.3	87.6	100.0%
Ratio=25%, Retain up to 720 Tokens						
Stage 1 only	48:10	1503.8	70.6	63.1	**87.5**	98.8%
Stage 1 + Stage 2 (ToDRE)	**44:18**	**1504.3**	**70.7**	**63.3**	**87.5**	**98.9%**
Ratio=10%, Retain up to 288 Tokens						
Stage 1 only	31:18	1458.6	70.4	**59.4**	**85.0**	95.8%
Stage 1 + Stage 2 (ToDRE)	**29:43**	**1469.3**	**70.5**	**59.4**	**85.0**	**96.0%**

Table 5: Ablation study on two-stage token compression. We evaluated the individual and combined effects of the proposed two-stage pruning pipeline under retention ratios of 25% and 10%.

5.2 Efficiency

As shown in Table 4, we compare FLOPs, peak memory, throughput, and accuracy across token pruning methods at a fixed 10% retention ratio. ToDRE achieves the highest throughput on POPE [28] (2.9 samples/s, 1.9× over vanilla LLaVA-NeXT-7B) while matching the lowest peak memory (13.6 GB) with FasterVLM and DivPrune [2]. Despite its efficiency, ToDRE also attains the best average accuracy (95.0%), exceeding the second-best method by 1.5%, demonstrating a strong balance among speed, memory, and performance. In addition, as discussed in Section 3.1, since most benchmarks require only short answers (small L), decoding-stage gains are modest. We expect larger benefits for long-form generation where visual tokens otherwise dominate computation.

5.3 Ablation Study

We conduct ablation studies to evaluate individual and combined contributions of the two stages in our framework. As shown in Table 5, applying Stage 2 only already reduces the overall inference time by 8.8% on LLaVA-NeXT-7B baseline (from 77:04 to 70:15), while maintaining a lossless average performance of 100.0%. The limited efficiency gain is expected, as Stage 2 only accelerates the latter part of inference, and most tasks involve generating very short outputs.

In contrast, applying Stage 1 only yields substantial time savings of 37.5% (48:10) and 59.4% (31:18), respectively, with minimal drops in performance. When incorporating both stages (Stage 1 + Stage 2), we observe consistent improvements: First, at the 25% ratio, performance improves from 98.8% to 98.9% with total time reduced (from 48:10 to 44:18). Second, at the 10% ratio, performance increases from 95.8% to 96.0%, with total time reduced (from 31:18 to 29:43). Overall, ToDRE reduces inference time by 42.5% and 61.4% at the 25% and 10% token retention ratios, respectively, while even improving performance

(up to +0.2% gain). These results show that Stage 2 complements the diversity-based Stage 1, improving the accuracy-efficiency trade-off across compression settings.

6 Conclusion

In this work, we systematically analyze redundancy in LVLM inference and identify two key inefficiencies: (1) redundant visual tokens that inflate intra-modal computation, and (2) tokens that contribute little cross-modal information during decoding. To address these inefficiencies, we propose ToDRE, a training-free, architecture-agnostic framework that first selects a maximally diverse subset of visual tokens via a greedy max-sum diversification algorithm, then removes all remaining visual tokens once cross-modal attention fades. Experiments on twelve image- and video-language benchmarks show that ToDRE prunes up to 90% of visual tokens while preserving 95.0% of the original performance, achieving 2.6× faster inference and 14.5% lower memory usage than uncompressed baselines.

Acknowledgment

This study is funded by the Ministry of Education, Singapore, under the Tier-2 project scheme with project number MOE-T2EP20123-0003.

We thank Yichao Gao for helpful discussions and valuable suggestions on the mathematical derivations in this paper.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: Proceedings of the International Conference on Learning Representations (2023)
7. Borodin, A., Jain, A., Lee, H.C., Ye, Y.: Max-sum diversification, monotone submodular functions and dynamic updates. arXiv preprint arXiv:1203.6397 (2012), `https://arxiv.org/abs/1203.6397`, accessed 28 June 2026

8. Cai, M., Yang, J., Gao, J., Lee, Y.J.: Matryoshka multimodal models. In: Workshop on Video-Language Models@ NeurIPS 2024 (2024)
9. Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., Wang, C., Bai, X.: Llava-kd: A framework of distilling multimodal large language models. arXiv preprint arXiv:2410.16236 (2024)
10. 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)
11. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
12. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. LMSYS Org Blog (2023), `https://lmsys.org/blog/2023-03-30-vicuna/`, accessed 28 June 2026
13. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Advances in neural information processing systems. vol. 35, pp. 16344–16359 (2022)
14. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 **abs/2306.13394** (2023), `https://arxiv.org/abs/2306.13394`, accessed 28 June 2026
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
16. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
17. 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)
18. Han, Y., Liu, X., Zhang, Z., Ding, P., Chen, J., Wang, D., Chen, H., Yan, Q., Huang, S.: Filter, correlate, compress: Training-free token reduction for mllm acceleration. arXiv preprint arXiv:2411.17686 (2025), aAAI 2026
19. 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)
20. 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)
21. Jiang, Y., Wu, Q., Lin, W., Yu, W., Zhou, Y.: What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. arXiv preprint arXiv:2501.02268 (2025)
22. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)

23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
24. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
25. Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., Zhang, L.: Tokenpacker: Efficient visual projector for multimodal llm. arXiv preprint arXiv:2407.02392 (2024)
26. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
27. 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. arXiv preprint arXiv:2403.18814 (2024)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (2023), https://arxiv.org/abs/2305.10355, accessed 28 June 2026
29. Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., Xie, P.: Not all patches are what you need: Expediting vision transformers via token reorganizations. In: Proceedings of the International Conference on Learning Representations (2022)
30. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
31. Lin, Z., Lin, M., Lin, L., Ji, R.: Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5334–5342 (2025)
32. 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)
33. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/ (2024), accessed 28 June 2026
34. Liu, T., Shi, L., Hong, R., Hu, Y., Yin, Q., Zhang, L.: Multi-stage vision token dropping: Towards efficient multimodal large language model. arXiv preprint arXiv:2411.10803 (2024)
35. Liu, X., Wang, Z., Chen, J., Han, Y., Wang, Y., Yuan, J., Song, J., Huang, S., Chen, H.: Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. arXiv preprint arXiv:2501.05179 (2025), accepted by AAAI 2026
36. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? In: European Conference on Computer Vision (2024), https://arxiv.org/abs/2307.06281, accessed 28 June 2026
37. 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. In: Advances in Neural Information Processing Systems. vol. 35, pp. 2507–2521 (2022)

38. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: Advances in Neural Information Processing Systems. vol. 36, pp. 46212–46244 (2023)
39. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. MIT press (2010)
40. Qiao, Y., Yu, Z., Guo, L., Chen, S., Zhao, Z., Sun, M., Wu, Q., Liu, J.: Vl-mamba: Exploring state space models for multimodal learning. arXiv preprint arXiv:2403.13600 (2024)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
42. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388 (2024)
43. Shukor, M., Cord, M.: Skipping computations in multimodal llms. arXiv preprint arXiv:2410.09454 (2024), accepted at NeurIPS 2024 Workshop RBFM
44. 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)
45. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
46. Team, I.: Internvl 2.0: Scaling open-source multimodal large language models. `https://internvl.github.io/blog/2024-07-02-InternVL-2.0/` (2024), accessed 28 June 2026
47. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
48. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. 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. arXiv preprint arXiv:2502.11494 (2025)
51. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: Advances in Neural Information Processing Systems. vol. 37, pp. 28828–28857 (2024)
52. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
53. Xie, J., Zhang, Y., Lin, M., Cao, L., Ji, R.: Advancing multimodal large language models with quantization-aware scale learning for efficient adaptation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10582–10591 (2024)
54. Yao, L., Li, L., Ren, S., Wang, L., Liu, Y., Sun, X., Hou, L.: Deco: Decoupling token compression from semantic abstraction in multimodal large language models. arXiv preprint arXiv:2405.20985 (2024)

55. 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. pp. 22128–22136 (2025)
56. Zhang, Q., Cheng, A., Lu, M., Zhuo, Z., Wang, M., Cao, J., Guo, S., She, Q., Zhang, S.: [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. arXiv e-prints pp. arXiv–2412 (2024)
57. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. In: International Conference on Machine Learning (2025)
58. Zhao, S., Wang, Z., Juefei-Xu, F., Xia, X., Liu, M., Wang, X., Liang, M., Zhang, N., Metaxas, D.N., Yu, L.: Accelerating multimodal large language models by searching optimal vision token reduction. arXiv preprint arXiv:2412.00556 (2024)
59. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: Benchmarking multi-task long video understanding. arXiv preprint arXiv:2406.04264 (2024)
60. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
61. Zhu, Y., Xie, C., Liang, S., Zheng, B., Guo, S.: Focusllava: A coarse-to-fine approach for efficient and effective visual token compression. arXiv preprint arXiv:2411.14228 (2024)