CATP: Contextually Adaptive Token Pruning for Efficient and Enhanced Multimodal In-Context Learning

Yanshu Li[1], **Jianjiang Yang**[2], **Zhennan Shen**[1], **Ligong Han**[3], **Haoyan Xu**[4], **Ruixiang Tang**[5]*

[1]Brown University
[2]University of Bristol
[3]MIT-IBM Watson AI Lab
[4]University of Southern California
[5]Rutgers University
yanshu_li1@brown.edu, ruixiang.tang@rutgers.edu

Abstract

Modern large vision-language models (LVLMs) convert each input image into a large set of tokens that far outnumber the text tokens. Although this improves visual perception, it also introduces severe image token redundancy. Because image tokens contain sparse information, many contribute little to reasoning but greatly increase inference cost. Recent image token pruning methods address this issue by identifying important tokens and removing the rest. These methods improve efficiency with only small performance drops. However, most of them focus on single-image tasks and overlook multimodal in-context learning (ICL), where redundancy is higher and efficiency is more important. Redundant tokens weaken the advantage of multimodal ICL for rapid domain adaptation and lead to unstable performance. When existing pruning methods are applied in this setting, they cause large accuracy drops, which exposes a clear gap and the need for new approaches. To address this, we propose Contextually Adaptive Token Pruning (CATP), a training-free pruning method designed for multimodal ICL. CATP uses two stages of progressive pruning that fully reflect the complex cross-modal interactions in the input sequence. After removing 77.8% of the image tokens, CATP achieves an average performance gain of 0.6% over the vanilla model on four LVLMs and eight benchmarks, clearly outperforming all baselines. At the same time, it improves efficiency by reducing inference latency by an average of 10.78%. CATP strengthens the practical value of multimodal ICL and lays the foundation for future progress in interleaved image-text settings.

Extended version — https://arxiv.org/abs/2508.07871

Introduction

Large vision-language models (LVLMs) have recently shown strong performance across a wide range of visual-language tasks (Liu et al. 2023a; Chen et al. 2024d; Wang et al. 2024). To equip LVLMs with new domain capabilities while reducing the high cost of multimodal data preparation and training, multimodal in-context learning has become an important extension (Doveh et al. 2024; Chen et al. 2024c). It allows a model to adapt on the fly by adding a small set of exemplars, known as in-context demonstrations (ICDs), directly into the prompt. This paradigm provides a simple path toward task generalization without any parameter updates, making it increasingly attractive for real-world use (Li et al. 2025a).

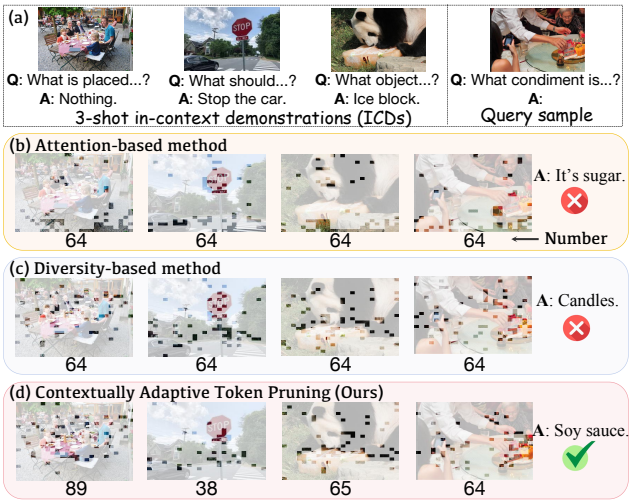

Figure 1. (a) Example of a 3-shot in-context sequence. (b-d) show the effects of three image token pruning methods on this sequence. Attention-based methods tend to keep tokens at the bottom of the image due to attention sinks in interleaved inputs. Diversity-based methods lack semantic guidance from the entire context. In contrast, our proposed method **CATP** accounts for the complex cross-modal interactions within the sequence, leading to superior results.

Although multimodal ICL offers a user-friendly and rapid solution, its input characteristics often impede the desired effects. As shown in Figure 1(a), every ICD and the subsequent query sample include an image, so the image token redundancy that is already a bottleneck in single-image tasks becomes even more acute. For instance, LLaVA-Next (Liu et al. 2024) converts each image into 576 tokens. On VizWiz (Gurari et al. 2018), a 2-shot ICL run requires 3.2 times the computation of single-image inference and an impressive 14.3 times that of text-only inference. The resulting latency and

*Corresponding Author.

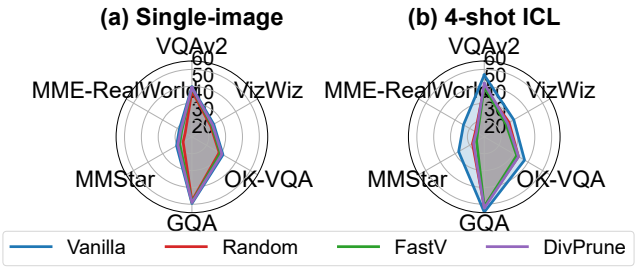

Figure 2. Performance comparison of three token pruning methods under a 77.8% per-image pruning ratio, shown for single-image versus 4-shot ICL across six benchmarks.

memory footprint conflict with the efficiency users expect. Moreover, the interleaved format introduces additional issues. Multimodal ICL involves several kinds of cross-modal interactions, and many uninformative tokens can complicate this process (Baldassini et al. 2024; Li et al. 2025c). They prevent the model from focusing on the visual regions in each image that match the paired text. They also make it harder to locate the key information provided by the user based on the query sample (Zhou et al. 2024; Chen et al. 2025).

To ease this serious redundancy, training-free image token pruning has become a promising direction. Its core idea is to define an importance criterion that measures the contribution of each image token during inference (Kim et al. 2022; Wen et al. 2025a). By retaining only the most informative tokens and removing or merging the rest, latency is effectively reduced with minimal accuracy loss. Existing pruning strategies fall mainly into two categories. (1) **Diversity-based** methods operate after the vision encoder and projector but before the decoder, when image tokens have been projected into the decoder's representation space yet have not interacted with texts (Wen et al. 2025b; Jeddi et al. 2025). These methods measure feature similarity among image tokens to penalize redundancy and encourage diversity, as exemplified by DivPrune (Alvar et al. 2025). (2) **Attention-based** methods act inside the LLM decoder, where visual and textual embeddings are processed together (Zhang et al. 2024a; Arif et al. 2025; Xu et al. 2025). They estimate the importance of each image token by the amount of attention it receives from other tokens, as exemplified by FastV (Chen et al. 2024b).

Despite being able to reduce up to 90% of the image tokens with less than a 5% drop in performance, the above two pruning methods are developed and evaluated only in single-image tasks and thus ignore the much more demanding multi-image scenarios, especially interleaved image-text ICL. When we adapt FastV and DivPrune to multimodal ICL, their shortcomings become clear, as shown in the visualization of Figure 1 and the results on LLaVA-Next-7B in Figure 2. Both methods incur greater performance degradation in 4-shot ICL compared to the single-image setting, and in many cases even underperform random pruning. This highlights a non-negligible gap between single-image tasks and multimodal ICL. To investigate this gap, we conduct further ICL experiments with both pruning methods under varied setups and obtain two key findings. First, in multimodal ICL, diversity-based methods fail to perform fine-grained pruning as they lack information from other images and texts, and thus cannot capture diverse cross-modal interactions. Second, while layer choice matters for attention-based pruning, attention shifts (Zhang et al. 2025) in every image bias all its variants toward text-adjacent tokens, and the bias compounds to reduce accuracy. To this end, we question that

"Can we identify the image tokens that contribute the most to building a complete and effective in-context sequence amid complex cross-modal interactions?"

To approach this target, we propose Contextually Adaptive Token Pruning (CATP), a training-free image token pruning method designed for multimodal ICL. CATP removes a ratio R of the total image tokens in an in-context sequence. Based on our analysis of existing methods, CATP is divided into two stages. Stage 1 is applied between the projector and the decoder, and is tailored to the characteristics of multimodal ICL. It evaluates the token importance in two dimensions. The first is semantic alignment with the paired text, which ensures that key visual regions are retained as a foundation for effective ICL. The second is feature diversity, which prevents the tokens that matter more to the entire sequence from being discarded. Stage 2 acts on the shallow decoder layers. It adopts a novel progressive adaptation strategy that first treats all ICD image tokens as an integral context. By combining semantic relevance with layer-wise attention differences, it precisely identifies the most important context tokens. In the following layer, it prunes the query image tokens based on their semantic relevance to this distilled context, ensuring effective ICL after pruning. Extensive experiments show that CATP is both effective and generalizable in multimodal ICL. It improves efficiency and performance at the same time and provides clear advantages over all existing baselines.

The contributions of this paper are summarized below:

- By testing two mainstream image token pruning methods in the multimodal ICL setting, we reveal that their effectiveness degrades due to the interleaved image-text nature of the input, which amplifies their limitations.
- We propose Contextually Adaptive Token Pruning (CATP), the first image token pruning method designed for multimodal ICL, filling the important gap. CATP is training-free and can capture multiple types of interaction to select the tokens most critical for overall ICL.
- Experiments on four LVLMs and eight benchmarks show that while all other baselines lead to performance degradation, CATP not only improves the efficiency of multimodal ICL but also improves its performance.

Related Work

Large Vision-language Models (LVLMs). The rapid rise of large language models (LLMs) gives birth to LVLMs (Li et al. 2024b,a; Chen et al. 2024d; Wang et al. 2024). A typical LVLM comprises **three** key modules: a vision encoder, a projector, and an LLM decoder. Although this architecture is generally effective, it has been found to be inefficient (Liu et al. 2023b; Li et al. 2024c). Input images usually occupy a large number of tokens, and these tokens carry much sparser information than text tokens, creating severe image token

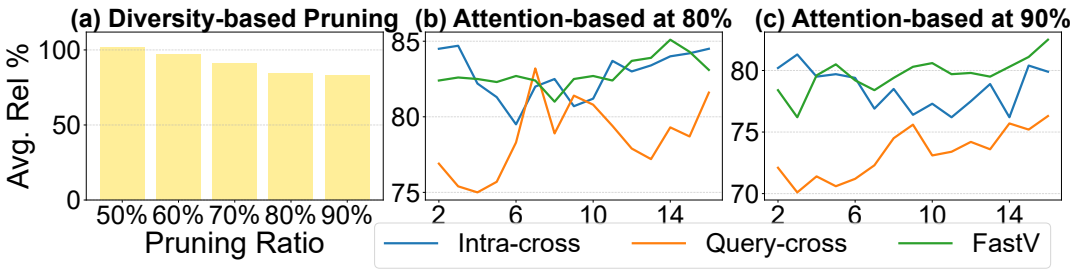

Figure 3. Average relative performance of pruning methods in 4-shot ICL: (a) showing diversity-based pruning at 50%-90% ratios; (b-c) showing three different attention-based pruning methods applied to various layers at 80% and 90% ratios.

redundancy (Hu et al. 2024; Xing et al. 2024). For example, LLaVA-1.5 (Li et al. 2024b) converts each image into 576 tokens, while LLaVA-Next (Liu et al. 2024) and Qwen2.5-VL (Bai et al. 2025) increase the count up to several thousand tokens per image to handle high-resolution inputs.

In-context Learning (ICL). ICL refers to a model's ability to solve new tasks by conditioning on a sequence of in-context demonstrations (ICDs) without updating any parameters (Brown et al. 2020; Dong et al. 2024; Wei et al. 2023). Because of its convenience, ICL has been widely used in NLP and extended to the multimodal setting. Current-generation LVLMs, including the InternVL and QwenVL series, already treat ICL as a core capability, which highlights its strong practical value (Chen et al. 2024d; Wang et al. 2024). However, the ideal efficiency of ICL is often reduced by image token redundancy, and the complex interactions in multimodal ICL make this problem even harder to address.

Image Token Pruning. Researchers have begun to tackle token redundancy by image token pruning. FastV (Chen et al. 2024b) first proposes to prune image tokens based on the attention they obtain in an early decoder layer. Following FastV, more attention-based methods have appeared (Zhang et al. 2024a; Arif et al. 2025). VTW (Lin et al. 2025) proposes a token withdrawal strategy, removing all image tokens after a chosen layer. FitPrune (Ye et al. 2025) further leverages both cross-attention and self-attention of image tokens in every layer to measure their importance. Meanwhile, diversity-based methods argue that semantic diversity among image tokens is a better importance criterion than attention scores (Wen et al. 2025b; Jeddi et al. 2025). DivPrune (Alvar et al. 2025) performs pruning before image tokens enter the decoder and frames it as a Max-Min Diversity Problem. Although all these methods are effective, they were designed for single-image tasks and overlook multimodal ICL. We find that their known limitations (Wen et al. 2025a; Zhang et al. 2025) become even more pronounced in this setting.

Method

Preliminary and Motivation

Consider an LVLM parameterized by θ and an n-shot in-context sequence ICS that contains n ICDs with both image and text, plus one query sample. We denote the sequence as:

$$ICS = \{(I_1, T_1), \ldots, (I_n, T_n), (I_q, T_q)\}, \quad (1)$$

where I_q and T_q are the image and text of the query sample. In the prefilling stage, each i-th image in ICS is first passed through a vision encoder f_θ^v and then a projector g_θ, producing a set of image token embeddings:

$$X_i^I = g_\theta(f_\theta^v(I_i)) \in \mathbb{R}^{S_i^I \times D}, \quad (2)$$

where D denotes the dimension of the LLM decoder ϕ^θ and S_i^I is the number of tokens produced from each image. This number varies across LVLMs and is usually user-configurable. **Diversity-based** pruning methods are typically applied at this stage to reduce S_i^I. Meanwhile, each i-th text segment is converted into a set of text token embeddings $X_i^T \in \mathbb{R}^{S_i^T \times D}$ by ϕ^θ. After inserting the image tokens at their corresponding positions, we obtain the input token sequence:

$$X = (X_{sys}, X_1^I, X_1^T, \ldots, X_n^I, X_n^T, X_{n+1}^I, X_{n+1}^T), \quad (3)$$

where X_{sys} denotes the system prompt and the query sample is indexed as n+1 for unified notation. The sequence X is then fed into ϕ^θ and processed by an N-layer decoder to produce the final answer. Each layer contains a multi-head attention module (MHA) and a feedforward MLP. The h-th head in the l-th layer maps the hidden states of X_{ICS} to queries $Q_{l,h}$, keys $K_{l,h}$, and values $V_{l,h}$ by linear transformations, and the head's attention weight matrix can be represented as:

$$\mathbf{A}_{l,h} = softmax\left(\frac{Q_{l,h} K_{l,h}^\top + \mathbf{M}}{\sqrt{D_k}}\right) \in \mathbb{R}^{S \times S} \quad (4)$$

where S is the length of X and $\mathbf{M}$ is the causal mask. D_k is the head dimension of ϕ^θ. $K_{l,h}$ and $V_{l,h}$ are stored in the KV cache during the prefilling stage to prepare for subsequent decoding. **Attention-based** pruning methods are applied here to leverage $\mathbf{A}_{l,h}$ to discard less informative image tokens, preventing them from entering later layers. Both methods aim to identify the tokens whose removal minimizes the adverse effects on θ's decoding, given a specific pruning ratio.

To further examine the gap revealed in Figure 2, we conduct additional 4-shot ICL evaluations on the six benchmarks using LLaVA-Next-7B and report the average relative performance (Avg. Rel) compared with the vanilla model. We first adapt DivPrune to ICL with pruning ratios ranging from 50% to 90%. As Figure 3(a) shows, it performs well at relatively low ratios and at 50% it even outperforms the vanilla model. When the ratio increases from 60% to 70%, its performance

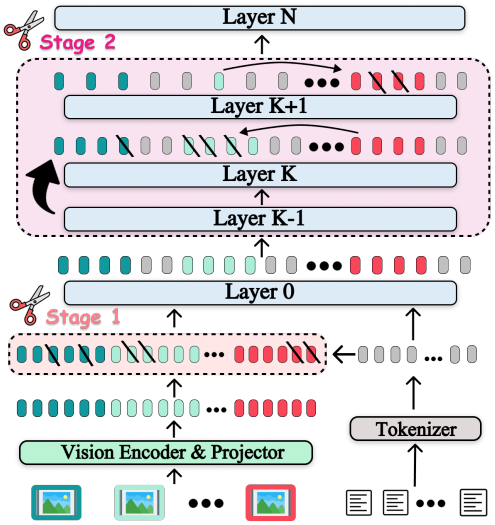

Figure 4. An overview pipeline of CATP.

drops sharply and continues to decline. These results indicate that diversity-based methods **fail to deliver fine-grained pruning in multimodal ICL**. In these methods, each image token interacts only with tokens from the same image, so no information from other images or texts is considered. Meanwhile, multimodal ICL involves much richer cross-modal interactions than single-image tasks, making this limitation more pronounced. Nonetheless, they remain useful as an **initial filter** before the decoder.

Next, we evaluate FastV by pruning 80% and 90% of image tokens at different decoder layers ranging from 2 to 16. FastV ranks tokens by the total attention they receive from all other tokens in a chosen layer. Considering the property of multimodal ICL, we introduce two alternatives: *intra-cross* measures the attention that each image token receives from all text tokens within its own image-text pair, and *query-cross* measures the attention that each image token receives from all tokens in the query sample. Intra-cross performs best in very shallow layers, such as layer 2, where the model is focused on feature perception and alignment. As the layer index increases, the effectiveness of intra-cross declines while FastV becomes superior. Query-cross shows a sharp rise in the shallow layers, roughly layers 7 to 10, indicating that the LVLM shifts to query-guided reasoning after perception.

However, none of the three methods reaches the desired performance in ICL. We attribute this to **attention shifts** that cause the model to focus excessively on image tokens positioned closer to the text segments (Zhang et al. 2025). In interleaved settings, these shifts accumulate and make the attention signal even less indicative of true importance, as illustrated in Figure 1(b). Hence, **statically relying on the attention assigned to image tokens is suboptimal** for pruning in multimodal ICL.

Contextually Adaptive Token Pruning

Overview To bridge the gap left by existing pruning methods in multimodal ICL, we propose Contextually Adaptive Token Pruning (CATP). Figure 4 outlines its pipeline. Given a target pruning ratio R, CATP retains only $(1 - R)$ of the total image tokens for decoding, which effectively reduces inference latency while keeping ICL performance nearly unaffected. Guided by the preceding analysis, CATP proceeds in two stages to deliver progressive and thorough pruning. Each stage removes $R/2$ of the total image tokens in X. Stage 1 is applied between the projector and the decoder. It measures the importance of image tokens using feature distribution diversity and alignment with their paired text. Stage 2 operates in two shallow decoder layers. It first prunes the ICD image tokens by combining inter-layer attention differences with semantic relevance, and then prunes the query sample's image tokens in the following layer.

Stage 1 Due to the multiple images in the input ICS, the converted token sequence X is lengthy, causing the decoder to be hampered during prefilling and disrupting the intended attention flow. Thus, Stage 1 of CATP is performed directly after the projector to prune the least informative image tokens before they reach the decoder, maximizing overall precision and efficiency. We design a targeted strategy to keep the image tokens entering the decoder relatively informative.

In multimodal ICL, each image is paired with text, so semantically aligned tokens are crucial for reasoning. However, the whole ICL process can depend more on features that interact with other ICDs or the query sample, such as complementary cross-sample cues. Thus, we measure the importance of each image token in this stage in two dimensions. The first dimension assesses the **alignment** between each image token and its paired text. The advantage of considering alignment at this stage is two-fold. After reducing the number of tokens, it promotes correct attention in the decoder by easing attention shifts (Li et al. 2025b). It also captures a key interaction in multimodal ICL that occurs within each image-text pair. This interaction is hard to isolate in the decoder as the forward pass merges semantics. The second assesses the **diversity** of the information carried by the image tokens. This keeps the retained tokens close to the feature distribution of the complete set and prevents tokens that contribute to the entire sequence from being removed. This matches the diversity-based methods, which are found to be suitable for initial filtering. To unify the two dimensions, we frame the task as a **weighted submodular maximization** problem.

Formally, for each i-th visual embedding set X_i^I, our goal is to select an optimal subset $Y_i^* \subseteq X_i^I$ that maximizes a composite objective function $\mathcal{F}(Y_i)$. This function is a weighted sum of the alignment and diversity scores. The optimization problem is defined as:

$$Y_i^* = \arg \max_{\substack{Y_i \subseteq X_i^I, \\ |Y_i| = \lfloor S_i^I \cdot (1-(R/2)) \rfloor}} \left(\mathcal{F}_{\text{div}}(Y_i) + \lambda_1 \cdot \mathcal{F}_{\text{align}}(Y_i) \right), \quad (5)$$

where λ_1 is a hyperparameter that balances the two objectives. The alignment function $\mathcal{F}_{\text{align}}(Y_i)$ quantifies how well the selected image tokens align with the semantics of their paired text segment X_i^T. We first compute a single vector representing the i-th text segment's semantics, $\bar{\mathbf{v}}_i$, by averaging all embeddings in X_i^T. Then, we obtain $\mathcal{F}_{\text{align}}(Y_i)$ using the similarity score between each image token $x_{i,j} \in X_i^I$ and $\bar{\mathbf{v}}_i$:

$$\mathcal{F}_{\text{align}}(Y_i) = \sum_{x_{i,j} \in Y_i} \hat{\text{sim}}(x_{i,j}, \bar{\mathbf{v}}_i), \tag{6}$$

where $\hat{\text{sim}}(\cdot)$ denotes the cosine similarity rescaled into the range $[0, 1]$. The diversity function $\mathcal{F}_{\text{div}}(Y_i)$ is modeled with a facility location objective that drives the tokens in Y_i to represent as diverse a set of features as possible. This function is a monotone submodular set function, measuring the total similarity between all tokens in the ground set X_i^I and their closest token in the chosen subset Y_i:

$$\mathcal{F}_{\text{div}}(Y_i) = \sum_{x \in X_i^I} \max_{y \in Y_i} \hat{\text{sim}}(y, x). \tag{7}$$

Since $\mathcal{F}_{\text{div}}(Y_i)$ is submodular and $\mathcal{F}_{\text{align}}(Y_i)$ is a modular function, their linear combination in Eq. 5 remains submodular. This property allows us to efficiently find a near-optimal pruning strategy using a standard greedy algorithm, making stage 1 a lightweight and effective pre-decoder filter.

Stage 2 After Stage 1, the decoder receives a pruned token sequence whose image tokens equal $(1 - R/2)$ of those in X. Only inside the decoder can we capture the complex interactions in multimodal ICL that go beyond image-text alignment via hidden states and attention patterns. Figure 3(b-c) shows the sharp rises of query-cross in the shallow layers, indicating that LVLM begins to rely more on the query sample to locate key context features at this depth, making query-context interactions more salient. Thus, we apply Stage 2 here. As static single-layer attention fails to measure importance, we propose a novel strategy, **progressive adaptation**.

We set the starting layer index of progressive adaptation to K. At this layer, we treat all ICDs as an integral context and prune their image tokens altogether until the overall ratio R is reached. Thus, unlike existing methods, the pruning ratio inside each ICD is not fixed. It adapts to the importance of each ICD to the context as a whole. Because not every ICD contributes equally to ICL and some are even detrimental, this strategy preserves post-pruning performance. We use the query sample as a lens to prune the context. First, we compute a focused query vector $\bar{\mathbf{v}}_{\text{q}}^K$ by applying average pooling to the hidden states of all tokens in the query sample at layer K. We then score each context image token c by comprehensively considering its relevance to the query and its dynamic growth in importance. This growth is measured by the change in attention it receives from the query sample's last token t_{ql} between layers:

$$\Delta\mathcal{A}(c) = \max\left(0, \mathbf{A}_K[idx_{ql}, idx_c] - \mathbf{A}_{K-1}[idx_{ql}, idx_c]\right), \tag{8}$$

where $\mathbf{A}_l$ is the l-th layer's attention weight matrix averaged over all heads. idx_{ql} and idx_c denote the token index of t_{ql} and c, respectively. The full importance score is defined as:

$$\mathcal{S}_{\text{context}}(c) = \Delta\mathcal{A}(c) + \lambda_2 \cdot \text{sim}(\mathbf{h}_c^K, \bar{\mathbf{v}}_{\text{q}}^K), \tag{9}$$

where $\mathbf{h}_c^K$ denotes the hidden states of c at layer K and λ_2 is a hyperparameter. We globally rank all context image tokens by this score and prune them in ascending order. This step simultaneously distills the large context, allowing the LVLM to conduct more targeted and in-depth reasoning.

Next, at layer $K + 1$ we perform context-guided pruning on the query sample's image tokens. We apply average pooling to the hidden states of the pruned context at this layer to obtain a distilled context vector $\bar{\mathbf{v}}_c^{K+1}$. To maintain a focused mechanism, the importance of each query image token q is only determined by its relevance to the distilled context:

$$\mathcal{S}_{\text{query}}(q) = \text{sim}(\mathbf{h}_q^{K+1}, \bar{\mathbf{v}}_{\text{c}}^{K+1}), \tag{10}$$

where $\mathbf{h}_q^{K+1}$ is the hidden states of q at layer $K + 1$. After pruning the query sample's image tokens with this score, CATP is complete. Only $1 - R$ of the total image tokens in X proceeds to layer $K + 2$ and all subsequent inference.

Experiments

Setup

Benchmarks. To validate CATP, we conduct extensive experiments on eight vision-language benchmarks. They include standard multimodal ICL evaluations (Li 2025): VQAv2 (Goyal et al. 2017), VizWiz (Gurari et al. 2018), OK-VQA (Marino et al. 2019), GQA (Hudson and Manning 2019), and HatefulMemes (Kiela et al. 2020). Three latest and challenging multimodal benchmarks are also used: MMStar (Chen et al. 2024a), MME-Realworld (Zhang et al. 2024b), and VL-ICL (Zong, Bohdal, and Hospedales 2025).

Baselines and models. We compare CATP with the vanilla model, random pruning, and seven strong baselines. In addition to the previously introduced FastV, FitPrune, VTW, and DivPrune, we also evaluate HiRED (Arif et al. 2025), SparseVLM (Zhang et al. 2024a), and PLPHP (Meng et al. 2025). For methods relying on user-specified hyperparameters, we test four configurations and report the **best** performance. We first evaluate each pruning method on LLaVA-Next-7B under a range of pruning ratios. We then report additional results at a fixed ratio on LLaVA-Next-13B, InternVL2.5-8B, and Qwen2.5VL-7B. In LLaVA-Next, every image in the input sequence is uniformly converted into 576 tokens, whereas in InternVL2.5-8B and Qwen2.5VL-7B, the number of tokens per image depends on its original resolution and ranges from 576 to 1280.

Implementation details. All experiments are conducted on NVIDIA H200 GPUs. For every benchmark, the validation samples serve as query samples. Each query is paired with **four** ICDs randomly retrieved from the training split, producing a 4-shot sequence. For LLaVA-Next-7B, InternVL2.5-8B, and Qwen2.5VL-7B, we set $K = 6$. For LLaVA-Next-13B, we set $K = 10$. Across all LVLMs, we set $\lambda_1 = 0.7$ and $\lambda_2 = 0.6$. As CATP operates without the full attention matrices, it is compatible with FlashAttention (Dao et al. 2022), which further boosts its efficiency.

Main Results

Table 1 reports the performance of CATP under three commonly used pruning ratios. These are the levels at which pruning methods typically deliver real efficiency gains. CATP achieves the best performance under all three settings. At both 66.7% and 77.8%, it enhances performance over the vanilla model. When the ratio reaches 89.9%, CATP incurs only a modest 1.4% drop. Meanwhile, CATP is effective

Method	VQAv2	VizWiz	OK-VQA	GQA	HatefulMemes	MMStar	MME-RealWorld	VL-ICL	Avg. Rel
LLaVA-Next-7B	*Upper Bound 100%*								
Vanilla	56.7	40.2	47.6	65.7	66.8	37.5	34.2	26.8	100.0%
LLaVA-Next-7B	*Pruning Ratio 66.7%*								
Random	52.3	37.4	44.9	63.4	64.5	30.9	28.8	22.7	90.5%
FastV (ECCV24)	52.1	35.8	43.7	63.0	64.1	27.9	26.9	19.8	86.4%
HiRED (AAAI25)	53.8	37.6	45.8	63.9	64.7	30.8	31.0	23.2	92.3%
FitPrune (AAAI25)	51.0	36.7	44.5	62.3	63.6	28.4	28.1	21.7	88.0%
VTW (AAAI25)	52.0	36.1	44.8	61.3	64.7	32.6	31.5	22.3	91.0%
DivPrune (CVPR25)	53.5	37.0	43.7	63.5	64.3	29.2	30.8	23.4	90.8%
SparseVLM (ICML25)	52.5	37.2	44.8	63.4	64.5	31.0	28.3	20.5	89.3%
PLPHP (2025.2)	54.7	39.5	46.9	65.5	66.3	36.2	33.4	25.0	97.5%
CATP (Ours)	**57.1**	**40.6**	**47.9**	**66.4**	**67.0**	**37.6**	**34.0**	**27.3**	**100.7%**
LLaVA-Next-7B	*Pruning Ratio 77.8%*								
Random	49.8	37.0	43.4	63.2	63.9	28.0	26.3	18.4	85.4%
FastV (ECCV24)	49.7	34.9	42.3	62.4	63.5	25.3	24.6	17.7	82.4%
HiRED (AAAI25)	52.4	36.9	44.5	62.7	64.0	28.5	28.9	21.6	88.8%
FitPrune (AAAI25)	50.1	35.6	43.7	62.0	63.2	27.3	27.0	18.2	84.7%
VTW (AAAI25)	50.9	33.6	41.7	61.5	62.7	30.2	28.9	21.8	86.9%
DivPrune (CVPR25)	51.7	36.0	43.5	63.3	63.4	27.3	26.0	20.9	86.3%
SparseVLM (ICML25)	49.5	35.8	42.7	62.6	64.0	26.8	25.4	18.3	83.9%
PLPHP (2025.2)	54.0	39.1	46.0	64.8	65.5	35.9	32.7	24.1	95.9%
CATP (Ours)	**56.8**	**40.0**	**47.8**	**65.9**	**66.7**	**37.2**	**34.4**	**27.0**	**100.1%**
LLaVA-Next-7B	*Pruning Ratio 89.9%*								
Random	46.3	34.6	42.3	60.1	61.5	25.8	23.9	16.2	79.9%
FastV (ECCV24)	46.0	33.5	41.8	59.3	61.7	23.5	23.2	16.5	78.4%
HiRED (AAAI25)	48.6	35.0	42.9	61.9	62.7	26.4	26.0	19.2	83.6%
FitPrune (AAAI25)	46.1	34.6	42.5	61.1	62.0	25.3	22.7	16.4	79.7%
VTW (AAAI25)	47.6	32.8	37.4	60.1	61.5	26.3	25.9	18.7	80.4%
DivPrune (CVPR25)	48.2	34.8	43.0	62.3	62.5	25.7	25.0	20.1	83.4%
SparseVLM (ICML25)	46.2	31.3	41.6	61.3	61.9	25.4	24.3	17.0	79.4%
PLPHP (2025.2)	53.4	39.0	45.6	65.0	65.2	35.8	32.1	23.3	95.0%
CATP (Ours)	**56.3**	**39.7**	**47.2**	**65.4**	**66.3**	**36.7**	**34.0**	**25.5**	**98.6%**

Table 1. 4-shot performance comparison of different pruning methods on LLaVA-Next-7B under three pruning ratios. **Avg. Rel** denotes the average percentage of performance relative to the vanilla model.

across diverse LVLMs and benchmarks, as shown in Tables 1 and 2. On four LVLMs and eight benchmarks, CATP consistently achieves the best performance. Notably, none of the seven SOTA baselines improve multimodal ICL, with some even performing worse than random pruning. In contrast, CATP provides performance gains by capturing the complex interactions in multimodal ICL and by selecting the image tokens that contribute most to the overall reasoning process.

Efficiency Analysis

To demonstrate the efficiency of CATP, we conduct extensive comparative experiments on four LVLMs, reporting TFLOPs, inference latency, and KV Cache size. Following FastV, we compute the FLOPs of the multi-head attention and feedforward network modules as $4nd^2 + 2n^2d + 3ndm$, where n is the number of image tokens, d is the hidden state size and m is the intermediate size of the FFN. For latency, we report the average total inference time per sequence. For KV Cache memory usage, we report the average GPU memory consumption during inference. Table 3 reports the results for each LVLM, averaged across eight benchmarks. CATP achieves the highest accuracy and shows excellent efficiency. It ranks second only to DivPrune in reducing FLOPs and KV Cache and delivers the largest cut in inference latency. As pointed out by (Wen et al. 2025a), latency is the most reliable indicator of token pruning efficiency because the execution cost of the method is not reflected in FLOPs or

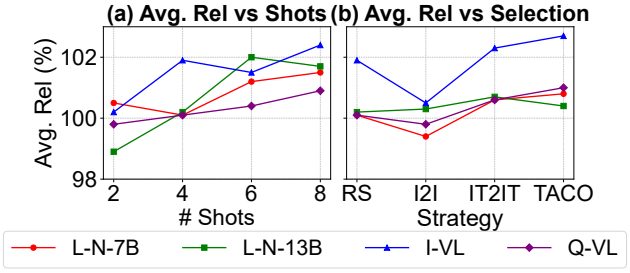

Figure 5. Performance of CATP on four LVLMs across diverse: (a) shot counts and (b) ICD selection strategies.

KV Cache. These results indicate that CATP brings real efficiency gains in multimodal ICL. Together with its unique performance boost, CATP is an outstanding pruning method with strong practical potential.

Ablation Study and Analysis

All ablation studies are under a pruning ratio of $R = 77.8\%$.

Adaptability to diverse in-context sequence configurations. In practical multimodal ICL, both the number of ICD shots and the ICD selection strategy vary across applications. Since sequence construction strongly affects performance (Gao, Fisch, and Chen 2021; Liu et al. 2021), we evaluate CATP under diverse configurations to assess its ro-

Method	VQAv2	VizWiz	OK-VQA	GQA	HatefulMemes	MMStar	MME-RealWorld	VL-ICL	Avg. Rel
LLaVA-Next-13B									
Vanilla	60.3	42.7	49.9	70.1	69.8	40.2	39.4	28.5	100%
Random	53.2	39.4	44.0	65.3	65.7	33.4	34.1	23.2	88.4%
FastV (ECCV24)	51.9	38.5	44.2	65.1	66.0	32.3	33.3	21.9	86.7%
DivPrune (CVPR25)	54.5	40.1	46.3	65.4	66.8	34.5	34.9	24.8	90.9%
PLPHP (2025.2)	58.7	42.3	49.3	69.8	69.8	39.3	38.5	23.0	96.4%
CATP (Ours)	**60.5**	**43.2**	**50.1**	**70.5**	**70.1**	**39.7**	**39.2**	**28.6**	**100.2%**
InternVL2.5-8B									
Vanilla	66.3	55.7	60.2	70.9	73.2	57.4	64.9	41.5	100%
Random	59.4	49.3	55.8	68.5	71.5	49.1	57.8	35.9	90.8%
FastV (ECCV24)	59.6	48.8	55.0	67.7	72.0	47.5	58.2	34.3	89.7%
DivPrune (CVPR25)	60.8	50.5	54.7	67.3	70.2	50.9	60.3	35.7	91.5%
PLPHP (2025.2)	65.8	50.0	56.5	70.3	73.4	56.5	64.3	40.6	97.2%
CATP (Ours)	**66.0**	**56.0**	**60.4**	**80.3**	**73.7**	**57.9**	**64.5**	**41.7**	**101.9%**
Qwen2.5VL-7B									
Vanilla	70.3	54.9	63.7	73.0	74.9	61.8	63.0	45.3	100%
Random	60.3	47.2	60.8	70.0	69.7	55.8	56.6	37.4	89.9%
FastV (ECCV24)	58.6	47.3	59.4	69.7	69.2	54.9	53.1	35.4	87.7%
DivPrune (CVPR25)	61.2	47.0	62.6	72.3	73.5	56.4	56.3	40.1	92.2%
PLPHP (2025.2)	67.3	53.6	63.5	72.6	74.0	59.7	61.4	44.6	98.0%
CATP (Ours)	**70.7**	**55.3**	**63.9**	**73.6**	**75.2**	**61.2**	**62.5**	**45.0**	**100.1%**

Table 2. 4-shot performance comparison of different pruning methods on LLaVA-Next-13B, InternVL2.5-8B, and Qwen2.5VL-7B at a fixed pruning ratio of 77.8% . **Avg. Rel** denotes the average percentage of performance relative to the vanilla model.

Method	FLOPs	Latency	KV Cache	Avg. Rel
LLaVA-Next-7B				
Vanilla	20.84	3.82s	1.12GB	100%
FastV	6.37 (69.4%↓)	3.25s (14.9%↓)	0.38GB (66.1%↓)	82.4%
DivPrune	4.19 (79.9%↓)	4.03s (5.5%↑)	0.30GB (73.2%↓)	86.3%
PLPHP	5.33 (74.4%↓)	3.49s (8.6%↓)	0.37GB (67.0%↓)	95.9%
CATP	4.38 (79.0%↓)	3.19s (16.5%↓)	0.35GB (68.8%↓)	**100.1%**
InternVL2.5-8B				
Vanilla	28.69	2.75s	460.8MB	100%
FastV	9.63 (66.4%↓)	2.57s (6.5%↓)	103.5MB (77.5%↓)	89.7%
DivPrune	6.75 (76.5%↓)	2.90s (5.5%↑)	82.3MB (82.1%↓)	91.5%
PLPHP	8.10 (71.8%↓)	2.55s (7.3%↓)	156.2MB (66.1%↓)	97.2%
CATP	7.03 (75.5%↓)	2.53s (8.0%↓)	97.6MB (78.8%↓)	**101.9%**
Qwen2.5VL-7B				
Vanilla	27.43	2.31s	215.0MB	100%
FastV	9.26 (66.2%↓)	2.11s (8.7%↓)	47.5MB (77.9%↓)	87.7%
DivPrune	6.18 (77.5%↓)	2.40s (3.9%↑)	37.6MB (82.5%↓)	92.2%
PLPHP	7.93 (71.1%↓)	2.19s (5.2%↓)	75.2MB (65.0%↓)	98.0%
CATP	6.48 (76.4%↓)	2.08s (10.0%↓)	41.7MB (80.6%↓)	**100.1%**

Table 3. 4-shot efficiency analysis of different pruning methods on four LVLMs. The results are averaged across 8 benchmarks at a ratio of 77.8%. **Avg. Rel** denotes the average percentage of performance relative to the vanilla model.

bustness. Figure 5(a) shows that CATP performs better as shot count increases, due to its ability to handle rising token redundancy and enhance reasoning by keeping only informative tokens. As LVLMs support longer contexts, CATP's benefits can grow accordingly. Its inference speedup also scales nonlinearly with shot count. Figure 5(b) reports results under four ICD selection strategies. RS denotes the random sampling used in our main experiments. I2I and IQ2IQ are similarity-based methods commonly used in modern retrieval-augmented generation (RAG) systems, based on image-only and joint image-text similarity, respectively. TACO (Li et al. 2025c) is a SoTA ICD selection model. CATP brings greater performance gains as the quality of the

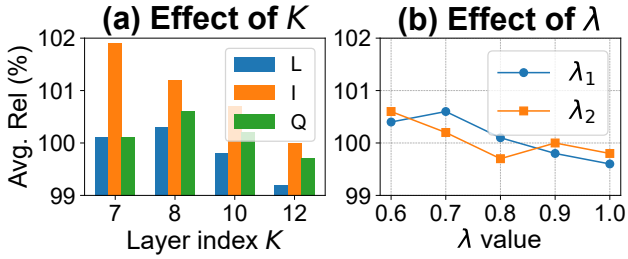

Figure 6. Average performance of CATP under different settings of K, λ_1, and λ_2.

sequence improves, suggesting that it can be integrated into existing systems to provide incremental benefits.

Impact of hyperparameters. Figure 6 shows how CATP's performance changes under different hyperparameter settings. Overall, CATP remains stable, which demonstrates strong robustness to hyperparameter choices. When K is set to the middle layers, performance decreases. This suggests that shallow layers contribute more to query-guided reasoning and that attention differences in these layers provide a clearer signal of token importance.

Conclusion

In this paper, we introduce CATP, a novel and training-free image token pruning method tailored to multimodal ICL. In the two-stage pruning process, CATP adaptively identifies the image tokens that are most important to the entire ICL process based on the input in-context sequence, and discards the rest. This leads to improvements in both performance and efficiency. CATP addresses an important gap in existing work on multimodal ICL and offers a reliable and efficient solution to enhance the capability of LVLMs when handling complex inputs. We believe that CATP provides a solid foundation and valuable insights for advancing LVLM capabilities.

Acknowledgments

We acknowledge the computing resources provided by NSF ACCESS.

References

Alvar, S. R.; Singh, G.; Akbari, M.; and Zhang, Y. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9392–9401.

Arif, K. H. I.; Yoon, J.; Nikolopoulos, D. S.; Vandierendonck, H.; John, D.; and Ji, B. 2025. HiRED: Attention-Guided Token Dropping for Efficient Inference of High-Resolution Vision-Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1773–1781.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Baldassini, F. B.; Shukor, M.; Cord, M.; Soulier, L.; and Piwowarski, B. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1539–1550.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37: 27056–27087.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024b. 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*, 19–35. Springer.

Chen, S.; Han, Z.; He, B.; Liu, J.; Buckley, M.; Qin, Y.; Torr, P.; Tresp, V.; and Gu, J. 2024c. Can Multimodal Large Language Models Truly Perform Multimodal In-Context Learning? arXiv:2311.18021.

Chen, W.; Li, L.; Yang, Y.; Wen, B.; Yang, F.; Gao, T.; Wu, Y.; and Chen, L. 2025. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8073–8082.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024d. 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*, 24185–24198.

Dao, T.; Fu, D. Y.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv:2205.14135.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Liu, T.; Chang, B.; Sun, X.; Li, L.; and Sui, Z. 2024. A Survey on In-context Learning. arXiv:2301.00234.

Doveh, S.; Perek, S.; Mirza, M. J.; Lin, W.; Alfassy, A.; Arbelle, A.; Ullman, S.; and Karlinsky, L. 2024. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. arXiv:2012.15723.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. arXiv:1802.08218.

Hu, Y.; Cheng, Y.; Lu, A.; Cao, Z.; Wei, D.; Liu, J.; and Li, Z. 2024. LF-ViT: Reducing spatial redundancy in vision transformer for efficient image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2274–2284.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.

Jeddi, A.; Baghbanzadeh, N.; Dolatabadi, E.; and Taati, B. 2025. arity-aware token pruning: Your vlm but faster. *arXiv preprint arXiv:2503.11549*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624.

Kim, S.; Shen, S.; Thorsley, D.; Gholami, A.; Kwon, W.; Hassoun, J.; and Keutzer, K. 2022. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 784–794.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. arXiv:2407.07895.

Li, Y. 2025. Advancing Multimodal In-Context Learning in Large Vision-Language Models with Task-aware Demonstrations. *arXiv preprint arXiv:2503.04839*.

Li, Y.; He, H.; Cao, Y.; Cheng, Q.; Fu, X.; and Tang, R. 2025a. M2IV: Towards Efficient and Fine-grained Multimodal In-Context Learning in Large Vision-Language Models. *arXiv preprint arXiv:2504.04633*.

Li, Y.; Yang, J.; Li, B.; and Tang, R. 2025b. CAMA: Enhancing Multimodal In-Context Learning with Context-Aware Modulated Attention. *arXiv preprint arXiv:2505.17097*.

Li, Y.; Yun, T.; Yang, J.; Feng, P.; Huang, J.; and Tang, R. 2025c. TACO: Enhancing Multimodal In-context Learning via Task Mapping-Guided Sequence Configuration. *arXiv preprint arXiv:2505.17098*.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26763–26773.

Lin, Z.; Lin, M.; Lin, L.; and Ji, R. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5334–5342.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llavanext: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Shen, D.; Zhang, Y.; Dolan, B.; Carin, L.; and Chen, W. 2021. What Makes Good In-Context Examples for GPT-3? arXiv:2101.06804.

Liu, Y.; Li, Z.; Li, H.; Yu, W.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Jin, L.; et al. 2023b. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2(5): 6.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. arXiv:1906.00067.

Meng, Y.; Li, K.; Huang, C.; Gao, C.; Chen, X.; Li, Y.; and Zhang, X. 2025. PLPHP: Per-Layer Per-Head Vision Token Pruning for Efficient Large Vision-Language Models. *arXiv preprint arXiv:2502.14504*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wen, Z.; Gao, Y.; Li, W.; He, C.; and Zhang, L. 2025a. Token Pruning in Multimodal Large Language Models: Are We Solving the Right Problem? *arXiv preprint arXiv:2502.11501*.

Wen, Z.; Gao, Y.; Wang, S.; Zhang, J.; Zhang, Q.; Li, W.; He, C.; and Zhang, L. 2025b. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Xu, R.; Wang, Y.; Luo, Y.; and Du, B. 2025. Rethinking Visual Token Reduction in LVLMs under Cross-modal Misalignment. arXiv:2506.22283.

Ye, W.; Wu, Q.; Lin, W.; and Zhou, Y. 2025. 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*, volume 39, 22128–22136.

Zhang, Q.; Liu, M.; Li, L.; Lu, M.; Zhang, Y.; Pan, J.; She, Q.; and Zhang, S. 2025. Beyond Attention or Similarity: Maximizing Conditional Diversity for Token Pruning in MLLMs. *arXiv preprint arXiv:2506.10967*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2024a. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Zhang, Y.-F.; Zhang, H.; Tian, H.; Fu, C.; Zhang, S.; Wu, J.; Li, F.; Wang, K.; Wen, Q.; Zhang, Z.; et al. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Zhou, C.; Zhang, M.; Chen, P.; Fu, C.; Shen, Y.; Zheng, X.; Sun, X.; and Ji, R. 2024. Vega: Learning interleaved image-text comprehension in vision-language large models. *arXiv preprint arXiv:2406.10228*.

Zong, Y.; Bohdal, O.; and Hospedales, T. 2025. VL-ICL Bench: The Devil in the Details of Multimodal In-Context Learning. arXiv:2403.13164.