Instruction-Guided Cross-Modal Clustering for Training-Free Visual Token Pruning in Vision-Language Models

Yunqian Yu, Biao Chen, Yunya Zhang, Tonglan Xie, Mengmeng Jing, Lin Zuo*

School of Information and Software Engineering,
University of Electronic Science and Technology of China
yuyunqian2022@163.com, chenbiao@std.uestc.edu.cn, lingerya2333@163.com, {tonglanxie, jingmeng1992}@gmail.com, linzuo@uestc.edu.cn

Abstract

Large vision-language models (LVLMs) have demonstrated remarkable capabilities in understanding multimodal data such as images and text. However, the number of visual tokens in these models often far exceeds that of textual tokens, resulting in substantial redundancy and high inference costs. Existing pruning methods primarily rely on either unimodal information or cross-modal attention mechanisms. The former often overlooks the semantic alignment between instructions and visual representations in the multimodal space, while the latter is prone to attention drift and dispersion, leading to significant performance degradation under high pruning ratios. All the above issues stem from the lack of effective textual guidance during the pruning process. To identify effective informational cues for guiding pruning, we conduct an in-depth analysis of the interaction between language instructions and visual features based on the cross-modal information bottleneck attribution (CIBA) method, revealing the presence of noun anchors. Based on this analysis, we propose the Instruction-Guided Cross-Modal Clustering Token Pruning (ICCTP) method, a plug-and-play, training-free pruning paradigm. Specifically, ICCTP first leverages global attention to retain a small set of visual tokens that preserve global context. It then extracts nouns from the instruction as clustering centers to perform cross-modal clustering over the remaining visual tokens. To balance semantic diversity and global relevance while reducing intra-cluster redundancy, we design an importance scoring mechanism. Finally, visual tokens within each cluster are pruned according to a specified pruning ratio. We evaluate ICCTP on multiple VLM architectures, including LLaVA-1.5-7B, LLaVA-1.5-13B, and LLaVA-NeXT-7B. Experimental results show that ICCTP maintains strong performance across various pruning rates without requiring retraining. Notably, even under an extreme setting where 94.4 % of visual tokens are removed, ICCTP retains 90.02% of the original accuracy while reducing TFLOPs by 82.36%.

Introduction

Large Vision-Language Models (LVLMs) (Liu et al. 2023a; Li et al. 2024, 2023a; Liu et al. 2024a; Bai et al. 2023) have demonstrated strong capabilities in multimodal tasks such as visual question answering and instruction following. These models typically encode images into dense patch-level visual token sequences, which are then fused with textual inputs via a multimodal alignment module. However, the number of visual tokens is often significantly larger than that of textual tokens—for example, a single image in LLaVA-NeXT-7B (Liu et al. 2024b) can produce up to 2880 visual tokens, while the corresponding text typically contains fewer than 100. This substantial imbalance leads to high computational overhead during inference.

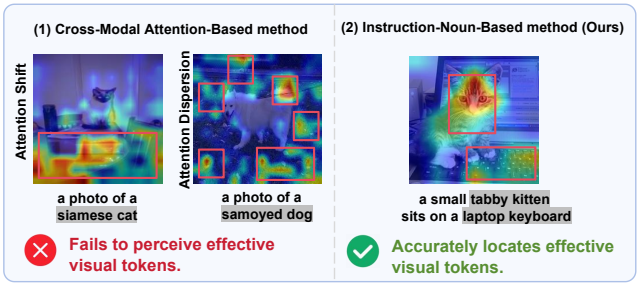

Figure 1: Attention visualization of cross-modal attention-based methods and our instruction-noun-based method.

In recent years, various studies have attempted to reduce inference costs through visual token pruning. Most methods either rely on unimodal visual cues (Zhang et al. 2024a) or use cross-modal attention weights (Zhang et al. 2024b) within the language model to estimate token importance. However, neither approach identifies effective semantic cues from the instruction to guide pruning. Specifically, unimodal methods often focus solely on global visual information and lack instruction-aware token selection mechanisms. In contrast, cross-modal attention-based pruning methods face two key issues, as shown in Figure 1: (1) attention shift caused by rotary position embeddings (RoPE), which biases the model toward lower image regions and leads to missed semantic content; and (2) attention dispersion, where attention spreads too evenly across visual tokens, weakening its discriminative capacity (Zhang et al. 2024a, 2025). These issues become especially detrimental under high pruning rates, leading to notable performance drops.

To address these limitations and derive effective pruning cues, we revisit semantic interactions between vision and text modalities from an information-theoretic perspective.

*Corresponding author.

We introduce a Cross-Modal Information Bottleneck Attribution (CIBA) method (Wang, Rudner, and Wilson 2023) to identify the most informative visual features under a language instruction. Our analysis reveals that nouns in the instruction serve as stable semantic anchors in the multimodal representation space, effectively guiding the structured aggregation and semantic alignment of visual tokens. This observation forms the core motivation of our pruning strategy.

Building on the insight into semantic anchors, we propose Instruction-Guided Cross-Modal Clustering Token Pruning (ICCTP), a plug-and-play, training-free pruning framework. ICCTP first employs global attention to retain a few visual tokens that preserve the image's contextual information. It then extracts nouns from the instruction as clustering centers in the multimodal embedding space to guide the clustering of the remaining visual tokens. Furthermore, we design an importance scoring mechanism that jointly considers the semantic similarity to the instruction and the global attention response, allowing the selection of key tokens within each cluster for pruning according to a specified ratio. The main contributions of our work are as follows:

- We propose a cross-modal information bottleneck attribution method, which reveals that nouns in language instructions serve as semantic anchors in the vision-language representation space, providing a stable and generalizable signal for visual token pruning.
- We propose a training-free, plug-and-play pruning framework, ICCTP, that integrates semantic relevance and attention responses to enable efficient and robust visual token pruning.
- We conduct extensive evaluations on ten datasets and several mainstream models, demonstrating that ICCTP remains robust and generalizable even under extreme pruning. It preserves 90.02% of performance while pruning 94.4% of visual tokens and cutting inference cost by 82.36%, outperforming prior methods.

Related Work

Vision-Language Models

The success of large language models (LLMs) (Brown et al. 2020; Dubey et al. 2024; Luo et al. 2024; Ouyang et al. 2022; Radford et al. 2019) has inspired a wave of research extending their reasoning capabilities to the multimodal domain, giving rise to vision-language models (VLMs) (Achiam et al. 2023; Team et al. 2023; Wang et al. 2024). Typically, a VLM consists of a vision encoder and a language generator, where visual features are projected into the language space via a modality alignment module (Bai et al. 2023; Liu et al. 2024a). While VLMs have achieved remarkable progress in tasks such as image captioning and visual question answering, they suffer from excessive computational overhead due to the redundancy of visual tokens. For example, LLaVA-1.5 (Liu et al. 2023a) encodes a 336×336 image into 576 tokens, while its high-resolution variant LLaVA-NeXT (Liu et al. 2024b) produces up to 2880 tokens per image. Improving inference efficiency under such conditions has become a critical research focus.

Token Pruning in VLMs

To reduce visual token redundancy, a variety of token pruning methods have been proposed, which can be broadly categorized into trainable and training-free approaches.

Trainable methods such as ELIP (Guo et al. 2023), LVPruning (Sun et al. 2025), and IVTP (Huang et al. 2024) utilize language supervision or attention-guided strategies to learn fine-grained pruning policies. However, these approaches often introduce additional training costs and architectural modifications, which limit their generalizability and deployment. In contrast, training-free methods aim to enhance inference efficiency without retraining. DivPrune selects diverse tokens based on unimodal visual information. FitPrune (Ye et al. 2024)formulates pruning as minimizing the divergence between self- and cross-modal attention before and after pruning. FastV (Chen et al. 2024) discards low-attention visual tokens after the second layer of the language model. SparseVLM (Zhang et al. 2024b) selects relevant visual tokens based on text-guided attention and prunes adaptively across layers. AdaptPrune (Luan et al. 2025) integrates attention scores and spatial-aware adaptive NMS for comprehensive token importance estimation. VisionZip (Yang et al. 2024) combines CLS-based dominant token selection with similarity-driven contextual token merging to achieve effective compression.

Subsequently, an increasing number of studies have relied on unimodal visual cues or internal cross-modal attention weights within language models to estimate the importance of visual tokens (Xing et al. 2024). However, our analysis reveals that these approaches tend to be unstable under high pruning ratios and fail to provide reliable guidance. In contrast, we adopt a cross-modal information bottleneck attribution approach, which reveals that instruction nouns serve as stable semantic anchors in the multimodal representation space. This insight enables a more efficient and interpretable guidance signal for token pruning.

Method

In this section, we first review the issue of token pruning in LVLMs. Then, we employ the Cross-Modal Information Bottleneck Attribution method to extract effective guidance signals for token pruning. Based on this analysis, we propose ICCTP, which uses instruction noun phrases to guide cross-modal clustering and selects an optimal token subset by maximizing semantic relevance and global consistency.

Preliminaries

Modern LVLMs are generally composed of three key components: a vision encoder E_v, a projector f_p, and a large language model (LLM) f_θ. The vision encoder, such as the CLIP model's image backbone, processes the input image X_v to produce a sequence of visual representations. Then, they are subsequently mapped by the projector into the embedding space of the LLM, forming the visual token set T_v, as shown in Equation (1):

$$T_v = f_p\left(E_v\left(X_v\right)\right) \in \mathbb{R}^{n \times d} \tag{1}$$

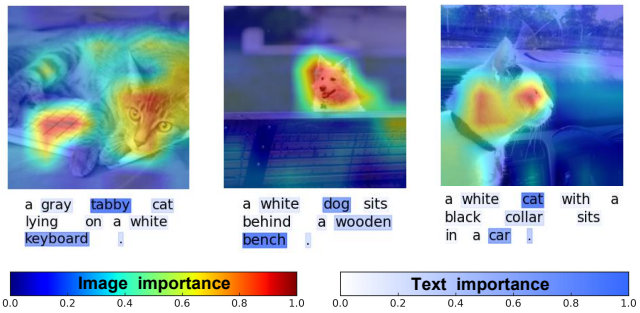

Figure 2: Visualization of the Cross-Modal Information Bottleneck Attribution method on the VQAv2 dataset.

The visual tokens are concatenated with textual tokens and passed to the LLM for joint reasoning. Note that the number of visual tokens is typically larger than the number of textual tokens, which greatly increases the computational cost during inference. To address this issue, recent studies have proposed token pruning methods that select a subset of visual tokens while preserving key features. However, unimodal visual methods struggle to identify task-relevant regions, resulting in low semantic relevance of retained tokens. In contrast, cross-modal attention-based methods suffer from attention shift and dispersion, leading to the mis-pruning of critical tokens and degraded performance.

Cross-Modal Information Bottleneck Attribution

To address the above issues, we leverage the Cross-Modal Information Bottleneck Attribution (CIBA) (Wang, Rudner, and Wilson 2023) method to analyze the semantic interactions between language instructions and visual features, thereby extracting more effective guidance.

As a preliminary, we introduce the traditional Information Bottleneck Attribution (IBA) method (Schulz et al. 2020), which applies the Information Bottleneck (Tishby, Pereira, and Bialek 2000) theory to neural network interpretability. By applying a bottleneck mask, IBA identifies the most informative regions of intermediate representations. The goal is to encode the input X into a latent representation Z that preserves information relevant to the target Y while minimizing redundancy with X. This process is formulated as shown in Equation (2):

$$\max_{\theta} I(Z, Y; \theta) - \beta I(Z, X; \theta), \tag{2}$$

where $I(\cdot, \cdot; \theta)$ represents the mutual information, and β is the hyperparameter that controls the trade-off between compression and fitting. Through this approach, we can effectively extract key information relevant to the task and eliminate noise. Although IBA performs well in unimodal tasks, it struggles to generalize to cross-modal scenarios due to the inherent heterogeneity between modalities. Visual and textual inputs reside in distinct representational spaces, creating a semantic gap that hinders direct correlation. Moreover, the one-way information flow in traditional IBA ignores bidirectional dependencies, often leading to attention drift and inaccurate attribution.

To address these limitations, we extend IBA to the cross-modal setting by introducing CIBA, which models the mutual dependence between visual and textual representations. By explicitly incorporating the embedding E'_m from the complementary modality, CIBA enhances attribution accuracy and better captures cross-modal semantic alignment. Its objective defined in Equation (3):

$$F_m(\theta_m) = I(Z_m, E'_m; \theta_m) - \beta_m I(Z_m, X_m; \theta_m), \tag{3}$$

where Z_m is the latent representation of input X_m, θ_m denotes the learnable parameters, and β_m balances compression and task relevance. To enable efficient optimization, we adopt a variational approximation, yielding the following empirical objective:

$$\begin{aligned}\hat{F}_m^{\text{emp}}(\theta_m) = &\frac{1}{N} \sum_{n=1}^{N} \int p(z_m \mid x_m^{(n)}; \theta_m)\, \mathrm{S}(e_{m'}^{(n)}, g_m(z_m))\, dz_m \\ &- \beta_m \,\mathrm{KL}\left(p(z_m \mid x_m^{(n)}; \theta_m) \,\|\, q(z_m)\right),\end{aligned} \tag{4}$$

where $\mathrm{S}(\cdot, \cdot)$ is the cosine similarity. The integrals are estimated via Monte Carlo sampling with reparameterization. The parameter set is defined as $\theta_m = \{\lambda_m, \sigma_m, \ell_m\}$, where λ_m is optimized per modality, and β_m, σ_m, and ℓ_m are modality-specific hyperparameters. It can be applied to each modality to retain the key features needed to predict the other, making cross-modal representation more efficient.

To better illustrate cross-modal interactions, we visualize the CIBA attribution results on the VQAv2 (Goyal et al. 2019) dataset, as shown in Figure 2. The top part shows that only a few image regions are informative for predicting language instructions, supporting visual token pruning. The bottom part highlights that nouns in the instruction (e.g., dog, wooden bench) are most relevant for predicting visual features, while other parts of speech contribute less.

Compared to existing methods, CIBA captures cross-modal attention regions more accurately by computing mutual information between visual features and language instructions, enabling the preservation of critical information during pruning. Through this theoretical framework, we are able to identify key information—nouns—within the language instructions, allowing for more rational pruning of visual tokens.

Instruction-Guided Cross-Modal Clustering Token Pruning

To address the bias in cross-modal guidance inherent in existing pruning methods, we leverage the theoretical foundation of cross-modal information bottleneck attribution to derive effective pruning cues and propose Instruction-Guided Cross-Modal Clustering Token Pruning (ICCTP). As shown in Figure 3, this plug-and-play, training-free approach consists of three key steps: (1) Global Token Preselection, (2) Instruction-Driven Cross-Modal Clustering, and (3) Redundancy-Aware Pruning Strategy. The pseudocode is provided in Algorithm 1.

Global Token Preselection(GTP): We first leverage the `[CLS]` token's attention from the visual encoder to estimate

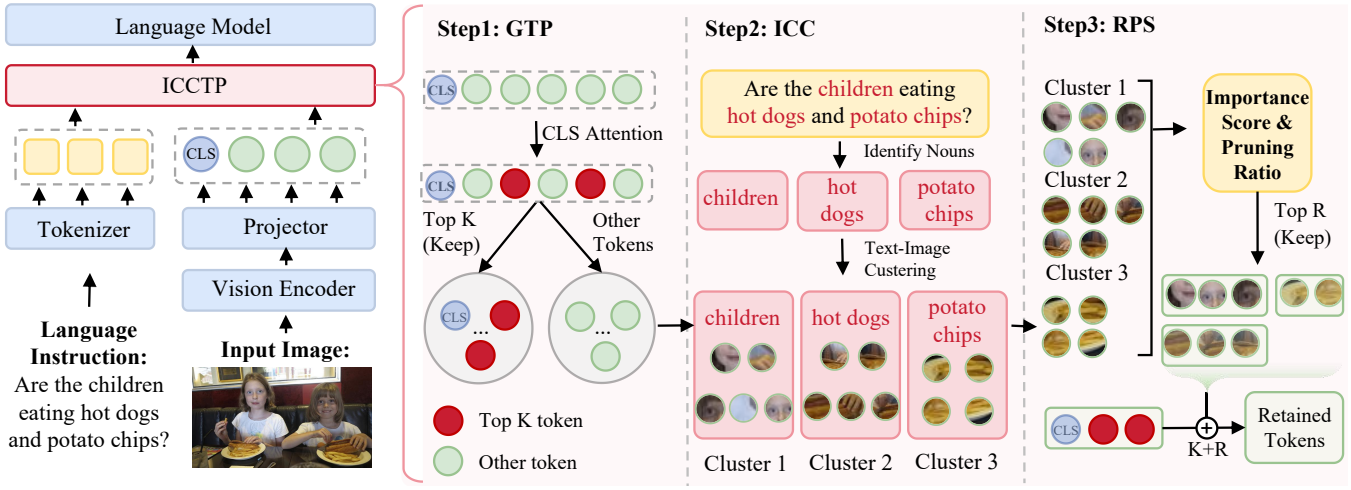

Figure 3: Illustration of the Instruction-Guided Cross-Modal Clustering Token Pruning method. GTP denotes Global Token Preselection; ICC denotes Instruction-Driven Cross-Modal Clustering; and RPS denotes Redundancy-Aware Pruning Strategy.

the global saliency of each visual token, thereby preserving those sensitive to scene-level context as reliable priors for subsequent cross-modal clustering and pruning.

Specifically, let the attention matrix produced by the visual encoder be $A \in \mathbb{R}^{h\times n\times n}$, where h is the number of attention heads and n is the number of visual tokens. We first average the attention scores across all heads to obtain the aggregated attention matrix:

$$\bar{A} = \frac{1}{h}\sum_{i=1}^{h} A^{(i)} \in \mathbb{R}^{n\times n} \tag{5}$$

We then take the first row of $\bar{A}$, which corresponds to the attention from the `[CLS]` token to all other visual tokens, as the saliency vector:

$$\mathbf{a} = \bar{A}_{0,:} = [a_1, a_2, \ldots, a_n], \tag{6}$$

where a_j denotes the attention weight from the `[CLS]` token to the j-th visual token.

To control the pruning ratio, we retain the top $k_g = \lfloor \rho_g \cdot n_a \rfloor$ tokens with the highest attention scores, where $\rho_g \in (0, 1)$ is the global retention ratio, and n_a is the number of visual tokens preserved under the overall pruning budget.

$$\mathcal{G} = \mathrm{TopK}_{k_g}(\mathbf{a}), \tag{7}$$

where $\mathcal{G}$ denotes the set of globally important visual tokens.

Instruction-Driven Cross-Modal Clustering(ICC): Although global token preselection retains key visual tokens, the lack of fine-grained textual guidance and the limitations of attention shift and dispersion, hinder the precise localization of task-relevant visual regions.

To overcome these limitations, we leverage the Cross-Modal Information Bottleneck Attribution (CIBA) method to identify nouns in instructions as stable visual semantic anchors. Based on this insight, we extract noun phrase embeddings from the instruction and treat them as clustering centers for guiding visual token grouping, which serves as the foundation for subsequent token pruning. The main steps are outlined as follows:

Let the remaining visual tokens after global preselection be $\mathcal{R} = \{v_i\}_{i=1}^{n_r}$, $v_i \in \mathbb{R}^d$, where n_r denotes the number of tokens to be clustered and d is the embedding dimension. We extract a set of noun phrases $\mathcal{N} = \{\mathcal{I}_j\}_{j=1}^{K}$ from the instruction text using SpaCy, a lightweight and efficient NLP toolkit, where $\mathcal{I}_j$ denotes the subword set of the j-th noun phrase. Each clustering center is computed as the average embedding of its tokens:

$$c_j = \frac{1}{|\mathcal{I}_j|}\sum_{w\in\mathcal{I}_j} \mathrm{Emb}(w), \quad j = 1, 2, \ldots, K. \tag{8}$$

For each visual token $v_i \in \mathcal{R}$, we compute its cosine similarity to all centers and assign it to the nearest cluster:

$$\pi(i) = \arg\max_{j\in\{1,\ldots,K\}} \cos\left(v_i, c_j\right), \tag{9}$$

resulting in a set of clusters:

$$C_j = \{v_i \in \mathcal{R} \mid \pi(i) = j\}. \tag{10}$$

This instruction-driven clustering explicitly aligns visual tokens with the key semantics of the instructions, mitigating the bias and noise issues inherent in conventional cross-modal attention and providing a task-relevant foundation for subsequent redundancy suppression and pruning.

Redundancy-Aware Pruning Strategy(RPS): Although cross-modal clustering explicitly aligns visual tokens with nouns, substantial redundancy may still exist within each cluster due to highly similar tokens. To address this, we design a redundancy-aware scoring mechanism that combines both semantic relevance and global saliency, thereby selecting only the most informative tokens in each cluster. For each cluster $C_j = \{v_i\}$, we define a joint importance score s_i:

$$s_i = \alpha \cdot \cos(v_i, c_j) + (1-\alpha)\cdot \tilde{a}_i, \tag{11}$$

where $\cos(v_i, c_j)$ measures the cosine similarity between visual token v_i and the noun center c_j, $\tilde{a}_i \in [0, 1]$ denotes

Algorithm 1: Instruction-Guided Cross-Modal Clustering Token Pruning (ICCTP)

Require: Visual tokens $V = \{v_i\}_{i=1}^{n}$, attention matrix A from the visual encoder, instruction tokens T, global keep ratio ρ_g, cluster keep ratio ρ_c, weighting factor α.

Ensure: Pruned visual token set $\hat{V}$.

Step 1: Global Token Preselection
Compute attention scores $\mathbf{a}$ from CLS token: $\mathbf{a} = \bar{A}_{0,:}$, where $\bar{A} = \frac{1}{h}\sum_{i=1}^{h} A^{(i)}$.
Select top-k_g tokens with $k_g = \lfloor \rho_g \cdot n_a \rfloor$: $\mathcal{G} = \mathrm{TopK}_{k_g}(\mathbf{a})$.
Set residual tokens $\mathcal{R} = V \setminus \mathcal{G}$.
Step 2: Instruction-Driven Cross-Modal Clustering
Extract noun phrases $\mathcal{N} = \{\mathcal{I}_j\}_{j=1}^{K}$ from T.
for $j = 1$ to K **do**
    Compute cluster centers: $c_j = \frac{1}{|\mathcal{I}_j|}\sum_{w \in \mathcal{I}_j} \mathrm{Emb}(w)$.
end for
Assign each $v_i \in \mathcal{R}$ to the nearest center: $\pi(i) = \arg\max_j \cos(v_i, c_j)$.
Form clusters $C_j = \{v_i \in \mathcal{R} \mid \pi(i) = j\}$.
Step 3: Redundancy-Aware Pruning
for $j = 1$ to K **do**
    for each $v_i \in C_j$ **do**
        Compute importance score: $s_i = \alpha \cdot \cos(v_i, c_j) + (1-\alpha)\cdot \tilde{a}_i$.
    end for
    Retain top-k_j tokens: $k_j = \max(1, \lfloor \rho_c \cdot |C_j| \rfloor)$.
end for
Output: $\hat{V} = \mathcal{G} \cup \bigcup_{j=1}^{K} \mathrm{TopK}_{k_j}\{v_i \in C_j \mid s_i\}$.

the global saliency of v_i derived from CLS signals, and $\alpha \in [0, 1]$ is referred to as the semantic-context trade-off factor, which balances the importance between semantic alignment and global context. Within each cluster C_j, we rank all tokens by s_i and retain the top k_j

$$k_j = \max\left(1, \lfloor \rho_c \cdot |C_j| \rfloor\right), \tag{12}$$

where $\rho_c \in (0, 1)$ is the cluster-level retention ratio. The final pruned set of visual tokens is defined as:

$$\hat{V} = \mathcal{G} \cup \bigcup_{j=1}^{K} \mathrm{TopK}_{k_j}\{v_i \in C_j \mid s_i\}, \tag{13}$$

where $\mathcal{G}$ is the set of globally important tokens preserved from Step (1). The final token set $\hat{V}$ is concatenated with the textual tokens I_t and fed into the LLM for joint reasoning. This strategy effectively reduces redundant tokens while maximally preserving visual tokens that are strongly aligned with the key nouns as well as global contextual information, thereby ensuring that the pruned representations retain robust multimodal reasoning capabilities.

Experiments

Experimental Setup

Datasets: To evaluate ICCTP, we conduct experiments on ten benchmark datasets covering key vision-language tasks. These include VQAv2 (Goyal et al. 2019), GQA (Hudson and Manning 2019), VizWiz (Gurari et al. 2018), ScienceQA (Lu et al. 2022), TextVQA (Singh et al. 2019), POPE (Li et al. 2023b), MME (Fu et al. 2023), MMBench (Liu et al. 2023b), MMBench-CN (Liu et al. 2023b), and MM-Vet (Yu et al. 2023).

Implementation Details: We apply ICCTP to various VLM architectures, including the LLaVA series (e.g., LLaVA-1.5-7B and LLaVA-1.5-13B(Liu et al. 2024a)) and the high-resolution LLaVA-NeXT-7B. Specifically, both LLaVA-1.5-7B and LLaVA-1.5-13B contain 576 visual tokens, while LLaVA-NeXT-7B dynamically introduces up to 2880 visual tokens. The hyperparameter configurations are discussed in detail in the ablation section. All experiments are conducted on a single RTX 4090 GPU.

Main Results

We evaluate ICCTP on LLaVA-1.5-7B and compare it with representative pruning methods, including ToMe (Bolya et al. 2023), FastV (Chen et al. 2024), SparseVLM (Zhang et al. 2024b), PruMerge+ (Cao, Paranjape, and Hajishirzi 2023), DART (Wen et al. 2025), and VisionZip (Yang et al. 2024). Table 1 presents the performance on multiple benchmarks with 128, 64, and 32 visual tokens retained. Figure 4 presents a improvement in inference efficiency.

With a pruning rate of 77.8%, ICCTP retains 98% of the original performance. On 7 out of 10 datasets, ICCTP achieves the best or comparable results: for example, 53.0% on VizWiz (vs. 50.0% originally), 69.0% on SQAIMG (vs. 66.8% originally), and 85.2% on POPE (compared to 83.2% for VisionZip). When the pruning rate increases to 88.9%, ICCTP retains 94.8% of the original performance. On most datasets, ICCTP continues to lead or match the strongest baselines—for instance, 82.8% on POPE, 5.8% higher than VisionZip, and outperforming DART by 1.7% on VQAv2.

Under an extreme pruning ratio of 94.4%, ICCTP retains approximately 90.0% of the original performance while reducing TFLOPs by 82.4%. Across most datasets, ICCTP continues to demonstrate a clear advantage. For example, it reaches 80.4% on POPE, surpassing VisionZip by 11.7%; on SQAIMG, it achieves 69.4%, outperforming all competing methods and even exceeding the unpruned model by 2.6%. Overall, ICCTP outperforms the latest VisionZip by 1.0%, DART by 0.8%, and SparseVLM by 7.5%, highlighting that even when only 5.6% of visual tokens are retained, our method can still effectively capture and preserve the most critical cross-modal semantics and global information.

ICCTP with Higher Resolution

As shown in Table 2 and Figure 4, ICCTP consistently outperforms all baselines across pruning settings, , demonstrating strong robustness and significant efficiency gains. At 77.8% pruning, it achieves 72.4% accuracy, only 0.7% below the full LLaVA-NeXT-7B, and even surpasses the full model by 1.1% on POPE and 10.7 on MME. At 88.9% and 94.4% pruning, ICCTP outperforms the latest methods by 2.8% and 3.6%, while reducing TFLOPs by 87.5% and 92.3%, respectively. In addition, we also evaluate inference latency and GPU memory usage in high-resolution settings.

Method	VQAv2	GQA	VizWiz	SQAIMG	TextVQA	POPE	MME	MMB	MMBCN	MMVet	Acc.
All 576 Tokens (100%)											
LLaVA-1.5-7B	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	31.1	63.1
128 Tokens (↓77.8%)											
ToMe (ICLR23)	63.0	52.4	50.5	59.6	49.1	62.8	1088.4	53.3	48.8	27.2	52.1
FastV (ECCV24)	61.8	49.6	51.3	60.2	50.6	59.6	1208.9	56.1	51.4	28.1	52.9
SparseVLM (ICML25)	73.8	56.0	51.4	67.1	54.9	80.5	1376.2	60.0	51.1	30.0	59.4
PruMerge+ (2024.05)	74.7	57.8	52.4	67.6	54.3	81.5	1420.5	61.3	54.7	28.7	60.4
DART (2025.02)	74.7	57.9	52.8	69.1	56.3	80.4	1408.7	60.7	57.3	30.9	61.1
VisionZip (CVPR25)	75.6	57.6	52.0	68.9	56.8	83.2	1432.4	62.0	56.7	32.6	61.7
ICCTP (Ours)	75.6	57.8	53.0	69.0	56.8	85.2	1440.0	61.6	55.8	31.6	**61.8**
64 Tokens (↓88.9%)											
ToMe (ICLR23)	57.1	48.6	50.2	50.0	45.3	52.5	922.3	43.7	38.9	24.1	45.7
FastV (ECCV24)	55.0	46.1	50.8	51.1	47.8	48.0	1019.6	48.0	42.7	25.8	46.6
SparseVLM (ICML25)	68.2	52.7	50.1	62.2	51.8	75.1	1221.1	56.2	46.1	23.3	54.7
PruMerge+ (2024.05)	67.4	54.9	52.9	68.6	53.0	77.4	1198.2	59.3	51.0	25.9	57.0
DART (2025.02)	71.3	54.7	53.5	69.3	54.7	73.8	1365.1	59.5	54.0	26.5	58.6
VisionZip (CVPR25)	72.4	55.1	52.9	69.0	55.5	77.0	1365.6	60.1	55.4	31.7	59.7
ICCTP (Ours)	73.0	55.6	52.5	68.9	55.5	82.8	1382.7	60.3	54.6	25.9	**59.8**
32 Tokens (↓94.4%)											
ToMe (ICLR23)	46.8	43.6	51.3	41.4	38.3	39.0	828.4	31.6	28.1	17.3	37.9
FastV (ECCV24)	43.4	41.5	51.7	42.6	42.5	32.5	884.6	37.8	33.2	20.7	39.0
SparseVLM (ICML25)	58.6	48.3	51.9	57.3	46.1	67.9	1046.7	51.4	40.6	18.6	49.3
PruMerge+ (2024.05)	54.9	51.1	52.8	68.5	50.6	70.9	940.8	56.8	47.0	21.4	52.1
DART (2025.02)	67.1	52.9	52.5	69.3	52.2	69.1	1273.3	58.5	50.0	25.0	56.0
VisionZip (CVPR25)	67.1	51.8	52.9	68.8	53.1	68.7	1247.4	57.7	50.3	25.5	55.8
ICCTP (Ours)	68.0	52.9	51.9	69.4	53.3	80.4	1299.3	58.0	47.1	22.2	**56.8**

Table 1: Performance comparison of different pruning methods on LLaVA-1.5-7B across various benchmarks.

Method	VQAv2	GQA	TextVQA	POPE	MME	Acc.
All 2880 Tokens (100%)						
LLaVA-NeXT-7B	81.2	62.9	59.6	86.3	1513.8	73.1
640 Tokens (↓77.8%)						
FastV	78.9	60.4	58.4	83.1	1477.3	70.9
SparseVLM	78.2	59.1	56.2	80.9	1456.3	69.4
VisionZip	79.2	60.1	58.5	82.2	1468.4	70.7
ICCTP (Ours)	79.7	62.8	56.0	87.4	1524.5	**72.4**
320 Tokens (↓88.9%)						
FastV	71.9	55.9	55.7	71.7	1282.9	63.9
SparseVLM	71.4	56.5	52.4	73.5	1342.7	64.2
VisionZip	74.2	58.1	55.3	75	1348.8	66.0
ICCTP (Ours)	72.8	60.7	50.4	85.9	1481.4	**68.8**
160 Tokens (↓94.4%)						
FastV	61.8	49.8	51.9	51.7	1079.5	53.8
SparseVLM	62.2	50.2	45.1	54.6	1167.1	54.1
VisionZip	67.3	54.3	54.7	59.4	1239.7	59.5
ICCTP (Ours)	65.5	53.8	45.1	84.6	1327.1	**63.1**

Table 2: Performance comparison of different pruning methods on LLaVA-NeXT-7B across various benchmarks.

Method	VQAv2	GQA	TextVQA	POPE	MME	Acc.
All 576 Tokens (100%)						
LLaVA-1.5-13B	80.0	63.3	61.2	86.0	1531.2	73.4
144 Tokens (↓75%)						
FastV	77.2	59.9	60.1	79.4	1493.5	70.3
SparseVLM	76.1	58.0	57.9	68.6	1499.5	67.1
FasterVLM	77.4	58.7	59.0	83.1	1467.0	70.3
ICCTP (Ours)	77.1	58.4	57.1	84.5	1479.5	**70.2**
58 Tokens (↓90%)						
FastV	70.3	54.9	55.6	67.3	1359.7	63.2
SparseVLM	68.3	54.4	52.6	62.6	1285.3	60.4
FasterVLM	73.1	56.0	57.4	74.7	1370.8	65.9
ICCTP (Ours)	73.4	56.5	55.2	81.9	1422.2	**67.6**
29 Tokens (↓95%)						
FastV	62.3	50.3	52.1	49.8	1165.7	54.6
FasterVLM	67.9	52.6	54.8	65.9	1267.1	60.9
ICCTP (Ours)	68.9	53.9	52.7	79.4	1319.4	**64.2**

Table 3: Performance comparison of different pruning methods on LLaVA-1.5-13B across various benchmarks.

At a 94.4% pruning rate, latency on the POPE and MME datasets drops from 681 ms to 208 ms and from 1000 ms to 366 ms, respectively, while memory usage decreases from 16.41 GB to 14.65 GB and from 16.22 GB to 14.55 GB.

ICCTP with LLaVA-1.5-13B

LLaVA-1.5-13B has 13B parameters and 576 visual tokens, making efficient pruning essential. As shown in Table 3 and Figure 4, ICCTP retains 95.6%, 92.1%, and 87.5% of the original performance under 75%, 90%, and 95% pruning. At 90% and 95% pruning, it outperforms existing methods by 1.7% and 3.3%, with TFLOPs reduced by 80.3% and 84.7%, respectively. These results highlight the method's effectiveness in accurate token selection and efficient inference.

Ablation Study

Component Effectiveness Analysis: To evaluate the effectiveness of the two core components of our method, we conduct an ablation study on LLaVA-1.5-7B under a pruning ratio of 88.9%, and assess performance on three standard benchmarks: TextVQA, POPE, and MME, as shown in Figure 5. We compare three strategies: (1) Random selection of visual tokens; (2) CLS: importance-based selection using CLS-attention; and (3) ICCTP (Ours).

The results show that random token selection severely disrupts the image representation structure, while incorporating visual attention to identify important tokens yields significant gains: +5.0% on TextVQA (Acc), +2.5% on POPE (F1-score), and +50.4 on MME, validating the effectiveness of

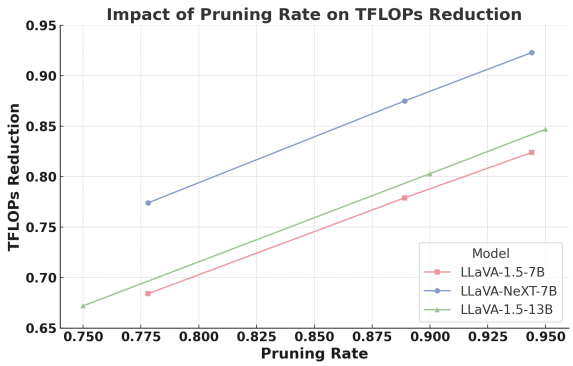

Figure 4: Comparison of TFLOPs reduction under different pruning rates across three models.

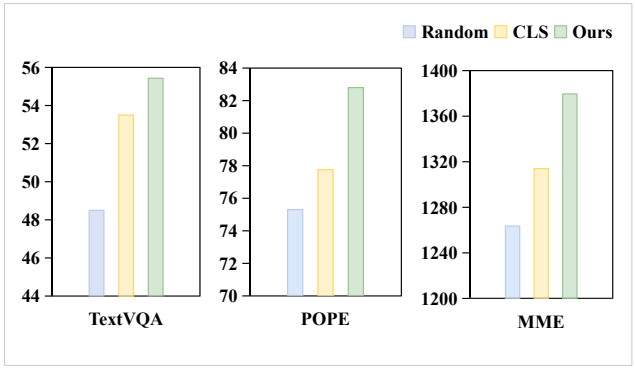

Figure 5: Ablation study of the core components.

preserving tokens with global contextual information.

Building upon this, our method further introduces a clustering-based pruning strategy guided by instruction-noun relevance, which significantly enhances semantic retention. Compared to the CLS-only strategy, ICCTP achieves additional improvements of +1.9%, +5.0%, and +65.6 on TextVQA, POPE, and MME, respectively, highlighting the importance of preserving tokens that are crucial for precise text perception and instruction grounding.

Analysis of the Global Retention Ratio ρ_g: We conduct experiments on LLaVA-1.5-7B with a pruning ratio of 88.9%, evaluating the effect of global retention ratio $\rho_g \in \{0.4, 0.6, 0.8\}$ across ten multimodal benchmarks. As shown in Figure 6, the model performs stably across different settings, with $\rho_g = 0.6$ achieving the best or near-best results on five datasets including GQA, TextVQA, and MME. This suggests that moderate global token retention, combined with instruction-guided selection, ensures robust performance under extreme pruning.

Analysis of the Semantic-Context Trade-Off Factor α: We evaluate the effect of the semantic-context trade-off factor $\alpha \in \{0.2, 0.3, 0.4\}$ on LLaVA-1.5-7B under a pruning ratio of 88.9% across ten multimodal benchmarks. As shown in Figure 7, the model exhibits overall stable performance, with $\alpha = 0.3$ performing better on POPE and MMBCN. This indicates that increasing the semantic weight benefits

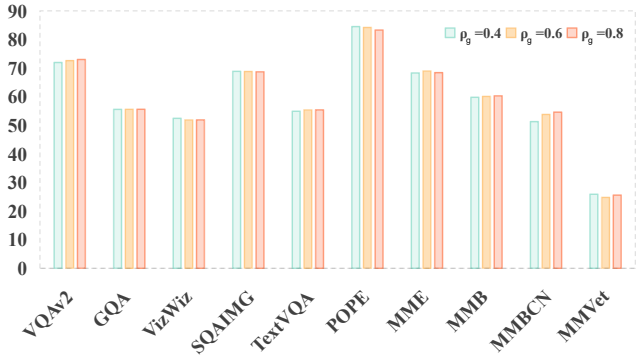

Figure 6: Global retention ratio analysis across ten datasets.

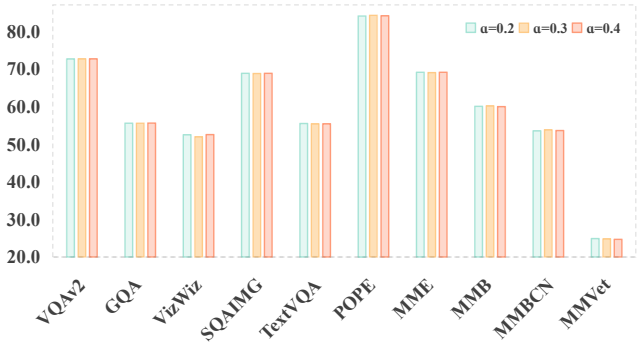

Figure 7: Analysis of the semantic-context trade-off factor across ten datasets.

complex reasoning and that $\alpha = 0.3$ achieves a good balance between semantic alignment and context preservation, effectively reducing clustering redundancy.

Conclusion

In this paper, we propose a novel instruction-centric visual token pruning framework, ICCTP. From the perspective of cross-modal information bottleneck attribution, we conduct an in-depth analysis of the semantic interaction between language instructions and visual features, revealing that nouns in instructions serve as stable semantic anchors within the cross-modal representation space. Building on this insight, we design a cross-modal clustering-based pruning strategy that leverages nouns as clustering centers, while incorporating global attention to preserve the overall visual context. This significantly enhances the semantic relevance and stability of the pruning process.

ICCTP is plug-and-play and requires no additional training, ensuring compatibility with mainstream LVLM architectures. It maintains over 90% of the original performance even under an aggressive pruning ratio of 94.4%, while substantially reducing inference costs. These results demonstrate the feasibility and generality of token pruning guided by informative linguistic cues. Furthermore, our method lays a solid foundation for future research on efficient representation compression in tasks such as fine-grained instruction understanding and multi-granular semantic guidance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62276054 and 62406060; in part by the Sichuan Science and Technology Program under Grant No. 2025ZNSFSC1500; and in part by the Xiaomi Young Scholar Program.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bolya, D.; Fu, C.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *Proc. ICLR*. Token merging for faster ViT inference without retraining.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, Q.; Paranjape, B.; and Hajishirzi, H. 2023. PuMer: Pruning and Merging Tokens for Efficient Vision Language Models. *arXiv preprint arXiv:2305.17530*. Text-informed pruning and modality-aware merging.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An Image Is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models (FastV). FastV introduces early-layer token ranking and pruning in LVLMs to significantly reduce FLOPs with minimal performance loss, arXiv:2403.06764.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*. First benchmark to evaluate MLLMs across perception and cognition via 14 subtasks.

Goyal, Y.; Khot, T.; Agrawal, A.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2019. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, 127(4): 398–414. CVPR 2017 challenge paper and IJCVi publication.

Guo, Y.; Zhang, H.; Wong, Y.; Nie, L.; and Kankanhalli, M. 2023. ELIP: Efficient Language–Image Pre-training with Fewer Vision Tokens. *arXiv preprint arXiv:2309.16738*. Vision token pruning and merging supervised by language outputs.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Huang, K.; Zou, H.; Xi, Y.; Wang, B.; Xie, Z.; and Yu, L. 2024. Instruction-Guided Visual Token Pruning for Large Vision-Language Models. In *Computer Vision – ECCV 2024, Lecture Notes in Computer Science*, 214–230. Springer International Publishing. Instruction-guided token pruning for LVLMs.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; and Jia, J. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. 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, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; Chen, K.; and Lin, D. 2023b. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Luan, B.; Zhou, W.; Feng, H.; Wang, Z.; Li, X.; and Li, H. 2025. Multi-Cue Adaptive Visual Token Pruning for Large Vision-Language Models. *arXiv preprint arXiv:2503.08019*. Integrates attention, spatial, similarity via adaptive NMS.

Luo, Y.; An, R.; Zou, B.; Tang, Y.; Liu, J.; and Zhang, S. 2024. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, 235–252. Springer.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Schulz, K.; Sixt, L.; Tombari, F.; and Landgraf, T. 2020. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Sun, Y.; Xin, Y.; Li, H.; Sun, J.; Lin, C.; and Batista-Navarro, R. 2025. LVPruning: An Effective yet Simple Language-Guided Vision Token Pruning Approach for Multi-modal Large Language Models. *arXiv preprint arXiv:2501.13652*. Language-guided vision token pruning for MLLMs.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37: 121475–121499.

Wang, Y.; Rudner, T. G.; and Wilson, A. G. 2023. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36: 16009–16027.

Wen, Z.; Gao, Y.; Wang, S.; Zhang, J.; Zhang, Q.; Li, W.; He, C.; and Zhang, L. 2025. Stop Looking for "Important Tokens" in Multimodal Language Models: Duplication Matters More. *arXiv preprint arXiv:2502.11494*. DART prunes tokens based on duplication rather than importance.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; and Lin, D. 2024. PyramidDrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. *arXiv preprint arXiv:2410.17247*. Visual redundancy reduction strategy for LVLMs with staged token dropping.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2024. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *CVPR 2025*. Informative token compression maintaining performance.

Ye, W.; Wu, Q.; Lin, W.; and Zhou, Y. 2024. Fit and Prune: Fast and Training-free Visual Token Pruning for Multi-modal Large Language Models. *arXiv preprint arXiv:2409.10197*. Training-free visual token pruning by minimizing attention divergence.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. *arXiv preprint arXiv:2308.02490*. Benchmark for holistic evaluation of generalist multimodal models.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2025. Beyond Text-Visual Attention: Exploiting Visual Cues for Effective Token Pruning in VLMs. arXiv:2412.01818.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024a. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. *arXiv e-prints*, arXiv–2412.

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