Global Compression Commander: Plug-and-Play Inference Acceleration for High-Resolution Large Vision-Language Models

Xuyang Liu[1*], Ziming Wang[2], Junjie Chen[1], Yuhang Han[3], Yingyao Wang[2], Jiale Yuan[2], Jun Song[2†], Siteng Huang[4], Honggang Chen[1, 5†]

[1] Sichuan University,
[2] Taobao & Tmall Group of Alibaba,
[3] Westlake University,
[4] Zhejiang University,
[5] Police Integration Computing Key Laboratory of Sichuan Province
liuxuyang@stu.scu.edu.cn, honggang_chen@scu.edu.cn

Abstract

Large vision-language models (LVLMs) excel at visual understanding, but face efficiency challenges due to quadratic complexity in processing long multi-modal contexts. While token compression can reduce computational costs, existing approaches are designed for single-view LVLMs and fail to consider the unique multi-view characteristics of high-resolution LVLMs with dynamic cropping. Existing methods treat all tokens uniformly, but our analysis reveals that global thumbnails can naturally guide the compression of local crops by providing holistic context for informativeness evaluation. In this paper, we first analyze dynamic cropping strategy, revealing both the complementary nature between thumbnails and crops, and the distinctive characteristics across different crops. Based on our observations, we propose "Global Compression Commander" (*i.e.*, **GlobalCom^2**), a novel plug-and-play token compression framework for HR-LVLMs. GlobalCom^2 leverages thumbnail as the "commander" to guide the compression of local crops, adaptively preserving informative details while eliminating redundancy. Extensive experiments show that GlobalCom^2 maintains over **90%** performance while compressing **90%** visual tokens, reducing FLOPs and peak memory to **9.1%** and **60%**.

Code — https://github.com/xuyang-liu16/GlobalCom2
Extended version — https://arxiv.org/abs/2501.05179

1 Introduction

By bridging visual encoders with large language models (LLMs) (Touvron et al. 2023; Yang et al. 2024), large vision-language models (LVLMs) (Liu et al. 2024a; Bai et al. 2023) have recently achieved remarkable progress. As LVLMs advance towards high-resolution image understanding (HR-LVLMs), dynamic cropping has emerged as a de facto standard, represents a high-resolution image as a global thumbnail combined with a set of local crops. While this enables models like LLaVA-NeXT (Liu et al. 2024b) and InternVL 3 (Zhu et al. 2025) to capture fine-grained details with vision encoding efficiency, it also introduces challenges with increased visual tokens and hierarchical visual context.

Figure 1: **Design philosophy of "global-to-local" guided token compression.** GlobalCom^2 evaluates the *information richness* of local crops from a global perspective to preserve informative regions while removing redundant ones.

To enhance the efficiency of LVLMs, recent efforts have increasingly adopted token compression approaches (Cha et al. 2024; Shang et al. 2025; Liu et al. 2025b,c,a; Wen et al. 2025c), which reduce visual tokens while preserving essential information. These architecture-agnostic methods achieve optimal efficiency-accuracy trade-offs and have proven effective for LVLM acceleration. However, these methods were primarily designed for traditional *single-view* architectures (the entire image in Figure 1 top). While the *multi-view* dynamic cropping approach enables more fine-grained visual understanding, it further increases the number of visual tokens and computational overhead.

However, directly applying existing token compression methods to HR-LVLMs overlooks ***three critical issues***: **(i) Global context neglect:** Current methods disregard the cru-

*This work was done during an internship at Alibaba.
†Corresponding author.

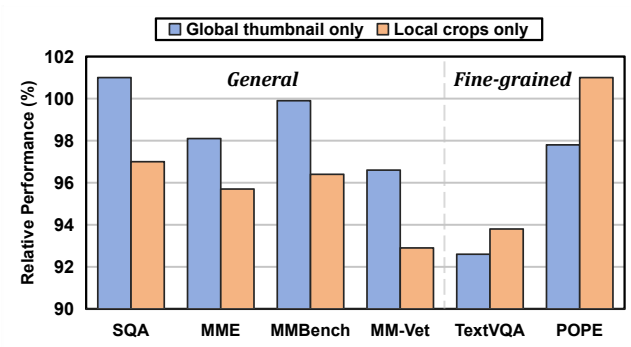

Figure 2: **Complementary roles of global thumbnail and local crops in HR-LVLMs with dynamic cropping.** Performance (%) denotes relative scores of LLaVA-NeXT-7B.

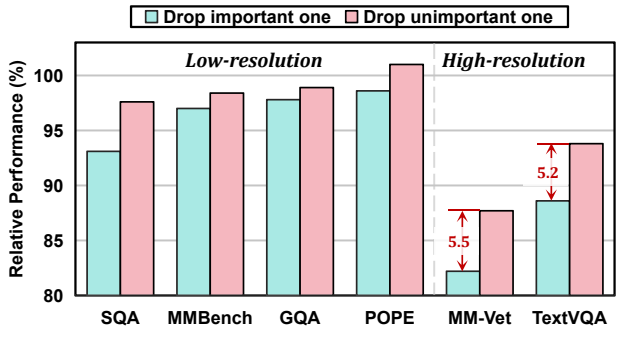

Figure 3: **Varying contributions of local crops.** Importance is quantified by the accumulated attention scores between thumbnail patches and `[CLS]` token within each crop.

cial role of global thumbnails in extracting holistic context and guiding visual understanding (Figure 2). **(ii) Information richness disparity:** They fail to account for varying information density across different crops (Figure 3), resulting in performance gaps exceeding **5%** on high-resolution tasks when comparing the removal of most versus least informative crops. **(iii) Content-agnostic positional bias:** Inner-LLM question-aware compression methods (Chen et al. 2024; Xing et al. 2025; Zhang et al. 2025) systematically allocate more tokens to later-positioned crops irrespective of their actual informative content (Figure 4), inducing severe multi-modal hallucinations under extreme compression (Table 1). These three oversights collectively result in **over-compression** of information-rich regions and disruption of the visual-semantic hierarchy (Figure 7).

To bridge this critical gap, we first conduct systematic analysis of dynamic cropping in Section 3, identifying ***two key observations***: ❶ Thumbnail and crop tokens serve complementary roles - thumbnails capture holistic context while crops provide fine-grained details, enabling global importance evaluation. ❷ Through global context, tokens from different crops exhibit varying informativeness, requiring differentiated compression to minimize information loss.

Building on these observations, we propose "**Global Compression Commander**" (**GlobalCom**^{2}), which follows a "***global-to-local***" guided token compression philosophy tailored for dynamic cropping-based HR-LVLMs. In Figure 1, GlobalCom^2 leverages holistic information from thumbnails to evaluate each crop's information richness, adaptively adjusting compression intensity. It performs token compression through comprehensive evaluation from both global and local perspectives, achieving differentiated compression across regions while preserving significant information. This approach can be integrated with existing question-aware methods. Notably, integrating this "global-to-local" design with FastV (Chen et al. 2024) and SparseVLM (Zhang et al. 2025) achieves significant improvements of **5.3%** and **5.2%** across benchmarks while alleviating multi-modal hallucination caused by positional bias.

To summarize, our main contributions are three-fold:

- **Systematic Dynamic Cropping Analysis:** We empirically analyze the hierarchical nature of dynamic cropping and thoroughly investigate existing token compression methods for HR-LVLMs, identifying fundamental causes of critical information over-compression.
- **Global-to-Local Compression Philosophy:** We propose GlobalCom^2, a training-free framework that adaptively adjusts compression intensity based on crop information richness assessment and preserves semantically important tokens across the entire visual hierarchy.
- **Superior Performance-Efficiency Trade-offs:** Extensive experiments on multiple HR-LVLMs with dynamic cropping demonstrate exceptional balance of GlobalCom^2, maintaining over 90% performance with 90% token reduction, while achieving substantial memory savings and throughput improvements.

2 Related Work

High-resolution LVLMs. LVLMs integrate vision encoders and LLM decoders via projectors for feature alignment (Liu et al. 2023; Dai et al. 2023). Early LVLMs (Liu et al. 2024a; Bai et al. 2023) resize images to fixed resolutions, causing shape distortion and detail loss. To address this, high-resolution LVLMs have emerged in three categories: **(i)** Hybrid resolution methods using dual visual encoders (Li et al. 2024; Luo et al. 2025). **(ii)** Native resolution methods with NaViT-style encoders (Wang et al. 2024; Guo et al. 2025). **(iii)** Dynamic cropping methods (Li et al. 2025; Zhu et al. 2025), which split images into regions for single-encoder processing. Dynamic cropping has gained widespread adoption (Lu et al. 2025; Chen et al. 2025) due to its vision encoding efficiency. However, increased visual tokens introduce computational challenges in inference speed and memory usage. Our work focuses on improving the efficiency of dynamic cropping-based HR-LVLMs.

Token Compression for LVLMs. Token compression, which aims to reduce the sequence length of tokens for computation efficiency, has been widely adopted for model acceleration (Rao et al. 2021; Liang et al. 2022; Bolya et al. 2023). For LVLMs, recent works have focused on training-free token compression for LVLM acceleration at

Figure 4: **Content-agnostic positional bias.** LLM attention-guided methods (*e.g.*, FastV) assign higher scores (bars) to later tokens, regardless of their content or input order (second row: sequential crops; third row: reversed crops).

two stages: **(i)** Vision Encoding stage (Shang et al. 2025; Yang et al. 2025): FasterVLM (Zhang et al. 2024) leverages `[CLS]` attention scores to prune visual tokens. **(ii)** LLM Pre-filling stage (Xing et al. 2025; Ye et al. 2025; Wen et al. 2025b): FastV (Chen et al. 2024) prunes tokens based on LLM self-attention, while SparseVLM (Zhang et al. 2025) uses cross-modal attention scores. Some methods combine both stages (Liu et al. 2024c; Han et al. 2024). However, these methods treat all visual tokens equally in a "flat" token space, ignoring HR-LVLMs' hierarchical structure where thumbnails and crops serve distinct roles.

Our work is the first to systematically quantify this oversight's severe consequences, revealing critical failure modes including uniform over-compression and positional bias. We address this by proposing a "global-to-local" compression framework that leverages hierarchical structure, representing a shift toward structure-aware token compression.

3 Analysis of Dynamic Cropping

3.1 Preliminary

We take LLaVA-NeXT, a widely-adopted HR-LVLM with dynamic cropping, as an example to conduct our analysis.

Model Architecture. Given input image and text, LLaVA-NeXT generates responses through: **(i)** Vision Encoding: ViT (Radford et al. 2021) converts pixels to embeddings via MLP projector. **(ii)** LLM Decoding: LLM processes concatenated tokens, generating responses auto-regressively.

Pre-processing. LLaVA-NeXT uses grid templates: $\{2 \times 2, 1 \times \{2,3,4\}, \{2,3,4\} \times 1\}$. Images of size $W \times H$ are scaled to $(336 \times a) \times (336 \times b)$, yielding $n = a \times b$ local crops $\mathbf{X}^L$ and a thumbnail $\mathbf{X}^G$. Total token length is $(1+n) \times N$.

Post-processing. LLaVA-NeXT removes padding tokens and adds boundary tokens to mark image regions, preserving aspect ratios while enhancing efficiency.

3.2 Stepping into Dynamic Cropping

HR-LVLMs process high-resolution images using dynamic cropping with thumbnails and crops. We analyze their characteristics to guide token compression design.

(I) Functions of Thumbnail and Crops: We begin with an exploratory experiment to investigate how thumbnail and crops contribute to image understanding in HR-LVLMs by using them separately as input to LLaVA-NeXT-7B.

Figure 2 demonstrates that using global thumbnail alone yields superior performance on general visual perception benchmarks (*e.g.*, SQA (Lu et al. 2022) and MMBench (Liu et al. 2024d)). This advantage stems from global thumbnails providing holistic visual information through complete ViT encoding. In contrast, using only local crops, where ViT independently encodes each crop, shows inferior performance on these general visual perception tasks. However, local crops excel in fine-grained perception tasks like TextVQA (Singh et al. 2019) (VQA^{T}, testing text-centric visual understanding) and POPE (Li et al. 2023) (testing model hallucination) by providing detailed visual features. Therefore, we identify that:

Observation ❶: Global thumbnails and local crops serve complementary functions in HR-LVLMs with dynamic cropping: the former acts as a "comprehensive visual extractor" for holistic representations, while the latter serves as a "detailed visual capturer" for fine-grained representations.

(II) Information Richness in Different Crops: Given that dynamic cropping leads to distinct visual representations across crops, we investigate the variance through both qualitative and quantitative perspectives.

Through visual inspection of the input image in Figure 5, we observe that the upper crops contain rich visual content (*e.g.*, football players), while the lower crops mainly show redundant grass areas. To quantitatively validate this observation, we analyze the attention scores between `[CLS]` and patch tokens, which have been shown to effectively indicate token importance in prior works (Liang et al. 2022; Han et al. 2024). As visualized in Figure 5, the attention distribution from CLIP-ViT's last layer shows significantly higher values in the upper regions, confirming that upper crops contain more semantic information from a global perspective.

Building upon this qualitative analysis, we conduct quantitative studies to validate our observations. Since attention scores with the `[CLS]` token effectively measure token-level semantic richness, we leverage these scores to assess crop importance by by computing the sum of partitioned attention scores within each crop region. To examine crop contributions in HR-LVLMs, we conduct experiments on LLaVA-NeXT-7B by selectively dropping local crops (*i.e.*, dropping the most/least important one). As shown in Figure 3, dropping the most versus least important crops leads to significant performance gaps across visual understanding tasks, with an average drop of **2.4%** across six benchmarks and particularly notable in VQA^{T} (**5.2%** gap). Based on the above analysis, we identify that:

Observation ❷: Local crops exhibit varying information richness in global context, leading to different contributions to the overall visual understanding of HR-LVLMs with dy-

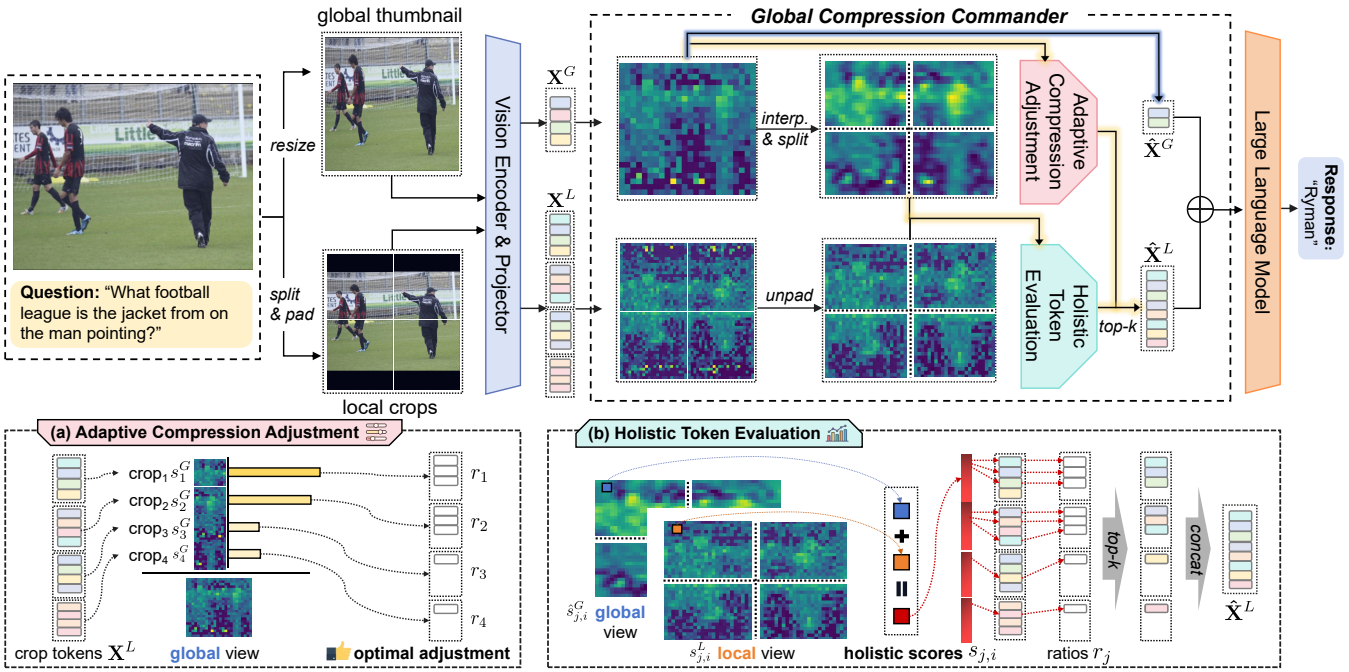

Figure 5: **Overall framework.** GlobalCom^2 guides token compression for HR-LVLMs through: 1) compressing thumbnail tokens (blue path), and 2) compressing crop tokens (yellow paths) by (a) adaptively adjusting compression intensity based on global visual richness, and (b) performing compression according to token informativeness from global and local perspectives.

namic cropping, with visually informative crops being particularly crucial for capturing fine-grained local details.

4 Global Compression Commander

HR-LVLMs with dynamic cropping increase token length (**5×** in LLaVA-NeXT (Liu et al. 2024b), **10×** in LLaVA-OV (Li et al. 2025)), making self-attention's quadratic complexity a bottleneck. Inspired by human vision's process of grasping scene gist before focusing on details, we propose "**Global Com**pression **Com**mander" (**GlobalCom**^{2}) which implements a "global-to-local" compression strategy.

In Figure 5, for thumbnail token compression, we identify tokens with essential holistic information. Given the `[CLS]` token's effectiveness as global image representation (Liang et al. 2022), GlobalCom^2 uses the last ViT layer's attention map to compute attention between each token and `[CLS]` (blue path in Figure 5). For the 1D token sequence $\mathbf{X}^G$ of length N, the importance score s_i^G for the i-th token is:

$$\mathbf{s}_i^G = \frac{\exp\left(\mathbf{q}^{\text{CLS}}\mathbf{K}_i^\top/\sqrt{D}\right)}{\sum_{i=1}^{N}\exp\left(\mathbf{q}^{\text{CLS}}\mathbf{K}_i^\top/\sqrt{D}\right)}, \tag{1}$$

where $\mathbf{q}^{\text{CLS}}$ and $\mathbf{K} \in \mathbb{R}^{N\times D}$ are query of `[CLS]` and keys of $\mathbf{X}^G$. Given a preset retention ratio R (%), GlobalCom^2 preserves the top-k ($k = R \times N$) tokens ranked by s^G:

$$\mathbf{X}^G \rightarrow \hat{\mathbf{X}}^G = \text{TopK}(\mathbf{X}^G, s^G, R \times N). \tag{2}$$

While thumbnail compression focuses on preserving holistic context, crop compression faces more complex challenges due to varying information densities across different crops. Following observation ❷, semantically rich crops should preserve more tokens for crucial details, while less informative ones can be compressed more aggressively.

GlobalCom^2 leverages the comprehensive visual knowledge from thumbnails to guide crop compression through a decoupled two-stage process: **(a)** Adaptive Compression Adjustment, which dynamically adapts the compression intensity for each crop based on its information richness, and **(b)** Holistic Token Evaluation, which assesses the informativeness of tokens from both local and global views.

4.1 Adaptive Compression Adjustment

As shown in bottom-left of Figure 5, GlobalCom^2 analyzes each local crop's semantic contribution and applies adaptive token compression accordingly.

For differentiated compression, we compute each crop's *information richness score* s_j^G by accumulating patch-to-`[CLS]` attention scores in its corresponding global thumbnail region: $s_j^G = \sum_{i\in\text{crop}_j} s_i^G$. We then normalize scores with $\tilde{s}_j = (s_j^G - \max(s_j^G))/\tau$ ($\tau = 10$) and compute relative importance weight σ_j via softmax:

$$\sigma_j = \frac{\exp(\tilde{s}_j)}{\sum_{l=1}^{n}\exp(\tilde{s}_l) + \epsilon}, \tag{3}$$

where $\epsilon = 10^{-8}$ prevents division by zero. The final retention ratio r_j for each crop is adjusted from the preset ratio $R(\%)$ based on its global content importance:

$$r_j = R \times \left(1 + \sigma_j - \frac{1}{n}\right), \tag{4}$$

Method	GQA	VizWiz	SQA	MMB	POPE	VQA^T	MME	MM-Vet	Average
Upper Bound, 2880 Tokens									
LLaVA-NeXT-7B	64.2	57.6	70.1	67.4	86.5	64.9	1519.0	43.9	100.0%
Ratio=50%, Retain up to 1440 Tokens									
FastV (ECCV24)	61.8	54.9	69.0	67.4	85.5	59.6	1490.3	37.6	95.5%
PDrop (CVPR25)	63.7	**57.9**	**69.2**	**67.7**	87.9	61.6	1499.6	37.5	97.4%
SparseVLM (ICML25)	63.7	57.2	68.3	67.6	87.9	60.5	1507.2	36.8	96.8%
FasterVLM (2024.12)	63.4	56.4	69.1	67.4	87.7	58.9	1533.3	39.6	97.3%
GlobalCom^2	**63.9**	56.5	68.5	67.6	**88.1**	**62.3**	**1552.9**	**40.4**	**98.5%**
Ratio=25%, Retain up to 720 Tokens									
FastV (ECCV24)	60.4	54.2	**68.8**	65.6	83.1	58.4	1477.3	35.4	93.4%
PDrop (CVPR25)	60.3	**56.8**	68.5	65.6	85.5	59.8	1473.7	31.1	93.3%
SparseVLM (ICML25)	59.9	56.0	67.5	65.6	85.0	58.3	1465.9	38.5	94.6%
FasterVLM (2024.12)	61.3	55.4	67.1	**66.0**	87.2	58.8	1454.6	37.8	94.8%
GlobalCom^2	**61.5**	55.7	68.1	65.9	**87.6**	**60.9**	**1493.5**	**40.7**	**96.7%**
Ratio=10%, Retain up to 288 Tokens									
FastV (ECCV24)	55.9	53.1	68.1	61.6	71.7	55.7	1282.9	27.2	85.4%
PDrop (CVPR25)	54.5	54.4	67.7	59.0	77.6	54.4	1262.1	24.0	84.3%
SparseVLM (ICML25)	56.3	52.1	68.5	60.0	80.1	53.9	1334.2	26.5	86.1%
PruMerge (ICCV25)	53.6	54.0	66.4	61.3	60.8	50.6	1149.3	25.5	80.6%
FasterVLM (2024.12)	56.9	52.6	66.5	61.6	83.6	56.5	1359.2	35.0	89.9%
GlobalCom^2	**57.1**	**54.6**	**68.7**	**61.8**	**83.8**	**58.4**	**1365.5**	**36.4**	**91.6%**

Table 1: **Comparisons with LLaVA-NeXT across image understanding benchmarks.** VQA^T (TextVQA), MME, MM-Vet are high-resolution benchmarks. "Average" shows mean performance across benchmarks, with **best** results highlighted.

where $\sigma_j - \frac{1}{n}$ is deviation from average importance, allocating more tokens to important content and fewer to unimportant, for adaptive content-aware compression.

Through this process, GlobalCom^2 allocates compression degrees (*i.e.*, $\{r_1, r_2, r_3, r_4\}$ in Figure 5) based on each crop's information richness from the global view.

4.2 Holistic Token Evaluation

After determining compression degrees for each crop, GlobalCom^2 evaluates token importance for preservation (Figure 5). For each crop, attention between patch tokens and `[CLS]` yields *local importance scores* $\{s_j^L\}_{j=1}^n$ from final-layer attention. These only capture within-crop importance, missing cross-region elements. Following observation ❶, GlobalCom^2 incorporates global thumbnail context by reshaping 1D attention scores s^G to 2D format and applying bilinear interpolation to match original dimensions, yielding $\{\hat{s}_j^G\}_{j=1}^n$ sub-maps per crop. The *holistic score* $s_{j,i}$ for the i-th token in the j-th crop is:

$$s_{j,i} = \alpha \hat{s}_{j,i}^G + (1-\alpha) s_{j,i}^L, \tag{5}$$

where we empirically set $\alpha = 0.5$ to give equal consideration to both information sources. The final compression:

$$\mathbf{X}_j^L \rightarrow \hat{\mathbf{X}}_j^L = \text{TopK}(\mathbf{X}_j^L, s_j, r_j \times N). \tag{6}$$

This holistic evaluation identifies globally significant tokens while preserving local details.

4.3 GlobalCom^2 without `[CLS]` Token

For HR-LVLMs without `[CLS]` token (*e.g.*, LLaVA-OneVision (Li et al. 2025) with SigLIP (Zhai et al. 2023)), we propose an alternative token informativeness measure for GlobalCom^2. Specifically, given a sequence of vision tokens $\mathbf{X} \in \mathbb{R}^{N \times d}$ after vision encoding, we first compute a global mean vector $\mathbf{g} \in \mathbb{R}^d$ through global average pooling over all tokens, and then calculate the cosine similarity between each patch token $\mathbf{x}_i$ and $\mathbf{g}$:

$$c_i = \cos(\mathbf{x}_i, \mathbf{g}) = \frac{\mathbf{x}_i \cdot \mathbf{g}}{\|\mathbf{x}_i\| \|\mathbf{g}\|}, \tag{7}$$

The informativeness score $s_i = -c_i$ is negatively correlated with the similarity, reflecting that tokens with greater difference from $\mathbf{g}$ carry more unique information. Specifically, tokens exhibiting low similarity to $\mathbf{g}$ represent distinctive and irreplaceable visual elements, while highly similar tokens typically correspond to redundant or common patterns. This scoring mechanism serves as an effective alternative to `[CLS]`-based scoring in Equation 2-5 for models without `[CLS]` token. It is adopted to evaluate both crop-level information richness and token-level importance in LLaVA-OneVision. We conduct comprehensive quantitative and qualitative analyses in Appendix to explore how to measure token informativeness without `[CLS]` token.

5 Experiments

5.1 Experimental Setting

We evaluate on LLaVA-NeXT (Liu et al. 2024b) and LLaVA-OneVision (Li et al. 2025) for evaluation. We compare with FastV (Chen et al. 2024), SparseVLM (Zhang et al. 2025), PDrop (Xing et al. 2025), PruMerge (Shang et al. 2025), FasterVLM (Zhang et al. 2024) at different retention ratios R. For fair comparison with multi-stage meth-

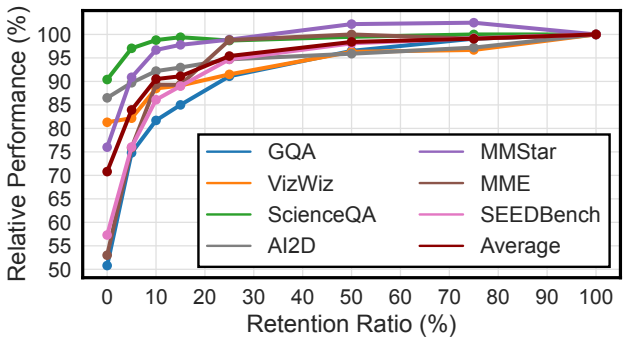

Figure 6: **Results of GlobalCom^2 on LLaVA-OneVision for image understanding.** GlobalCom^2 achieves 90.5% average performance with only 10% visial tokens.

ods, we use "equivalent token count" reflecting the average percentage of vision tokens retained across all LLM layers.

5.2 Main Results

Results on LLaVA-NeXT. Table 1 compares GlobalCom^2 with existing methods with LLaVA-NeXT, which demonstrates its three key advantages: **(i) Superior performance:** GlobalCom^2 consistently outperforms all baselines, being the only method that maintains **>90%** of the original performance across all retention ratios. **(ii) Extreme compression robustness:** While baseline methods suffer significant degradation at $R = 10\%$, GlobalCom^2 maintains robust performance and achieves the best results across all benchmarks, outperforming the second-best method by **1.7%** on average. Notably, FastV and PDrop demonstrate severe deterioration on POPE, exhibiting clear multi-modal hallucination due to positional bias from attention-guided token selection (Figure 7). **(iii) High-resolution excellence:** GlobalCom^2 demonstrates exceptional performance on high-resolution benchmarks (*e.g.*, VQA$^{\text{T}}$, MME, MM-Vet). Our "global-to-local" guided compression preserves both global semantics and local details, outperforming baselines that suffer from over-compression.

Results on LLaVA-OneVision. Figure 6 further presents GlobalCom^2's performance across benchmarks under varying retention ratios R on LLaVA-OneVision. Generally, model performance correlates with R, with more aggressive compression leading to degradation. Vision-centric tasks (*e.g.*, GQA, VizWiz, MME, SEED) show substantial drops with reduced visual tokens, while SQA maintains robust performance even with minimal tokens, suggesting language understanding dominates in scientific reasoning. Notably, GlobalCom^2 preserves **90.5%** performance at $R = 10\%$ while consuming only **35.4%** of the original GPU memory, all without any training overhead, demonstrating its effective and efficient token compression.

5.3 Ablation Study and Analysis

Ablation on Adaptive Compression Adjustment. Table 2 compares four settings: (a) "Uniform" baseline with $R = 25\%$ across all crops, and three adaptive strategies: (b) "$\mathbf{n}_{\text{top-}k}$" adjusting compression based on top-k most informative tokens per crop ($k = 25\% \times N$), (c) "Softmax (max)" applying softmax over maximum token importance score s_j^G in each crop's thumbnail, and (d) "Softmax (sum)" (our choice) computing softmax over sum of token importance scores s_j^G in each crop's thumbnail region. All adaptive strategies outperform uniform compression, with "Softmax (sum)" achieving the best performance. While $\mathbf{n}_{\text{top-}k}$ and "Softmax (max)" focus on strongest visual features per crop without considering global importance, "Softmax (sum)" adjusts compression based on each crop's overall importance to the entire image, preserving more semantic information and helping the LLM capture finer visual details.

Method	SQA	POPE	VQA$^{\text{T}}$	MME	MM-Vet	Avg.
Upper Bound, 2880 Tokens						
Vanilla	70.1	86.5	64.9	1519.0	43.9	100.0%
Ratio=25%, Retain up to 720 Tokens						
Uniform	67.1	87.2	60.1	1454.6	37.8	94.2%
$\mathbf{n}_{\text{top-}k}$	67.4	87.3	59.8	1471.5	35.7	94.5%
Softmax (max)	67.3	87.2	60.3	1462.6	38.4	94.7%
Softmax (sum)	**67.6**	**87.4**	**60.6**	**1473.3**	**39.6**	**95.6%**

Table 2: **Effects of different adjustment strategies.** "Uniform" performs no compression adjustment, while the other three strategies enable adaptive compression adjustment.

Method	SQA	POPE	VQA$^{\text{T}}$	MME	MM-Vet	Avg.
Upper Bound, 2880 Tokens						
Vanilla	70.1	86.5	64.9	1519.0	43.9	100.0%
Ratio=25%, Retain up to 720 Tokens						
Local only	67.6	87.4	60.6	1473.3	39.6	95.6%
Global only	67.9	86.4	60.2	1488.5	37.8	94.7%
Global and Local	**68.1**	**87.6**	**60.9**	**1493.5**	**40.7**	**96.7%**

Table 3: **Effects of different token evaluation metrics.** For the i-th token in crop j, "Global" and "Local" refer to importance scores $\hat{s}_{j,i}^G$ and $s_{j,i}^L$ in Equation (5), respectively.

Ablation on Holistic Token Evaluation. Table 3 compares different token evaluation strategies. While both strategies are effective, each has limitations: Local-only evaluation excels at fine-grained tasks (VQA$^{\text{T}}$, POPE) but underperforms on general perception benchmarks (MME, SQA) due to missing global context. Global-only evaluation maintains general perception but may overlook crucial local details. GlobalCom^2 achieves optimal performance by combining them, leveraging their complementary strengths.

Combination with Question-aware Methods. Figure 8 explores combining GlobalCom^2 with question-aware methods FastV and SparseVLM, enabling joint consideration of *textual relevance and visual importance*. Using our Adaptive Compression Adjustment strategy, we assign optimal compression intensities (*i.e.*, r_j for the j-th crop) per crop based on its visual information richness within the global context before applying FastV/SparseVLM's token evaluation metrics. Under extreme compression ($R = 10\%$), incorporating GlobalCom^2 yields average improvements of **5.3%** and **5.2%** for FastV and SparseVLM. Notably, GlobalCom^2 sig-

Figure 7: **Visualization of different token compression methods.** Gray masks indicate discarded tokens, where other methods exhibit significant *over-compression* issues, while GlobalCom^2 preserves both global important and local detailed information.

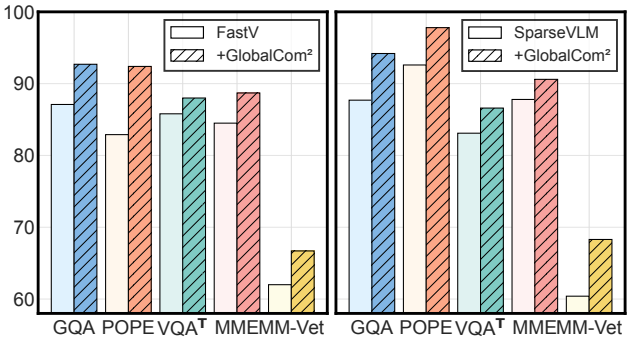

Figure 8: **Combination with question-aware methods.** "+GlobalCom^2" indicates the application of our Adaptive Compression Adjustment strategy under $R = 10\%$.

nificantly boosts performance on POPE, with improvements of **8.2** for FastV and **4.5** for SparseVLM, confirming our strategy's effectiveness in *mitigating positional bias* by preventing over-compression of important regions.

5.4 Efficiency Analysis

Table 4 compares inference efficiency among SparseVLM, FasterVLM and GlobalCom^2. SparseVLM requires explicit attention scores in LLM and is naturally incompatible with FlashAttention (Dao et al. 2022), leading to higher memory costs (Wen et al. 2025a). Instead, FasterVLM and our GlobalCom^2 enable efficient computation before LLM decoding for efficient computation. As a plug-and-play solution, GlobalCom^2 achieves superior performance-efficiency trade-offs, maintaining **90%** of the original performance while dramatically reducing peak memory usage by **40%** and boosting inference throughput by $1.8\times$.

5.5 Compression Visualizations

Figure 7 compares compression processes of different methods at extreme compression setting of $R = 10\%$ on VQA$^{\text{T}}$,

Method	TFLOPs↓	Memory↓	Throughput↑	Performance↑
Upper Bound, 2880 Tokens				
Vanilla	41.7	23.0	3.8	100%
Ratio=10%, Retain up to 288 Tokens				
SparseVLM	5.4 (↓87.1%)	24.2 (↑5.2%)	5.9 (1.6×)	85.7%
FasterVLM	**3.8** (↓90.9%)	**13.6** (↓40.1%)	**6.7** (1.8×)	89.5%
GlobalCom^2	**3.8** (↓90.9%)	13.9 (↓40.0%)	**6.7** (1.8×)	**90.8%**

Table 4: **Efficiency comparisons.** "Memory": peak GPU memory; "Throughput": POPE samples/second; "Performance": average score on eight multi-modal benchmarks.

revealing over-compression issues in baselines: **(i) Positional Bias:** FastV exhibits clear positional bias, allocating more tokens (>3×) to later-positioned crops regardless of visual content. **(ii) Uniform Compression:** FasterVLM applies uniform compression across all crops, treating them as equally important. It fails to preserve critical information in some regions while retaining redundancy in others. In contrast, GlobalCom^2 globally assesses crop informativeness and adaptively adjusts compression ratios, preserving crucial details while eliminating redundancy.

6 Conclusion

Token compression has achieved significant progress in accelerating LVLM inference. When applying existing methods to HR-LVLMs with dynamic cropping, these methods treat global thumbnails and local crops uniformly, overlooking their inherent characteristics. Through analyzing HR-LVLMs with dynamic cropping, we reveal distinct roles between thumbnails and crops, and observe varying information densities across crops. Based on these findings, we propose GlobalCom^2, a plug-and-play token compression framework that operates on a "global-to-local" guided principle, adaptively preserving informative regions while minimizing redundancy. Experiments show that GlobalCom^2 achieves superior performance and efficiency across benchmarks, significantly outperforming existing baselines.

Acknowledgments

This work was supported in part by the Chengdu Science and Technology Program (No. 2025-YF12-00006-RC) , Police Integration Computing Key Laboratory of Sichuan Province (No. JWRH202502002), and the Open Fund of Key Laboratory of the Ministry of Education on Artificial Intelligence in Equipment (No. 2024-AAIE-KF04-03).

References

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*.

Bolya, D.; Fu, C.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2023. Token Merging: Your ViT But Faster. In *Proceedings of the International Conference on Learning Representations*.

Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2024. Honeybee: Locality-Enhanced Projector for Multimodal LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13817–13827.

Chen, G.; Li, Z.; Wang, S.; Jiang, J.; Liu, Y.; Lu, L.; Huang, D.-A.; Byeon, W.; Le, M.; Rintamaki, T.; et al. 2025. Eagle 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*.

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. In *Proceedings of the European Conference on Computer Vision*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. C. H. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Proceedings of the Advances in Neural Information Processing Systems*.

Dao, T.; Fu, D. Y.; Ermon, S.; Rudra, A.; and Ré, C. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Proceedings of the Advances in Neural Information Processing Systems*.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Han, Y.; Liu, X.; Zhang, Z.; Ding, P.; Wang, D.; Chen, H.; Yan, Q.; and Huang, S. 2024. Filter, correlate, compress: Training-free token reduction for mllm acceleration. *arXiv preprint arXiv:2411.17686*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 292–305.

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*.

Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In *Proceedings of the International Conference on Learning Representations*.

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*, 26286–26296.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Proceedings of the Advances in Neural Information Processing Systems*.

Liu, T.; Shi, L.; Hong, R.; Hu, Y.; Yin, Q.; and Zhang, L. 2024c. Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model. *arXiv preprint arXiv:2411.10803*.

Liu, X.; Gui, X.; Zhang, Y.; and Zhang, L. 2025a. Mixing Importance with Diversity: Joint Optimization for KV Cache Compression in Large Vision-Language Models. *arXiv preprint arXiv:2510.20707*.

Liu, X.; Wang, Y.; Ma, J.; and Zhang, L. 2025b. Video Compression Commander: Plug-and-Play Inference Acceleration for Video Large Language Models. *arXiv preprint arXiv:2505.14454*.

Liu, X.; Wen, Z.; Wang, S.; Chen, J.; Tao, Z.; Wang, Y.; Jin, X.; Zou, C.; Wang, Y.; Liao, C.; et al. 2025c. Shifting ai efficiency from model-centric to data-centric compression. *arXiv preprint arXiv:2505.19147*.

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. 2024d. MMBench: Is Your Multi-modal Model an All-Around Player? In *Proceedings of the European Conference on Computer Vision*, 216–233.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.; Zhu, S.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *Proceedings of the Advances in Neural Information Processing Systems*, 2507–2521.

Lu, X.; Chen, Y.; Chen, C.; Tan, H.; Chen, B.; Xie, Y.; Hu, R.; Tan, G.; Wu, R.; Hu, Y.; et al. 2025. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4145–4155.

Luo, G.; Zhou, Y.; Zhang, Y.; Zheng, X.; Sun, X.; and Ji, R. 2025. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. In *Proceedings of the International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; and Hsieh, C. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Proceedings of the Advances in Neural Information Processing Systems*, 13937–13949.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2025. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

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.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

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? In *Findings of the Association for Computational Linguistics: ACL 2025*, 15537–15549.

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*.

Wen, Z.; Wang, S.; Zhou, Y.; Zhang, J.; Zhang, Q.; Gao, Y.; Chen, Z.; Wang, B.; Li, W.; He, C.; et al. 2025c. Efficient multi-modal large language models via progressive consistency distillation. *arXiv preprint arXiv:2510.00515*.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2025. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. arXiv:2407.10671.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ye, W.; Wu, Q.; Lin, W.; and Zhou, Y. 2025. Fit and Prune: Fast and Training-free Visual Token Pruning for Multimodal Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11941–11952.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. *arXiv preprint arXiv:2412.01818*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; and Zhang, S. 2025. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *Proceedings of the International Conference on Machine Learning*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.