Don't Just Chase "Highlighted Tokens" in MLLMs: Revisiting Visual Holistic Context Retention

Xin Zou[1,2], **Di Lu**[1,†], **Yizhou Wang**[1], **Yibo Yan**[1,2], **Yuanhuiyi Lyu**[1,2], **Xu Zheng**[1,3], **Linfeng Zhang**[4], **Xuming Hu**[1,2*]

[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
[3] INSAIT, Sofia University "St. Kliment Ohridski"
[4] Shanghai Jiao Tong University

https://github.com/obananas/HoloV

Abstract

Despite their powerful capabilities, Multimodal Large Language Models (MLLMs) suffer from considerable computational overhead due to their reliance on massive visual tokens. Recent studies have explored token pruning to alleviate this problem, which typically uses text-vision cross-attention or [CLS] attention to assess and discard redundant visual tokens. In this work, we identify a critical limitation of such attention-first pruning approaches, *i.e.*, they tend to preserve semantically similar tokens, resulting in pronounced performance drops under high pruning ratios. To this end, we propose HoloV, a simple yet effective, plug-and-play visual token pruning framework for efficient inference. Distinct from previous attention-first schemes, HoloV rethinks token retention from a holistic perspective. By adaptively distributing the pruning budget across different spatial crops, HoloV ensures that the retained tokens capture the global visual context rather than isolated salient features. This strategy minimizes representational collapse and maintains task-relevant information even under aggressive pruning. Experimental results demonstrate that our HoloV achieves superior performance across various tasks, MLLM architectures, and pruning ratios compared to SOTA methods. For instance, LLaVA1.5 equipped with HoloV preserves 95.8% of the original performance after pruning 88.9% of visual tokens, achieving superior efficiency-accuracy trade-offs.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated outstanding capabilities [82, 12] in tasks such as image captioning [35, 61, 14], visual question answering [24, 99, 36], and video understanding [34, 64, 79]. However, these models [43, 78, 38] typically require converting visual inputs into long sequence representations (*i.e.*, visual tokens), which increases the computational complexity and cost of inference [97], especially for high-resolution images [41] and multi-frame videos [57], where redundant visual information further exacerbates the computational overhead.

To address this challenge, researchers have introduced token pruning strategies [49, 13, 98, 87] that aim to retain the highlighted visual tokens as well as prune others for accelerating MLLM's inference. These methods typically define importance criteria for tokens, such as attention scores [13, 19] or gradient information [59, 58], to quantify the significance of visual tokens, and less important tokens are pruned during the inference phase, which balances speed and performance, but with limitations.

*Corresponding author, †Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

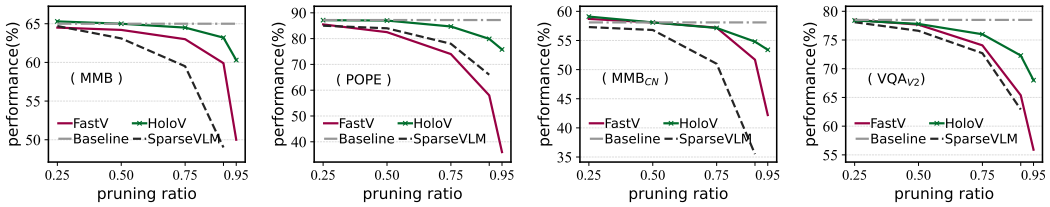

Figure 2: Relationship between performance and pruning ratios of different baseline methods. As the token pruning ratio grows, the performance of these attention-first strategies degrades dramatically, while HoloV maintains the substantial performance even at 90% and 95% of the pruning ratios.

As shown in Fig. 1, FastV [13] is an intuitive solution that ranks visual tokens based on attention distributions across different layers, and then prunes the bottom $R\%$ of tokens based on the computational budget, thus reducing visual token redundancy. Subsequently, more work has followed this paradigm [91, 98, 4], designing different strategies to prune redundant visual tokens via cross-modal (*i.e.*, text-vision) attention from LLMs. Besides, there are vision-centric pruning methods [77, 25, 94, 66, 88] (*e.g.*, FasterVLM [93]) that presume those visual tokens with low correlation to the [CLS] token in ViT [17], or those exhibit duplicated features tokens [20] to be redundant.

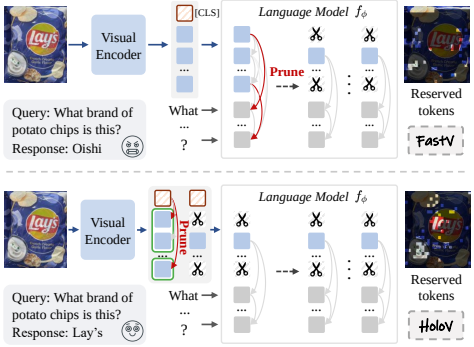

Figure 1: Snapshots of FastV and our HoloV.

Although these pruning methods can recognize the inefficiency of visual tokens in MLLMs, they are not consistently effective. As shown in Fig. 2, the performance decreases significantly as the pruning ratio increases. In our argument, this occurs because these approaches implicitly assume that *visual tokens with high attention correspond to higher informativeness*, which disregards the spatial-semantic relations of the visual scene, *i.e.*, they tend to retain tokens from localized salient regions where attention is drawn to, rather than those conducive to holistic semantic comprehension. Thus, at a high pruning ratio, such methods would only retain homologous tokens with higher scores. In a complex scene with multiple objects, retaining only "highlighted tokens" may sever relative positional and semantic connectivity information or lose key tokens associated with the subject, leading to a dramatic performance degradation. Besides, the attention mechanism introduces systematic biases [80, 81], *i.e.*, the position encoding mechanism of transformer-based MLLMs may introduce spatial priors, those in upper and lower areas visual tokens usually being assigned higher attention weights as shown in Fig. 3 right. This bias can distort the semantic contributions of the visual scene, leading the model to produce incorrect or logically contradictory inferences, or even hallucinations [100, 103]. Drawing inspiration from the above discussion, we raise the following question: *"How to locate and preserve those not highlighted but critical to visual holistic understanding tokens?"*

Cognitive science research suggests that the human visual system forms a complete semantic understanding by integrating local features with global scene cues [70, 2, 63] (*e.g.*, background textures and spatial layouts). In MLLMs, we analyzed the text-mapping relationships of different visual tokens through the strategy in [60]. As shown in Fig. 3 left, the objects in a scene could be represented by a

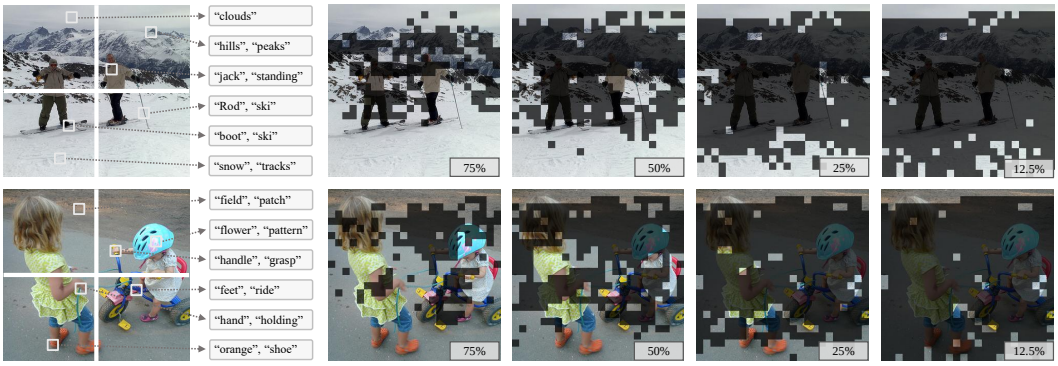

Figure 3: LEFT - Examples of textual semantics corresponding to visual tokens from scattered crops. RIGHT - Sparsification visualization examples of FastV, where retention ratios are tagged in the pics.

small number of scattered tokens, and the semantic relationships between those tokens from different regions facilitate the overall understanding, *e.g.*, *"snow", "ski", "hills"* are kind of self-explanatory. Motivated by this insight, we propose HoloV, which explicitly balances overall semantic connectivity and contextual attention during visual token pruning, addressing the critical limitation of redundancy in attention-first strategies. Our analysis demonstrates the importance of preserving visual holistic context, offering a new perspective on efficient visual token pruning in MLLMs. Through extensive experiments on diverse benchmarks and MLLM architectures, we demonstrate that HoloV consistently surpasses existing state-of-the-art token pruning approaches, achieving up to 88.9% token reduction while preserving about 96% of the original performance. Besides, HoloV is model-agnostic and easily integrable into a wide range of MLLMs, making it well-suited for practical deployment.

2 Related Work

2.1 MLLMs and Their Challenges

The recent remarkable success of Large Language Models (LLMs) [62, 95, 72, 18, 56] has spurred the trend of applying their strong capabilities to multimodal comprehension tasks, fostering the development of MLLMs [1, 69]. Leveraging open-source LLMs such as LLaMA families [72, 73, 18], MLLMs [6, 46, 47] have demonstrated enhanced adaptability across a range of visual understanding tasks, leading to a more profound ability to interpret the world. While this empowers LLMs with the capability of visual perception, the incorporation of lengthy visual tokens significantly escalates the computational burdens. Moreover, studies have shown that existing MLLMs still suffer from certain visual deficiencies [71, 32] and some hallucinations [29, 28]. Some work mitigates these issues by increasing the resolution of input images or videos [55, 86], but this further exacerbates the computational overhead. For example, LLaVA-1.5 [48] encodes a 336-resolution image into 576 visual tokens, while LLaVA-NeXT [47] doubles the resolution and generates 2,880 tokens. LLaVA-OneVision [37] represents an image using 7,290 visual tokens, and Video-LLaVA [44] faces even higher costs, as it must process numerous visual tokens from multiple frames during inference. These visual tokens occupy a large portion of the context window of their LLMs. In this work, we conducted experiments and analysis on these representative models to verify HoloV's applicability.

2.2 Visual Redundancy Identification

In MLLMs, visual redundancy identification facilitates the distillation of visual tokens with high informativeness for faster inference. There are two main research directions: a) Vision-centric strategies analyze the image's structure and feature distribution to discard less relevant visual tokens [13, 77]. Existing approaches include spatial-similarity clustering (*e.g.*, TokenLearner [65]), dynamic pruning based on attention scores [25, 89, 84], and using information bottleneck or entropy metrics during the prefilling stage to estimate background redundancy. b) Instruction-centric strategies typically use cross-modal attention analysis or gradient accumulation to identify redundant tokens [49, 101, 68]. Tokens with low attention or negligible gradient impact are deemed redundant [26]. Building on this, some studies explore learned importance scoring, training a lightweight end-to-end model to predict each patch's "instruction relevance," enabling even finer-grained pruning [31, 75, 91]. As the existence of language bias in LLM may cause hallucinations, we use a vision-centric scheme.

2.3 Visual Token Compression and Pruning

The inclusion of visual information in MLLMs introduces long token sequences, leading to high computation and memory costs. For example, mini-Gemini-HD [41] generates 2880 tokens from high-definition images, creating inference bottlenecks. To address this, research has focused on token compression and pruning techniques in Vision Transformers [10] and MLLMs [27]. Methods like LLaMA-VID [40] and DeCo [90] address this by modifying models and adding training, which increases computational costs. ToMe [11] reduces tokens without training but disrupts early cross-modal interactions [83]. LLaVA-PruMerge [66] selectively retains key tokens while merging less critical ones based on key similarity. FasterVLM [93] utilizes `[CLS]` attention scores from the visual encoder to re-rank and retain top visual tokens. FastV [13] and SparseVLM [98] focus on token selection using attention scores or cross-modal guidance, but overlook the role of token duplication and lack Flash-Attention [16, 15]. Our proposed HoloV maintains hard acceleration compatibility (*e.g.*, Flash-Attention), and effectively retains visual holistic context during aggressive pruning.

3 Preliminary and Motivation

3.1 Preliminary

Architecture of MLLMs. Given an MLLM $\mathcal{M}_\theta^{\text{MLLM}}$ parameterized by θ, with a general architecture consisting of a text embedding layer, a vision encoder, a vision-text interface module, a text decoder consisting of L number of transformer layers, and an affine layer which predicts the distribution of the next token. For an image-grounded text generation task, given a textual query x and an input image v, $\mathcal{M}_\theta^{\text{MLLM}}$ first extracts vision features of v by the vision encoder, and then converts them into visual tokens z_v by MLP or Q-Former [76] modules. Aligned vision tokens z_v are concatenated with the query x as input to the text decoder, and finally decoded into a textual response y autoregressive, which is formulated as: $y_t \sim p_\theta(\cdot|v, x, y_{<t}) \propto \textit{softmax}(f_\theta(\cdot|v, x, y_{<t}))$, where y_t indicates the t^{th} token, $y_{<t}$ is the token sequence generated up to the time step t, and f_θ is the logit distribution.

Attention mechanism. Considering the computational burden associated with the length of visual tokens in MLLMs, many studies have followed the paradigm of using attention scores to evaluate the redundancy of visual tokens. Specifically, transformer-based MLLMs typically utilize causal self-attention [5] to perform computation as: $\text{Self-attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\mathbf{Q} \cdot \mathbf{K}^\top / \sqrt{d_k}\right) \cdot \mathbf{V}$, where d_k is the dimension of $\mathbf{K}$, the result of $\text{softmax}\left(\mathbf{Q} \cdot \mathbf{K}^\top / \sqrt{d_k}\right)$ is known as the attention matrix. In this work, we focus on the attention received by visual tokens from the visual [CLS] token.

3.2 Information Redundancy in Highlighted Tokens

When token selection is based exclusively on attention scores, the model tends to retain similar clusters, resulting in information redundancy. As shown in Fig. 4 left, adjacent tokens with similar visual features frequently receive comparable attention scores, especially in regions characterized by flat backgrounds or repetitive textures. Their spatial proximity leads these tokens to capture overlapping features, making it hard to distinguish those not highlighted yet informative tokens.

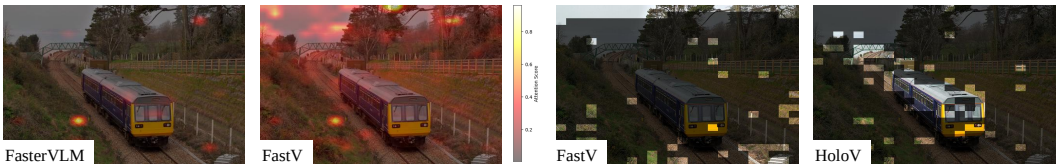

Figure 4: LEFT - Distribution map of visual token attention. RIGHT - Visualization cases of FastV and HoloV. HoloV retains contextual tokens with rich semantics, while FastV contains much redundancy.

Positional Bias. To further investigate attention-based token pruning methods, we take FastV as an example and visualize the distribution of the retained visual tokens. As illustrated in Fig. 4 right, the attention scores for image tokens present a consistent pattern: tokens located at the beginning and end of the sequence tend to have higher attention and are thus more likely to be preserved during pruning, leading to a positional bias. We extend our analysis by conducting statistics on samples from the text-based VQA task using the VQA V2 [23] dataset. Notably, even though these samples originate from a different task, the attention distributions of image tokens at the same layer remain highly similar, revealing recurring patterns. While the overall shape of the distributions varies slightly across layers, the set of tokens receiving relatively high attention remains stable. We suggest that this phenomenon occurs because all visual tokens are processed with text tokens in the same manner during decoding, leading to positional bias of text shift to the visual modality, *e.g.*, boundary positions of text usually imply important information, but for images, targets are mostly located in the center.

Attention Dispersion. In addition to positional bias, we further analyze the phenomenon of attention dispersion, i.e., a small subset of similar tokens receives the majority of attention, while most tokens are assigned low attention scores [93]. Specifically, we compute the cumulative distribution of visual tokens sorted by their attention scores, as shown in Fig. 5. The curves of last-token attention [13] and equi last attn with identical position embedding are noticeably less steep than that for [CLS] attention. It is evident that compared to [CLS] attention, text-vision attention tends to be dispersed over more visual tokens, *e.g.*, the top 20% of visual tokens account for only 40% of the total attention.

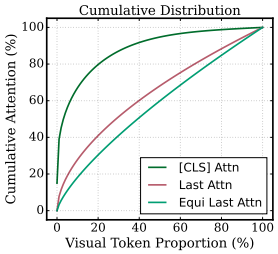

Figure 5: Cumulative distribution of different attentions.

3.3 Holistic Context Trumps Local Duplicates

Based on our previous analysis, attention-first token pruning methods suffer from over-localization due to positional bias and attention dispersion, *i.e.*, over-reliance on attention scores disrupts spatial-semantic relationships, *e.g.*, breaking occlusion hierarchies in multi-object interactions. Thus, our key insight is that visual token importance should be evaluated through global contextual cohesion, *i.e.*, jointly considers holistic context and local saliency rather than isolated attention magnitudes.

To further validate our hypothesis, we devised a straightforward holistic context retention strategy, *i.e.*, pruning visual tokens through random masks to retain visual information from different regions. As shown in Fig. 6 up, compared with FastV, this random strategy outperforms on more than half of the benchmarks, which demonstrates the significance of preserving holistic context for visual understanding. On the VQA text dataset, however, the random strategy failed, possibly because random pruning discards some salient fine-grained information. This result also suggests that local saliency is indispensable, especially for densely packed elements within small regions.

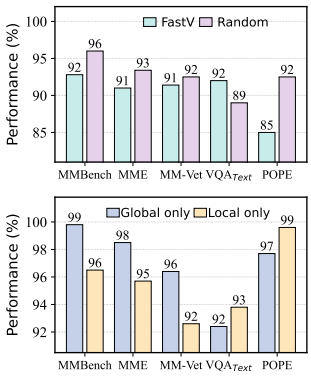

Figure 6: UP - FastV v.s. Random strategy. DOWN - Performance comparison of the thumbnail and local crops as inputs.

In addition, we conducted an exploratory experiment to investigate how holistic context contributes to visual understanding in MLLMs. Specifically, we use the global thumbnail and multiple local crops as visual input separately [47], and evaluate performance on the two settings against various benchmarks. As shown in Fig. 6 down, with only the global thumbnail yields strong results on general visual perception benchmarks such as MMBench [53], MME [21], and MM-Vet [92], highlighting the inherent role of holistic context in guiding general visual understanding. On the contrary, using only local crops leads to poor performance in these general perception tasks but excels in fine-grained perception benchmarks such as TextVQA [67] and POPE [42], which suggests that local duplicated saliency can offer fine-grained visual information for semantic understanding.

4 Methodology

Building on the above analysis, we propose HoloV, which better preserves the holistic context of images for visual understanding. By removing redundant visual tokens before the LLM decoder, our approach could make MLLMs inference faster than methods that prune tokens within the LLM. An overview of our approach is depicted in Fig. 7. In what follows, we elaborate on how our HoloV guides overall visual token compression under a high pruning ratio to keep semantic completeness.

4.1 HoloV Framework

To address the pivotal question raised in Sec. 1 for effective and efficient visual token pruning, we propose HoloV framework, which leverages crop-wise adaptive allocation to decentralize attention over those non-highlighted but heterogeneous tokens. Fig. 7 illustrates the core idea of HoloV.

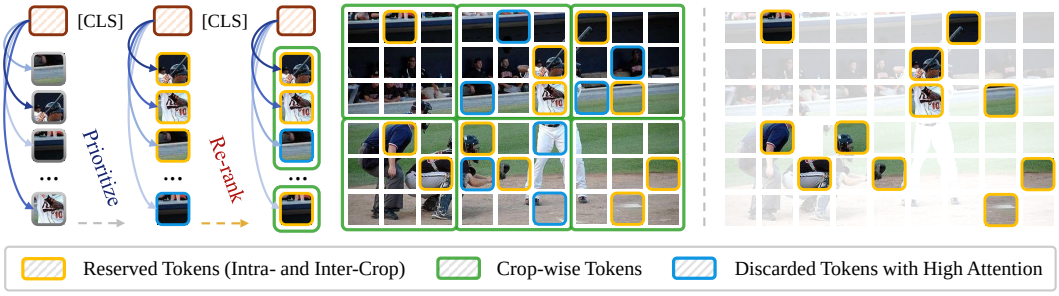

Figure 7: Illustration of HoloV. We re-rank highlighted visual tokens for holistic context retention.

Based on our findings about the positional bias, We first rearrange visual tokens into local crops. Let the total number of image tokens be N_v, which is evenly partitioned into $\mathcal{C}$ crops. This enables the model to maintain spatial granularity and gather statistics both locally and globally. Given the

normalized embeddings $\mathbf{Z}_v^c \in \mathbb{R}^{M\times d}$ in c-th crop, we first compute intra-crop similarity matrix $\mathbf{S}^c$ as

$$\mathbf{S}^c = (\mathbf{1} - \mathbf{I}_M) \odot \mathbf{Z}_v^c \mathbf{Z}_v^{c\top}, \tag{1}$$

where $\odot$ denotes Hadamard product, and $\mathbf{I}_M$ is the identity matrix masking self-similarities. Then, we capture intra-crop diversity by the variance of semantic distribution, the formula is as follows

$$\mathcal{V}_i^c = \frac{1}{M-1} \sum \left(\mathbf{S}_{i,j}^c - \mu_i^c\right)^2, \tag{2}$$

where a high value of $\mathcal{V}_i^c$ indicates that i-th token has diverse connections with others, the visual semantics expressed by the informative token is essential within the crop. To obtain holistic attention, we establish a balanced scoring mechanism combining contextual diversity and attention saliency. Specifically, we merge variance $\mathcal{V}^c$ and [CLS] attention $\mathcal{A}^c$ in the crop using adaptive scaling:

$$\mathcal{H}^c = \gamma_c \mathcal{V}^c + \mathcal{A}^c, \text{ where } \gamma_c = \mathbb{E}[\|\mathcal{A}^c\|]/\mathbb{E}[\|\mathcal{V}^c\|]. \tag{3}$$

Adaptive holistic token allocation. To preserve overall scene semantics and spatial diversity, we compute a crop-level priority score by averaging token scores within each crop. The total quota for selected image tokens T' is dynamically allocated to crops according to their normalized crop-level importance. The allocation to each crop is discrete and capped, ensuring spatial coverage while preventing over-concentration on specific regions. We resolve rounding and overflow through an iterative reallocation procedure, so that crops with excess quota donate surplus tokens to those with remaining capacity, according to their crop-level scores.

We compute crop importance weights via

$$w_c = (\frac{1}{M}\sum_{t=1}^{M} \mathcal{H}_t^c)^\tau / \sum_{c'=1}^{\mathcal{C}} (\frac{1}{M}\sum_{t=1}^{M} \mathcal{H}_t^{c'})^\tau, \tag{4}$$

where τ controls the sharpness of allocation. Thus, initial quota $q_c = \lfloor w_c \hat{N}_v \rfloor$, where $\hat{N}_v$ denotes the number of retained tokens. When the allocated tokens overflow or fall short, we redistribute residual tokens. For overflow, the quota is changed by $q_c = \min(q_c + \Delta_c, M), \Delta_c \propto w_c \cdot (M - q_c)$, while for fall short, we allocate the remaining quota to the crop with the highest weight. In this way, HoloV adaptively adjusts its compression degree according to the informativeness of different crops.

Top-k visual token selection. Within each crop, select visual tokens by maximizing:

$$\text{argmax}_{\Omega_c \subset \{1,\dots,M\}} \sum \mathcal{H}^c, \text{ subject to } |\Omega_c| = q_c\,, \tag{5}$$

which ensures both crop-wise local saliency and global relevance. We retain top-k visual tokens in each crop, where k is determined by the quota q_c in the allocation. By performing token pruning before the LLM decoder, we dynamically adjust the number of visual tokens as input to the language model based on the actual computational budget, thus accelerating the MLLM inference.

4.1.1 Fast Visual Context Refetching

Motivated by the attention sinks [96], and information loss during visual token pruning, we further propose visual context refetching to fast supplement the visual holistic context. Specifically, we treat pruned tokens as supplementary evidence, re-injecting them into the MLLM through Feed Forward Network (FFN) as "key-value memory" at the middle trigger layer. This *refetch* mechanism occurs when the model exhibits high uncertainty during inference, achieving effective and efficient visual information replenishment. Limited by space, the details can be found in Appendix D.

4.2 Theoretical Analysis

To further justify the trustworthiness of our proposed HoloV, we provide a theoretical analysis of it. Under Assumption 1, for any pruned token, there exists a retained token that is sufficiently close in the embedding space, with bounded context variance. By leveraging the *Lipschitz continuity* [8] of the transformer layer, we can bound the semantic difference between the outputs on the original and pruned token sets. The residual error introduced by the scoring threshold is also controlled. Combining these components, we obtain the stated upper bound. More details are in Appendix C.

Table 1: Performance comparison of various methods across different benchmarks. Results are shown for different pruning ratios, with accuracy and average performance highlighted. Best results in **blue**.

Methods	GQA	MMB	MMB_{CN}	MME	POPE	SQA	VQA_{V2}	VQA_{Text}	VizWiz	Average
Upper Bound, 576 Tokens	61.9	64.7	58.1	1862	85.9	69.5	78.4	58.2	50.0	100%
LLaVA-1.5 7B	*Retain 192 Tokens* (↓ **66.7%**)									
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	88.5%
FastV (ECCV24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	90.5%
MustDrop (2024.11)	58.2	62.3	55.8	1787	82.6	69.2	76.0	56.5	51.4	97.2%
LLaVA-PruMerge (ICCV25)	54.3	59.6	52.9	1632	71.3	67.9	70.6	54.3	50.1	91.4%
PDrop (CVPR25)	57.1	63.2	56.8	1766	82.3	68.8	75.1	56.1	51.1	96.7%
FiCoCo-V (2025.03)	58.5	62.3	55.3	1732	82.5	67.8	74.4	55.7	51.0	96.1%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	74.9	47.4	50.1	94.6%
VisionZip (CVPR25)	**59.3**	64.5	57.3	1767	86.4	68.9	**76.8**	57.3	**51.6**	98.1%
SparseVLM (ICML25)	57.6	62.5	53.7	1721	83.6	69.1	75.6	56.1	50.5	96.1%
DART (EMNLP25)	58.9	63.6	57.0	**1856**	82.8	69.8	76.7	57.4	51.1	98.5%
HoloV (Ours)	59.0	**65.4**	**58.0**	1820	**85.6**	**69.8**	76.7	**57.4**	50.9	**99.2%**
LLaVA-1.5 7B	*Retain 128 Tokens* (↓ **77.8%**)									
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	80.4%
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	51.3	85.4%
MustDrop (2024.11)	56.9	61.1	55.2	1745	78.7	68.5	74.6	56.3	**52.1**	95.7%
LLaVA-PruMerge (ICCV25)	53.3	58.1	51.7	1554	67.2	67.1	68.8	54.3	50.3	89.4%
PDrop (CVPR25)	56.0	61.1	56.6	1644	82.3	68.3	72.9	55.1	51.0	94.9%
FiCoCo-V (2025.03)	57.6	61.1	54.3	1711	82.2	68.3	73.1	55.6	49.4	94.9%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	73.4	46.1	51.3	93.1%
VisionZip (CVPR25)	57.6	63.4	56.7	1768	84.7	68.8	75.6	56.8	52.0	97.2%
SparseVLM (ICML25)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	93.8%
DART (EMNLP25)	**57.9**	63.2	**57.0**	**1845**	80.1	69.1	**75.9**	56.4	51.7	97.5%
HoloV (Ours)	57.7	**63.9**	56.5	1802	**84.0**	**69.8**	75.5	**56.8**	51.5	**98.0%**
LLaVA-1.5 7B	*Retain 64 Tokens* (↓ **88.9%**)									
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	70.1%
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	76.7%
MustDrop (2024.11)	53.1	60.0	53.1	1612	68.0	63.4	69.3	54.2	51.2	90.1%
LLaVA-PruMerge (ICCV25)	51.9	55.3	49.1	1549	65.3	68.1	67.4	54.0	50.1	87.7%
PDrop (CVPR25)	41.9	33.3	50.5	1092	55.9	68.6	69.2	45.9	50.7	77.5%
FiCoCo-V (2025.03)	52.4	60.3	53.0	1591	76.0	68.1	71.3	53.6	49.8	91.5%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	69.7	44.2	50.2	89.4%
VisionZip (CVPR25)	55.1	60.1	**55.4**	1690	77.0	69.0	72.4	**55.5**	**52.9**	94.5%
SparseVLM (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	87.3%
DART (EMNLP25)	**55.9**	60.6	53.2	**1765**	73.9	**69.8**	72.4	54.4	51.6	93.9%
HoloV (Ours)	55.3	**63.3**	55.1	1715	**80.3**	69.5	**72.8**	55.4	52.8	**95.8%**

4.3 Computational Complexity

As language instructions are much shorter than visual tokens, we focus on the FLOPs contributed by visual tokens. Let n denote the number of visual tokens, d the hidden size, and m the FFN intermediate size (with SwiGLU). For the prefill stage, the FLOPs per transformer layer can be approximated as $an^2d + bnd^2 + cndm$, where a, b, and c are constants. If the token count is reduced by a ratio R ($\hat{n} = (1 - R)n$), the FLOPs reduction ratio is:

$$F = 1 - \frac{a\hat{n}^2 d + b\hat{n}d^2 + c\hat{n}dm}{an^2 d + bnd^2 + cndm}. \tag{6}$$

For large n, the quadratic term dominates, so $F \approx 1 - (1 - R)^2 = 2R - R^2$. Thus, the reduction is slightly better than linear in R. In the decode stage (with KV cache), the complexity becomes linear in n, and the FLOPs per layer are $bd^2 + (bd + cdm)n$, so the reduction is nearly proportional to R. HoloV speeds up inference by pruning ahead of the LLM to avoid KV cache inefficiency.

5 Experiments

5.1 Experimental Setup

Benchmarks. We conducted experiments on several widely used visual understanding benchmarks. For image understanding task, we performed experiments on ten widely used benchmarks, including GQA [30], MMBench (MMB) and MMB-CN [53], MME [21], POPE [42], VizWiz [9], SQA (ScienceQA) [54], VQA_{V2} (VQA V2) [23], VQA_{Text} (TextVQA) [67], and MM-Vet [92]. Video

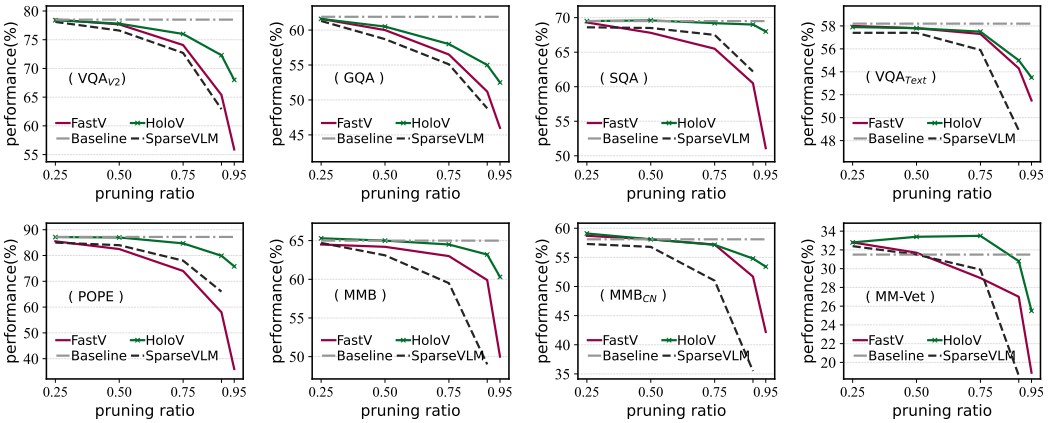

Figure 8: Comparison of different methods across multiple benchmarks under varying pruning ratios.

QA benchmarks include MSVD-QA and MSRVTT-QA [85]. All experiments on these benchmarks follow the default settings. More details of the benchmarks are provided in Appendix A.1.

Comparison methods. We compare our approach with several representative methods for accelerating multi-modal language models (MLLMs) via token reduction, including ToMe [11], FastV [13], SparseVLM [98], HiRED [4], LLaVA-PruMerge [66], PDrop [83], MustDrop [49], FasterVLM [93], GlobalCom^2[52], VisionZip [88], DART [81]. These baselines employ diverse strategies such as token merging, attention-based pruning, adaptive allocation, and hierarchical retention to improve efficiency by reducing redundant tokens. Each method offers a unique perspective on balancing computational cost and model performance. More details of baselines are provided in Appendix A.2.

5.2 Main Results

General-purpose benchmarks. We evaluate the performance of HoloV on general-purpose datasets, *i.e.*, GQA, MM-Vet, MME, MMBench, SQA, and VizWiz. As shown in Tab. 1, HoloV consistently outperforms competing approaches at different pruning ratios, *e.g.*, HoloV removes up to 88.9% of visual tokens with only a 4.2% performance drop, and 77.8% with just 2% on average. Further, we show more results under varying pruning ratios, as shown in Fig. 8, the performance of FastV and SparseVLM drops dramatically under high pruning ratios, while HoloV maintains robust performance with relatively minor losses at all pruning ratios on SQA and MMBench. On MMBench$_{CN}$ and MM-Vet, HoloV even achieves higher than baseline (unpruned) scores at pruning ratios of 25%, 50%, and 75% (MM-Vet), then the score slowly drops as the pruning ratio increases. For VizWiz evaluation, the result in Fig. 9 indicates that HoloV can consistently obtain performance improvements at different pruning ratios, even at 95%, which means HoloV effectively retains visual holistic semantics.

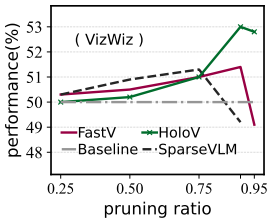

Figure 9: Performance of different methods on VizWiz under varying pruning ratios.

Hallucination benchmarks validation. We conduct the hallucination evaluations on POPE and MME benchmarks, with results on LLaVA-1.5-7B presented in Tab. 1, where the proposed HoloV shows robust capabilities, and the performance significantly exceeds the results of the compared SOTA methods, *e.g.*, with a pruning rate of 88.9%, HoloV achieves 80.3% accuracy compared to 76% for the second runner-up on POPE, and achieved desirable performance on MME evaluation, compared to other comparative approaches.

5.3 HoloV with Higher Resolution

For further comprehensive evaluation, we also evaluated HoloV for LLaVA-NeXT on different benchmarks mentioned above, with comparison to current SOTA approaches. LLaVA-NeXT introduces a new image processing method, leading to dynamic lengths of visual embeddings for various image inputs. Thus, during the evaluation, 320 visual tokens has been kept (from up to 2880 raw tokens). As shown in Table 3, the evaluation results of all various benchmarks show that HoloV obtained the highest score on almost every track, and has an average of 95. 6%, much higher than the current SOTA of 93.3%.

Table 3: Performance comparison of various methods across different benchmarks. Results are shown for different pruning ratios, with accuracy and average performance highlighted. Best results in **blue**.

Methods	GQA	MMB	MMB_{CN}	MME	POPE	SQA	VQA_{V2}	VQA_{Text}	VizWiz	Average
Upper Bound, 2880 Tokens	64.2	67.4	60.6	1851	86.5	70.1	81.8	64.9	57.6	100%
LLaVA-NeXT 7B	*Retain 320 Tokens* (↓ **88.9%**)									
FastV (ECCV24)	55.9	61.6	51.9	1661	71.7	62.8	71.9	55.7	53.1	88.0%
LLaVA-PruMerge (ICCV25)	53.6	61.3	55.3	1534	60.8	66.4	69.7	50.6	54.0	85.6%
PDrop (CVPR25)	56.4	63.4	56.2	1663	77.6	67.5	73.5	54.4	54.1	90.9%
MustDrop (2024.11)	57.3	62.8	55.1	1641	82.1	68.0	73.7	**59.9**	54.0	92.2%
FasterVLM (ICCV25)	56.9	61.6	53.5	1701	83.6	66.5	74.0	56.5	52.6	91.1%
HiRED (AAAI25)	59.3	64.2	55.9	1690	83.3	66.7	75.7	58.8	54.2	93.3%
SparseVLM (ICML25)	56.1	60.6	54.5	1533	82.4	66.1	71.5	58.4	52.0	89.7%
GlobalCom^2 (2025.3)	57.1	61.8	53.4	1698	83.8	67.4	76.7	57.2	54.6	92.2%
DART (EMNLP25)	61.7	65.3	**58.2**	1710	**84.1**	68.4	79.1	58.7	**56.1**	93.9%
HoloV (Ours)	**61.7**	**65.3**	57.5	**1738**	83.9	**68.9**	**79.5**	58.7	55.3	**95.6%**

Table 4: Real inference comparison on POPE. Experiments adopt 66.7% and 90% pruning ratios.

Methods	Time	Prefill	Latency	Mem.	Acc.	Time	Prefill	Latency	Mem.	Acc.
Upper Bound, 576 Tokens	49:41	0.5ms	0.334s	19.0G	100.%	49:41	0.5ms	0.334s	19.0G	100.%
LLaVA-1.5-7B	*Retain 192 Tokens* (↓ **66.7%**)					*Retain 58 Tokens* (↓ **90%**)				
FastV (ECCV24)	35:34	0.5ms	0.239s	16.0G	75.4%	30:41	0.5ms	0.206s	15.6G	66.8%
MustDrop (2024.11)	32:30	0.5ms	0.273s	15.6G	96.2%	29:40	0.6ms	0.199s	14.5G	87.1%
FasterVLM (ICCV25)	**30:09**	0.5ms	**0.202s**	15.6G	100.%	25:08	0.5ms	**0.168s**	14.5G	92.5%
HiRED (AAAI25)	30:08	0.6ms	0.210s	15.7G	96.4%	**25:03**	0.6ms	0.168s	14.5G	92.7%
SparseVLM (ICML25)	40:51	0.6ms	0.251s	15.8G	97.3%	31:28	0.6ms	0.212s	14.6G	92.3%
HoloV (Ours)	31:02	0.5ms	0.208s	**15.6G**	**99.7%**	27:36	0.5ms	0.176s	**14.5G**	**95.7%**

Besides, on video understanding benchmarks, HoloV maintains close to the original performance, significantly outperforming FasterVLM and FastV, as shown in Table 2. This demonstrates the value of HoloV when it comes to high-resolution visual input.

Table 2: Video QA Evaluations of different methods with 50% of visual tokens retained. HoloV beats SOTA.

Methods	MSVD-QA		MSRVT-QA		Avgerge	
	Acc.	Score	Acc.	Score	Acc.	Score
Video-ChatGPT 7B	64.9	3.3	49.3	2.8	57.1	3.1
Video-LLaVA 7B	70.2	3.9	57.3	3.5	63.8	3.7
FastV (ECCV24)	71.0	3.9	55.0	3.5	63.0	3.7
FasterVLM (ICCV25)	70.5	3.9	56.2	3.5	63.4	3.7
DART (EMNLP25)	71.0	4.0	**56.7**	3.6	58.0	3.7
HoloV (Ours)	**71.0**	**4.0**	56.5	**3.6**	**63.7**	**3.7**

5.4 Efficiency Analysis

To assess the efficiency of HoloV, we compare total inference time, prefill time, end-to-end latency, GPU memory usage, and accuracy on LLaVA-1.5-7B. As shown in Tab. 4, under a 90% pruning ratio, HoloV achieves a 42.7% reduction in inference time and a 42.8% decrease in latency, with only a 4.3% drop in accuracy, similarly under 66.7% pruning ratio. Compared to FastV and SparseVLM, HoloV uses less memory and runs faster. Although FasterVLM offers slightly quicker inference, HoloV improves accuracy by 3.0%, demonstrating a better balance between efficiency and performance.

5.5 Ablation Analysis of Crop Numbers

Partition granularity does not affect pruning efficiency: retained visual tokens are determined by pruning quotas, and the quota per crop, *i.e.*, calculated dynamically via intra-crop visual token informativeness, leaves total pruning quotas unchanged. For high-resolution images, dynamic crop number adjustment is beneficial: using fewer crops for high-detail areas and more for low-detail regions. Specifically, Table 5 shows results when total crops vary from 4 to 16, where the values represent percentages relative to original performance. We observe no significant performance impact from varying crop numbers.

Table 5: Ablation of different crop numbers.

Methods	#=4	#=8	#=12	#=16
Upper Bound	100%	100%	100%	100%
LLaVA-1.5-7B	*Token Pruning Rate = 66.7%*			
HoloV (Ours)	95.1%	**96.7%**	96.1%	94.9%
LLaVA-1.5-7B	*Token Pruning Rate = 77.8%*			
HoloV (Ours)	94.5%	**95.1%**	94.6%	94.8%
LLaVA-1.5-7B	*Token Pruning Rate = 88.9%*			
HoloV (Ours)	89.3%	89.3%	90.0%	**91.2%**

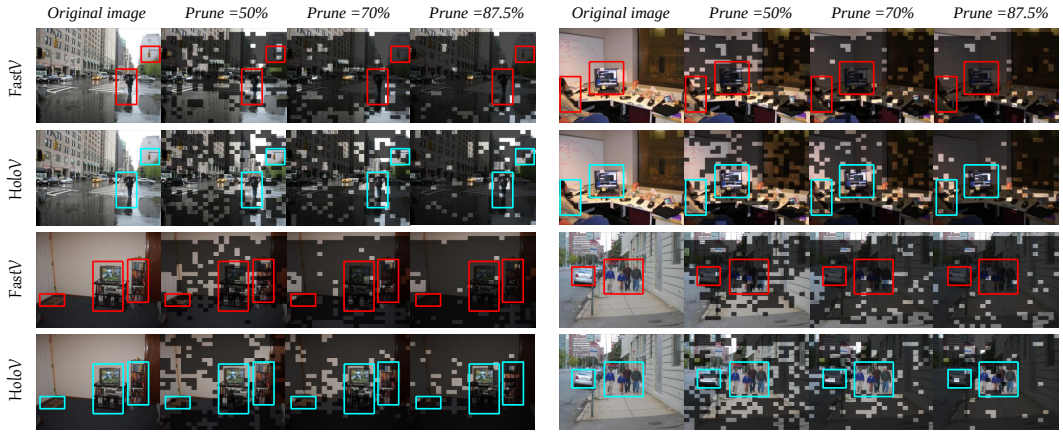

Figure 10: The case comparison between FastV and HoloV from the GQA. It presents original images alongside their pruned versions at pruning rates of 50%, 70%, and 87.5%. The bounding boxes highlight specific regions and objects across images, where HoloV well preserves the pivotal tokens.

5.6 Visualization Analysis

Further, we visualize retained visual patches under different pruning rates. As shown in Fig. 10, black areas indicate discarded tokens, while colored regions show key semantic areas aligned with text. Compared to FastV, HoloV preserves more relevant visual cues even under high pruning (e.g., 87.5%), effectively filtering out redundant visual tokens while keeping critical objects. This supports better cross-modal alignment, allowing pivotal holistic tokens for visual overall understanding.

5.7 HoloV with Qwen Architecture

To verify the architectural generalization of HoloV beyond LLaVA-based models, we conduct experiments on the Qwen2.5-VL-7B [7] architecture. As shown in Tab. 6, HoloV demonstrates strong generalization capability across this architecture, consistently outperforming the text-visual attention-based FastV at various reduction ratios, highlighting its robustness and adaptability to different model designs. Notably, it achieves average performance retention rates of 94.6%, 92.7%, and 90.5% at 66.7%, 77.8%, and 88.9% token pruning rates respectively, significantly higher than FastV's 92.3%, 89.2%, and 84.3% performance. These results show that our proposed holistic pruning strategy effectively generalizes across different MLLM architectures.

Table 6: Comparative Experiments on Qwen2.5-VL-7B.

Methods	MMB	MME	POPE	SQA	VQA_{Text}	Avg.
Upper Bound	82.8	2304	86.1	84.7	84.8	100%
Qwen2.5-VL-7B	*Token Pruning Rate = 66.7%*					
FastV (ECCV24)	75.7	2072	82.2	78.5	77.9	92.3%
HoloV (Ours)	**78.3**	**2093**	**85.0**	**79.8**	**78.9**	**94.6%**
Qwen2.5-VL-7B	*Token Pruning Rate = 77.8%*					
FastV (ECCV24)	74.9	2036	80.7	78.0	69.0	89.2%
HoloV (Ours)	**76.5**	**2043**	**82.3**	**79.8**	**70.3**	**92.7%**
Qwen2.5-VL-7B	*Token Pruning Rate = 88.9%*					
FastV (ECCV24)	69.2	1940	78.6	77.4	60.3	84.3%
HoloV (Ours)	**72.4**	**2006**	**80.7**	**79.5**	**61.8**	**90.5%**

6 Conclusion

We present HoloV, a holistic token pruning framework that addresses two critical limitations of attention-based visual compression: 1) semantic fragmentation from over-pruning non-salient regions, and 2) static importance estimation ignoring token interdependencies. The core innovation lies in variance-modulated dynamic scoring and capacity-constrained allocation, which preserve holistic context. Extensive experiments validate our method's effectiveness in maintaining both perceptual details and abstract spatial reasoning capabilities under aggressive token reduction.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); CAAI-Ant Group Research Fund; Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Yael Adini, Dov Sagi, and Misha Tsodyks. Context-enabled learning in the human visual system. *Nature*, 415(6873):790–793, 2002. 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 32

[4] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models in resource-constrained environments. *arXiv preprint arXiv:2408.10945*, 2024. 2, 8, 26

[5] Vaswani Ashish. Attention is all you need. *Advances in neural information processing systems*, 30:I, 2017. 4

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 10

[8] Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto Gonzalez Sanz. Pay attention to your loss: understanding misconceptions about lipschitz neural networks. *Advances in Neural Information Processing Systems*, 35:20077–20091, 2022. 6

[9] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 7, 25

[10] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

[11] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. 3, 8, 26

[12] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, 2024. 1

[13] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 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*, pages 19–35, 2024. 1, 2, 3, 4, 8, 26, 27

[14] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1

[15] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[16] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[19] Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. *arXiv preprint arXiv:2412.13180*, 2024. 1

[20] Zhanzhou Feng and Shiliang Zhang. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing*, 32:4156–4169, 2023. 2

[21] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023. 5, 7, 25

[22] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021. 32

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 4, 7, 25

[24] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023. 1

[25] Yuhang Han, Xuyang Liu, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. Rethinking token reduction in mllms: Towards a unified paradigm for training-free acceleration. *arXiv preprint arXiv:2411.17686*, 2024. 2, 3

[26] Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipvl: Efficient large vision-language models with dynamic token sparsification and kv cache compression. *arXiv preprint arXiv:2410.08584*, 2024. 3

[27] Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pages 214–230. Springer, 2024. 3

[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 3

[29] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 3

[30] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 7, 25

[31] Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. Fopru: Focal pruning for efficient large vision-language models. *arXiv preprint arXiv:2411.14164*, 2024. 3

[32] Yifan Jiang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, Jay Pujara, et al. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592, 2024. 3

[33] Shibo Jie, Yehui Tang, Ning Ding, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Memory-space visual prompting for efficient vision-language fine-tuning. In *Forty-first International Conference on Machine Learning*, 2024. 32

[34] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1

[35] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506, 2023. 1

[36] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025. 1

[37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[38] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 32

[40] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[41] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1, 3

[42] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*, 2023. 5, 7, 25

[43] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 26

[44] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3

[45] Hanxiao Liu, Andy Brock, Karen Simonyan, and Quoc Le. Evolving normalization-activation layers. *Advances in Neural Information Processing Systems*, 33:13539–13550, 2020. 32

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[47] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 5, 26

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 26

[49] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024. 1, 3, 8, 26

[50] Xiang Liu, Zhenheng Tang, Hong Chen, Peijie Dong, Zeyu Li, Xiuze Zhou, Bo Li, Xuming Hu, and Xiaowen Chu. Can llms maintain fundamental abilities under kv cache compression? *arXiv preprint arXiv:2502.01941*, 2025. 32

[51] Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *arXiv preprint arXiv:2502.00299*, 2025. 32

[52] Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, Siteng Huang, and Honggang Chen. Compression with global guidance: Towards training-free high-resolution mllms acceleration. *arXiv preprint arXiv:2501.05179*, 2025. 8, 26

[53] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 5, 7, 25, 27

[54] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 7, 25

[55] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3

[56] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer, 2025. 3

[57] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 1

[58] Junzhu Mao, Yang Shen, Jinyang Guo, Yazhou Yao, and Xiansheng Hua. Efficient token compression for vision transformer with spatial information preserved. *arXiv preprint arXiv:2503.23455*, 2025. 1

[59] Junzhu Mao, Yang Shen, Jinyang Guo, Yazhou Yao, Xiansheng Hua, and Hengtao Shen. Prune and merge: Efficient token compression for vision transformer with spatial information preserved. *IEEE Transactions on Multimedia*, 2025. 1

[60] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 2

[61] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023. 1

[62] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[63] Marius V Peelen, Li Fei-Fei, and Sabine Kastner. Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251):94–97, 2009. 2

[64] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 1

[65] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in neural information processing systems*, 34:12786–12797, 2021. 3

[66] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 2, 3, 8, 26

[67] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 5, 7, 25

[68] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. *arXiv preprint arXiv:2409.10994*, 2024. 3

[69] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[70] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996. 2

[71] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 3

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[73] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[74] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004. 32

[75] Dezhan Tu, Danylo Vashchilenko, Yuzhe Lu, and Panpan Xu. Vl-cache: Sparsity and modality-aware kv cache compression for vision-language model inference acceleration. *arXiv preprint arXiv:2410.23317*, 2024. 3

[76] Shakti N Wadekar, Abhishek Chaurasia, Aman Chadha, and Eugenio Culurciello. The evolution of multimodal model architectures. *arXiv preprint arXiv:2405.17927*, 2024. 4

[77] Ao Wang, Fengyuan Sun, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. [cls] token tells everything needed for training-free efficient mllms. *arXiv preprint arXiv:2412.05819*, 2024. 2, 3

[78] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[79] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1

[80] Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*, 2025. 2

[81] Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. 2, 8, 26

[82] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. 1

[83] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024. 3, 8, 26

[84] Bingxin Xu, Yuzhang Shang, Yunhao Ge, Qian Lou, and Yan Yan. freepruner: A training-free approach for large multimodal model acceleration. *arXiv preprint arXiv:2411.15446*, 2024. 3

[85] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM international conference on Multimedia*, pages 1645–1653, 2017. 8, 25, 26

[86] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 3

[87] Yibo Yan, Guangwei Xu, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. *arXiv preprint arXiv:2509.23883*, 2025. 1

[88] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024. 2, 8, 26

[89] Te Yang, Jian Jia, Xiangyu Zhu, Weisong Zhao, Bo Wang, Yanhua Cheng, Yan Li, Shengyuan Liu, Quan Chen, Peng Jiang, et al. Enhancing instruction-following capability of visual-language models by reducing image redundancy. *arXiv preprint arXiv:2411.15453*, 2024. 3

[90] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv:2405.20985*, 2024. 3

[91] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 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, pages 22128–22136, 2025. 2, 3

[92] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024. 5, 7, 25

[93] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024. 2, 3, 4, 8, 26

[94] Renshan Zhang, Yibo Lyu, Rui Shao, Gongwei Chen, Weili Guan, and Liqiang Nie. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv preprint arXiv:2407.14439*, 2024. 2

[95] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3

[96] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pages arXiv–2406, 2024. 6

[97] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 1

[98] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 1, 2, 3, 8, 26

[99] Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. Lova3: Learning to visual question answering, asking and assessment. *Advances in Neural Information Processing Systems*, 37:115146–115175, 2024. 1

[100] Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*, 2024. 2

[101] Yuke Zhu, Chi Xie, Shuang Liang, Bo Zheng, and Sheng Guo. Focusllava: A coarse-to-fine approach for efficient and effective visual token compression. *arXiv preprint arXiv:2411.14228*, 2024. 3

[102] Xin Zou, Chang Tang, Xiao Zheng, Zhenglai Li, Xiao He, Shan An, and Xinwang Liu. Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3550–3559, 2023. 32

[103] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025. 2

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are clearly stated in the abstract and the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion on the limitations of our work is stated in the paragraph E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our work is motivated by an interesting experimental phenomenon and proposes methods based on this observation, which improves the baseline by a large margin. There are no assumptions and no following proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper includes the implementation details in the experiment section and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the dataset URL and code URL as full submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specific experiment settings in Section 5 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We don't need to conduct such an evaluation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specific experiment settings in Section 5.4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We conducted the research in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The discussion on both potential positive societal impacts and negative societal impacts is stated in Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects, so no related details are included.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The research described in the paper does not involve study participants or human subjects, thus questions regarding potential risks, disclosure, or IRB approvals are not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The paper does not mention the usage of LLMs as a significant or original component of the core methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Contents of Technical Appendices

∞ Technical Appendices and Supplements

In this appendix, we first provide the details of the experimental setup, including information about the datasets, model architectures, and comparison methods. Then, we offer a more detailed computational complexity and theoretical analysis, along with more visualizations and insights.

A Detailed Experiment Settings

A.1 Benchmarks and Metrics

We conducted experiments on several widely used visual understanding benchmarks. For image understanding task, we performed experiments on ten widely used benchmarks, including GQA [30], MMBench (MMB) and MMB-CN [53], MME [21], POPE [42], VizWiz [9], SQA (ScienceQA) [54], VQA_{V2} (VQA V2) [23], VQA_{Text} (TextVQA) [67], and MMVet [92].

GQA [30] The GQA benchmark is composed of three main components: scene graphs, questions, and images. The image section encompasses not only the images themselves but also their spatial features and the attributes of all objects within the images. The questions in GQA are specifically crafted to assess the model's ability to comprehend visual scenes and engage in reasoning about different aspects of the images.

MMBench [53]. MMBench provides a comprehensive evaluation of a model's performance across multiple dimensions. It is structured into three levels of ability dimensions. The first level (L-1) focuses on two core abilities: perception and reasoning. Building on this foundation, the second level (L-2) includes six sub-abilities, further elaborating the model's capabilities. At the third level (L-3), the evaluation becomes more granular, encompassing 20 specific ability dimensions, thus ensuring a detailed and multi-faceted analysis of the model's performance.

MME [21]. The MME benchmark is another holistic evaluation framework, designed to thoroughly assess various facets of a model's performance. It includes 14 distinct subtasks, each targeting specific perceptual and cognitive abilities of the model. By employing carefully crafted instruction-answer pairs and maintaining concise instruction designs, the benchmark minimizes issues such as data leakage and unfair evaluation, ensuring a fair and reliable performance assessment.

POPE [42]. POPE focuses on evaluating the degree of Object Hallucination in models. It reformulates hallucination evaluation by prompting the model with specific binary questions regarding the presence of objects in images. Key metrics such as Accuracy, Recall, Precision, and F1 Score are utilized to measure the hallucination level across three different sampling strategies, providing a robust and precise evaluation of the model's object detection and hallucination behavior.

ScienceQA [54]. ScienceQA spans many domains, including natural sciences, language sciences, and social sciences. Questions are categorized within each domain according to topics, categories, and skills, which results in 26 topics, 127 categories, and 379 skills. This hierarchical categorization facilitates a thorough and diverse range of scientific questions, enabling an in-depth evaluation of the model's multimodal understanding, multi-step reasoning abilities, and interpretability.

VQA-V2 [23]. VQA-V2 is designed to evaluate a model's visual perception capabilities through open-ended questions. It consists of 265,016 images representing a wide variety of real-world scenes and objects, providing rich visual contexts for the associated questions. Each question is accompanied by 10 ground truth answers provided by human annotators, enabling a comprehensive evaluation of the model's ability to answer questions accurately and effectively.

TextVQA [67]. TextVQA focuses on the integration of text within images, evaluating the model's ability to comprehend and reason about both the visual and textual information present. The benchmark includes a series of visual question-answering tasks where the model must not only interpret the visual content but also read and understand the embedded text in order to respond correctly.

MMVet [92]. MMVet is designed to assess a model's ability to solve complex tasks by leveraging various core vision-language capabilities. It defines six core vision-language capabilities and examines 16 distinct integrations of these capabilities. This allows for a nuanced evaluation of how well models integrate and utilize multiple vision-language abilities to solve tasks.

MSVD-QA [85]. The MSVD-QA benchmark is derived from the Microsoft Research Video Description (MSVD) dataset and consists of 1970 video clips paired with approximately 50.5K question-

answer pairs. The questions span a wide range of topics and aspects related to the video content, making it suitable for video question-answering and video captioning tasks. The questions fall into five categories: what, who, how, when, and where, providing a comprehensive set of queries for model evaluation.

MSRVTT-QA [85]. MSRVTT-QA includes 10,000 video clips and 243,000 question-answer pairs. One of its primary challenges lies in understanding and reasoning about video content, which involves both visual and temporal aspects. To answer questions accurately, models must effectively integrate and process these components. Similar to MSVD-QA, the tasks in MSRVTT-QA are categorized into five question types: what, who, how, when, and where, allowing for detailed performance evaluation across multiple dimensions.

A.2 Backbones and Baselines

Models. We evaluate HoloV using various open-source MLLMs. For image understanding tasks, experiments are conducted on the LLaVA family, including LLaVA-1.5[2] [48] and LLaVA-NeXT[3] [47], with the latter used to validate performance on high-resolution images. For video understanding tasks, we use Video-LLaVA [43] as the baseline model. Following the settings reported in their paper.

We analyze multiple representative methods for accelerating MLLM inference through visual token pruning. These methods share the goal of improving efficiency by reducing redundant visual tokens.

ToMe [11] merges similar tokens in visual transformer layers through lightweight matching techniques, achieving acceleration without requiring additional training.

LLaVA-PruMerge [66] combines pruning and merging strategies by dynamically removing less important tokens using sparse `CLS`-visual attention and clustering retained tokens based on key similarity.

FastV [13] focuses on early-stage token pruning by leveraging attention maps, effectively reducing computational overhead in the initial layers.

HiRED [4] allocates token budgets across image partitions based on `CLS` token attention, followed by the selection of the most informative tokens within each partition, ensuring spatially aware token reduction.

PDrop [83] adopts a progressive token-dropping strategy across model stages, forming a pyramid-like token structure that balances efficiency and performance.

FasterVLM [93] evaluates token importance via `CLS` attention in the encoder and performs pruning before interaction with the language model, streamlining the overall process.

MustDrop [49] integrates multiple strategies, including spatial merging, text-guided pruning, and output-aware cache policies, to reduce tokens across various stages.

GlobalCom^2 [52] introduces a hierarchical approach by coordinating thumbnail tokens to allocate retention ratios for high-resolution crops while preserving local details.

SparseVLM [98] ranks token importance using cross-modal attention and introduces adaptive sparsity ratios, complemented by a novel token recycling mechanism.

VisionZip [88] evaluates token importance via attention in the encoder and clustering retained tokens based on key similarity.

DART [81] introduces a duplication-aware token reduction method that selects a small subset of pivot tokens, calculates cosine similarity between pivot tokens and remaining tokens, retains those with the lowest duplication to pivots, achieving significant acceleration while maintaining performance and good compatibility with efficient attention operators. These methods collectively highlight diverse approaches to token reduction, ranging from attention-based pruning to adaptive merging, offering complementary solutions for accelerating MLLMs.

[2] https://huggingface.co/liuhaotian/llava-v1.5-7b
[3] https://huggingface.co/liuhaotian/llava-v1.6-vicuna-7b

Table 7: Fine-grained comparison MMBench [53] between FastV and HoloV at high pruning ratios.

Category (dev)	Vanilla (576 Tokens)	FastV ↓ **90%** (58 Tokens)	HoloV ↓ **90%** (58 Tokens)	FastV ↓ **75%** (144 Tokens)	HoloV ↓ **75%** (144 Tokens)
Action Recognition	90.7	85.2	85.3	87.0	89.7
Attribute Comparison	50.0	50.0	53.9	52.3	48.7
Attribute Recognition	79.7	68.9	71.7	77.0	79.7
Celebrity Recognition	79.8	76.8	74.7	78.8	78.8
Function Reasoning	75.9	72.2	83.9	75.9	83.9
Future Prediction	45.0	30.0	58.3	40.0	58.3
Identity Reasoning	93.3	86.7	97.5	95.6	97.7
Image Emotion	78.0	58.0	68.7	78.0	76.0
Image Quality	35.8	22.6	38.8	28.3	40.1
Image Scene	96.2	90.4	91.5	96.2	97.1
Image Style	77.4	73.6	71.7	77.4	77.4
Image Topic	83.3	80.6	92.9	83.3	83.3
Nature Relation	41.7	39.6	49.4	37.5	37.5
Object Localization	39.5	35.8	23.3	37.0	38.3
OCR	59.0	59.0	81.8	59.0	84.4
Physical Property Reasoning	50.7	60.3	49.3	53.3	58.0
Physical Relation	33.3	41.7	32.7	41.7	41.7
Social Relation	88.4	53.5	75.8	72.1	75.7
Spatial Relationship	17.8	17.8	18.5	17.8	18.5
Structured Image-Text Understanding	26.9	30.8	21.8	28.2	21.9

A.3 Reproducibility

Implementaion Details. All of our experiments are conducted on Nvidia A800-80G GPU. The implementation was carried out in Python 3.10, utilizing PyTorch 2.1.2, and CUDA 11.8. All baseline settings follow the original paper. We set $num_{crop} = [1024/N]$, where N denotes the number of retained visual tokens, thus the smaller the quota, the more crops there will be for visual holistic context retention.

B More Sparsification Visualization

We conduct a detailed visualization of retained visual patches across varying pruning rates to illustrate the effectiveness of HoloV. As depicted in Fig. 11, 12, 13, the black regions represent discarded visual tokens, whereas the colored areas highlight key semantic zones that align with textual descriptions, demonstrating how HoloV strategically preserves informative content. Compared to FastV, a representative attention-based pruning method, HoloV exhibits superior capability in retaining relevant visual cues even at extremely high pruning ratios, such as 87.5%. This is achieved through its holistic pruning strategy, which prioritizes spatial-semantic diversity over isolated attention scores. By dynamically allocating pruning budgets across different image crops, HoloV effectively filters out redundant tokens while safeguarding critical objects and their contextual relationships. For instance, in complex scenes with multiple interacting elements, HoloV ensures that tokens corresponding to both focal objects and their surrounding environmental cues are preserved, whereas FastV tends to over-concentrate on high-attention regions, leading to loss of contextual coherence. This enhanced preservation of visual holistic understanding facilitates more accurate cross-modal alignment between visual features and language tokens, enabling MLLMs to maintain robust semantic reasoning capabilities even under aggressive token reduction. The visualization not only validates the superiority of HoloV's design philosophy but also provides empirical evidence of its ability to balance efficiency and semantic integrity in visual token pruning.

B.1 MMBench Finegrained Results

As shown in Table 7, in the MMBench [53] fine-grained comparison between FastV [13] and HoloV at 90% and 75% pruning ratios, significant performance improvements are evident with HoloV in several categories. Specifically, HoloV shows enhanced outcomes in Action Recognition, Attribute Recognition, Future Prediction, Identity Reasoning, Image Emotion, Image Quality, and Image Scene. These results underline HoloV's ability to retain crucial visual information for complex understanding and response capabilities within dynamic environments.

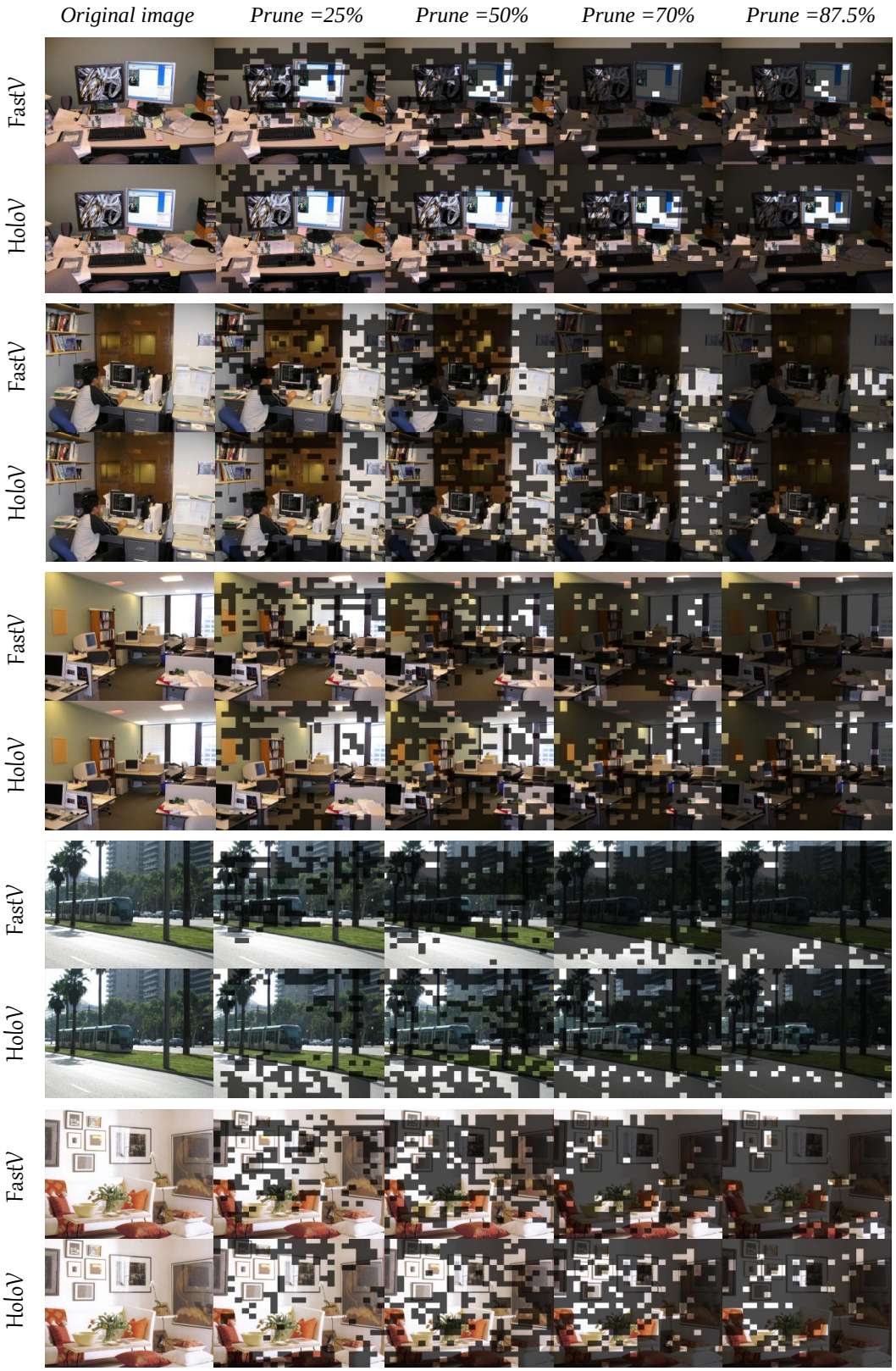

Figure 11: The case comparison between FastV and HoloV from the GQA. It presents original images alongside their pruned versions at pruning rates of 25%, 50%, 70%, and 87.5%. The bounding boxes highlight specific regions and objects across images, where HoloV well preserves the pivotal tokens.

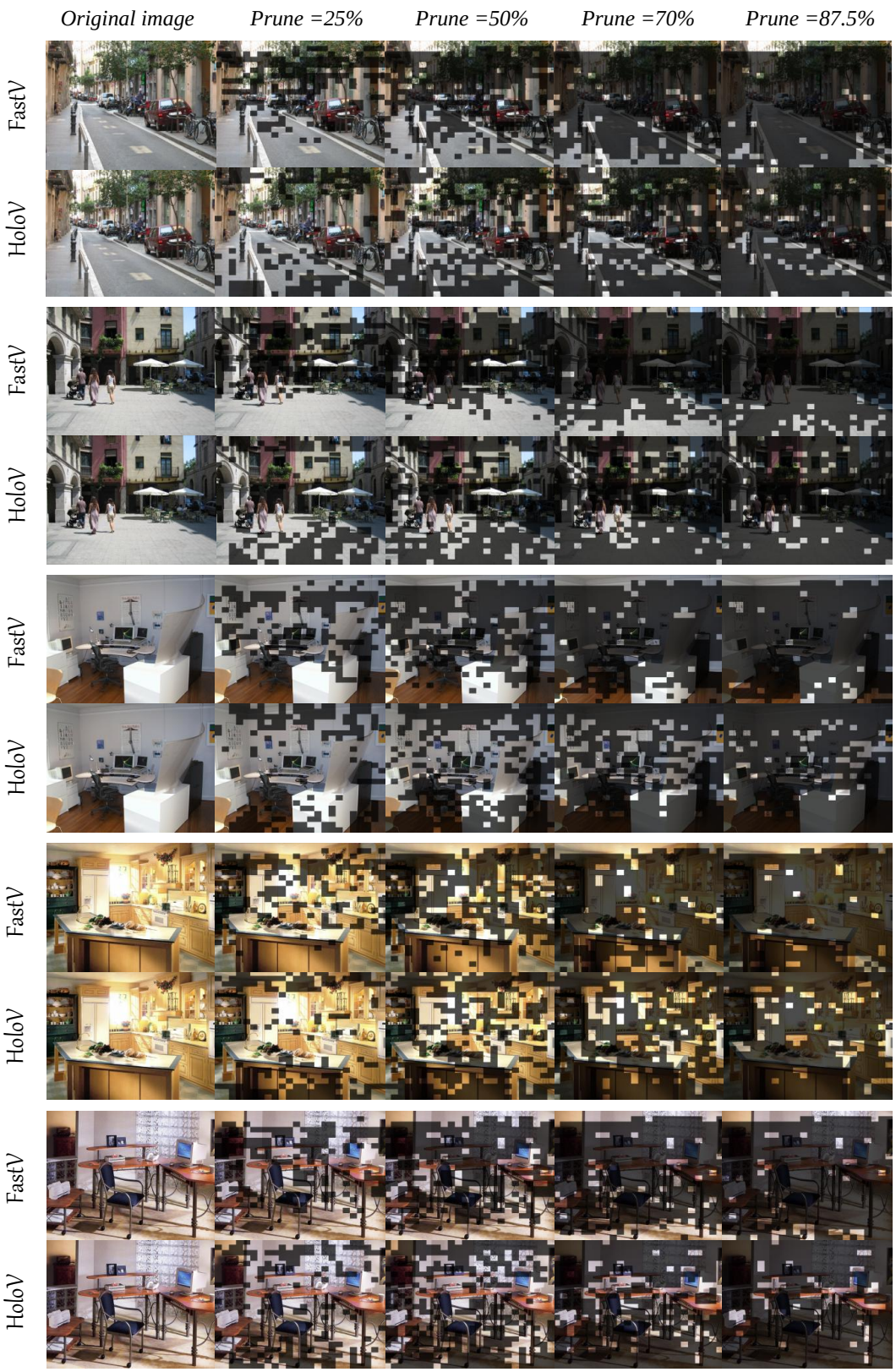

Figure 12: The case comparison between FastV and HoloV from the GQA. It presents original images alongside their pruned versions at pruning rates of 25%, 50%, 70%, and 87.5%. The bounding boxes highlight specific regions and objects across images, where HoloV well preserves the pivotal tokens.

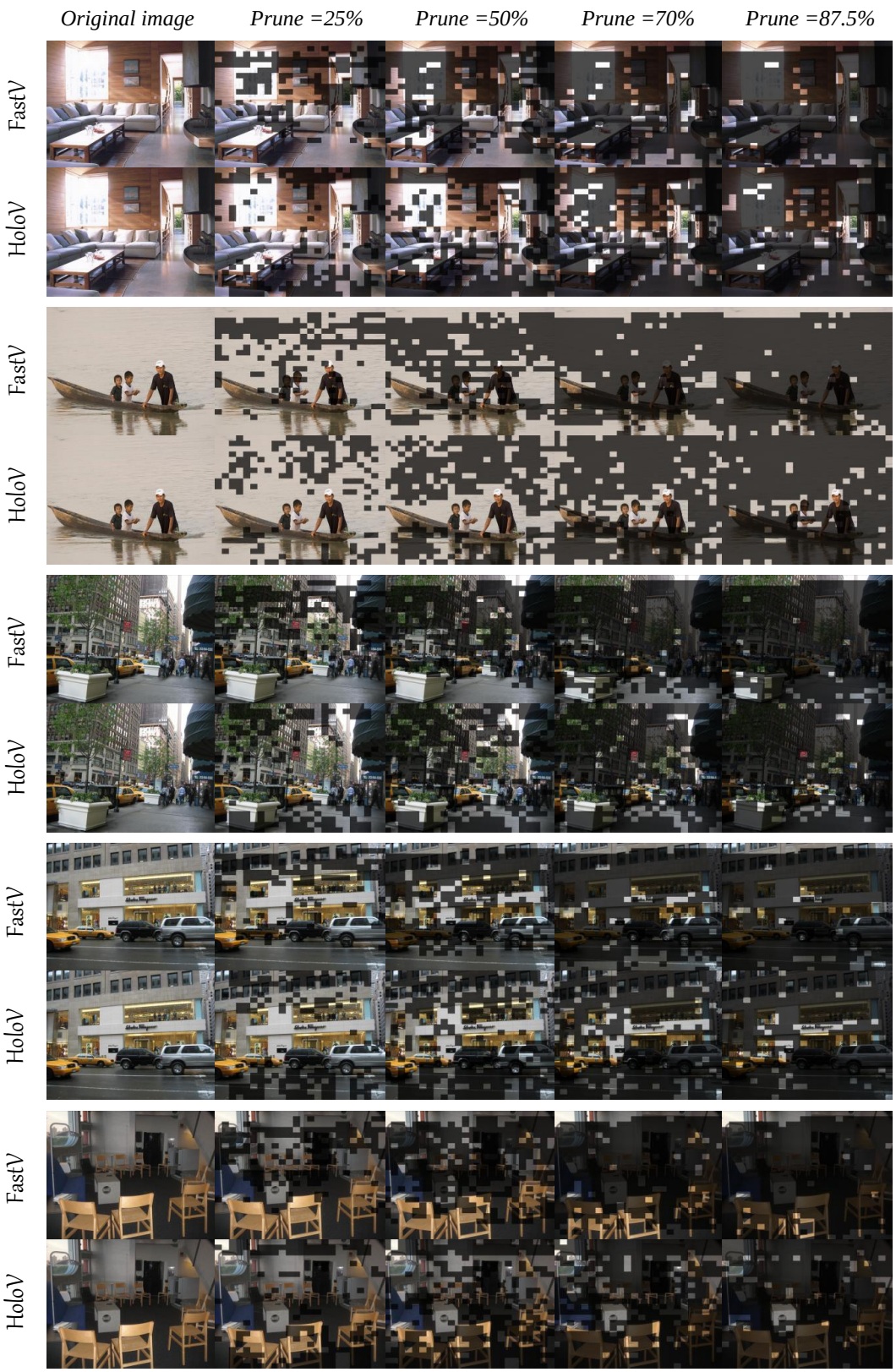

Figure 13: The case comparison between FastV and HoloV from the GQA. It presents original images alongside their pruned versions at pruning rates of 25%, 50%, 70%, and 87.5%. The bounding boxes highlight specific regions and objects across images, where HoloV well preserves the pivotal tokens.

C Theoretical Analysis of HoloV

To further justify the trustworthiness of our proposed HoloV, we provide a theoretical analysis of it.

Assumption 1 (Contextual Stability) *Let $\mathcal{X}_v$ be the original visual tokens set, and $\mathcal{R}_v \subseteq \mathcal{X}_v$ the retained visual tokens subset, We assume the following:*

***(C1)**. For any pruned visual token $x_j \in \mathcal{X}_v \setminus \mathcal{R}_v$, there exists $x_i \in \mathcal{R}_v$ with:*

$$d(x_i, x_j) \geq \epsilon \;\; \textit{and} \;\; \mathbb{V}\mathrm{ar}(d(x_i, \mathcal{N}(x_j))) \leq \delta \;,$$

where d means distance function like cosine similarity, $\mathcal{N}(x_j)$ denotes x_j's local context neighbors.

***(C2)**. For $\mathcal{H}(x_i) = \gamma\mathcal{V}(x_i) + \mathcal{A}(x_i)$ satisfies $\mathcal{H}(x_i) \geq \gamma$ for all retained tokens $x_i \in \mathcal{R}_v$*

Lemma C.1 (Token Coverage Guarantee) *Under **(A1)**, for any pruned token x_j, there exists $x_i \in \mathcal{R}$ such that:*

$$\|x_i - x_j\| \leq \sqrt{2(1-\epsilon)}\|x_j\| + \sqrt{\delta}$$

Proof C.1 *From the cosine similarity bound, there have $x_i^\top x_j \geq \epsilon\|x_i\|\|x_j\|$. Using the variance constraint:*

$$\mathbb{E}[(x_i^\top x_k - \mu)^2] \leq \delta, \quad \forall x_k \in \mathcal{N}(x_j)$$

where $\mu = \mathbb{E}[x_i^\top x_k]$. Combining via the triangle inequality:

$$\begin{aligned}\|x_i - x_j\|^2 &= \|x_i\|^2 + \|x_j\|^2 - 2x_i^\top x_j \\ &\leq 2B^2 - 2\epsilon B^2 + \sqrt{\delta} \\ &= 2(1-\epsilon)B^2 + \sqrt{\delta}\end{aligned}$$

The lemma shows that pruned tokens can be approximated by retained tokens in Euclidean space.

Theorem C.1 (Semantic Preservation) *Let f be a transformer layer with Lipschitz constant L. For input embeddings $\mathcal{X}_v$ and pruned set $\mathcal{R}_v$ satisfying **(C1)-(C2)**:*

$$\|f(\mathcal{X}_v) - f(\mathcal{R}_v)\| \leq L\left[\sqrt{2(1-\epsilon)}B + \sqrt{\delta}\right] + \eta(B, \gamma)$$

where $\eta(B, \gamma) = \mathcal{O}\left(B^2/\gamma\right)$ is the residual error from the scoring threshold.

Proof C.2 *Decompose the error into three components: 1) **Geometric distortion**: Bounded by Lemma C.1 2) **Context variance**: Controlled by $\sqrt{\delta}$ 3) **Scoring residual**:*

For any $x \in \mathcal{X}_v \setminus \mathcal{R}_v$ with $\mathcal{S}(x) < \gamma$:

$$\mathcal{V}^c + \mathcal{A}^c < \gamma \Rightarrow \mathcal{V}(x) < \gamma - \mathcal{A}(x)$$

Using Cauchy-Schwarz inequality:

$$\eta \leq \frac{1}{\gamma}\sum_{x \notin \mathcal{R}_v} \|W_V x\|^2 \leq \frac{CB^2}{\gamma}$$

Combining terms via the triangle inequality completes the proof.

This theorem guarantees that, even after pruning, the semantic difference between the outputs of the transformer for the original.

Corollary 1 (Dynamic Allocation Optimality) *The token allocation in Section 4 achieves:*

$$\max_{\{k_p\}} \sum_{p=1}^{P} \log\left(\sum_{t=1}^{k_p} \mathcal{S}_{pt}\right) \quad \textit{s.t.} \quad \sum_p k_p = N_{\textit{target}}$$

with approximation ratio $1 - 1/e$ when using greedy selection.

Proof C.3 *The allocation problem is equivalent to maximizing a monotone submodular function. Greedy algorithms provide* $(1 - 1/e)$*-approximation guarantees [74] for such problems.*

This corollary shows that your token allocation strategy is not only efficient but also theoretically near-optimal.

This theoretical framework demonstrates that HoloV: 1) Preserves semantic relationships through bounded geometric distortion. 2) Context variance is controlled via stability-aware pruning. 3) Token allocation is provably near-optimal, balancing efficiency and effectiveness.

D Fast Visual Context Refetching

D.1 Preliminary: Reformulation of FFN

Vanilla FFN comprises two fully connected layers with non-linear activation in between. We suppose $\boldsymbol{x} \in \mathbb{R}^d$ as an input token of the FFN, and FFN function can be formulated as

$$\mathrm{FFN}(\boldsymbol{x}) = \phi\left(\boldsymbol{x}\boldsymbol{W}_1\right)\boldsymbol{W}_2^\top, \tag{7}$$

where ϕ is activation function like ReLU or SiLU [45], and $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d\times D}$ are the weight matrices, in usual $D = 4d$. Peculiarly, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ can be rewritten as

$$\boldsymbol{W}_1 = (\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_D), \boldsymbol{W}_2 = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D), \tag{8}$$

where $\boldsymbol{k}_i, \boldsymbol{v}_i \in \mathbb{R}^d$ denote entries of key and value, respectively. As a result, the FFN can be reformulated as

$$\mathrm{FFN}(\boldsymbol{x}) = \sum \phi\left(\langle \boldsymbol{x}, \boldsymbol{k}_i\rangle\right) \cdot \boldsymbol{v}_i \ . \tag{9}$$

Thus, the FFN function can be construed as using input $\boldsymbol{x}$ as a query to measure similarity with keys, find matching values, and gather values by similarity, which works like a key-value memory storing the factual knowledge as found in previous studies [22, 33].

D.2 FFN with Visual Context Refetching

We propose visual context refetching (VCR), *i.e.*, reinjecting pruned visual information into the middle layer of the text decoder during elevated uncertainty during reasoning. This strategy treats pruned visual tokens as anchors to recalibrate off-target predictions and reduces uncertainties in *object, attribute, relationship* tokens. The reason we call this pattern of reinjecting visual evidence VCR is that the model finds and refreshes key visual memories based on the hidden states in this process. In particular, inspired by the fact that FFN executes analogous retrieval from its key-value memory, we consider VCR to serve as a simplified and efficient information re-retrieval process. Given a hidden token $\boldsymbol{x} \in \mathbb{R}^d$ and dimension-aligned vision tokens $\boldsymbol{z}_v$, FFN with visual context refetching at l-th layer can be written as follows

$$\mathrm{FFN}^{(l)}(\boldsymbol{x} \propto \boldsymbol{z}_v) = \alpha\underline{\Delta} + (1-\alpha)\,\mathrm{FFN}^{(l)}(\boldsymbol{x}), \tag{10}$$

where $\boldsymbol{z}_v = (\boldsymbol{z}_{v,1}, \ldots, \boldsymbol{z}_{v,N_v}) \in \mathbb{R}^{d\times N_v}$, $x \propto \boldsymbol{z}_v$ denotes execute VCR $\underline{\Delta}$ from $\boldsymbol{x}$ to visual features $\boldsymbol{z}_v$, and $\alpha \in [0, 1]$ denotes injection ratio of visual memory through the FFN layer which proportional to image complexity. Specifically, instead of performing retrieval via cross-attention layers as in previous approaches [39, 3, 102], we consider a simple retrieval process for VCR as,

$$\underline{\Delta}(\boldsymbol{z}_v \mid \boldsymbol{x}) = \sum\nolimits_{i=1}^{N_v} \phi(\langle \boldsymbol{x}, \boldsymbol{z}_{v,i}\rangle) \cdot \boldsymbol{z}_{v,i}. \tag{11}$$

From the perspective of FFN, VCR works by treating $\boldsymbol{x}$ as a query, and $\langle \boldsymbol{z}_{v,i} : \boldsymbol{z}_{v,i}\rangle$ as new key-value entries (visual evidence) to supplement vision-related information in the hidden states. In this information re-retrieval process, MemVCR does not introduce any parameters that need to be trained. Notably, since the size of key-value memory D in FFN typically far exceeds the number of visual tokens N_v (for instance, $D = 11008$ in LLaMA-7B and $N_v = 256$ for ViT-L/14, $N_v \ll D$), the computation of VCR is negligible. Thus, VCR operation is more efficient than the cross-attention mechanism with quadratic complexity [50, 51].

D.3 Further Efficiency Analysis

As shown in Fig. 14, we conduct efficiency evaluation on LLaVA-NeXT 7B at 95% pruning ratio, where we also introduce baseline (unpruned Vanilla) and FastV (95% pruned) for comparison. We evaluate these approaches using QA pairs from GQA, and the output length has been set to 1. During evaluation, an A800 80GB GPU has been used, and the average FLOPs, VMemory usage and throughput has been calculated, shown in Fig. 14. HoloV reduces over 90% of FLOPs requirement, 37% lower than FastV, and its VMemory usage is at the lowest level, while keeping throughput at 5.2 per second, 2.16x and 1.13x faster than baseline and FastV, respectively.

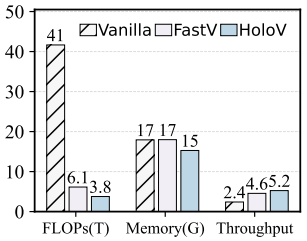

Figure 14: Inference efficiency comparison between FastV and HoloV.

E Impact Statement

This paper presents HoloV, a visual token pruning framework for MLLMs, and discusses its potential societal impacts. On the positive side, HoloV enhances the accessibility of multimodal technologies by reducing computational overhead, making advanced applications like medical image analysis and autonomous driving more feasible in resource-constrained environments such as edge devices or underserved regions. Its efficiency also contributes to energy sustainability by lowering the energy consumption of MLLM inference, aligning with global efforts to mitigate the environmental impact of AI. Additionally, by preserving holistic visual context instead of relying solely on attention-based "highlighted tokens," HoloV may reduce biases in model outputs, improving fairness in diverse scenarios like visual reasoning involving underrepresented communities. The framework's plug-and-play design further accelerates its integration into real-world systems, driving innovations in education, accessibility tools, and emergency response to enhance societal resilience.

However, the work also carries potential negative implications. The reduced computational barriers enabled by HoloV could facilitate misuse, such as the creation of deepfakes or misinformation, particularly in regions with limited regulatory oversight. While aiming to mitigate attention-based biases, the framework's crop-wise token allocation might inadvertently reinforce other biases if training data lacks diversity, potentially disadvantaging underrepresented groups. Moreover, the focus on inference efficiency might lead developers to prioritize speed over model interpretability, raising concerns about accountability in "black-box" deployments for high-stakes tasks like healthcare diagnostics. Lastly, over-reliance on post-hoc pruning could deter investments in more equitable training data or architectural improvements, potentially accumulating technical debt and masking foundational issues in MLLM development.

Limitations and Future Work. HoloV demonstrates robust performance in preserving holistic visual context but faces two key limitations: its dependence on fixed spatial crop partitioning may hinder fine-grained semantic capture in complex scenes, and minor accuracy declines persist even at high pruning ratios (e.g., 4.2% drop when pruning 88.9% visual tokens). To address these, future work could prioritize adaptive crop, sparse attention, multi-modality extensions (*e.g.*, 3D data), and integration with hallucination mitigation, while optimizing for edge computing energy efficiency.