

MR-Pruner: Training-free Multi-resolution Visual Token Pruning for Multi-modal Large Language Models

Seunghoon Han Hyewon Lee Soyoung Park Jong-Ryul Lee* Sungsu Lim*
 Chungnam National University, South Korea

{tmdgns129, noweyh927}@g.cnu.ac.kr, sypark1452@o.cnu.ac.kr, {jongryul.lee, sungsu}@cnu.ac.kr

Abstract

Large Language Models (LLMs) extended to multi-modal inputs have led to Multi-modal LLMs (MLLMs) that perform strongly on vision-language tasks. Recent MLLMs adopt multi-resolution inputs to capture both global context and local details, but this substantially increases visual tokens and computational cost. Existing pruning methods reduce redundancy but are designed for single-resolution settings, overlooking the characteristics of multi-resolution tokens. We observe two key properties: tokens from different resolutions follow distinct distributions of information content, and tokens across resolutions exhibit mutual complementarity, such that pruning one type can often be compensated by the other. Based on this observation, we propose Multi-Resolution Token Pruning method (MR-Pruner), a training-free, graph-based pruning framework for multi-resolution MLLMs. MR-Pruner incorporates three components—Intra-resolution, Cross-resolution Token Scoring, and Informativeness-aware Token Pruning—that adaptively allocate pruning ratios and facilitate information propagation across resolutions. Experiments on eight benchmarks show that MR-Pruner achieves superior efficiency–performance trade-offs. For example, when only 10% of the visual tokens are retained, it leads to an average performance degradation of 3.6%. For reproducibility, the source code is available at <https://github.com/gooriie/MR-Pruner>.

1. Introduction

Large Language Models (LLMs) [4, 7, 25, 26] have achieved remarkable progress in recent years, exhibiting strong performance across a broad spectrum of natural language tasks. Building upon these advances, researchers have extended LLMs to multi-modal inputs, leading to the development of Multi-modal Large Language Models (MLLMs) [2, 13, 24, 29, 32]. These models are particularly

*Corresponding authors.

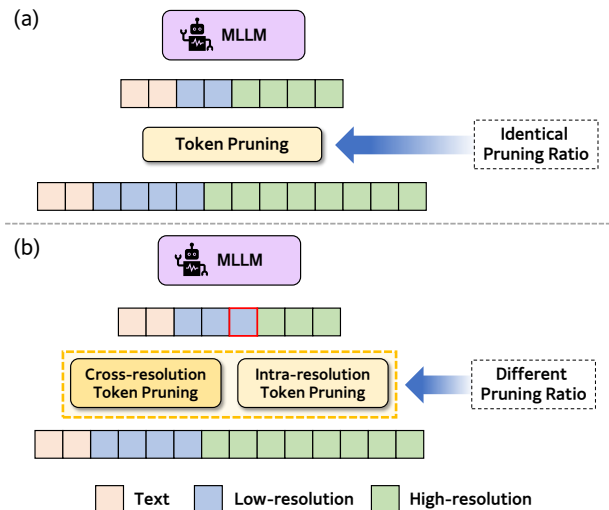


Figure 1. (a) Overview of existing single-resolution pruning methods, where low- and high-resolution tokens share the identical pruning strategy. (b) Overview of MR-Pruner, which applies distinct pruning strategies and ratios to different resolutions.

impactful in diverse and complex vision-language tasks.

In these tasks, visual features are extracted by a vision encoder and projected into the semantic space of LLMs to be jointly processed with text. However, although this paradigm effectively equips LLMs with visual grounding, extracting features from entire images poses significant challenges in capturing fine-grained details. To mitigate these, recent works [6, 17, 20, 31] have explored processing images at higher resolutions and partitioning them into multiple regions. Such designs enable MLLMs to capture both global context and local details. However, this comes at a considerable computational cost: the number of visual tokens, which already greatly exceeds textual inputs, increases further as multiple resolutions are increased.

To alleviate this inefficiency, token pruning has been proposed to discard redundant visual tokens while retaining those most relevant to the task. Existing approaches [3, 15,

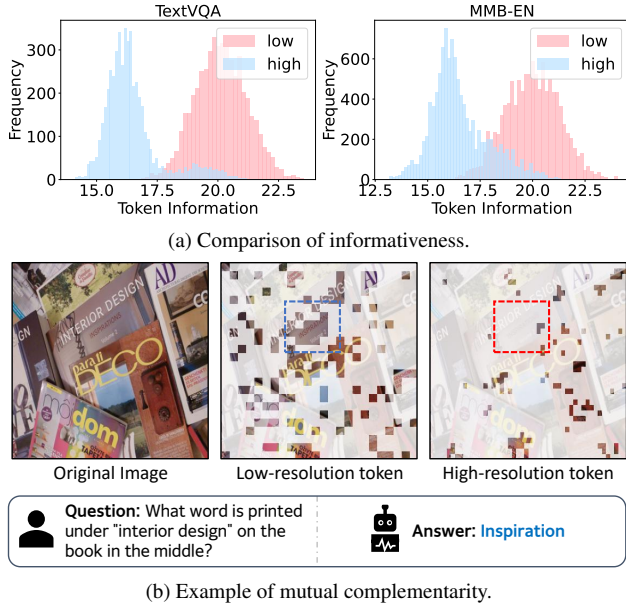


Figure 2. Properties of visual tokens from different resolution.

[21, 28] typically estimate token importance and remove tokens deemed less significant. These methods are typically designed for single-resolution settings and tend to overlook the distinct characteristics of visual tokens extracted from multiple resolutions. We observe that the tokens extracted from different resolutions exhibit two key properties: (1) Distinct informativeness distributions. Tokens from different resolutions show different distributions of informativeness, which is typically quantified by metrics such as attention weights or the L2-norm of tokens [5, 12]. As shown in Figure 2a, across both the TextVQA [23] and MMB-EN [18] datasets, low-resolution tokens generally exhibit higher information content than high-resolution tokens. (2) Mutual complementarity. Tokens from different resolutions also exhibit a complementary nature. As illustrated in Figure 2b, even when only one type of token is retained—either high- or low-resolution—the model can still answer correctly in certain cases. This suggests that tokens from different resolutions can effectively compensate for each other.

Based on these observations, we introduce two pruning strategies that are tailored to efficiently handle complex multi-resolution settings: (1) Informativeness-aware pruning ratio. Since tokens from different resolutions contain varying levels of information, we adaptively assign pruning ratios according to their informativeness. Tokens from resolutions with relatively higher information content are preserved more aggressively, while tokens from less informative resolutions are pruned more heavily. (2) Cross-resolution preservation. If a region is pruned from one resolution, the corresponding area is preserved in the other resolution to prevent the simultaneous loss of token information.

To realize these strategies, we propose Multi-Resolution Token Pruner (MR-Pruner), a lightweight, training-free, graph-based token pruning method designed as a flexible plug-in module for multi-resolution MLLMs. The framework is constructed of three components: Intra-resolution Token Scoring, Cross-resolution Token Scoring, and Informativeness-aware Token Pruning. In Intra-resolution Token Scoring, we model the relationships between tokens within the same resolution by constructing an Intra-resolution Token Graph. Through information propagation within this graph, the importance of tokens can be measured without explicitly relying on resolution differences, which provides a basis for identifying representative tokens within each resolution. In Cross-resolution Token Pruning, we explicitly model the relationships between tokens from different resolutions by constructing a Cross-resolution Token Graph. This graph enables information propagation from pruned tokens in one resolution to their corresponding regions in the other, thereby ensuring that complementary information is retained across resolutions. In Informativeness-aware Token Pruning, we exploit the distinct distributions of informativeness across resolutions. By assigning pruning ratios adaptively according to token informativeness, the framework preserves a larger proportion of tokens from more informative resolutions.

To evaluate the effectiveness of MR-Pruner, we integrate it into LLaVA-NeXT [17], which is a widely-used famous multi-resolution MLLM, and conduct experiments on eight real-world benchmark datasets. Our observations show that MR-Pruner substantially reduces the computational cost of MLLMs, while achieving consistently superior performance compared to existing state-of-the-art methods. In addition, MR-Pruner demonstrates faster or comparable throughput compared to existing baselines, thereby validating its efficiency as well as its effectiveness. For example, in the 90% pruning scenario, MR-Pruner achieves an average degradation of 3.6% in performance.

The main contributions of our work can be summarized as follows:

- We propose MR-Pruner, a novel training-free and plug-in framework for multi-resolution MLLMs that leverages graph-based token pruning methods.
- We introduce three components—Intra-resolution, Cross-resolution Token Scoring, and Informativeness-aware Token Pruning—to address distributional differences and mutual complementarity across resolutions.
- We design a Cross-information Propagation mechanism that propagates information between tokens across resolutions, thereby preventing the simultaneous removal of complementary tokens.
- We conduct extensive experiments on eight real-world benchmark datasets, demonstrating that MR-Pruner achieves superior efficiency-performance trade-offs.

2. Related Works

2.1. Multi-modal Large Language Models

Large Language Models (LLMs) such as GPT-3 [4] and LLaMA [25] have shown strong generalization in text-only tasks, inspiring their extension to multi-modal settings. Early MLLMs, including BLIP-2 [13], MiniGPT-4 [32], Qwen-VL [2], and Flamingo [1], aligned visual features from frozen vision encoders with LLMs through lightweight projection or cross-attention modules, enabling effective performance on tasks like visual question answering and captioning. These approaches established a simple but powerful pipeline that could scale with larger backbones and instruction datasets. To handle fine-grained reasoning, recent models increased image resolution and adopted multi-resolution inputs. LLaVA [16] introduced visual instruction tuning, while LLaVA-NeXT [17] partitioned images into regions for multi-resolution processing. InternVL [6] further scaled this design with diverse aspect ratios and higher resolution patches. Although effective, these approaches substantially expand the number of visual tokens, creating severe redundancy and high computational cost. This motivates token pruning as an input-level compression strategy, which directly reduces the visual sequence length while preserving semantic grounding.

2.2. Token Pruning

Token pruning has been widely studied in Vision Transformers (ViTs) [8] to accelerate inference. In MLLMs, the presence of more layers and higher computational demands compared to standard ViTs makes token pruning particularly effective. To this end, existing methods proposed strategies to eliminate redundant or unnecessary visual tokens. For example, DynamicViT [21], EViT [15], and Evo-ViT [28] reduce the number of tokens in ViTs, but require additional training to preserve performance after pruning. Since this leads to significant extra computational costs, recent studies have explored training-free approaches. ToMe [3] merges highly similar token pairs through weighted averaging to remove redundancy, while FastV [5] discards tokens with lower scores based on the cross-attention weights from intermediate layer outputs. PruMerge [22] exploits the sparsity of attention between class and visual tokens to adaptively select and merge informative tokens. Zero-TPrune [27] regards the attention weights as a graph to retain the most important tokens, and G-Prune [12] models the relationships between visual tokens as a graph to estimate their importance. However, these methods only consider single-resolution tokens and thus fail to capture the characteristics of multi-resolution tokens. To address this limitation, this paper focuses on a training-free, graph-based token pruning approach that incorporates the properties of multi-resolution tokens.

3. Method

In this section, we describe MR-Pruner, a graph-based token pruning method designed for multi-resolution MLLMs. As illustrated in Figure 3, MR-Pruner consists of three components: Intra-resolution Token Scoring, Cross-resolution Token Scoring, and Informativeness-Aware Token Pruning. The Intra-resolution Token Scoring measures the importance of tokens within the identical resolution token type through Intra-information Propagation in the Intra-resolution Token Graph. The Cross-resolution Token Scoring estimates the importance of tokens within the different resolution token types by leveraging Cross-information Propagation in the Cross-resolution Token Graph. Finally, Informativeness-Aware Token Pruning leverages the differences among tokens extracted from different resolution to dynamically adjust their pruning ratios. The overall procedure of MR-Pruner is presented in Algorithm 1, and we provide further details in the following subsection.

3.1. Intra-resolution Token Scoring

In order to model the relationships among visual tokens of identical resolution through a Intra-resolution Token Graph, we follow the graph construction strategy of the prior method. However, in contrast to the prior approach, information propagation within the Intra-resolution Token Graph is executed only once instead of iteratively, thereby reducing computational overhead.

Intra-resolution Token Graph. Suppose we are given a set of visual tokens $\mathbf{X} \in \mathbb{R}^{N \times d}$, where N and d denote the number and dimension of visual tokens. We construct an Intra-resolution Token Graph for tokens that share the identical resolution. The set of visual tokens \mathbf{X} consists of a high-resolution subset $\mathbf{X}^H \in \mathbb{R}^{N_h \times d}$ and a low-resolution subset $\mathbf{X}^L \in \mathbb{R}^{N_l \times d}$, where N_h and N_l denote the number of high-resolution tokens and low-resolution tokens, respectively, and their sum equals N . To model the relationships among similar visual tokens, we construct an adjacency matrix $\mathbf{A}^H \in \mathbb{R}^{N_h \times N_h}$ as follows:

$$\mathbf{A}_{ij}^H = \begin{cases} \cos(\mathbf{X}_i^H, \mathbf{X}_j^H), & \cos(\mathbf{X}_i^H, \mathbf{X}_j^H) \geq s_a, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where i and j denote the indices of visual tokens, $\cos(\cdot, \cdot)$ represents the cosine similarity, and s_a is the threshold used to sparsify the Intra-resolution Token Graph. The adjacency matrix $\mathbf{A}^L \in \mathbb{R}^{N_l \times N_l}$ for the low-resolution subset is constructed using the same procedure.

Intra-information Propagation. To evaluate token importance, we utilize information propagation within the Intra-resolution Token Graph. Although the importance of a token can be estimated from its own informativeness, leveraging information from neighboring tokens yields a more reliable assessment. Specifically, the constructed Intra-

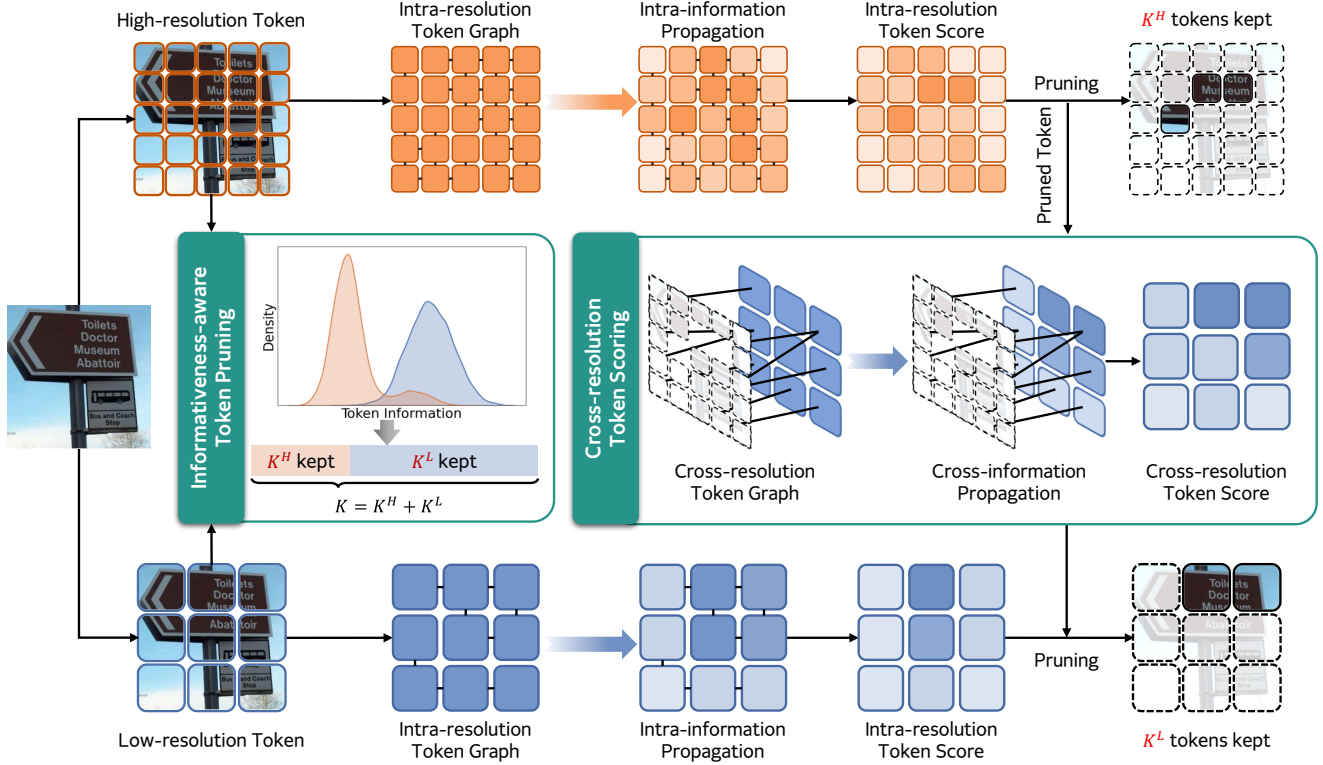


Figure 3. The overview of MR-Pruner. MR-Pruner can be seamlessly integrated into multi-resolution MLLMs as a plug-in method to reduce redundant visual tokens. It identifies token importance through Intra-resolution and Cross-resolution Token Scoring, and applies Informativeness-aware Token Pruning to adaptively determine pruning ratios.

resolution Token Graph enables information propagation between tokens of the same resolution to measure their informativeness. To define the informativeness $I^H \in \mathbb{R}^{N_h \times 1}$ of each token, we apply the L2-norm of its feature, formulated as follows:

$$I_i^H = \sqrt{\sum_{k=1}^d X_{ik}^{H^2}}, \quad (2)$$

where X_{ik}^H represents the k -th feature component of the i -th token. The initialized token information is then propagated through the graph, and the intra-resolution token score vector $S^H \in \mathbb{R}^{N_h \times 1}$ within the graph is computed as follows:

$$S^H = A^H I^H, \quad (3)$$

where S^H reflects not only the information of the token itself but also that of its neighbors, and it serves as the basis for pruning. The intra-resolution token score vector $S^L \in \mathbb{R}^{N_l \times 1}$ for the low-resolution subset is computed in the same way.

3.2. Informativeness-aware Token Pruning

Informativeness-aware Pruning Ratio. Prior works pruned tokens uniformly without considering resolution.

However, as shown in Figure 2a, token informativeness varies across resolutions, which implies that incorporating resolution-aware pruning strategies could be beneficial for performance preservation. To this end, we compute the average information of each token type and assign the pruning ratio proportionally. Let \bar{I}^H and \bar{I}^L denote the mean informativeness scores of the high-resolution and low-resolution token types, respectively. Given the total number of tokens to retain, denoted as K , the informativeness-aware pruning ratio is determined as:

$$K^H = K \cdot \frac{\bar{I}^H}{\bar{I}^H + \bar{I}^L}, \quad K^L = K - K^H, \quad (4)$$

where K^H and K^L represent the number of preserved tokens for the high-resolution and low-resolution token types, respectively. In this way, token types with higher informativeness scores retain more tokens, while those with lower scores are pruned more aggressively. This strategy allows the pruning process to better reflect the inherent information distribution across resolutions.

Token Pruning. Let S^H denote the final scores of high-resolution tokens and S^L those of low-resolution tokens. We define \bar{S}^L as the combined score of S^L and the cross-

resolution token score introduced in Section 3.3. Based on these definitions, the pruning process is as follows:

$$\begin{aligned} R^H &= \operatorname{argmax}_{\mathcal{I} \subset \{1, 2, \dots, N_h\}, |\mathcal{I}|=K^H} \sum_{i \in \mathcal{I}} S_i^H, \\ R^L &= \operatorname{argmax}_{\mathcal{I} \subset \{1, 2, \dots, N_l\}, |\mathcal{I}|=K^L} \sum_{i \in \mathcal{I}} \bar{S}_i^L, \end{aligned} \quad (5)$$

where R^H and R^L denote the indices of the retained tokens for each type. The pruning process is performed sequentially rather than simultaneously. The high-resolution tokens are pruned first, and the low-resolution tokens are pruned afterwards.

3.3. Cross-resolution Token Scoring

Cross-resolution Token Graph. By giving high scores to low-resolution tokens corresponding to pruned high-resolution tokens, the model can better exploit the mutual complementarity between tokens. To this end, we propagate the information of the pruned high-resolution tokens to semantically similar low-resolution tokens, thereby enhancing their importance scores. Specifically, given the indices R^H of the remaining high-resolution tokens after pruning, the pruned tokens $\mathbf{X}^{H, \text{pruned}} \in \mathbb{R}^{|P^H| \times d}$ are defined as follows:

$$P^H = \{1, 2, \dots, N_h\} \setminus R^H, \quad (6)$$

$$\mathbf{X}^{H, \text{pruned}} = \{\mathbf{X}_i^H \mid i \in P^H\}. \quad (7)$$

We build a Cross-resolution Token Graph to capture the relationships between the pruned high-resolution tokens and the low-resolution tokens, the adjacency matrix $\mathbf{A}^C \in \mathbb{R}^{N_l \times |P^H|}$ is generated as follows:

$$\mathbf{A}_{ij}^C = \begin{cases} \cos(\mathbf{X}_i^L, \mathbf{X}_j^{H, \text{pruned}}), & \cos(\mathbf{X}_i^L, \mathbf{X}_j^{H, \text{pruned}}) \geq s_c, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where i and j denote the indices of visual tokens, and s_c is the threshold for sparsifying Cross-resolution Token Graph.

Cross-information Propagation. To enhance the importance of low-resolution tokens associated with regions similar to the pruned high-resolution tokens, we perform Cross-information Propagation, propagating the information from the pruned tokens. From the informativeness \mathbf{I}^H obtained in Section 3.1, we propagate only the portion $\mathbf{I}^{H, \text{pruned}} \in \mathbb{R}^{|P^H| \times 1}$ corresponding to the pruned tokens to the low-resolution tokens. The cross-resolution token score vector $\mathbf{S}^C \in \mathbb{R}^{N_l \times 1}$ is computed as follows:

$$\mathbf{S}^C = \mathbf{A}^C \mathbf{I}^{H, \text{pruned}}, \quad (9)$$

where $\mathbf{I}^{H, \text{pruned}}$ is a pruned version of \mathbf{I}^H . The cross-resolution token score vector propagated from the pruned high-resolution tokens is integrated with the intra-resolution

Algorithm 1 MR-Pruner: Multi-Resolution Token Pruning

Input: High-resolution tokens $\mathbf{X}^H \in \mathbb{R}^{N_h \times d}$, Low-resolution tokens $\mathbf{X}^L \in \mathbb{R}^{N_l \times d}$, Number of tokens to retain K , similarity threshold s_a, s_c , balancing factor α .

Output: Indices of selected tokens R^H, R^L .

```

1: for  $r \in \{H, L\}$  do
2:    $N \leftarrow \begin{cases} N_h & \text{if } r = H \\ N_l & \text{if } r = L \end{cases}$ 
3:   for  $i, j = 1$  to  $N$  do
4:      $\mathbf{A}_{ij}^r \leftarrow \cos(\mathbf{X}_i^r, \mathbf{X}_j^r) \cdot \mathbf{1}\{\cos(\mathbf{X}_i^r, \mathbf{X}_j^r) \geq s_a\}$ 
5:   end for
6:   for  $i = 1$  to  $N$  do
7:      $\mathbf{I}_i^r \leftarrow \|\mathbf{X}_i^r\|_2$ 
8:   end for
9:    $\mathbf{S}^r \leftarrow \mathbf{A}^r \mathbf{I}^r$ 
10: end for
11:  $K^H \leftarrow K \cdot \frac{\bar{\mathbf{I}}^H}{\bar{\mathbf{I}}^H + \bar{\mathbf{I}}^L}$ 
12:  $K^L \leftarrow K - K^H$ 
13:  $R^H \leftarrow \operatorname{argmax}_{\mathcal{I} \subset \{1, 2, \dots, N_h\}, |\mathcal{I}|=K^H} \sum_{i \in \mathcal{I}} S_i^H$ 
14:  $P^H \leftarrow \{1, 2, \dots, N_h\} \setminus R^H$ 
15:  $\mathbf{X}^{H, \text{pruned}} \leftarrow \{\mathbf{X}_i^H \mid i \in P^H\}$ 
16: for  $i = 1$  to  $N_l$ ,  $j = 1$  to  $|P^H|$  do
17:    $\mathbf{A}_{ij}^C \leftarrow \cos(\mathbf{X}_i^L, \mathbf{X}_j^{H, \text{pruned}}) \cdot \mathbf{1}\{\cos(\mathbf{X}_i^L, \mathbf{X}_j^{H, \text{pruned}}) \geq s_c\}$ 
18: end for
19:  $\mathbf{I}^{H, \text{pruned}} \leftarrow \{\mathbf{I}_i^H \mid i \in P^H\}$ 
20:  $\mathbf{S}^C \leftarrow \mathbf{A}^C \mathbf{I}^{H, \text{pruned}}$ 
21:  $\bar{\mathbf{S}}^L \leftarrow \mathbf{S}^L + \alpha \cdot \mathbf{S}^C$ 
22:  $R^L \leftarrow \operatorname{argmax}_{\mathcal{I} \subset \{1, 2, \dots, N_l\}, |\mathcal{I}|=K^L} \sum_{i \in \mathcal{I}} \bar{S}_i^L$ 
23: return  $R^H, R^L$ 

```

token score vector to derive the final low-resolution token score vector. At this stage, we allow the impact of the cross-resolution token score vector on the overall token importance to be adjusted as follows:

$$\bar{\mathbf{S}}^L = \mathbf{S}^L + \alpha \cdot \mathbf{S}^C. \quad (10)$$

The final score vector reflects both the informativeness of the low-resolution tokens and the borrowed informativeness from the pruned high-resolution tokens. We select the important low-resolution tokens based on this score vector.

4. Experiments

4.1. Datasets

We conduct experiments on eight widely-used MLLMs benchmark datasets to evaluate the effectiveness and efficiency of MR-Pruner. Specifically, we report results on GQA [11] for real-world visual reasoning, VQA2.0 [10] for general visual question answering, MME [9] for comprehensive multi-modal evaluation, POPE [14] for object

Table 1. Comparison with existing methods on MLLMs benchmark datasets. The **bold** text indicates the best performance for each dataset at the same pruning ratio. Throughput denotes the number of processed samples per second (samples/sec).

Method	Pruning Ratio	GQA	VQA 2.0	MME	POPE	MMB-EN	MMB-CN	TextVQA	SQA-IMG	Throughput
<i>Upper Bound Model</i>										
LLaVA-NeXT-8B	0%	65.38	82.70	1587.72	87.84	72.08	67.18	65.41	73.43	1.46
<i>Single-resolution Pruning Methods</i>										
Random	50%	64.95	81.61	1605.43	86.47	70.27	63.83	58.21	73.48	2.25 (1.54×)
	70%	64.22	80.17	1576.33	84.98	69.42	61.34	49.48	73.53	2.36 (1.61×)
	90%	60.55	74.23	1475.07	79.63	61.77	51.46	31.73	72.43	2.54 (1.74×)
ToMe	50%	65.07	81.82	1566.60	87.56	70.88	64.43	59.07	72.52	0.59 (0.41×)
	70%	64.07	80.56	1564.36	87.33	68.21	61.91	52.19	70.88	0.64 (0.44×)
	90%	59.72	76.36	1453.13	84.29	61.77	53.14	38.36	69.98	0.72 (0.49×)
FastV	50%	65.11	82.51	1604.14	87.51	71.82	65.91	65.15	72.85	2.07 (1.41×)
	70%	64.34	81.83	1600.83	87.08	68.35	62.56	63.08	71.50	2.19 (1.50×)
	90%	60.20	77.21	1488.16	83.01	67.23	56.91	53.53	69.41	2.25 (1.54×)
G-Prune	50%	65.25	82.54	1623.27	87.76	71.91	66.15	65.17	73.53	2.13 (1.45×)
	70%	64.37	81.91	1604.86	87.69	70.19	63.74	63.87	72.58	2.26 (1.54×)
	90%	61.40	77.51	1456.14	84.49	67.27	58.59	59.31	71.74	2.31 (1.58×)
<i>Multi-resolution Pruning Method</i>										
MR-Pruner	50%	65.31	82.62	1597.31	87.91	72.16	65.98	64.76	73.87	2.24 (1.53×)
	70%	64.88	82.10	1595.79	87.80	70.96	65.12	64.13	73.62	2.35 (1.60×)
	90%	62.32	78.47	1530.87	85.97	68.04	60.14	60.90	72.68	2.52 (1.72×)

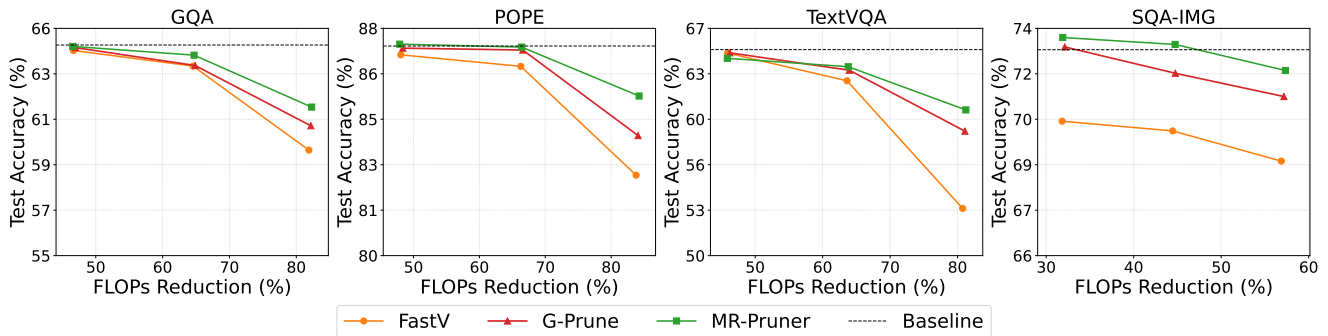


Figure 4. Performance comparison of MR-Pruner and other pruning methods under different FLOPs reductions.

hallucination, MMB-EN [18] and MMB-CN [18] for multi-lingual multi-modal understanding in English and Chinese, TextVQA [23] for text-oriented VQA, and SQA-IMG [19] for science question answering with image-based reasoning. These datasets cover diverse tasks, allowing us to comprehensively validate the robustness of MR-Pruner. All experiments are conducted the default settings and evaluation metrics of each dataset to ensure a fair comparison.

4.2. Implementation Details

We propose MR-Pruner as a plug-in module that can be seamlessly integrated into pre-trained MLLMs without any

additional training. As our method is designed for multi-resolution MLLMs, we employ the LLaVA-NeXT-8B [17] model. To ensure fair comparison and implementation, we employ LMMs-Eval [30], a widely-used toolkit that enables unified evaluation of MLLMs across diverse datasets and models. For the specific experimental setup, we set the threshold s_a for constructing the Intra-Resolution Token Graph to 0.5 and the threshold s_c for constructing the Cross-Resolution Token Graph to 0.5. The balancing factor α , which regulates the contribution of the cross-information token score, is set to 0.3. All experiments are conducted on a single NVIDIA H100 80G GPU.

Table 2. Comparison with existing methods in extreme pruning scenario (95% pruning ratio).

Method	GQA	MMB-EN	MME	TextVQA
FastV	55.10	59.01	1301.74	49.10
G-Prune	56.42	59.11	1299.95	52.15
MR-Pruner	56.81	62.29	1312.11	54.47

Table 3. Effect of each component in MR-Pruner.

IR	CR	IA	GQA	POPE	TextVQA	SQA-IMG
✗	✗	✗	60.87	82.21	43.26	70.94
✗	✓	✓	61.08	83.94	54.15	71.36
✓	✗	✓	61.90	85.11	60.21	72.23
✓	✓	✗	61.26	84.47	59.53	71.94
✓	✓	✓	62.32	85.97	60.90	72.68

4.3. Experimental Results

Main results. Table 1 presents a comparative evaluation of MR-Pruner against four existing pruning methods—Random, ToMe, FastV, and G-Prune—on LLaVA-NeXT-8B under pruning ratios of 50%, 70%, and 90%. Across all methods, including our own, pruning 50% or 70% of visual tokens retains competitive performance despite a substantial reduction in input token size. However, performance deteriorates considerably as the pruning ratio increases. ToMe and FastV generally outperform random pruning across most benchmarks. In particular, FastV achieves the best result among all methods on the TextVQA dataset, with only a 0.26% accuracy drop under the 50% pruning setting. This suggests that its pruning strategy based on attention scores is highly effective in preserving task-relevant tokens. G-Prune, on the other hand, shows the strongest performance on MME at 50% and 70% pruning as well as on MMB-CN at 50% pruning. We attribute this to its iterative information propagation, which enables more accurate identification of salient tokens. MR-Pruner consistently achieves the best results across most datasets and pruning levels. Notably, it attains 73.87 and 73.62 on SQA-IMG at 50% and 70% pruning, respectively, and preserves robust performance on more challenging benchmarks such as MME and MMB-CN at 90%, reaching 1530.87 and 60.14. These results demonstrate that the Cross-resolution Token Scoring and Informativeness-aware Token Pruning effectively safeguard performance in multi-resolution scenarios. MR-Pruner also achieves the highest throughput among existing pruning methods, excluding random pruning. We attribute this advantage to its ability to reduce the temporal overhead of the iterative information propagation required by G-Prune, while efficiently exploiting the mutual complementarity across different resolutions at lower cost.

Figure 4 shows the performance changes of the baseline model and our method with respect to FLOPs reduc-



Figure 5. Visualization of the effect of Cross-resolution Token Scoring. The left shows the intra-resolution token scores, the middle presents the final scores obtained by combining intra- and cross-resolution token scores, and the right highlights the difference between the final and intra-resolution scores.

tion across four benchmark datasets. As expected, all methods exhibit performance degradation as the pruning ratio increases. However, the trend of degradation varies significantly across methods. FastV shows the steepest decline, particularly on TextVQA where accuracy drops below 53% at 90% pruning. G-Prune demonstrates more robust performance, consistently outperforming FastV across all benchmarks, especially at higher pruning levels. Notably, MR-Pruner achieves the best results, maintaining accuracy closest to the baseline across different pruning ratios. These results indicate that the proposed method provides strong resilience to performance degradation under reduced FLOPs.

Extreme pruning scenario. We further evaluate all methods under an extreme pruning ratio of 95%, as summarized in Table 2. In this challenging scenario, both FastV and G-Prune suffer from severe performance drops across benchmarks. For instance, FastV records only 49.10 on TextVQA and 55.10 on GQA, while G-Prune achieves a slight improvement with 52.15 and 56.42, respectively. In contrast, MR-Pruner consistently exhibits the best performance in all datasets. In particular, MR-Pruner demonstrates stronger performance retention than existing methods on MMB-EN, MME, and TextVQA. Although a performance drop is inevitable when only 5% of the visual tokens are retained, the results confirm that our method removes tokens more effectively than previous approaches. These results highlight the robustness of our multi-resolution pruning strategy, demonstrating its ability to preserve critical visual information even when only a small fraction of tokens is retained.

Ablation study. As shown in Table 3, we evaluate the contribution of each component: Intra-resolution Token Scoring (IR), Cross-resolution Token Scoring (CR), and Informativeness-aware Token Pruning (IA). Specifically, removing IR means that token importance within the same resolution is measured solely by the L2-Norm without intra-resolution propagation. Removing CR means eliminating cross-resolution propagation, so that no information is transferred across different resolutions and pruning is performed solely based on token importance within the same resolution. Finally, removing IA implies that pruning ratios

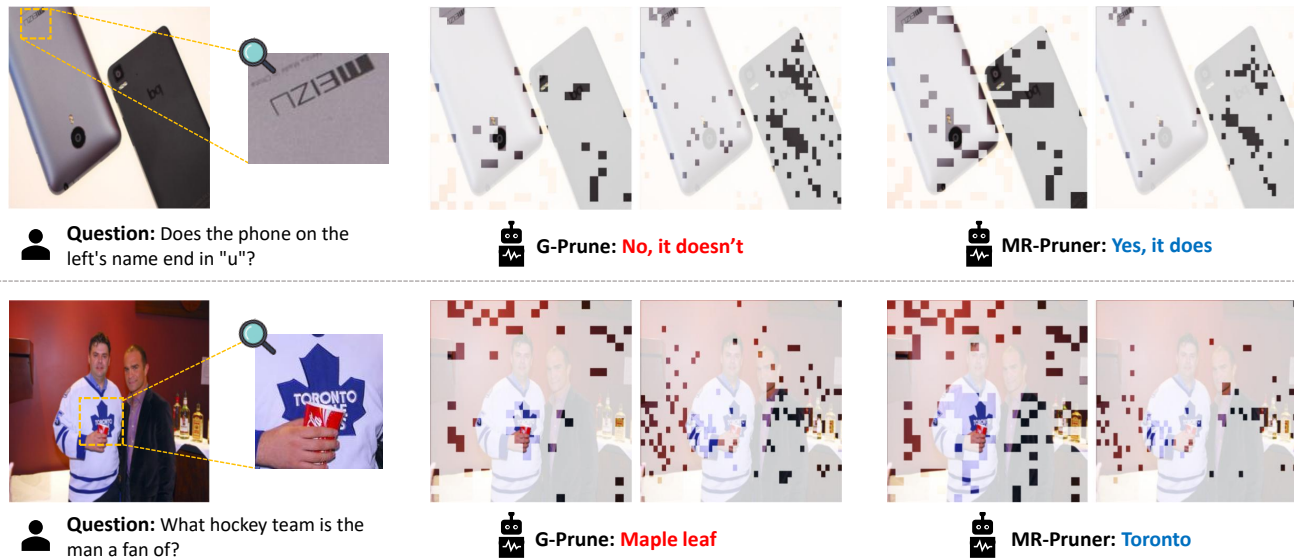


Figure 6. Examples of token pruning results of G-Pruner and MR-Pruner. The left side shows the pruning results of low-resolution tokens, while the right side presents the pruning results of high-resolution tokens.

are no longer adaptively adjusted according to informativeness across resolutions, but are instead applied uniformly. When all proposed components are removed, we obtain the worst performance, indicating that relying solely on the L2-norm of tokens is insufficient for effective pruning. Furthermore, performance degradation is also evident when either IR or CR is removed, suggesting that information propagation within the same resolution as well as across different resolutions is essential. Specifically, we observe that propagating information from pruned high-resolution tokens is beneficial for selecting informative low-resolution tokens. Removing IA similarly results in performance drops, demonstrating that appropriate pruning ratios across tokens from different resolutions are crucial for maintaining performance. Finally, the best results are achieved when all proposed components are combined, confirming the effectiveness of our method in multi-resolution settings.

Visualization. As shown in Figure 5, we qualitatively evaluate the effectiveness of Cross-resolution Token Scoring through real-world examples, illustrating its impact in practice. To answer the given question, tokens corresponding to the word (“data”) must be preserved. The intra-resolution token score for the region is not low but does not rank the highest. When the cross-resolution token score is incorporated, the final score of this region increases noticeably. To make this effect clearer, we visualize the score difference between the two settings, where the region exhibits a significant increase. These results indicate that Cross-resolution Token Scoring effectively guides the pruning process by retaining critical tokens.

Case study. To further illustrate the effectiveness of MR-Pruner, we present qualitative examples comparing it with G-Pruner. As shown in Figure 6, both G-Pruner and MR-Pruner prune certain high-resolution tokens that are necessary for answering the question. However, G-Pruner also removes the corresponding low-resolution tokens, leading the model to fail in producing the correct answer. In contrast, MR-Pruner preserves these regions in the low-resolution tokens, enabling the MLLMs to correctly respond. This result highlights the effectiveness of Cross-resolution Token Scoring in leveraging the mutual complementarity between tokens of different resolutions in multi-resolution settings.

5. Conclusions

In this work, we present MR-Pruner, a novel training-free, graph-based token pruning framework for multi-resolution MLLMs. Unlike prior single-resolution methods, MR-Pruner accounts for both the distinct informativeness and complementarity of tokens across resolutions. By combining Intra- and Cross-resolution Token Scoring with Informativeness-aware Pruning, it adaptively allocates pruning ratios and propagates information to preserve critical visual content. Experiments on eight benchmark datasets show that MR-Pruner achieves superior efficiency–performance trade-offs. Under extreme pruning scenarios and qualitative analyses, it also proves more robust and effective than existing methods. For future work, we aim to extend MR-Pruner to more than two visual resolutions and develop robust strategies for extreme pruning scenarios (e.g., 95%) in highly constrained environments.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-25435830).

References

- [1] 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. 3
- [2] 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*, 1(2):3, 2023. 1, 3
- [3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *Proceedings of International Conference on Learning Representations*, 2023. 1, 3
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 3
- [5] 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 *Proceedings of European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2, 3
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 1, 3
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 1
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations*, 2021. 3
- [9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, King Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 5
- [10] 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. 5
- [11] 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. 5
- [12] Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4075–4083, 2025. 2, 3
- [13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3
- [14] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 5
- [15] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *Proceedings of International Conference on Learning Representations*, 2022. 1, 3
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [17] 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. 1, 2, 3, 6
- [18] 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 *Proceedings of European conference on computer vision*, pages 216–233. Springer, 2024. 2, 6
- [19] 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. 6
- [20] Yiwei Ma, Zhibin Wang, Xiaoshuai Sun, Weihuang Lin, Qiang Zhou, Jiayi Ji, and Rongrong Ji. Inf-llava: Dual-perspective perception for high-resolution multimodal large language model. *arXiv preprint arXiv:2407.16198*, 2024. 1
- [21] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances*

- in *neural information processing systems*, 34:13937–13949, 2021. 2, 3
- [22] 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. 3
- [23] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2, 6
- [24] 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. 1
- [25] 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. 1, 3
- [26] 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. 1
- [27] Hongjie Wang, Bhishma Dedhia, and Niraj K Jha. Zero-prune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16070–16079, 2024. 3
- [28] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2964–2972, 2022. 2, 3
- [29] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwa403, 2024. 1
- [30] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics*, 2025. 6
- [31] Yipeng Zhang, Yifan Liu, Zonghao Guo, Yidan Zhang, Xuesong Yang, Chi Chen, Jun Song, Bo Zheng, Yuan Yao, Zhiyuan Liu, et al. Llava-uhd v2: an mllm integrating high-resolution feature pyramid via hierarchical window transformer. *arXiv e-prints*, pages arXiv–2412, 2024. 1
- [32] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of International Conference on Learning Representations*, 2024. 1, 3