Lifting the Veil on Visual Information Flow in MLLMs: Unlocking Pathways to Faster Inference

Hao Yin Gunagzong Si Zilei Wang*
University of Science and Technology of China
{yinhnavi, guangzongsi}@mail.ustc.edu.cn, zlwang@ustc.edu.cn

Abstract

Multimodal large language models (MLLMs) improve performance on vision-language tasks by integrating visual features from pre-trained vision encoders into large language models (LLMs). However, how MLLMs process and utilize visual information remains unclear. In this paper, a shift in the dominant flow of visual information is uncovered: (1) in shallow layers, strong interactions are observed between image tokens and instruction tokens, where most visual information is injected into instruction tokens to form cross-modal semantic representations; (2) in deeper layers, image tokens primarily interact with each other, aggregating the remaining visual information to optimize semantic representations within visual modality. Based on these insights, we propose Hierarchical Modality-Aware Pruning (HiMAP), a plug-and-play inference acceleration method that dynamically prunes image tokens at specific layers, reducing computational costs by approximately 65% without sacrificing performance. Our findings offer a new understanding of visual information processing in MLLMs and provide a state-of-the-art solution for efficient inference. The code is available at `https://github.com/ustc-hyin/HiMAP`.

1. Introduction

Multimodal large language models (MLLMs) [3, 4, 16, 22, 45] have emerged as an advanced architecture that integrates visual and textual information, demonstrating exceptional performance across various tasks. Compared to traditional multimodal models [14, 15, 17], MLLMs achieve superior information fusion and complex semantic understanding by utilizing large language models (LLMs) [1, 37, 44] to process visual features. However, the mechanisms of information interaction within these models remain underexplored. This study poses two critical questions: (1) To what extent do image tokens influence model predictions? (2) How is visual information processed within the model?

With respect to the first question, we developed three metrics based on saliency scores to quantify the impact of system tokens, image tokens, and instruction tokens on prediction outcomes. Experimental results indicated that the importance of image tokens was minimal, only equivalent to 0.03% of that of instruction tokens, despite image tokens comprising a significant portion of the model input.

With respect to the second question, saliency analysis of the attention matrices reveals strong interactions between image tokens and instruction tokens in shallow layers, while interactions among image tokens become more significant in deeper layers. This result intuitively reveals that as the model depth increases, the dominant flow of visual information within MLLMs undergoes a shift. Based on this, we propose the following hypothesis.

Phased Processing of Visual Information

$\mathcal{H}_1$: *In shallow layers, image tokens primarily interact with instruction tokens, injecting most visual information into instruction tokens to establish a cross-modal semantic representation for subsequent computations.*

$\mathcal{H}_2$: *In deeper layers, interactions among image tokens are enhanced, consolidating the residual visual information, thereby refining the semantic representation within visual modality.*

Fig. 1 provides a detailed elaboration of our hypothesis. Two experiments were conducted to validate the aforementioned hypothesis. (1) By blocking the information interaction between image and instruction tokens in specific layers, we observed that perturbations in shallow layers significantly degraded model performance, confirming that image tokens inject most visual information into instruction tokens. (2) We compared the significance of *visual-textual* and *intra-visual information flows* at various model depths, discovering that perturbations to *intra-visual information flow* in deeper layers led to more pronounced prediction deviations, thereby

*Corresponding Author

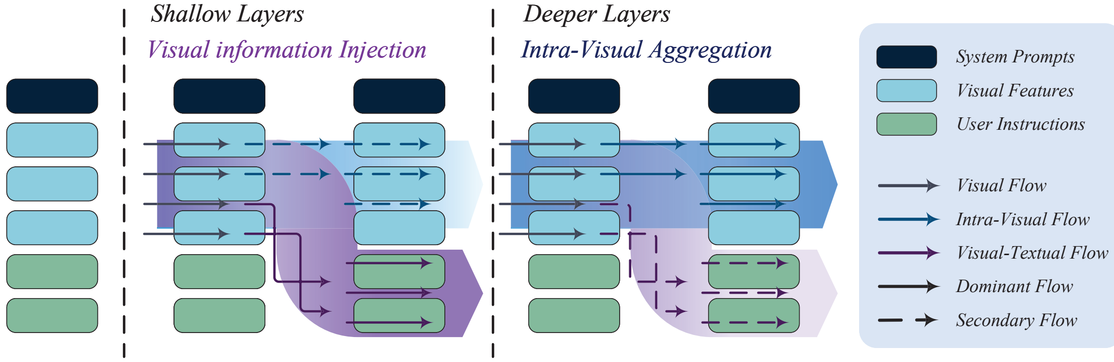


Figure 1. Illustration of our hypothesis. In shallow layers, image tokens inject most of the visual information into instruction tokens, establishing a cross-modal semantic representation for subsequent computations. In deeper layers, image tokens aggregate the residual visual information, refining the semantic representation within the visual modality.

validating the interaction among image tokens to aggregate the residual visual information. These results support our hypothesis, indicating that MLLMs process visual information differently at varying depths.

Despite their substantial computational cost, image tokens contribute minimally to prediction results. To address this issue, we propose a method for pruning image tokens to accelerate inference. Based on insights into internal information interactions within MLLMs, we introduce Hierarchical Modality-Aware Pruning (HiMAP), a plug-and-play technique that effectively streamlines the computational process by focusing the model on the most influential image tokens. HiMAP dynamically ranks the importance of image tokens according to the dominant visual information flow at different depths and applies pruning strategies in specified layers. By reducing the computational overhead of both self-attention modules and feed-forward networks modules, HiMAP reduce FLOPs by over 65%. Experimental results demonstrate that HiMAP can reduce inference latency by about 50% while maintaining model performance.

In summary, our contributions are fourfold: (1) Conducting an in-depth analysis of the phenomenon where image tokens have minimal impact on prediction outcomes in MLLMs; (2) Identifying latent patterns in the interactions between visual and textual modalities within MLLMs; (3) Introducing HiMAP, a plug-and-play technique that reduces inference latency in MLLMs while maintaining performance; (4) Validating the efficacy of HiMAP across a diverse range of vision-language tasks.

2. Inefficient Contribution of image tokens

This section aims to highlight the limited contribution of image tokens to model predictions. Section 2.1 outlines the three token categories used as inputs in MLLMs, along with their respective processing mechanisms. Section 2.2 evaluates the impact of various modalities on prediction outcomes through metrics derived from saliency scores. The results of the quantitative analysis reveal that the visual modality contributes substantially less than other modalities.

2.1. Preliminaries

This section introduces how MLLMs process different tokens when generating output. Typically, these models follow a transformer decoder architecture [38], predicting responses autoregressively [5] based on a given image-question pair.

Before being fed into the transformer decoder, multimodal information (including images and text) is converted into sequence embeddings. For images, a common approach involves extracting visual features using pre-trained encoders, such as CLIP-VIT [31]. To align the dimensions of these visual features with the embedding size of LLMs and ensure semantic consistency, additional linear transformations or cross-attention modules are introduced. For text, natural language is tokenized into discrete units, and corresponding text embeddings are generated through embedding lookup. In this paper, ”image tokens” and ”text tokens” refer to both the discrete units of visual and textual data as well as the embeddings derived from them.

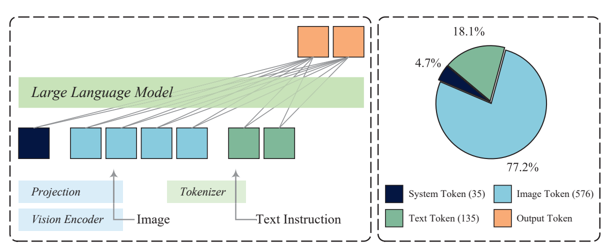


Figure 2. Distribution of Input Sequence Tokens. Image tokens constitute 77% of the total input tokens, nearly double the combined total of system and instruction tokens. This highlights a considerable computational overhead associated with image tokens.

After preprocessing the image and text tokens into a unified embedding space, these tokens are input into the transformer decoder to generate output tokens. During this decoding process, the input tokens are categorized into three types: (1) **system prompts**, which provide general information for controlling the behavior of MLLMs; (2) **image tokens**, derived from features learned by pre-trained visual encoders; and (3) **user instructions**, which specify requests or questions related to the given images. The index sets of system, image, and instruction tokens are denoted by $\mathcal{S}$, $\mathcal{V}$, and $\mathcal{I}$, respectively. Comprehensive index set of all input tokens is represented as $\mathcal{X}$, where $\mathcal{X} = \mathcal{S} \cup \mathcal{V} \cup \mathcal{I}$.

Fig. 2 illustrates the workflow of MLLMs and the distribution of sequence lengths for three types of input tokens. The sequence length for image tokens is 576, which is nearly twice the combined length of the system and instruction tokens. This suggests that the computational load associated with image tokens in MLLMs is significantly higher.

2.2. Visual Modality Impact Assessment

This subsection quantitatively evaluates the impact of the visual modality on prediction outcomes. We employ the saliency technique [34], a widely used interpretability tool, to highlight key token interactions within the attention mechanism. Following established practices, we utilize Taylor expansion [27] to compute saliency scores for each element of the attention matrix:

$$I_l = \left| \sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|. \tag{1}$$

Here, $A_{h,l}$ represents the attention matrix value for the h-th attention head in the l-th layer, x denotes the input, and $\mathcal{L}(x)$ is the loss function of the task, e.g., the cross-entropy objective for question-answering tasks. The saliency matrix I_l for the l-th layer is obtained by averaging across all attention heads. The significance of information flow from the j-th token to the i-th token in MLLMs is represented by $I_l(i,j)$. To illustrate the contributions of different modalities to prediction outcomes, three quantitative metrics based on I_l are introduced as following.

S_{sys}, which measures the importance of information flow from system tokens to other tokens:

$$S_{sys} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{S}} I_l(i,j); \tag{2}$$

S_{img}, which measures the importance of information flow from image tokens to other tokens:

$$S_{img} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{V}} I_l(i,j); \tag{3}$$

S_{ins}, which measures the importance of information flow from instruction tokens to other tokens:

$$S_{ins} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{X}} \sum_{j \in \mathcal{I}} I_l(i,j). \tag{4}$$

These three metrics enable a systematic observation of the information flow intensity from different modalities across various layers, facilitating the evaluation of their contributions to prediction outcomes.

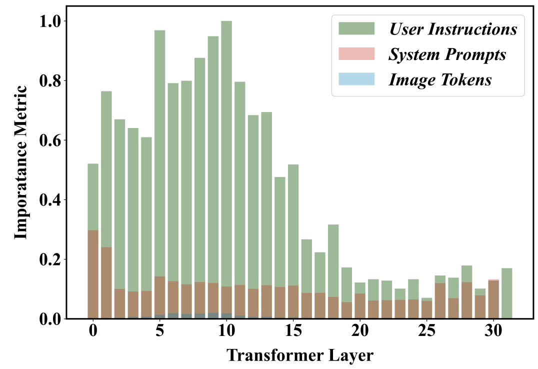


Figure 3. Contributions of different modalities to prediction outcomes across layers. The contribution of visual modality is significantly lower than textual modality.

Our experiments with the LLaVA-v1.5-7B model on A-OKVQA and Sci-VQA datasets revealed interesting insights into the impact of different modalities on prediction outcomes. As shown in Fig. 3, instruction tokens play a pivotal role in shaping predictions, exerting the most significant influence. Conversely, image tokens demonstrate a comparatively minor impact. Considering prior findings on the sparsity of attention to image tokens [23, 28, 33, 36], this limited impact may stem from the redundancy inherent in the image signals provided to the model. Additional experimental results are available in Sec. 8.

3. Shift in dominant flow of visual information

This section provides a detailed analysis of how MLLMs process visual information. In Sec. 3.1, two importance metrics are introduced to intuitively characterize the flow of visual information within MLLMs. The quantitative results support the following hypotheses:

- $\mathcal{H}_1$: *In shallow layers, image tokens inject visual information into instruction tokens, facilitating cross-modal semantic representations for subsequent computations.*
- $\mathcal{H}_2$: *In deeper layers, image tokens consolidate residual visual information, refining the semantic representation within the visual modality.*

In Sec. 3.2 and Sec. 3.3, these hypotheses are validated through information flow perturbation experiments.

3.1. Hypothesis Driven by saliency Scores

This subsection seeks to uncover the underlying patterns of visual information interaction through attention mechanism in MLLMs. We continue to use $I_l(i,j)$ from Equation (1) to represent the significance of information flow from the

j-th token to the i-th token. To clarify the visual information flow in MLLMs, we introduce two new quantitative metrics based on $I_l(i,j)$, with a particular focus on the information interaction involving image tokens. The metrics are defined as follows.

S_{vv}, measuring the importance of information flow among image tokens:

$$S_{vv} = \frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}}\sum_{i\in\mathcal{V}} I_l(i,j). \tag{5}$$

S_{vt}, measuring the importance of information flow from image tokens to instruction tokens:

$$S_{vt} = \frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{I}}\sum_{i\in\mathcal{V}} I_l(i,j). \tag{6}$$

S_{vv} and S_{vt} are utilized to analyze the mechanisms of visual information processing in MLLMs. Specifically, S_{vt} quantifies the extent of information injection from image tokens to instruction tokens, whereas S_{vv} measures the degree of information aggregation among image tokens. We define attention interactions among image tokens as *intra-visual information flow* and those between image and instruction tokens as *visual-textual information flow*.

Results and Analysis. We conducted experiments using the LLaVA-v1.5-7B models on the A-OKVQA and Sci-VQA datasets. As shown in Fig. 4, the significance of two information flows changes rapidly across model depths: (1) In the shallow layers (i.e., layers 1–3), the importance of the *visual-textual information flow* (S_{vt}) is substantially higher than that of the *intra-visual information flow* (S_{vv}). (2) Conversely, in the deeper layers (i.e., layers 8–16), the *intra-visual information flow* (S_{vv}) becomes predominant. Additional experimental results are detailed in Sec. 9.

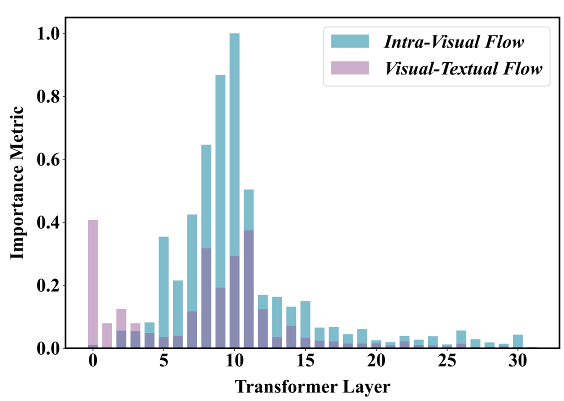


Figure 4. Importance of *intra-visual flow* and *visual-textual flow* across layers. Dominant flow of visual information shifts as model depth increases.

Proposed Hypothesis. Based on the observed shifts in dominant visual information flows, we hypothesize a phased processing mechanism for visual information in MLLMs. In the shallow layers, image tokens predominantly interact with instruction tokens, channeling the majority of visual information into these tokens to establish a cross-modal semantic representation. In the deeper layers, interactions among image tokens intensify, consolidating residual visual information to refine the semantic representation within the visual modality. This hypothesis is illustrated in Fig. 1.

3.2. Shallow Layers: Visual Information Injection

In this section, we validate the first part of our hypothesis. We propose that injecting visual information into instruction tokens depends on the the information flow from image tokens to instruction tokens, facilitated by the attention mechanism. By manipulating attention layers and disrupting the *visual-textual information flow*, we aim to confirm the presence of this injection process and its effect on predictions.

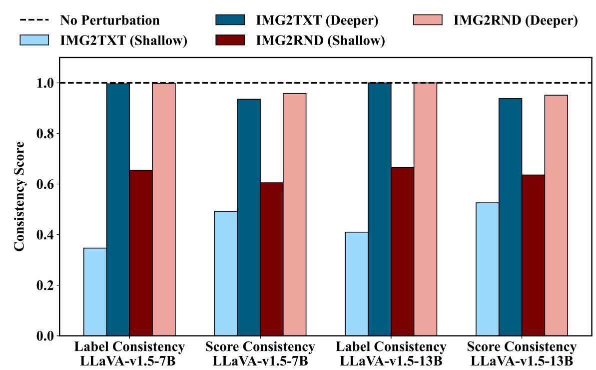


Figure 5. Disrupting *visual-textual flow* versus disrupting *visual-random flow* within the first or last 5 layers. Disrupting *visual-textual flow* in the first 5 layers has the most substantial effect, highlighting shallow-layers information injection from image tokens to instruction tokens.

Implementation Details. To disrupt the *visual-textual information flow*, we block the interaction between image and instruction tokens by modifying the attention matrix A. Specifically, we set $A_l(i,j)$ to 0 for $i \in \mathcal{I}$ and $j \in \mathcal{V}$ in the attention matrix A_l of the l-th layer. This modification prevents the instruction tokens from receiving information from the image tokens in the l-th layer.

Evaluation Metrics. Inspired by the loyalty metrics [40], we design the following metrics to assess the impact of disrupting *visual-textual information flow*. (1) **Label Consistency**: evaluates how consistent the prediction outcomes are before and after disruption. (2) **Score Consistency**: applies the Jaccard similarity to compare the top-5 predicted tokens before and after disruption, capturing broader changes in prediction results. A lower consistency score indicates a greater impact on prediction outcomes.

Results and Analysis. We conducted experiments with the LLaVA-v1.5-7B models on Sci-VQA and A-OKVQA datasets. As illustrated in Fig. 5, performance dropped significantly when disruptions occurred in the first five layers, but this effect diminished with increasing network depth. Conversely, disrupting the flow of information from image tokens to random tokens had only a minor impact on performance. These findings confirm that visual information is integrated into instruction tokens in the shallow layers.

3.3. Deeper Layers: Intra-Visual Aggregation

This section further validates the second part of the hypothesis, which posits that in deeper layers, enhanced interactions between image tokens lead to aggregation of residual visual information. To investigate this, we manipulated the attention layers to separately disrupt *intra-visual* and *visual-textual information flows*. By comparing the effects of these two disruptions on prediction outcomes, we confirm changes in the underlying visual processing mechanisms.

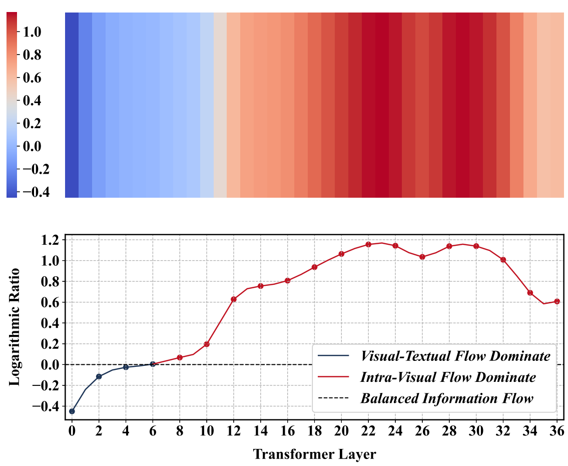


Figure 6. The values of D_l for every two layers in LLaVA-v1.5-13B. In deeper layers, $D_l > 0$, indicating that disruptions in intra-visual flow lead to greater prediction biases, thus validating the aggregation of residual visual information through interactions between image tokens.

Implementation Details. We modified the attention matrix A to block interactions between image tokens, thereby disrupting *intra-visual information flow*. Specifically, we set $A_l(i, j)$ to 0 for $i, j \in \mathcal{V}$ in the attention matrix A_l of the l-th layer, thereby preventing information interactions among image tokens within that layer. The disruption in *visual-textual information flow* is consistent with the procedure described in Sec. 3.2.

Evaluation Metrics. The prediction biases resulting from disruptions in *visual-textual* and *intra-visual information flows* are denoted as $E_{vt,l}$ and $E_{vv,l}$, respectively, where l refers to the l-th layer. To quantify the relative impact of these two disruptions on prediction outcomes, we introduce Bias Ratio D_l:

$$D_l = \log(E_{vv,l}/E_{vt,l}). \tag{7}$$

This metric represents the logarithmic ratio of prediction biases caused by two distinct disruptions in visual information flow. When $D_l > 0$, it indicates that *intra-visual information flow* dominates in the l-th layer. Conversely, when $D_l < 0$, it suggests that *visual-textual information flow* prevails in the l-th layer. Additional experimental details are available in Sec. 10 and Sec. 11.

Results and Analysis We conducted experiments using the LLaVA-v1.5-13B model on the Sci-VQA and A-OKVQA datasets. As shown in Fig. 6, the results, averaged across both datasets, reveal that in the deeper layers, D_l approaches 1.2, signifying the aggregation of residual visual information. Conversely, in the shallow layers, D_l drops to -0.5, indicating the integration of visual information from image tokens into instruction tokens.

3.4. Hypothesis Discussion

In Sec. 3.2, we establish that image tokens predominantly transfer visual information to instruction tokens in the shallow layers. In contrast, as detailed in Sec. 3.3, image tokens aggregate residual visual information in the deeper layers. Furthermore, Sec. 2.2 highlights the inefficiency of image tokens in influencing prediction outcomes. Based on these findings, *we propose that not all image tokens are necessary at every layer*. Specifically, once the majority of visual information has been injected into instruction tokens in the shallow layers, many image tokens lose their significance. This explains why the *visual-textual information flow* has minimal influence on prediction outcomes in the subsequent layers. Similarly, in the deeper layers, after aggregating the residual visual information, most remaining image tokens become redundant.

4. Hierarchical Modality-Aware Pruning

Given the relatively minor contribution of image tokens to prediction outcomes, coupled with their substantial computational cost in MLLMs, we propose a dynamic pruning method for image tokens. This approach effectively reduces computational overhead during inference without compromising model performance.

4.1. Hierarchical image token Pruning

Leveraging the phased processing mechanism of visual information in MLLMs, we propose Hierarchical Modality-Aware Pruning (HiMAP), a technique designed to accelerate inference by dynamically pruning image tokens at varying network depths.

Fig. 7 illustrates the overall framework of HiMAP, which includes two core components: shallow-layer pruning module and deeper-layer pruning module. Each module features an importance ranking function f_ϕ and two parameters: the filtering layer K and the filtering ratio $R\%$. At the K-th layer of MLLMs, the ranking function f_ϕ accepts a set of image tokens as input and ranks them according to a predefined importance criterion ϕ. After ranking, image tokens deemed to have the lowest importance in the bottom $R\%$ are pruned in subsequent layers, thus optimizing the utilization of computational resources.

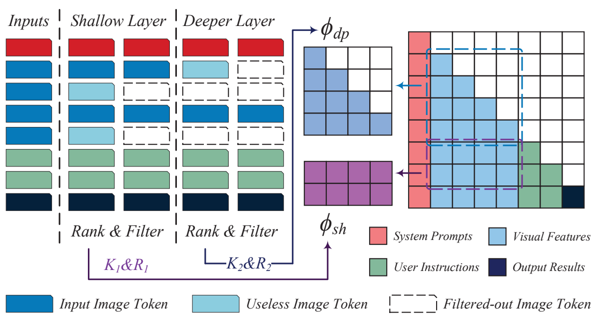


Figure 7. Illustration of Hierarchical Modality-Aware Pruning (HiMAP). In shallow layers, HiMAP ranks image tokens at the K_1-th layer based on the importance criterion ϕ_{sh}, removing the image tokens in the bottom $R_1\%$. In deeper layers, HiMAP ranks the remaining image tokens at the K_2 layer according to the importance criterion ϕ_{dp}, filtering out those in the bottom $R_2\%$.

In shallow layers, image tokens primarily interact with instruction tokens, injecting most visual information into instruction tokens. Consequently, we define the importance criterion ϕ_{sh} in shallow-layer pruning module as the sum of attention scores from the given image token v to all instruction tokens, represented as:

$$\phi_{sh}(v) = \sum_{i \in \mathcal{I}} A_{K_1}(i, v). \tag{8}$$

Here, K_1 denotes the filtering layer in shallow-layer pruning module. This criterion quantifies the influence of image tokens on instruction tokens, thereby guiding the pruning of image tokens in shallow layers.

In deeper layers, interactions among image tokens are enhanced, consolidating the residual visual information. Thus, we define the importance criterion ϕ_{dp} in deeper-layer pruning module as the sum of attention scores from all other image tokens to the given image token v, expressed as:

$$\phi_{dp}(v) = \sum_{i \in \mathcal{V}} A_{K_2}(v, i) \tag{9}$$

Here, K_2 denotes the filtering layer in deeper-layer pruning module. This criterion evaluates the information interaction between image tokens, directing the pruning of image tokens in deeper layers.

4.2. Computation Cost Estimation

We estimate the computational cost of the multi-head attention (MHA) and feed-forward neural network (FFN) modules in terms of Floating Point Operations Per Second (FLOPs). For a transformer layer, given that the input contains n image tokens, the hidden layer dimension is d, and the intermediate layer dimension of the FFN is m, the FLOPs for this layer can be represented as $\Omega(n) = 4nd^2 + 2n^2d + 2ndm$.

For the entire model, assuming there are L layers in total, the shallow-layer pruning module reduces the number of image tokens from n to $n_1 = (1 - R_1\%) \cdot n$ at the K_1-th layer. The deeper-layer pruning module further reduces the number of image tokens to $n_2 = (1 - R_2\%) \cdot n_1$. Thus, the theoretical FLOPs reduction rate η related to image tokens can be calculated using the following formula:

$$\begin{aligned} \eta = 1 - &\frac{K_1 \cdot \Omega(n) + (K_2 - K_1) \cdot \Omega(n_1)}{L \cdot \Omega(n)} \\ &- \frac{(L - K_2) \cdot \Omega(n_2)}{L \cdot \Omega(n)} \end{aligned} \tag{10}$$

5. Experiments

In this section, we showcase HiMAP's performance across various visual language benchmarks. Sec. 5.1 provides an overview of the evaluation tasks and parameter configurations. Sec. 5.2 highlights HiMAP's exceptional results in terms of both inference accuracy and speed. Lastly, Sec. 5.3 presents ablation studies that validate the effectiveness of HiMAP's individual components.

5.1. Experimental Setup

The performance of the HiMAP method is rigorously examined across four essential tasks: short-answer QA, multiple-choice QA, image captioning, and natural QA. These tasks encompass structured question-answering and open-ended generation, offering a comprehensive evaluation of MLLMs utilizing HiMAP method.

- **Short-answer QA:** This task utilizes datasets such as VQAv2 [12], TextVQA [35], MME [10], and POPE [18] to assess the model's ability to generate concise answers from multimodal information.
- **Multiple-choice QA:** Using ScienceQA [24] and A-OKVQA [32] datasets, this task tests reasoning and decision-making in multiple-choice scenarios.
- **Image Captioning:** Nocaps [2] and Flickr30k [29, 30] datasets are employed to measure the model's ability to generate natural language descriptions of images. Performance is evaluated using the CIDEr score [39].
- **Natural QA:** This task involves experiments with the LLaVA-Bench [22] and MM-Vet [43] datasets. GPT-4 [1]

Model	Method	TFLOPs	FLOPs Ratio	VQAv2	T-VQA	POPE	MME	S-VQA	A-OKVQA
LLaVA-7B	Baseline	2.98	100%	**78.3**	58.2	**86.4**	**1749.9**	67.9	76.6
	FastV	1.56	54%	78.1	**58.3**	84.9	1742.6	**68.1**	**77**
	HiMAP	**0.73**	**24%**	**78.6**	**58.4**	**86.2**	**1785.1**	**68.3**	**77.2**
LLaVA-13B	Baseline	5.81	100%	79.8	61.4	**87.2**	1794.4	**71.6**	**82**
	FastV	3.09	53%	**79.9**	**61.4**	84.8	**1796.3**	71.3	81.3
	HiMAP	**1.36**	**23%**	**80.2**	**61.7**	**86.5**	**1809.4**	**72.1**	**81.4**
QwenVL-7B	Baseline	3.6	100%	78.4	**60.8**	**84.5**	**1782.6**	68	75.7
	FastV	1.9	53%	**78.5**	58.3	82.7	1767.2	**68.2**	**75.3**
	HiMAP	**0.89**	**25%**	**78.8**	**61.3**	**83.7**	**1798.3**	**68.5**	**75.9**
InternVL-7B	Baseline	2.71	100%	**79.4**	**57.1**	**86.9**	**1812.2**	**70.9**	**79.6**
	FastV	1.39	52%	79	56.8	85.2	1802.1	70.5	79.1
	HiMAP	**0.56**	**20%**	**79.6**	**57.1**	**86.5**	**1821.3**	**71.7**	**80.1**

Table 1. **Performance of HiMAP on Short-answer QA and Multiple-choice QA Tasks**. The parameters are set as $K_1 = 2, R_1 = 50\%$, $K_2 = 8, R_2 = 75\%$. The evaluation metric used is **Accuracy**. The **highest** score for each configuration is highlighted in **red**, while the **lowest** computational cost is marked in **green**.

is used to assess the correctness of the generated answers, evaluating the model's capability to handle open-ended question-answering in complex multimodal scenarios.

Experiments were done with mainstream MLLMs, including LLaVA-v1.5-7B [21], LLaVA-v1.5-13B, QwenVL-Chat-7B [4], and InternVL-v1.0-7B [7, 8]. For structured question-answering tasks such as Short-answer QA and Multiple-choice QA, HiMAP adopts an aggressive parameter configuration ($K_1 = 2, R_1 = 50\%, K_2 = 8, R_2 = 75\%$) to effectively prune irrelevant image tokens, significantly enhancing inference speed. Conversely, for open-ended generation tasks like Image Captioning and Natural QA, HiMAP applies a more conservative parameter configuration ($K_1 = 2, R_1 = 50\%, K_2 = 15, R_2 = 75\%$) to ensure that the model captures comprehensive image information, leading to high-quality responses.

We primarily compared our method to FastV [6], which calculates the average attention score received by a token from all other tokens as its importance criterion, denoted as ϕ_{attn}. For a fair comparison, we adopted the optimal settings recommended for FastV: $K = 2$ and $R = 50\%$ for LLaVA-v1.5-7B, and $K = 3$ and $R = 50\%$ for LLaVA-v1.5-13B.

5.2. Results and Analysis

As shown in Sec. 4.2, HiMAP achieves outstanding accuracy in both Short-answer QA and Multiple-choice QA tasks. In **Short-answer QA** task, HiMAP surpasses baseline accuracy on VQAv2, TextVQA, and MME datasets, demonstrating its superior precision and stability in pruning image tokens. Even on the POPE benchmark, where performance declines for both HiMAP and FastV, HiMAP exhibits a notably smaller impact.

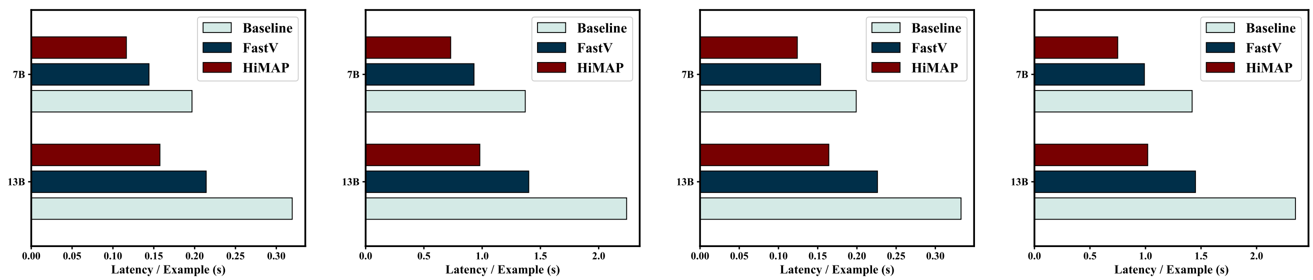
Similarly, in **Multiple-choice QA** task, HiMAP outperforms the baseline. By contrast, applying FastV to the InternVL-v1.0-7B and LLaVA-v1.5-13B models results in significant performance degradation, further underscoring HiMAP's advantages. Additional experimental results for **Short-answer QA** tasks can be found in Sec. 12.1 and Sec. 12.2.

Model	Method	Ratio	Nocaps	Flickr30k
LLaVA-7B	Baseline	100%	**78.8**	**50.9**
	FastV	54%	78.6	50.6
	HiMAP	**34%**	**78.7**	**51.3**
LLaVA-13B	Baseline	100%	**82.8**	**53.6**
	FastV	53%	81.9	53.1
	HiMAP	**33%**	**83.7**	**53.8**

Table 2. **Performance of HiMAP on Image Captioning task**. The parameters for HiMAP are set as $K_1 = 2, R_1 = 50\%, K_2 = 15$, and $R_2 = 75\%$. The evaluation metric used is the **CIDEr** score. The **highest** score for each configuration is highlighted in **red**, while the **lowest** computational cost is marked in **green**.

As shown in Sec. 5.2, HiMAP performs comparably to the baseline in **Image Captioning** task, whereas FastV results in a noticeable decline in model performance. This clearly demonstrates that HiMAP is better at preserving the model's ability to generate long-text descriptions compared to FastV.

Tab. 3 presents the experimental results of HiMAP and FastV on **Natural QA** task. Based on GPT-4's evaluation, HiMAP's responses significantly surpass those of FastV, even exceeding the baseline. This highlights HiMAP's ability to maintain robust open-domain QA performance in com-



(a) ScienceQA Dataset (b) A-OKVQA Dataset (c) Nocaps Dataset (d) Flickr30k Dataset

Figure 8. **Comparison of real-world inference speeds between HiMAP and FastV.** The experiment was conducted using LLaVA-v1.5 model family on a server equipped with a single 80GB A800 GPU.

plex scenarios. For case studies on LLaVA-Bench, please see Sec. 12.3.

Model	Method	Ratio	LV-BM	MM-Vet
LLaVA-7B	Baseline	100%	**65.7**	**33.4**
	FastV	54%	62.4	31.2
	HiMAP	**34%**	**66.5**	**33.7**
LLaVA-13B	Baseline	100%	**73.5**	**37.4**
	FastV	53%	71.7	35.5
	HiMAP	**33%**	**74.5**	**37.4**

Table 3. **Performance of HiMAP on Natural QA task.** The parameters for HiMAP are set as $K_1 = 2, R_1 = 50\%, K_2 = 15, R_2 = 75\%$. LV-BM refers to LLaVA-Bench dataset. The **highest** score for each configuration is highlighted in **red**, while the **lowest** computational cost is marked in **green**.

As shown in Sec. 4.2, for **Short-answer QA** and **Multiple-choice QA** tasks, HiMAP adopts aggressive parameter settings, specifically $K_1 = 2, R_1 = 50\%, K_2 = 8, R_2 = 75\%$. These settings achieve approximately a 75% reduction in FLOPs without sacrificing model performance, significantly decreasing computational overhead during inference.

For **Image Captioning** and **Natural QA** tasks, as illustrated in Sec. 5.2 and Tab. 3, HiMAP employs more conservative parameter configurations, namely $K_1 = 2, R_1 = 50\%, K_2 = 15, R_2 = 75\%$. Despite the more restrained settings, HiMAP still achieves a roughly 70% reduction in computational costs, significantly outperforming FastV's 45% reduction rate.

Fig. 8 compares the real-world inference speed of HiMAP and FastV. The experiment was conducted on a server equipped with a single 80GB A800 GPU. HiMAP achieved a 60% reduction in inference time, significantly surpassing FastV's 35% improvement. Details on GPU memory usage and more comprehensive evaluations of inference speed can be found in Sec. 12.4.

5.3. Ablation Study

We conducted ablation experiments using the LLaVA-v1.5-7B model, as summarized in Tab. 4. Pruning visual tokens in shallow layers based on intra-visual information flow caused a significant performance drop, confirming their role in integrating visual information with instruction tokens for cross-modal semantic representations. Similarly, pruning tokens in deeper layers, guided by visual-textual information flow, also degraded performance, supporting their function in refining semantic representations within the visual modality. For additional ablation experiments, please refer to Sec. 13.

SHL-PM	DPL-PM	Ratio	ScienceQA	A-OKVQA
✗	✗	**100%**	67.9	76.6
img2txt	✗	**54%**	68.3	77.1
img2img	✗	**54%**	<u>62.4</u>	71.7
img2txt	*img2txt*	**34%**	<u>67.1</u>	<u>76.2</u>
img2txt	*img2img*	**34%**	**68.3**	**77.2**

Table 4. **Ablation Study for the Importance Criteria**. **SHL-PM** and **DPL-PM** refer to the shallow-layer pruning module and deeper-layer pruning module, respectively. The **best** model performance is highlighted in **blue**, while degraded performance is underlined.

6. Conclusion

In this paper, we propose a hypothesis regarding visual information processing in MLLMs, suggesting that image tokens inject most visual information into instruction tokens in shallow layers while consolidate the remaining visual information in deeper layers. Results from information flow perturbation experiments confirm this hypothesis for the LLaVA-v1.5 series models. Building on these insights, we introduce Hierarchical Modality-Aware Pruning, a plug-and-play method that dynamically prunes image tokens at specific layers to improve inference speed. This method not only reaffirms our hypothesis but also demonstrates significant potential for practical applications.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62176246. This work is also supported by Anhui Province Key Research and Development Plan (202304a05020045), Anhui Province Natural Science Foundation (2208085UD17) and National Natural Science Foundation of China under Grant 62406098.

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. 1, 6

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 6

[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. 1

[4] 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. 1, 7

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[6] 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. *arXiv preprint arXiv:2403.06764*, 2024. 7

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 7

[8] 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. 7

[9] 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. 1

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 6

[11] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*, 2023. 1

[12] 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, 2017. 6

[13] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. 1

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[15] 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. 1

[16] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[17] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 1

[18] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[19] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. *arXiv preprint arXiv:2405.05803*, 2024. 1

[20] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023. 1

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 7

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 6

[23] Haogeng Liu, Quanzeng You, Xiaotian Han, Yongfei Liu, Huaibo Huang, Ran He, and Hongxia Yang. Visual anchors are strong information aggregators for multimodal large language model, 2024. 3

[24] Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 6

[25] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. 2

[26] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. 2

[27] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019. 3

[28] Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers, 2024. 3

[29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[30] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. 6

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[32] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 6

[33] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models, 2024. 3

[34] Karen Simonyan. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[35] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. 6

[36] Lin Song, Yukang Chen, Shuai Yang, Xiaohan Ding, Yixiao Ge, Ying-Cong Chen, and Ying Shan. Low-rank approximation for sparse attention in multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2024. 3

[37] 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

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[40] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023. 4, 1

[41] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*, 2024. 1

[42] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 1

[43] 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. *arXiv preprint arXiv:2308.02490*, 2023. 6

[44] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 1

[45] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1