

AVG-LLaVA: An Efficient Large Multimodal Model with Adaptive Visual Granularity

Zhibin Lan^{1,3*}, Liqiang Niu², Fandong Meng², Wenbo Li^{1,3}, Jie Zhou², Jinsong Su^{1,3,4†}

¹School of Informatics, Xiamen University, China,

²Pattern Recognition Center, WeChat AI, Tencent Inc, China,

³Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan (Xiamen University), Ministry of Culture and Tourism, China,

⁴Shanghai Artificial Intelligence Laboratory, China

lanzhibin@stu.xmu.edu.cn, jssu@xmu.edu.cn

Abstract

Recently, large multimodal models (LMMs) have achieved significant advancements. When dealing with high-resolution images, dominant LMMs typically divide them into multiple local images and a global image, leading to a large number of visual tokens. In this work, we introduce AVG-LLaVA, an LMM that can adaptively select the appropriate visual granularity based on the input image and instruction. Specifically, we first apply the multiple pooling layers to obtain visual tokens at different granularities. Then we propose a visual granularity router, which includes a Transformer layer, an MLP layer, and a voter layer, used to select the appropriate visual granularity based on the image and instruction. Furthermore, we put forward RGLF, a novel training paradigm that aims at aligning the granularity predicted by the router with the preferences of the LMM, without the need for additional manually annotated data. Extensive experiments and analysis show that AVG-LLaVA achieves superior performance across 11 benchmarks, as well as significantly reduces the number of visual tokens and speeds up inference (e.g., an 85.3% reduction in visual tokens and a $2.53\times$ increase in inference speed on the AI2D benchmark). Our code and model can be found at <https://github.com/DeepLearnXMU/AVG-LLaVA>.

1 Introduction

Recently, the field of artificial intelligence (AI) has witnessed a significant advancement in large multimodal models (LMMs) (OpenAI, 2023b; Zhu et al., 2023; Dai et al., 2023; Liu et al., 2023b, 2024a), marking a further step toward artificial general intelligence (AGI). Most existing LMMs follow the structure of LLaVA (Liu et al., 2023b, 2024a), which includes a vision encoder to embed

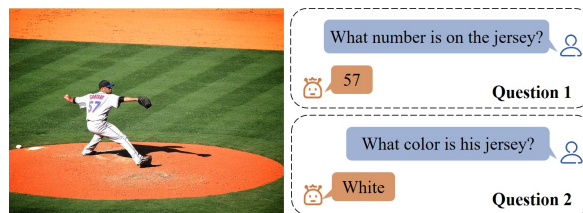


Figure 1: An example of VQA from MSCOCO (Lin et al., 2014). Notably, responding to Question 1 necessitates fine-grained visual information, whereas responding to Question 2 requires only coarse-grained visual information.

input images into visual tokens and a connector to map them into the word embedding space. Subsequently, these visual tokens are fed into a large language model (LLM) (Touvron et al., 2023; OpenAI, 2023a; Chiang et al., 2023) for multimodal understanding and reasoning (Li et al., 2023b; Liu et al., 2023a; Zhang et al., 2024; Lin et al., 2025), alongside the word embeddings.

Due to the limitations imposed by the fixed aspect ratio (e.g., 1:1) and low resolution (e.g., 336×336) used by visual encoders (e.g., CLIP-ViT (Radford et al., 2021)), early LMMs face challenges in processing high-resolution images with different aspect ratios. To deal with this limitation, dominant models, such as LLaVA-NeXT (Liu et al., 2024b), dynamically divide each input high-resolution image into multiple local images. These local images are encoded separately, and their tokens are then concatenated with the tokens of the original global image. This approach will lead to longer visual token sequences, such as 2880 visual tokens for a 672×672 image. However, in practice, such fine-grained visual information is not always necessary, and in some cases, coarse-grained visual information can even be more beneficial for model predictions. For instance, as shown in Figure 1, when the model is asked to recognize the number on the jersey, it requires relatively fine-grained visual information. In contrast, determining the color

* Work was done when Zhibin Lan was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

† Corresponding author.

of the jersey only necessitates coarse-grained visual information.

In this paper, we propose *Adaptive Visual Granularity LLaVA* (AVG-LLaVA), an LMM that can adaptively select the appropriate visual granularity based on the input image and instruction. The basic intuition behind our model is that humans only scrutinize images carefully when answering difficult questions; otherwise, a brief glance is sufficient.

As displayed in Figure 2, AVG-LLaVA extends LLaVA-NeXT with a *visual granularity scaler* and a *visual granularity router*. The visual granularity scaler performs multiple rounds of pooling on visual tokens, each time halving the number of visual tokens, thus obtaining a series of visual features with different granularities. The visual granularity router adaptively selects the appropriate visual granularity features based on the input multi-granularity visual features and text features. By doing so, for images and instructions that do not require fine-grained details, the number of visual tokens can be reduced, which not only speeds up inference but also may improve performance. This performance enhancement likely stems from the reduction of redundant information, as selecting appropriate visual granularity makes it easier for the model to answer questions based on images effectively.

Besides, we observe that it is challenging to train the visual granularity router directly through visual instruction tuning (Liu et al., 2023b). This may be because the router cannot learn the distinctions between different visual granularities from visual instruction tuning, making it difficult to learn how to select the most appropriate visual granularity based on the image and instruction. To deal with this issue, we propose a novel training paradigm called *Ranking Granularity* based on *LMM Feedback* (RGLF). This paradigm aligns router probabilities of multiple granularities with LMM preferences by a ranking loss (Hopkins and May, 2011; Liu et al., 2022), effectively aiding the router in distinguishing between different visual granularities and selecting the appropriate one.

We further evaluate AVG-LLaVA on 11 benchmarks including tasks from various types (e.g., general VQA and text-oriented VQA, etc.). Extensive experimental results show that AVG-LLaVA can effectively reduce the number of visual tokens and improve inference speed (e.g., an 85.3% reduction in visual tokens and a $2.53\times$ increase in inference

speed on the AI2D (Kembhavi et al., 2016) benchmark) while achieving better performance under the same base LLM.

2 Related Work

High-Resolution LMMs. Large language models (LLMs) such as GPT-4 (OpenAI, 2023a), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2023) have achieved significant success in language understanding and generation, driving the development of LMMs that integrate vision encoders with LLMs and leverage visual instruction data for fine-tuning. However, early LMMs (Li et al., 2023b; Zhang et al., 2023; Liu et al., 2023b) rely on fixed-resolution (e.g., 336×336) CLIP-ViT to process images, which limits their ability to capture high-resolution image details.

To perceive images with higher resolutions, Qwen-VL (Bai et al., 2023) increases the input resolution of the visual encoder to 448×448 and introduces an additional training stage. Along this line, both Vary (Wei et al., 2023) and Mini-Gemini (Li et al., 2024a) include two vision encoders: one is an additional introduced high-resolution vision encoder, and the other is the original low-resolution vision encoder. Unlike the methods mentioned above, SPHINX (Lin et al., 2023) and Monkey (Li et al., 2024b) enlarge the input image to a high resolution, and then divide it into a fixed number of local images, which are individually encoded using an image encoder to obtain local image tokens. Subsequently, the original global image tokens are concatenated with all local image tokens to feed into the LLM. Furthermore, LLaVA-NeXT (Liu et al., 2024b) enumerates various resolutions and adaptively selects the one that most closely matches the input image resolution. Although these methods can achieve better performance, they significantly increase the number of visual tokens, as the computational complexity scales quadratically with the number of input tokens, resulting in higher inference costs.

Vision Token Reduction for LMMs. Recently, several methods are proposed to reduce the visual tokens for LMMs, including visual token compression and pruning. For example, LLaVA-UHD (Guo et al., 2024) adopts a QFormer-like (Dai et al., 2023) structure to compress visual tokens, while MG-LLaVA (Zhao et al., 2024) employs a convolution layer to compress high-resolution visual features. In contrast, CrossGET (Shi et al., 2023) in-

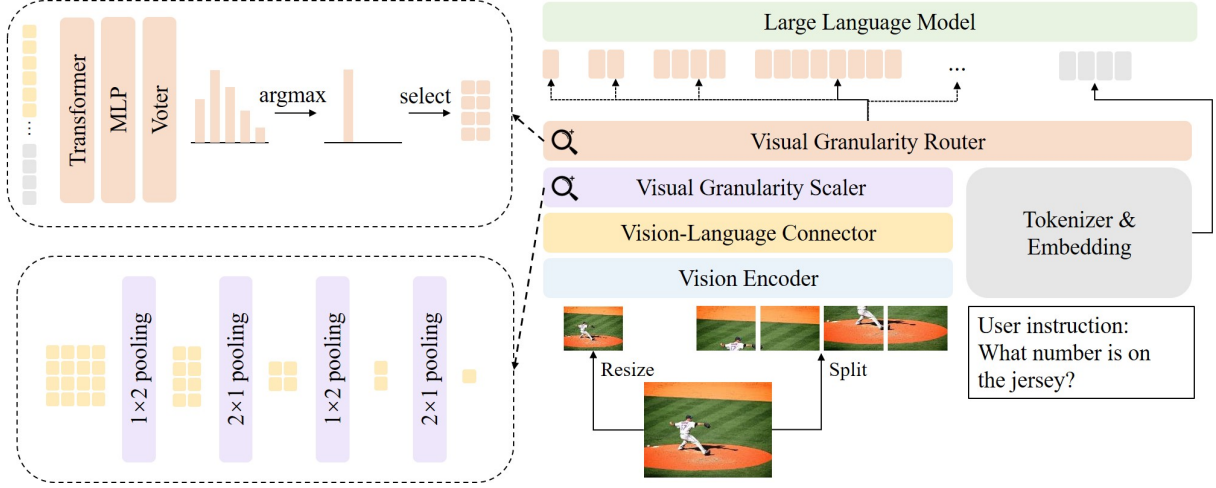


Figure 2: The architecture of AVG-LLaVA. AVG-LLaVA additionally introduces two modules based on LLaVA-NeXT: (1) Visual granularity scaler. This module consists of multiple pooling layers that progressively increase the granularity of visual features, thereby reducing the number of visual tokens; (2) Visual granularity router. This module includes a Transformer layer, an MLP layer, and a voter layer, which are used to select the appropriate granularity of visual features based on the input multi-granularity visual tokens and instruction tokens.

introduces a cross-modal token for leveraging cross-modal information to make decisions on token selection and merging. LLaVA-PruMerge (Shang et al., 2024) employs the similarity between the class token and other tokens as a key criterion for pruning and merging vision tokens.

Furthermore, FastV (Chen et al., 2024b) finds that most image tokens receive inefficient attention after the second decoder layer, and thus prunes half of the image tokens. Similarly, VTW (Lin et al., 2024b) adopts a more aggressive strategy to prune all visual tokens at a certain layer. Unfortunately, despite the above methods effectively reducing the number of visual tokens, they often lead to a certain degree of decline in model performance. More recently, LLaVA- M^3 (Cai et al., 2024) obtains multi-granularity visual features by merging visual tokens through pooling, enabling manual control of the tradeoff between inference cost and performance.

Significantly different from aforementioned methods, our model is a dynamic neural network (Han et al., 2022) that can adaptively select the appropriate visual granularity based on the input image and instruction, improving model performance while reducing the number of visual tokens.

3 Our Model

3.1 Model Architecture

As shown in Figure 2, in addition to the visual encoder, visual-language connector, and LLM, AVG-

LLaVA introduces two additional modules on top of LLaVA-NeXT: the visual granularity scaler and the visual granularity router. The key components will be elaborated in the following.

High-Resolution Image Encoding. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we follow common practice (Liu et al., 2024b) to divide it into multiple smaller local images $\mathbf{I}_{local} \in \mathbb{R}^{H_v \times W_v \times 3}$. Here, H_v and W_v are the resolution that the vision encoder is originally trained for. Then, these local images are individually encoded into a $H_p \times W_p$ grid of visual tokens $\mathbf{X}_{local} \in \mathbb{R}^{H_p \times W_p \times C}$ by the image encoder, where C is the dimension of the visual encoder. To preserve the global context information of the input image, we resize the original image to $H_v \times W_v$ and encode it as global visual tokens. Finally, we map both global and local visual tokens to the word embedding space through an MLP-based vision-language connector.

Visual Granularity Scaler. This module follows the design of spatial pyramid pooling (He et al., 2015; Cai et al., 2024), sequentially stacks 1×2 and 2×1 average pooling layers, thereby obtaining visual features at multiple granularities and preserving the spatial information. In this work, we consider CLIP-ViT-L-336 (Radford et al., 2021) as the visual encoder, and thus each image is encoded into 24×24 grid of visual tokens. Then, these visual tokens are fed into the visual granularity scaler, obtaining visual tokens with a grid of 24×12 , 12×12 , 12×6 and 6×6 , respectively. In this way, we can

obtain visual tokens of different granularities in a fine-to-coarse manner without training.

Visual Granularity Router. Different visual granularity features can be considered as different experts, so the Mixture of Experts (MoE) (Shazeer et al., 2017; Komatsuzaki et al., 2023; Lin et al., 2024a; Zhang et al., 2025) structure is particularly well-suited for selecting the appropriate visual granularity. Unlike the previous MoE studies that use linear layers as routers, we propose a multi-layer structure as illustrated in Figure 2 to select the appropriate visual granularity based on the input image and the instruction.

Specifically, when dealing with an image, we first flatten and concatenate its visual tokens of all granularities to form multi-granularity visual tokens $\bar{\mathbf{X}}_v = [\mathbf{X}_v^1; \mathbf{X}_v^2; \dots; \mathbf{X}_v^N]$, where $\mathbf{X}_v^i$ represents the visual tokens of the i -th granularity, and N is the number of visual granularities¹. Then, these visual tokens are concatenated with the filtered instruction tokens $\bar{\mathbf{X}}_{instruct}$ to serve as the input for the visual granularity router. Here, $\bar{\mathbf{X}}_{instruct}$ is obtained by calculating the cosine similarity between the original instruction tokens $\mathbf{X}_{instruct}$ and the visual tokens with original granularity $\mathbf{X}_v$, retaining the top- k most relevant ones. Afterwards, we apply a single Transformer (Vaswani et al., 2017) layer to facilitate the fusion of visual tokens at different granularities with instruction tokens. Subsequently, an MLP is applied to each token for predicting the appropriate visual granularity, resulting in the logits $\mathbf{Z}_{out} \in \mathbb{R}^{L \times N}$, where L is the number of both visual and instruction tokens. To vote for the most appropriate visual granularity, we use a learnable weight matrix (Voter) $\mathbf{W} \in \mathbb{R}^{1 \times L}$ to aggregate the logits predicted by all tokens, yielding the final logits $\mathbf{Z}_{final} \in \mathbb{R}^{1 \times N}$. Finally, we use softmax to calculate the probability distribution of visual granularities, where the visual tokens corresponding to the granularity with the highest probability are fed into the LLM.

3.2 Multi-stage Training

We provide a detailed description of the training procedures for AVG-LLaVA, which consists of two stages. The first stage endows the model with the

¹To simplify the explanation, we use a single image as an example. In practice, we include a global image and multiple local images, and each image will go through the following steps. The final result will be obtained by averaging the results of all the images.

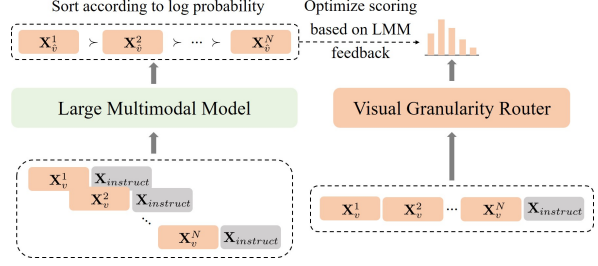


Figure 3: The overview of RGLF. Visual tokens of each granularity are concatenated with instruction tokens and then processed by the LMM to estimate the corresponding rewards. Visual granularity router optimizes the granularity selection based on the feedback from the LLM.

ability to perceive and process multi-granularity visual information, while the second stage enables the model to select the appropriate granularity based on the image and instructions.

Stage 1: Multi-Granularity Visual Instruction Tuning. In this stage, we use the high-quality visual instruction data to train the visual encoder, vision-language connector, and LLM, enabling them to perceive and process visual features of N different granularities. Specifically, we perform next-token prediction using visual features of different granularities and apply the cross-entropy loss only to the answering part, formulated as

$$\mathcal{L}_1 = -\frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \log P(x_t | \mathbf{X}_v^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}), \quad (1)$$

where $\mathbf{X}_a$ are the answer tokens before the current prediction token x_t , and T is the length of answer tokens.

Stage 2: Ranking Granularity Based on LLM Feedback. Then, we introduce the visual granularity router into the model training, where all other modules are frozen, and only the router is trained. This stage allows the model to select the appropriate visual granularity based on the input image and instruction. Intuitively, a straightforward approach to training the router is visual instruction fine-tuning. However, we find that the router trained with this method performs poorly. This could be due to the difficulty of visual instruction fine-tuning in effectively enabling the router to learn the differences between different visual granularities.

To address the above issue, we propose RGLF, as illustrated in Figure 3, where the router is trained

with a ranking loss, utilizing the feedback from the LMM fine-tuned with multi-granularity visual instructions as the ranking criterion. Concretely, for the given image and instructions, we let the LMM predict answers using visual tokens of different granularity and calculate their respective log probabilities. Then, based on these log probabilities, we sort $\mathbf{X}_v^1, \mathbf{X}_v^2, \dots, \mathbf{X}_v^N$ in a descending order to obtain $\mathbf{X}_{\hat{v}}^1, \mathbf{X}_{\hat{v}}^2, \dots, \mathbf{X}_{\hat{v}}^N$. Given the visual tokens $\mathbf{X}_{\hat{v}}^i$ of the i -th granularity, we directly consider those tokens $(\mathbf{X}_{\hat{v}}^1, \mathbf{X}_{\hat{v}}^2, \dots, \mathbf{X}_{\hat{v}}^{i-1})$ ranked above it as positive examples and the remaining tokens $(\mathbf{X}_{\hat{v}}^{i+1}, \mathbf{X}_{\hat{v}}^{i+2}, \dots, \mathbf{X}_{\hat{v}}^N)$ as negative ones. Afterwards, we use the router to give scores (log probability) s_i for each $\mathbf{X}_{\hat{v}}^i$:

$$s_i = \log P(g_i | \bar{\mathbf{X}}_v, \bar{\mathbf{X}}_{instruct}), \quad (2)$$

where g_i denotes the i -th granularity predicted by the router based on multi-granularity visual tokens $\bar{\mathbf{X}}_v$ and filtered instruction tokens $\bar{\mathbf{X}}_{instruct}$. Since we expect the router to assign higher probabilities to more appropriate visual granularities, the ranking loss is defined as follows:

$$\mathcal{L}_{rank} = \sum_{i=1} \sum_{j>i} \max(0, s_j - s_i + \lambda_{ij}), \quad (3)$$

where λ_{ij} is the log probability difference between the answers predicted by the LLM using visual tokens of the i -th and j -th granularities:

$$\lambda_{ij} = \frac{j-i}{|T|} \sum_{t=1}^T (\log P(x_t | \mathbf{X}_{\hat{v}}^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}) - \log P(x_t | \mathbf{X}_{\hat{v}}^j, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t})). \quad (4)$$

When the preference of $\mathbf{X}_{\hat{v}}^j$ is only slightly worse than $\mathbf{X}_{\hat{v}}^i$, the margin will be small. Conversely, when $\mathbf{X}_{\hat{v}}^j$ is significantly worse than $\mathbf{X}_{\hat{v}}^i$, the margin will correspondingly increase. In this way, we can dynamically adjust the margin to obtain adaptively penalty degrees between different pairs.

In addition to aligning with the LMM preference ranking, it is also desirable for the router to select the optimal visual granularity. Therefore, we add a cross-entropy loss to let the router learn the prediction of granularity with the highest log probability from the LMM, defined as follows:

$$k = \arg \max_i \sum_{t=1}^T \log P(x_t | \mathbf{X}_v^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}), \quad (5)$$

$$\mathcal{L}_{ce} = -\log P(g_k | \bar{\mathbf{X}}_v, \bar{\mathbf{X}}_{instruct}). \quad (6)$$

Finally, the total loss is defined as the weighted

sum of two losses:

$$\mathcal{L}_2 = \mathcal{L}_{rank} + \alpha \mathcal{L}_{ce}, \quad (7)$$

where α is the hyperparameter used to maintain the balance between the ranking loss $\mathcal{L}_{rank}$ and cross-entropy loss $\mathcal{L}_{ce}$.

4 Experiments

4.1 Settings

Training Datasets. In the first training stage, since the real user interaction data used for visual instruction fine-tuning in LLaVA-NeXT are not open-sourced, we opt to extract 200K samples from the ALLaVA (Chen et al., 2024a) dataset as a substitute. Although LLaVA-NeXT replaces TextVQA (Singh et al., 2019) with DocVQA (Mathew et al., 2021) and SynDog-EN (Kim et al., 2022), the TextVQA has already been included in the training data of most existing LMMs. Consequently, we choose to retain it to ensure a fair comparison with other models.² In total, the visual instruction fine-tuning data we use contains 1M image-text pairs.

Implementation Details. Note that in this work, we focus on investigating the effectiveness of adaptive visual granularity selection in reducing the number of visual tokens and improving model performance, rather than building a state-of-the-art model. Therefore, we use LLaVA-NeXT (Liu et al., 2024b) as the base LMM, where the visual encoder is CLIP ViT-L/14, and the LLM is Vicuna-7B (Chiang et al., 2023). We set the filtered instruction token number k to 32 and the cross-entropy loss weight α to 0.1.³ In the first stage, the learning rates for the visual encoder and other modules are set to 2×10^{-5} and 1×10^{-5} , respectively, with a batch size of 128. In the second stage, the learning rate for the visual granularity router is set to 1×10^{-3} , with a batch size of 128. More details of the training process are provided in Appendix A.3.

Evaluations. We evaluate our model on three kinds of benchmarks: (1) **general VQA benchmarks:** GQA (Hudson and Manning, 2019), SciQA-Img (Lu et al., 2022), and VizWiz (Gurari et al., 2018); (2) **text-oriented VQA benchmarks:** TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and

²Our data recipe follows Open-LLaVA-NeXT (Lin and Long, 2024).

³The impact of these two hyperparameters on model performance is discussed in Appendix A.2.

Model	LLM	General VQA			Text-oriented VQA			
		GQA	ScienceQA	VizWiz	TextVQA	ChartQA	DocVQA	AI2D
Standard-resolution LMMs								
InstructBLIP (Dai et al., 2023)	Vicuna-7B	49.2	60.5	34.5	-	-	-	-
IDEFICS-9B (Team, 2023)	LLaMA-7B	38.4	-	35.5	25.9	-	-	-
Qwen-VL (Bai et al., 2023)	Qwen-7B	59.3	67.1	35.2	63.8	<u>65.7</u>	65.1	62.3
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	57.5	68.2	38.9	61.6	66.3	62.6	57.7
InternVL-Chat (Chen et al., 2023)	Vicuna-7B	62.9	-	52.5	57.0	-	-	-
mPLUG-Owl2 (Ye et al., 2023)	LLaMA2-7B	56.1	68.7	54.5	58.2	-	-	-
MQT-LLaVA (Hu et al., 2024)	Vicuna-7B	61.6	67.6	53.1	-	-	-	-
LLaVA-1.5 (Liu et al., 2024a)	Vicuna-7B	62.0	66.8	50.0	58.2	-	-	-
High-resolution LMMs								
SPHINX-2k (Lin et al., 2023)	LLaMA2-7B	<u>63.1</u>	70.6	44.9	61.2	-	-	-
TextMonkey (Liu et al., 2024c)	Qwen-VL-7B	-	-	-	65.9	58.2	64.3	-
Mini-Gemini-HD (Li et al., 2024a)	Vicuna-7B	-	-	-	68.4	-	-	-
MG-LLaVA (Zhao et al., 2024)	Vicuna-7B	62.7	70.4	60.0	58.4	40.8	44.6	64.1
LLaVA-NeXT (Liu et al., 2024b)	Vicuna-7B	64.2	70.1	57.6	64.9	54.8	<u>74.4</u>	66.6
LLaVA-NeXT- M^3 (Cai et al., 2024)	Vicuna-7B	-	72.5	-	63.1	59.0	<u>72.6</u>	<u>66.7</u>
AVG-LLaVA	Vicuna-7B	63.0	<u>71.1</u>	<u>59.8</u>	<u>67.1</u>	66.3	74.6	67.3

Table 1: Comparison with LMMs of the same size on general VQA benchmarks and text-oriented VQA benchmarks. The best results are marked in bold, and the second best results are underlined. Since MG-LLaVA is trained on significantly more data across two stages, we retrain it using the same data as ours for a fair comparison. We also explore the impact of additional two-stage training on the performance of LLaVA-NeXT using the same instruction fine-tuning data in Appendix A.4.

AI2D (Kembhavi et al., 2016); and (3) **general multimodal benchmarks**: MME (Fu et al., 2023), MMB (Liu et al., 2023c), MMB^{CN} (Liu et al., 2023c), POPE (Li et al., 2023c), and MMMU (Yue et al., 2023).

4.2 Main Results

General VQA Benchmarks. The results in Table 1 show that AVG-LLaVA outperforms all standard-resolution LMMs on the general VQA benchmarks and achieves comparable performance to other high-resolution LMMs. Although it does not achieve the best results, it is important to note that AVG-LLaVA uses fewer visual tokens compared to other high-resolution models, and this comparison will be detailed in Section 4.3.

Text-oriented VQA Benchmarks. In this category of benchmarks, as illustrated in Table 1, except for TextVQA, AVG-LLaVA outperforms all other comparison models. Back to TextVQA, AVG-LLaVA achieves the second-best performance, only trailing behind Mini-Gemini-HD. Notably, Mini-Gemini-HD utilizes more than twice the amount of data during the pretraining and approximately 1.5 times the amount of data during the visual instruction fine-tuning compared to AVG-LLaVA.

General Multimodal Benchmarks. Compared to traditional VQA datasets, this type of bench-

marks cover a broader range of evaluation aspects, requiring models to possess more complex perception and reasoning capabilities. As summarized in Table 2, except for MME, AVG-LLaVA surpasses all other models across the remaining benchmarks, exhibiting superior overall performance and highlighting its adaptability and effectiveness across various disciplines. Specifically, AVG-LLaVA outperforms the second-best model by 6.1, 1.9, and 1.2 on MME^C , MMB, and MMB^{CN} , respectively. Moreover, AVG-LLaVA’s performance on the POPE and MMMU benchmarks demonstrates its ability to reduce hallucinations and perform complex reasoning.

4.3 Computational Efficiency

To validate the effectiveness of dynamic visual granularity selection, we compare AVG-LLaVA with LLaVA-NeXT in terms of visual token number and inference speed across multiple benchmarks. Specifically, for each type of benchmarks, we select three benchmarks for comparison, and report the reduction in the number of visual tokens per grid and the actual speedup during inference.

As shown in Table 3, except for text-intensive VQA benchmarks that require very fine-grained visual information, such as TextVQA and ChartVQA, AVG-LLaVA significantly reduces the number of visual tokens and improves inference speed across

Model	LLM	MME	MME ^C	MMB	MMB ^{C/N}	POPE	MMMU
<i>Standard-resolution LLMs</i>							
InstructBLIP (Dai et al., 2023)	Vicuna-7B	1084.0	229.0	-	-	-	30.6
Qwen-VL-Chat (Bai et al., 2023)	Qwen-7B	1487.6	<u>360.7</u>	60.6	-	-	-
InternVL-Chat (Chen et al., 2023)	Vicuna-7B	1525.1	-	-	-	86.4	-
mPLUG-Owl2 (Ye et al., 2023)	LLaMA2-7B	1450.2	-	64.5	-	-	-
MQT-LLaVA (Hu et al., 2024)	Vicuna-7B	1434.5	353.6	64.3	-	84.4	34.8
LLaVA-1.5 (Liu et al., 2023b)	Vicuna-7B	1510.7	-	64.3	58.3	<u>87.3</u>	-
<i>High-resolution LLMs</i>							
SPHINX-2k (Lin et al., 2023)	LLaMA2-7B	1470.6	326.8	65.9	-	87.2	-
OtterHD-8B (Li et al., 2023a)	Fuyu-8B	1223.4	331.4	58.3	-	86.0	-
Mini-Gemini-HD (Li et al., 2024a)	Vicuna-7B	1546.0	319.0	65.8	-	-	36.8
MG-LLaVA (Zhao et al., 2024)	Vicuna-7B	1561.1	325.4	67.4	48.4	86.9	35.3
LLaVA-NeXT (Liu et al., 2024b)	Vicuna-7B	1519.0	332.0	67.4	<u>60.6</u>	86.5	35.8
LLaVA-NeXT-M ³ (Cai et al., 2024)	Vicuna-7B	-	-	<u>68.0</u>	-	87.2	34.0
AVG-LLaVA	Vicuna-7B	<u>1557.4</u>	366.8	69.9	61.8	87.4	37.4

Table 2: Comparison with LLMs of the same size on general multimodal benchmarks.

Metric	General VQA			Text-oriented VQA			MLLM Benchmarks		
	GQA	ScienceQA	VizWiz	TextVQA	ChartQA	AI2D	MME	MMB	MMMU
Token Per Grid ↓	80.0%	26.4%	54.9%	92.3%	99.1%	14.7%	69.3%	30.0%	29.9%
Speed ↑	1.14×	1.77×	1.41×	1.04×	0.97×	2.53×	1.19×	1.87×	1.79×

Table 3: Comparisons of AVG-LLaVA and LLaVA-NeXT in terms of the number of visual tokens and actual inference speed, both of which are tested on 8 V100 GPUs with a batch size of 1. AVG-LLaVA can reduce the number of visual tokens by up to 85.3% and is up to $2.53\times$ faster than LLaVA-NeXT.

other benchmarks. Particularly, on the AI2D benchmark, AVG-LLaVA achieves better performance than LLaVA-NeXT while using only 14.7% of visual tokens, and the inference speed increases by $2.53\times$.⁴ Notably, even with the addition of two extra modules, there is no significant slowdown in inference speed on the ChartVQA benchmark when using a comparable number of visual tokens. Moreover, AVG-LLaVA only increases the number of parameters by 1.66% compared to LLaVA-NeXT.

4.4 Routing Visualization

To further understand the differences in the granularity selection of AVG-LLaVA across different benchmarks, we visualize the proportion of visual tokens selected at each granularity level for all benchmarks. Figure 4 shows the visualization results, it is evident that different tasks tend to favor different visual granularity, which is consistent with our expectations. In the case of text-intensive benchmarks like TextVQA, ChartQA, and DocVQA, the model requires fine-grained visual information, so the router predominantly selects the finest visual granularity. On the other hand, for benchmarks involving object-level questions, such

as AI2D and MMMU, the model may find it easier to answer correctly by utilizing coarse-grained visual information. Although the 72 and 288-token granularities are seldom selected, their inclusion helps the model progressively learn and differentiate between various levels of visual granularity (see the ablation study in Section 4.5).

4.5 Ablation Study

In order to validate the effectiveness of our designed modules and training paradigm, we conduct the following ablation analysis.

Adaptive Visual Granularity vs. Fixed Visual Granularity. We first delve into the proposed adaptive visual granularity router and report results in Table 4(a). It is clear that, compared to fixed visual granularity, adaptive visual granularity shows significant improvement on ScienceQA, MME, and MMB. It is worth noting that, in addition to performance improvement, adaptive visual granularity can also significantly reduce the number of visual tokens and increase the model’s inference speed, as reported in Section 4.3.

Router Granularity Selection vs. Random Granularity Selection. In Table 4(b), we replace the granularity selected by the router with randomly-

⁴We also present qualitative results in Appendix A.5 and illustrate the effectiveness of adaptive visual granularity.

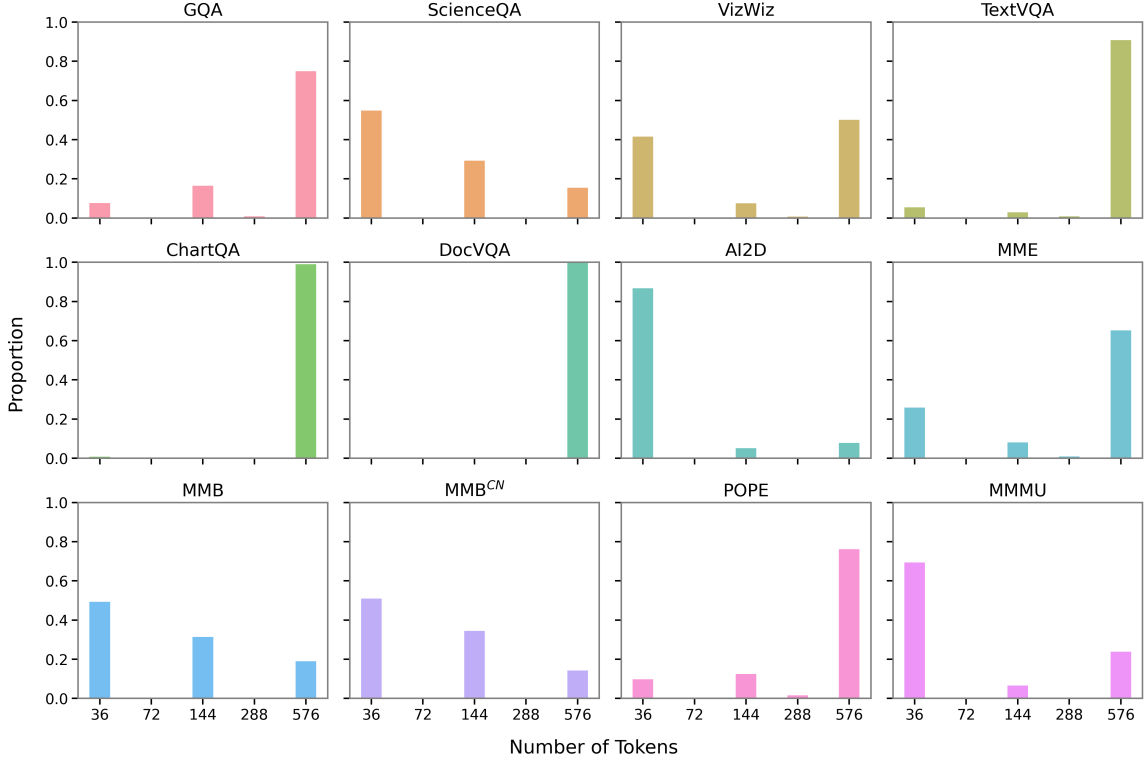


Figure 4: Visualization of the proportion for different granularity visual tokens.

Ablated Setting	Ablated Details	Original Value	→	Changed Value	ScienceQA	ChartQA	MME	MMB
AVG-LLaVA					71.1	66.3	1557.4	69.9
Architecture	(a) Visual Granularity	Adaptive		Fixed	70.0	66.4	1554.5	68.7
	(b) Granularity Selection	Router		Random	69.7	56.8	1535.7	67.9
	(c) Router Input	Image + Instruction		Image	70.1	53.9	1525.2	69.0
	(d) Granularity Range	{36, 72, 144, 288, 576}		{36, 144, 576}	69.8	65.3	1547.7	66.3
Training	(e) Router Training	Feedback from LMM		Visual Instruction Fine-tuning	70.5	50.9	1514.8	68.6
	(f) Ranking Loss	✓		✗	70.1	64.8	1534.6	68.6
	(g) Cross-entropy Loss	✓		✗	70.2	66.3	1550.8	69.4

Table 4: Ablation results on multiple benchmarks.

selected granularity. The results show that visual granularity router can indeed select a relatively appropriate granularity based on the input image and instruction, thereby significantly enhancing model performance.

Impact of Router Input. The instruction plays a crucial role in granularity selection. To validate this, we remove the instruction from the router input. As shown in Table 4(c), a clear performance degradation rises when solely using image as input (e.g., -12.4 on ChartQA), illustrating the importance of choosing granularity based on input image and instruction.

Impact of Granularity Range. In Section 4.4, we observe that granularities with 72 and 288 visual tokens are rarely selected, therefore we re-

move the visual tokens of these two granularities. As shown in Table 4(d), this change leads to a decrease in model performance, proving that introducing these granularities benefits the model’s progressive learning to utilize features of different visual granularities and distinguish among various visual granularities.

Impact of Router Training Methods. We directly train the router using visual instructions fine-tuning with the cross-entropy loss function. Unlike our original approach where the router is directly supervised by the LMM feedback, this variant computes the loss on the LMM and backpropagates the gradient to the router using the Gumbel-Softmax technique (Jang et al., 2017). The results in Table 4(e) show that the LLM feedback allows the router

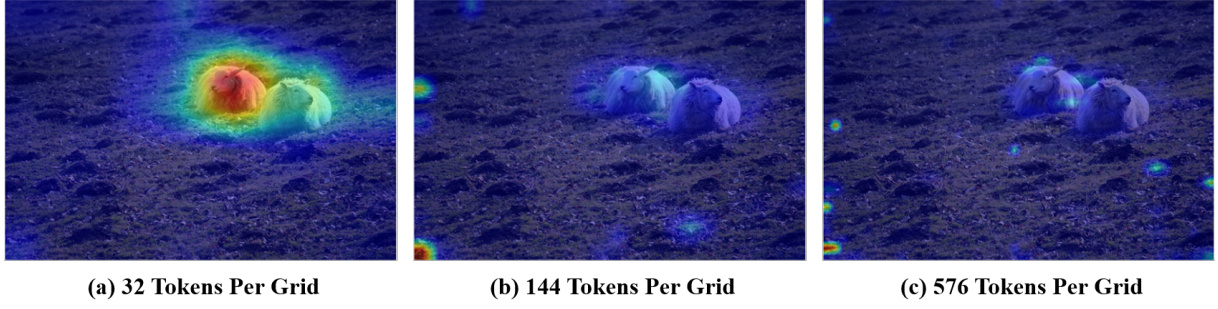


Figure 5: Attention maps of AVG-LLaVA at different visual granularities. The instruction is “*How many sheep are there? Answer the question with a single word.*”. The number of visual tokens per grid selected by the router is 32.

to better distinguish the advantages and disadvantages of different granularities, thereby enabling it to select an appropriate granularity.

Importance of Ranking Granularity. In Table 4(f) and Table 4(g), we remove the cross-entropy loss and ranking loss during the second stage, respectively. The results indicate that both types of loss are beneficial to model training and are complementary to each other, between which the ranking loss is more crucial. This underscores the necessity to train the router by ranking granularity based on LMM feedback.

4.6 Attention Map Visualization

To further understand how the appropriate granularity benefits the model in generating better answers, we visualize the attention map between the generated tokens and the visual tokens. The attention weights are calculated by accumulating the attention scores between image tokens and generated tokens across all layers and heads. As shown in Figure 5, when the instruction is “*How many sheep are there? Answer the question with a single word.*” the attention weights for the visual granularity selected by the router are mostly assigned to the two sheep, while the attention weights for other visual granularities are dispersed across the background. This means that selecting the appropriate visual granularity results in a more distinct attention map characterized by reduced background noise and enhanced focus on relevant regions, thereby improving model performance.

5 Conclusion

In this work, we propose AVG-LLaVA, an LMM that can adaptively select appropriate visual granularity based on input image and instruction. Besides, we introduce RGLF, which aligns router-predicted probabilities of multiple granularities

with LMM preferences by a ranking loss, effectively helping the model learn to distinguish between different granularities. Experimental results show that AVG-LLaVA not only exhibits superior performance across 11 benchmarks, but also significantly reduce the number of visual tokens and speed up inference in tasks that do not require fine-grained information. In future work, we aim to develop different visual granularity scaling networks to obtain richer visual granularity and integrate the two-stage training into a single stage to improve efficiency.

Limitations

While AVG-LLaVA has achieved good results, there is still considerable potential to be further explored. On text-intensive benchmarks, the model tends to select the finest-grained visual tokens, which may be due to the pooling directly reducing half of the tokens, resulting in significant differences in granularity size. Designing a more suitable granularity scaling network to provide richer visual granularities may help alleviate this issue. Besides, the two-stage training introduces additional overhead, which could be alleviated by interleaving multi-granularity visual instruction fine-tuning and router training within a single stage.

Acknowledgments

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *CoRR*.
- Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. [Matryoshka multimodal models](#). *CoRR*, abs/2405.17430.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. [Allava: Harnessing gpt4v-synthesized data for A lite vision-language model](#). *CoRR*, abs/2402.11684.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024b. [An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models](#). *CoRR*, abs/2403.06764.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. [Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). *CoRR*, abs/2312.14238.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. [MME: A comprehensive evaluation benchmark for multimodal large language models](#). *CoRR*, abs/2306.13394.
- Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zhanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. [Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images](#). In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, volume 15141 of *Lecture Notes in Computer Science*, pages 390–406. Springer.
- Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. [Vizwiz grand challenge: Answering visual questions from blind people](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3608–3617. Computer Vision Foundation / IEEE Computer Society.
- Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2022. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7436–7456.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(9):1904–1916.
- Mark Hopkins and Jonathan May. 2011. [Tuning as ranking](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1352–1362. ACL.
- Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. 2024. [Matryoshka query transformer for large vision-language models](#). *CoRR*, abs/2405.19315.
- Drew A. Hudson and Christopher D. Manning. 2019. [GQA: A new dataset for real-world visual reasoning and compositional question answering](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images](#). In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 235–251. Springer.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. [Ocr-free document understanding transformer](#). In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.
- Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie,

- Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. [Sparse upcycling: Training mixture-of-experts from dense checkpoints](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Bo Li, Peiyuan Zhang, Jingkan Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023a. [Otterhd: A high-resolution multi-modality model](#). *CoRR*, abs/2311.04219.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. [BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024a. [Mini-gemini: Mining the potential of multi-modality vision language models](#). *CoRR*, abs/2403.18814.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. [Monkey: Image resolution and text label are important things for large multi-modal models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773.
- Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024a. [Moe-llava: Mixture of experts for large vision-language models](#). *CoRR*, abs/2401.15947.
- Chen Lin and Xing Long. 2024. [Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community](#). <https://github.com/xiaoachen98/Open-LLaVA-NeXT>.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Yujie Lin, Ante Wang, Moye Chen, Jingyao Liu, Hao Liu, Jinsong Su, and Xinyan Xiao. 2025. [Investigating inference-time scaling for chain of multi-modal thought: A preliminary study](#). *arXiv preprint arXiv:2502.11514*.
- Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2024b. [Boosting multimodal large language models with visual tokens withdrawal for rapid inference](#). *CoRR*, abs/2405.05803.
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. [SPHINX: the joint mixing of weights, tasks, and visual embeddings for multi-modal large language models](#). *CoRR*, abs/2311.07575.
- Bingshuai Liu, Chenyang Lyu, Zijun Min, Zhanyu Wang, Jinsong Su, and Longyue Wang. 2023a. [Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models](#). *arXiv preprint arXiv:2312.01714*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Yixin Liu, Pengfei Liu, Dragomir R. Radev, and Graham Neubig. 2022. [BRIO: bringing order to abstractive summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2890–2903. Association for Computational Linguistics.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. [Mmbench: Is your multi-modal model an all-around player?](#) *CoRR*, abs/2307.06281.
- Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024c. [Textmonkey: An ocr-free large multimodal model for understanding document](#). *CoRR*, abs/2403.04473.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. [Chartqa: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. [Docvqa: A dataset for VQA on document images](#). In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.
- OpenAI. 2023a. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- OpenAI. 2023b. [Gpt-4v\(ision\) system card](#).
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. [Llava-prumerge: Adaptive token reduction for efficient large multimodal models](#). *CoRR*, abs/2403.15388.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Dachuan Shi, Chaofan Tao, Anyi Rao, Zhendong Yang, Chun Yuan, and Jiaqi Wang. 2023. [Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers](#). *CoRR*, abs/2305.17455.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards VQA models that can read](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- IDEFICS Research Team. 2023. [Introducing idefics: An open reproduction of state-of-the-art visual language model](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. [Vary: Scaling up the vision vocabulary for large vision-language models](#). *CoRR*, abs/2312.06109.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. [mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration](#). *CoRR*, abs/2311.04257.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2023. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). *CoRR*, abs/2311.16502.
- Liang Zhang, Ziyao Lu, Fandong Meng, Hui Li, Jie Zhou, and Jinsong Su. 2025. Advancing smoe for continuous domain adaptation of mllms: Adaptive router and domain-specific loss. In *Proceedings of ACL*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. [Llama-adapter: Efficient fine-tuning of language models with zero-init attention](#). *CoRR*, abs/2303.16199.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. [Multi-modal chain-of-thought reasoning in language models](#). *Trans. Mach. Learn. Res.*, 2024.
- Xiangyu Zhao, Xiangtai Li, Haodong Duan, Haian Huang, Yining Li, Kai Chen, and Hua Yang. 2024. [Mg-llava: Towards multi-granularity visual instruction tuning](#). *CoRR*, abs/2406.17770.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigpt-4: Enhancing vision-language understanding with advanced large language models](#). *CoRR*, abs/2304.10592.

A Appendix

A.1 Visual Granularity Selection Algorithm

In algorithm 1, we provide the detailed process of the router’s granularity selection.

A.2 Hyperparameter Analysis

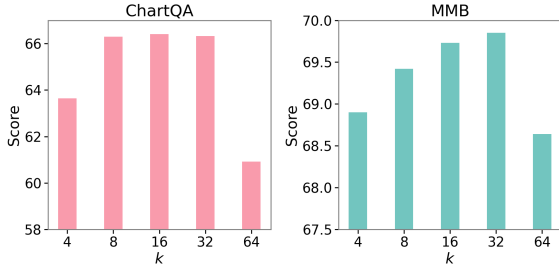


Figure 6: Influence of the filtered instruction token number k on model performance, measured on ChartQA and MMB benchmarks.

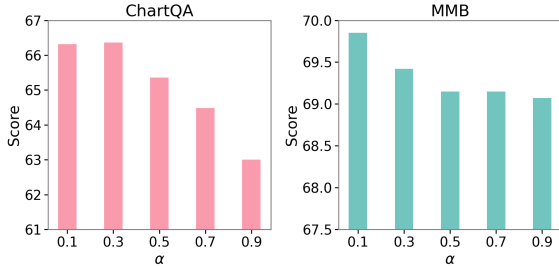


Figure 7: Influence of the cross-entropy loss weight α on model performance, measured on ChartQA and MMB benchmarks.

We experimentally explore the influence of the filtered instruction token number k and the cross-entropy loss weight α on model performance. As shown in Figure 6, the model performance is significantly affected when k is too small or too large. This may be due to the fact that too few instruction tokens provide insufficient text information, while too many tokens will introduce more noise. Figure 7 indicates that our approach is relatively robust to α and setting a smaller α is able to consistently enhance model performance, making our training method easy to apply.

A.3 Training Details

We list the training hyperparameters for two stages in Table 5. Our setup mainly refers to LLaVA-NeXT (Liu et al., 2024b).

A.4 Impact of Multiple Training Epochs on the Same Dataset

Since we conduct two additional training stages on the LLaVA-NeXT, we also investigate the impact

Hyperparameter	Stage 1	Stage 2
Data size	1M	1M
Batch size	128	128
Lr	1e-5	1e-3
Vision encoder Lr	2e-5	-
Lr schedule	cosine decay	
Lr warmup ratio	0.03	
Weight decay	0	
Epoch	1	
Optimizer	AdamW	
DeepSpeed stage	3	3
GPU	8 × H800	8 × H800
Training cost (#Hours)	65	14

Table 5: Training hyperparameters of AVG-LLaVA.

of training LLaVA-NeXT for two extra epochs on the same multimodal instruction-tuning data.

From Tables 6 and 7, we can observe that although three additional two epoch trainings result in improvements on 7 benchmarks (e.g., ChartQA and DocVQA), there is a considerable performance decline on 6 benchmarks (e.g., TextVQA and MMB^{CN}). This indicates that repeated training cannot improve the performance on all benchmarks. Besides, AVG-LLaVA performs better than LLaVA-NeXT-Extra on 9 benchmarks, is slightly worse on 2 benchmarks, and has a significant speed improvement, indicating that the advantage of AVG-LLaVA does not simply stem from repeated training.

A.5 Qualitative Evaluation

As shown in Figures 8 and 9, we compare the visual granularity selected by the router and other visual granularities. In Figure 8, given the user-provided image and instruction, AVG-LLaVA selects the coarsest visual granularity through the router. It can be observed that compared to other granularities, the model’s response with the coarsest granularity does not vary significantly. However, in Figure 9, with the given image and instruction, AVG-LLaVA selects the finest visual granularity. We find that coarser visual granularities could not generate a reasonably accurate poster description. These two examples demonstrate that AVG-LLaVA can adaptively select the appropriate visual granularity based on the image and instruction, thereby reducing the number of visual tokens, accelerating inference, and even improving model performance.

Figure 10 further shows several conversations between users and AVG-LLaVA.

Algorithm 1 Visual Granularity Selection Algorithm

Require: Multi-granularity visual tokens $\bar{\mathbf{X}}_v = [\mathbf{X}_v^1; \mathbf{X}_v^2; \dots; \mathbf{X}_v^N]$, Instruction tokens $\mathbf{X}_{instruct}$, Visual tokens of original granularity $\mathbf{X}_v$

Ensure: Selected granularity visual tokens $\mathbf{X}_v^{selected}$.

- 1: Obtain the filtered instruction tokens $\bar{\mathbf{X}}_{instruct} = \text{Top-}k(\text{cosine_sim}(\mathbf{X}_{instruct}, \mathbf{X}_v))$
 - 2: Concatenate $\bar{\mathbf{X}}_v$ and $\bar{\mathbf{X}}_{instruct}$ to form the input for the router
 - 3: Apply a Transformer layer to facilitate token fusion $\mathbf{Z}_{fusion} = \text{Transformer}([\bar{\mathbf{X}}_v; \bar{\mathbf{X}}_{instruct}])$.
 - 4: Use an MLP to predict logits for each token $\mathbf{Z}_{out} = \text{MLP}(\mathbf{Z}_{fusion})$
 - 5: Aggregate the logits using a learnable weight matrix $\mathbf{Z}_{final} = \mathbf{W}\mathbf{Z}_{out}$.
 - 6: Compute the probability distribution using softmax $\mathbf{P} = \text{softmax}(\mathbf{Z}_{final})$.
 - 7: Identify the granularity with the highest probability $selected = \arg \max(\mathbf{P})$.
 - 8: **return** $\mathbf{X}_v^{selected}$
-

Model	General VQA			Text-oriented VQA			
	GQA	ScienceQA	VizWiz	TextVQA	ChartQA	DocVQA	AI2D
LLaVA-NeXT	64.2	70.1	57.6	64.9	54.8	74.4	66.6
LLaVA-NeXT-Extra	64.6	69.9	58.3	63.9	66.3	75.1	65.3
AVG-LLaVA	63.0	71.1	59.8	67.1	66.3	74.6	67.3

Table 6: Results on general VQA benchmarks and text-oriented VQA benchmarks. LLaVA-NeXT-Extra refers to training for two extra epochs on the same multimodal instruction-tuning data.

A.6 Visualization of Router Selection for Different Instructions

As shown in Figure 11, we input the same image with different instructions and then visualize the selected visual granularity on the image, i.e., the number of patches. As can be seen, even for the same image, the router selects different visual granularities for different instructions. When asking about the color of the car, the model does not require such fine-grained visual information, whereas when asking whether there is a cat, the model requires finer-grained visual information.

Model	MME	MME ^C	MMB	MMB ^{CN}	POPE	MMMU
LLaVA-NeXT	1519.0	332.0	67.4	60.6	86.5	35.8
LLaVA-NeXT-Extra	1524.7	330.0	67.8	57.0	87.4	34.8
AVG-LLaVA	1557.4	366.8	69.9	61.8	87.4	37.4

Table 7: Results on general multimodal benchmarks.



User: Why this image is unusual?

AVG-LLaVA (36 selected by router): The image is unusual because it features a small white dog with wings, giving the appearance of a dog flying in the sky. This is not a natural occurrence, as dogs do not have the ability to fly. The image is likely a digitally manipulated or artistically created image, designed to evoke a sense of wonder and fantasy. The dog's wings and the fact that it is flying in the sky make the scene extraordinary and surreal, which is why it stands out as an unusual image.

AVG-LLaVA (72): The image is unusual because it features a dog with wings, which is not a natural occurrence. Dogs are mammals and do not have the biological ability to fly like birds. The image is likely a digitally manipulated or artistically created image, designed to evoke a sense of wonder or fantasy. It is important to note that while the image is visually striking, it does not represent reality and should not be taken as an actual depiction of a dog with wings.

AVG-LLaVA (144): The image is unusual because it features a small white dog with wings, which is not a natural occurrence. Dogs do not have the ability to fly, and it is highly unlikely to see a dog with wings in a real-life scenario. The image is likely a digitally manipulated or artistically created image, designed to evoke a sense of wonder, fantasy, or humor.

AVG-LLaVA (288): The image is unusual because it features a dog with wings, which is not a natural occurrence. Dogs do not have the ability to fly like birds or other animals with wings. This creates a whimsical and fantastical scene that is not typically seen in reality. The dog's wings are likely added digitally or through some form of artistic manipulation, making the image stand out as an imaginative and creative representation.

AVG-LLaVA (576): The image is unusual because it features a dog with wings, which is not a common sight. Dogs are typically known for their four legs and lack of wings, making this depiction of a dog with wings in the sky an imaginative and unexpected scene. The image could be a creative representation or an artistic interpretation, rather than a realistic portrayal of a dog.

Figure 8: Example of the finest visual granularity selected by the router compared to other visual granularities. The number in parentheses indicates the number of visual tokens. In this example, the responses from the model using both fine-grained and coarse-grained visual tokens show little difference, indicating that the model can adopt coarse visual granularity to significantly reduce the number of visual tokens and accelerate inference.



User: Please describe this poster in detail.

AVG-LLaVA (36): The image is a promotional poster for the anime "Imax 90", which is written in large, bold letters at the top of the image. The title is followed by the phrase "IMAX 90", indicating the name of the anime and possibly the episode or series title. The background of the poster is a vibrant yellow, which contrasts with the black and white illustration of a character. This character is depicted with a determined expression, holding a sword in their right hand and a shield in their left, suggesting a warrior or combatant theme. The text and illustration are arranged in such a way that the title and character are the focal points, drawing the viewer's attention to the central theme of the anime.

AVG-LLaVA (72): The image is a promotional poster for the anime "Maximilian", which is set to release on April 19th. The central figure in the poster is a young girl with pink hair, who is holding a megaphone. The text on the poster, written in Japanese, reads "IMAX MAXIMILIAN" and "4.19", indicating the title of the anime and its release date. The background of the poster is a vibrant yellow, adorned with black and white lines that add a dynamic visual element to the overall design. The girl's position and the megaphone suggest that she might be a character of significance in the anime, possibly a protagonist or a key figure. The use of bold colors and dynamic lines creates a sense of excitement and anticipation for the upcoming release.

AVG-LLaVA (144): The image is a vibrant promotional poster for the anime "Impossible Spy Family". The central figure is a young girl with pink hair, who is holding a megaphone. She is dressed in a red jacket and a white shirt, adding a pop of color to the scene. The background is a striking yellow, adorned with black text that reads "IMAX" and "SPY FAMILY", indicating the title of the anime. The text is written in a bold, sans-serif font, making it stand out against the yellow backdrop. The overall design of the poster is simple yet eye-catching, with the girl's bright hair and the bold text creating a dynamic visual appeal.

AVG-LLaVA (288): The image is a vibrant promotional poster for the anime "Maximum Spice!", which is set to premiere on April 19th. The central figure in the poster is a young girl with pink hair, who is holding a green microphone. She is dressed in a red jacket and a white shirt, adding a pop of color to the scene. The background is a bright yellow, adorned with Japanese text that reads "MAXIMUM SPICE!", "April 19th", and "スパイシーな家族", which translates to "Spicy Family". The text is written in a bold, black font, making it stand out against the yellow backdrop. The overall design of the poster suggests an exciting and spicy adventure is about to begin on April 19th.

AVG-LLaVA (576 selected by router): The image is a promotional poster for the movie "Spy x Family". The central figure is a young girl with pink hair, who is holding a green megaphone. She is dressed in a red jacket and is smiling, giving off a cheerful vibe. The background is a vibrant yellow, adorned with Japanese text and a black and white striped pattern. The text on the poster reads "IMAX", "SPY x FAMILY", "CODE: WHITE", and "EXPERIENCE IT IN IMAX", indicating that the movie is being promoted as an IMAX experience. The poster also includes the release date "APRIL 19", suggesting when the movie will be released. The overall design of the poster is eye-catching and colorful, likely intended to attract the attention of potential viewers.

Figure 9: Example of the coarsest visual granularity selected by the router compared to other visual granularities. The number in parentheses indicates the number of visual tokens. In this example, when coarse-grained visual tokens are used, the model generates incorrect descriptions. This suggests that the model should select fine visual granularity for the image and instructions in order to achieve better accuracy.



User: What is this place?

AVG-LLaVA: The place depicted in the image is likely a hotel, as indicated by the sign above the entrance. The architecture and the presence of the hotel sign suggest that this is a commercial area, possibly in a city or town center. The style of the building, with its brickwork and the design of the street, could imply that this location is in Europe, as such designs are common in many European cities. The overcast sky and the wet ground suggest that the photo was taken on a day with recent rainfall. The presence of the hotel, along with the parked cars and the residential-looking street, indicates that this area is a blend of commercial and possibly residential zones.



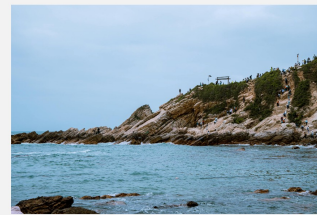
User: Why this image is funny?

AVG-LLaVA: The image is funny because it is an unexpected and absurd situation. Typically, one would not expect to see a llama, which is a large domesticated South American camelid, inside a car, especially sticking its head out of the window as if it were a human passenger. The juxtaposition of a llama behaving in a human-like manner in a typical urban setting, such as a traffic jam, is what makes the image humorous. It is a lighthearted and whimsical take on the unexpected and the anthropomorphism of animals, which often appeals to a sense of humor.



User: Tell me about this place.

AVG-LLaVA: The Tower Bridge is a combined bascule and suspension bridge in London, built between 1886 and 1894. It is a symbol of London and a popular tourist attraction. The bridge crosses the River Thames close to the Tower of London and has become an iconic symbol of London. The bridge is a bascule bridge, which means it can be raised to allow ships to pass. The suspension cables are anchored to the ground at either end of the bridge, and the towers are built on piers. The bridge is a testament to Victorian engineering and is still in operation today, although it is now primarily used for pedestrian traffic. The Tower Bridge is a testament to the engineering prowess of the Victorian era and has become a symbol of London's rich history and ongoing development.



User: What are the things I should be cautious about when I visit here?

AVG-LLaVA: When visiting this rocky coastline with a large group of people, there are several things to be cautious about. Firstly, the rocky terrain can be slippery and uneven, so it's essential to be mindful of your footing and wear appropriate footwear to avoid accidents. Secondly, the presence of a large body of water nearby can pose risks such as strong currents, sudden changes in weather, or the possibility of falling into the water. It's crucial to be aware of the local weather conditions and any potential hazards in the water. Additionally, the large number of people in the area might lead to overcrowding, which can make it difficult to navigate and enjoy the scenery. It's important to respect the natural environment and other visitors by maintaining a safe distance and being mindful of your actions. Lastly, be aware of the potential for tides to change rapidly, which can affect the safety of the area and the people in it.

Figure 10: More Examples of conversations between users and AVG-LLaVA.



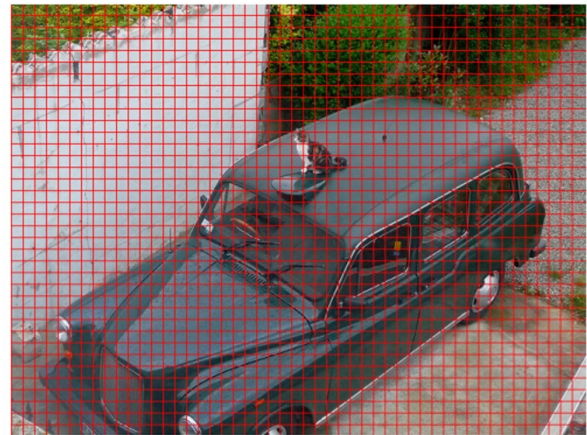
What color is this car?
Answer the question with a
single word.



Green



(a)



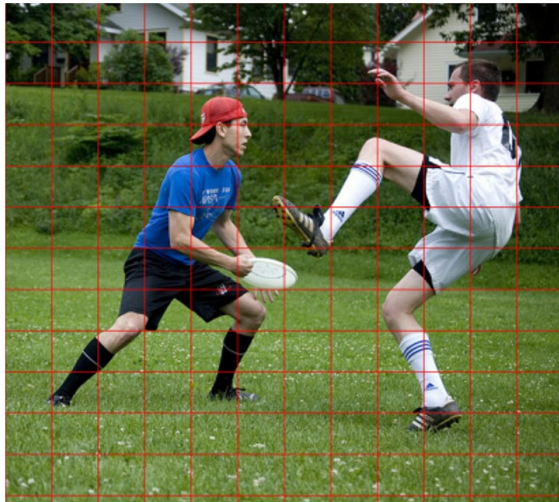
Is there a cat in this picture?
Answer the question with a
single word.



Yes



(b)



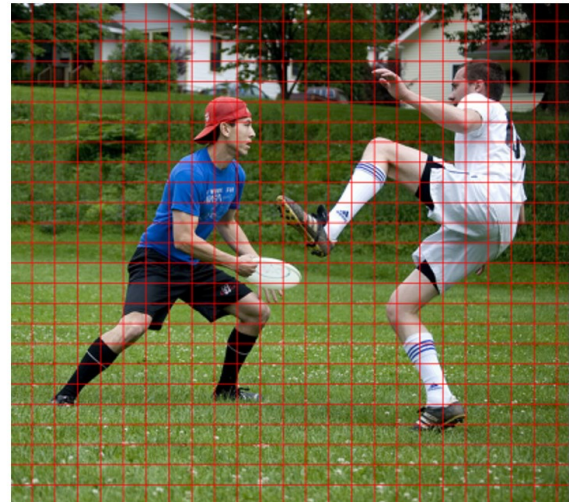
How many people are there in
the picture?
Please answer briefly



Two



(a)



What is the person on the left
holding?
Please answer briefly



Frisbe



(b)

Figure 11: Visualization of granularity selection using different instructions.