

AVG-LLaVA: A LARGE MULTIMODAL MODEL WITH ADAPTIVE VISUAL GRANULARITY

Anonymous authors

Paper under double-blind review

ABSTRACT

Recently, when dealing with high-resolution images, dominant large multimodal models (LMMs) usually divide them into multiple local images and one global image, which will lead 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. This approach not only reduces the number of visual tokens and speeds up inference, but also improves the overall model performance. Specifically, we introduce the following modules based on LLaVA-NeXT: (a) a visual granularity scaler that includes multiple pooling layers to obtain visual tokens with different granularities; (b) 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 propose 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).¹

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., 2023a; 2024a), marking a further step toward artificial general intelligence (AGI). Most existing LMMs follow the structure of LLaVA (Liu et al., 2023a; 2024a), which includes a vision encoder to embed 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, 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)), earlier 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

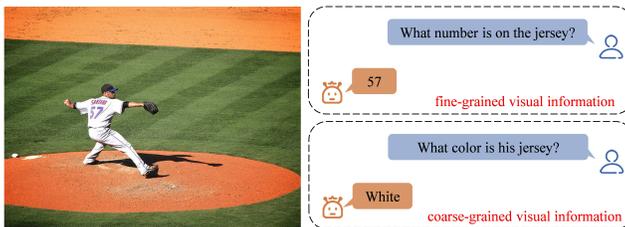


Figure 1: An example of VQA that requires information at different visual granularities. The image is from MSCOCO (Lin et al., 2014) dataset.

¹We will release our source code and model upon the acceptance of our paper.

054 visual token sequences, such as 2880 visual tokens for a 672×672 image. However, in practice, such
055 fine-grained visual information is not always necessary, and in some cases, coarse-grained visual
056 information can even be more beneficial for model predictions. For instance, as shown in Figure 1,
057 when the model is asked to recognize the number on the jersey, it requires relatively fine-grained
058 visual information. In contrast, determining the color of the jersey only necessitates coarse-grained
059 visual information.

060 In this paper, we propose *Adaptive Visual Granularity LLaVA (AVG-LLaVA)*, an LMM that can adap-
061 tively select the appropriate visual granularity based on the input image and instruction. The basic
062 intuition behind our model is that humans only scrutinize images carefully when answering difficult
063 questions; otherwise, a brief glance is sufficient. As displayed in Figure 2, AVG-LLaVA extends
064 LLaVA-NeXT with a *visual granularity scaler* and a *visual granularity router*. The visual granular-
065 ity scaler performs multiple rounds of pooling on visual tokens without training, each time halving
066 the number of visual tokens, thus obtaining a series of visual features with different granularities.
067 The visual granularity router adaptively selects the appropriate visual granularity features based on
068 the input multi-granularity visual features and text features. By doing so, for images and instructions
069 that do not require fine-grained details, the number of visual tokens can be reduced, which not only
070 speeds up inference but also improves performance. This performance enhancement likely stems
071 from the reduction of redundant information, as selecting the appropriate visual granularity makes
072 it easier for the model to answer questions based on images effectively.

073 Besides, we observe that it is challenging to train the visual granularity router directly through visual
074 instruction tuning (Liu et al., 2023a). This may be because the router cannot learn the distinctions
075 between different visual granularities from visual instruction tuning, making it difficult to learn how
076 to select the most appropriate visual granularity based on the image and instruction. To deal with
077 this issue, we propose a novel training paradigm called *Ranking Granularity* to align *LMM Feedback*
078 (*RGLF*). This paradigm aligns router probabilities of multiple granularities with LMM preferences
079 by a ranking loss (Hadsell et al., 2006; Hopkins & May, 2011; Liu et al., 2022), effectively aiding
080 the router in distinguishing between different visual granularities and selecting the appropriate one.

081 We further evaluate AVG-LLaVA on 11 benchmarks including tasks from various types (e.g., general
082 VQA and text-oriented VQA, etc.). Extensive experimental results show that AVG-LLaVA can
083 effectively reduce the number of visual tokens and improve inference speed (e.g., an 85.3% reduction
084 in visual tokens and a 2.53× increase in inference speed on the AI2D benchmark) while achieving
085 better performance under the same base LLM.

086 2 RELATED WORK

087 **Large Multimodal Models** LLMs such as GPT-4 (OpenAI, 2023a), LLaMA (Touvron et al.,
088 2023), and Gemini (Team et al., 2023) have achieved significant success in language understanding
089 and generation. Benefiting from this, multimodal large models (LMMs) have garnered widespread
090 attention. Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), and LLaMA-adapter (Zhang
091 et al., 2023) integrate a frozen visual encoder and trainable modules into a LLM, extending it into
092 a LMMs. These models are then fine-tuned using plain image-text pairs, enabling them to process
093 and perceive visual content. To further improve instruction-following abilities and response quality,
094 LLaVA (Liu et al., 2023a) fine-tunes the entire model using visual instruction data generated by
095 GPT-4. However, since these LMMs rely on CLIP-ViT to process images at a fixed resolution (e.g.
096 336×336), it hinders the LMMs from perceiving image details at higher resolutions.

097 **High-Resolution LMMs** To perceive images with higher resolutions, Qwen-VL (Bai et al., 2023)
098 increases the input resolution of the visual encoder to 448×448 and introduces an additional training
099 stage. Along this line, both Vary (Wei et al., 2023) and Mini-Gemini (Li et al., 2024a) include two
100 vision encoders: one is an additional introduced high-resolution vision encoder, and the other is the
101 original low-resolution vision encoder. Unlike the methods mentioned above, SPHINX (Lin et al.,
102 2023) and Monkey (Li et al., 2024b) enlarge the input image to a high resolution, and then divide
103 it into a fixed number of local images, which are individually encoded using an image encoder to
104 obtain local image tokens. Subsequently, the original global image tokens are concatenated with all
105 local image tokens to feed into the LLM. Furthermore, LLaVA-NeXT (Liu et al., 2024b) enumer-
106 ates various resolutions and adaptively selects the one that most closely matches the input image
107

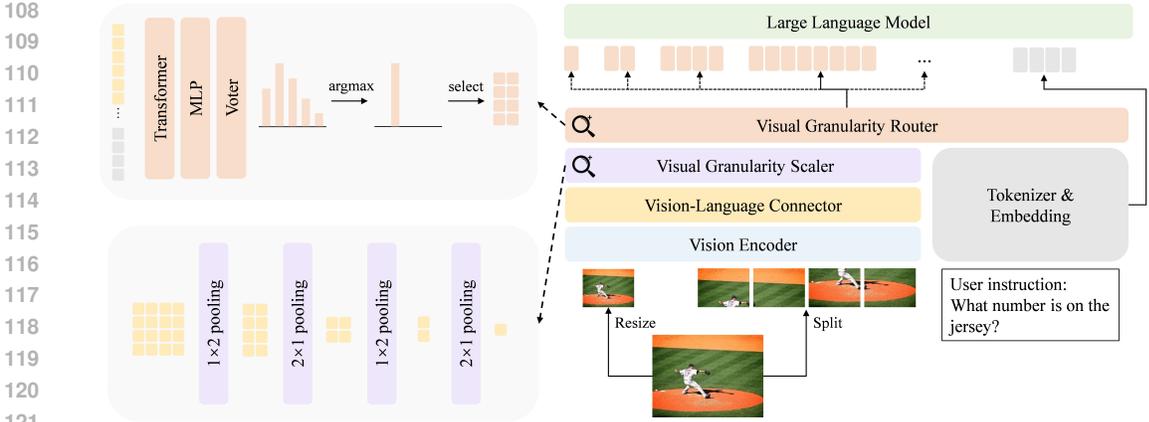


Figure 2: The architecture of AVG-LLaVA. AVG-LLaVA additionally introduces two modules based on LLaVA-NeXT: (1) Visual granularity scaler. This module consists of multiple pooling layers that progressively increase the granularity of visual features, thereby reducing the number of visual tokens; (2) 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.

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 pruning and merging. For example, CrossGET (Shi et al., 2023) 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. Unlike the aforementioned methods, 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 these methods, our model can adaptively select the appropriate visual granularity based on the input image and instruction, improving model performance while reducing 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 $\mathbf{I}_{local} \in \mathbb{R}^{H_v \times W_v \times 3}$, where 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, W_v) and encode it as global visual tokens. Finally, we map both global visual tokens and local visual tokens to the word embedding space through an MLP-based vision-language connector.

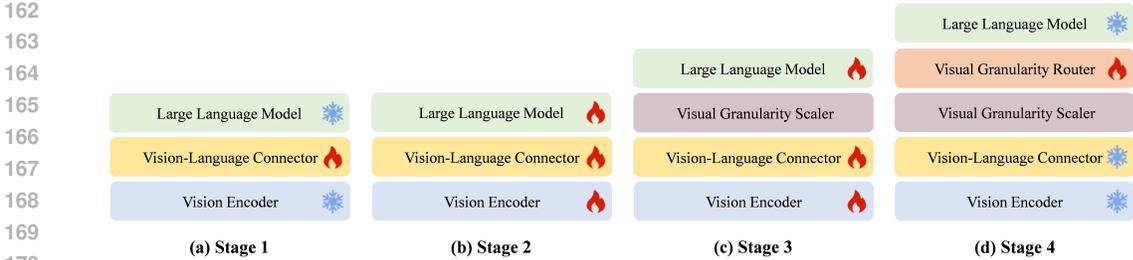


Figure 3: The overview of multi-stage training. Note that visual granularity scaler does not contain any parameters that require training.

Visual Granularity Scaler This module 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) 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 granularities². 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 to predict 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 each visual granularity, where the visual tokens corresponding to the granularity with the highest probability are fed into the LLM.

3.2 MULTI-STAGE TRAINING

To effectively train our model, we carefully design a multi-stage training strategy, which consists of four stages, as illustrated in Figure 3.

Stage 1: Pretraining During this stage, we only pretrain the vision-language connector on a plain image-caption dataset. Formally, we define the following cross-entropy loss for the next token prediction:

$$\mathcal{L}_1 = - \sum_{t=1}^T \log P(x_t | \mathbf{X}_v, \mathbf{X}_{c, < t}), \tag{1}$$

²To 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.

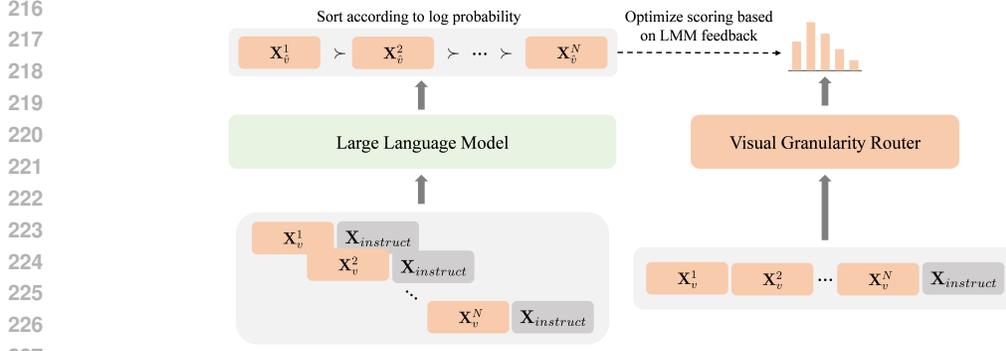


Figure 4: The overview of RGLF. Visual tokens of each granularity are concatenated with instruction tokens and then processed by the LMM to estimate the corresponding rewards. Visual granularity router optimizes the score (log probability) of each granularity based on the feedback from the LMM.

where $\mathbf{X}_{c,<t}$ are the caption tokens before the current prediction token x_t , and T is the length of target text tokens. Note that we fix the parameters of the vision encoder and large language model, both of which have been pre-trained on large-scale data. Through this stage of training, the image features can be aligned with the pre-trained LLM word embeddings.

Stage 2: Visual Instruction Tuning In the second stage, we jointly train the visual encoder, vision-language connector, and LLM on high-quality visual instruction data. In this way, the LLM can be converted into an LMM, which is able to complete various multimodal tasks. Specifically, we perform next-token predictions with the following cross-entropy loss only on the answering part

$$\mathcal{L}_2 = - \sum_{t=1}^T \log P(x_t | \mathbf{X}_v, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}), \quad (2)$$

where \mathbf{X}_a denotes the answer tokens.

Stage 3: Multi-Granularity Visual Instruction Tuning Following the previous stages, we introduce the visual granularity scaler. As described in Section 3.1, this module does not contain trainable parameters and thus does not need to be trained. Therefore, we use the same data as in the stage 2 to train the visual encoder, vision-language connector, and LLM, enabling them to perceive and process visual features of N different granularities. Formally, the loss at this stage is formulated as

$$\mathcal{L}_3 = - \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \log P(x_t | \mathbf{X}_v^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}). \quad (3)$$

Stage 4: Ranking Granularity to Align LMM Feedback Lastly, 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 to use the same visual instruction fine-tuning method as in previous stages. 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 4, 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 \mathbf{X}_v^i 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 $\mathbf{X}_v^1, \mathbf{X}_v^2, \dots, \mathbf{X}_v^N$. Given the visual tokens \mathbf{X}_v^i of the i -th granularity, we directly consider those tokens $(\mathbf{X}_v^1, \mathbf{X}_v^2, \dots, \mathbf{X}_v^{i-1})$ ranked above it as positive examples and the remaining

tokens ($\mathbf{X}_v^{i+1}; \mathbf{X}_v^{i+2}; \dots; \mathbf{X}_v^N$) as negative ones. Afterwards, we use the router to give scores (log probability) s_i for each \mathbf{X}_v^i :

$$s_i = \log P(g_i | \bar{\mathbf{X}}_v, \bar{\mathbf{X}}_{instruct}), \quad (4)$$

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:

$$\mathcal{L}_{rank} = \sum_{i=1} \sum_{j>i} \max(0, s_j - s_i + \lambda_{ij}), \quad (5)$$

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

$$\lambda_{ij} = \frac{j-i}{|T|} \sum_{t=1}^T (\log P(x_t | \mathbf{X}_v^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}) - \log P(x_t | \mathbf{X}_v^j, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t})). \quad (6)$$

When the preference of \mathbf{X}_v^j is only slightly worse than \mathbf{X}_v^i , the margin will be small. Conversely, when \mathbf{X}_v^j is significantly worse than \mathbf{X}_v^i , the margin will correspondingly increase. In this way, we can dynamically adjust the margin to obtain adaptive 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:

$$k = \arg \max_i \sum_{t=1}^T \log P(x_t | \mathbf{X}_v^i, \mathbf{X}_{instruct}, \mathbf{X}_{a,<t}), \quad (7)$$

$$\mathcal{L}_{ce} = -\log P(g_k | \bar{\mathbf{X}}_v, \bar{\mathbf{X}}_{instruct}). \quad (8)$$

In summary, the total loss is defined as the weighted sum of two losses:

$$\mathcal{L}_A = \mathcal{L}_{rank} + \alpha \mathcal{L}_{ce}, \quad (9)$$

where α is the hyperparameters 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 Note that in this work, we mainly focus on investigating the effectiveness of automatic visual granularity selection in reducing the number of visual tokens and improving model performance. Therefore, during the first stage, we also use CC-595K image-text pairs for model training, as implemented in LLaVA-NeXT (Liu et al., 2024b). In the subsequent training stages, we also hope to use the same data as LLaVA-NeXT. However, the real user interaction data used for visual instruction fine-tuning in LLaVA-NeXT are not open-sourced, so we opt to extract 200K samples from 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.³ In total, the visual instruction fine-tuning data we use contains 1M image-text pairs.

Implementation Details Following Liu et al. (2024b), we use CLIP ViT-L/14 as the visual encoder, Vicuna-7B (Chiang et al., 2023) as the LLM. We set the filtered instruction token number k to 32 and the cross-entropy loss weight α to 0.1. The impact of these two hyperparameters on model performance is discussed in Section 4.6. In our first two stages, we adopt the same training settings as LLaVA-NeXT for pre-training and visual instruction fine-tuning. In the third 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 fourth 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.2.

³Our data recipe follows Open-LLaVA-NeXT (Lin & Long, 2024).

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.

Model	LLM	General VQA			Text-oriented VQA			
		GQA	ScienceQA	VizWiz	TextVQA	ChartQA	DocVQA	AI2D
<i>Standard-resolution LMMs</i>								
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	-	-	-
<i>High-resolution LMMs</i>								
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	-	-	-
LLaVA-NeXT (Liu et al., 2024b)	Vicuna-7B	64.2	70.1	<u>57.6</u>	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	72.6	<u>66.7</u>
AVG-LLaVA	Vicuna-7B	63.0	<u>71.1</u>	59.8	<u>67.1</u>	66.3	74.6	67.3

Table 2: Comparison with LMMs of the same size on general multimodal benchmarks. The best results are marked in bold, and the second best results are underlined.

Model	LLM	MME	MME ^C	MMB	MMB ^{CN}	POPE	MMMU
<i>Standard-resolution LMMs</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., 2023a)	Vicuna-7B	1510.7	-	64.3	58.3	<u>87.3</u>	-
<i>High-resolution LMMs</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	<u>1546.0</u>	319.0	65.8	-	-	36.8
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	1557.4	366.8	69.9	61.8	87.4	37.4

Evaluations We evaluate our model on three kinds of benchmarks: (1) **general VQA benchmarks**: GQA (Hudson & 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 AI2D (Kembhavi et al., 2016); and (3) **general multimodal benchmarks**: MME (Fu et al., 2023), MMB (Liu et al., 2023b), MMB^{CN} (Liu et al., 2023b), 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 other high-resolution LMMs on VizWiz. Although it does not achieve the best results on GQA and ScienceQA, 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 result, 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 benchmark covers a broader range of evaluation aspects, requiring models to possess more complex perception and reasoning capabilities. As summarized in Table 2, AVG-LLaVA surpasses all other models, exhibiting superior overall performance and highlighting its adaptability and effectiveness across various disciplines. Specifically, AVG-LLaVA outperforms the second best model by 9.4 and 6.1 on MME and MME^C, respectively, and by 1.9 and 1.2 on 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

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× faster than LLaVA-NeXT.

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×

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 benchmark, 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 other benchmarks. Particularly, on the AI2D benchmark, AVG-LLaVA achieves better performance than LLaVA-NeXT while using only 14.7% of the visual tokens, and the inference speed increased 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 5 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. Additionally, we observe that the granularities with 72 and 288 visual tokens are rarely selected. However, we find that incorporating these granularities of visual tokens helps the model progressively learn to utilize visual tokens with different granularity and allows it to more fully understand the differences between various visual granularities. Further verifications are detailed 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

⁴We also present qualitative results in Appendix A.3 and illustrate the effectiveness of adaptive visual granularity.

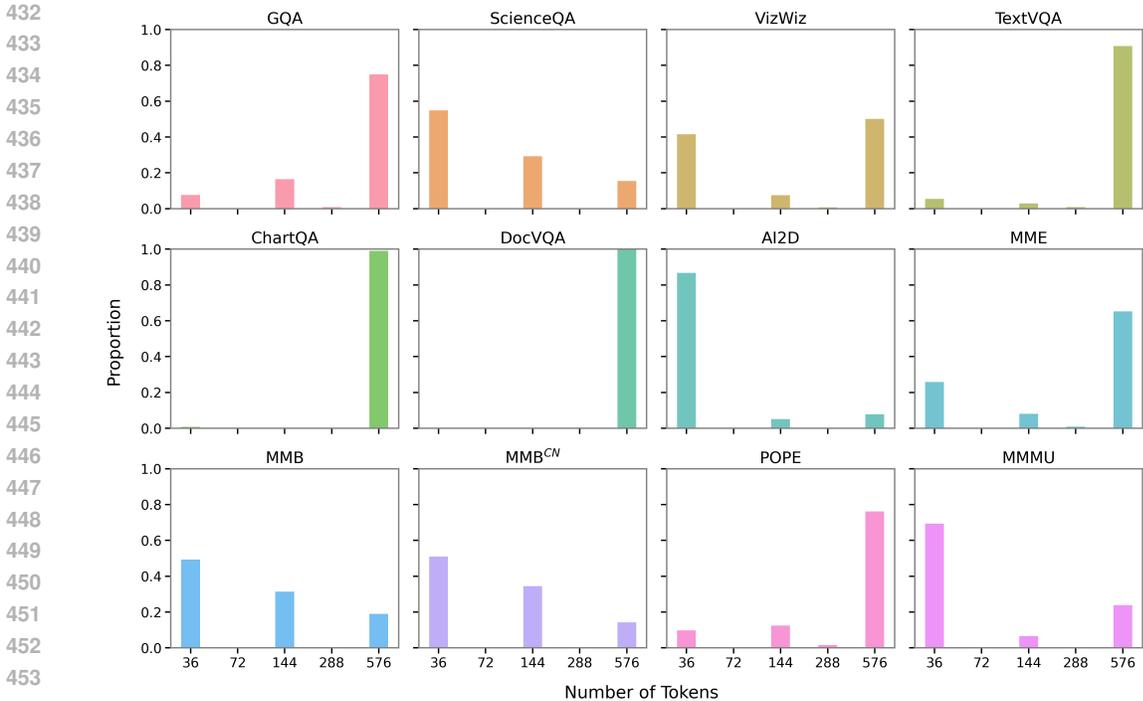


Figure 5: Visualization of the proportion for different granularity visual tokens.

Table 4: Ablation results on multiple benchmarks.

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

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-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 remove 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, applying the same loss function as in Stage 2. Unlike our original approach where the router is directly supervised by 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 direct feedback from the LMM allows the router 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 fourth 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 to align LMM feedback.

4.6 HYPERPARAMETER ANALYSIS

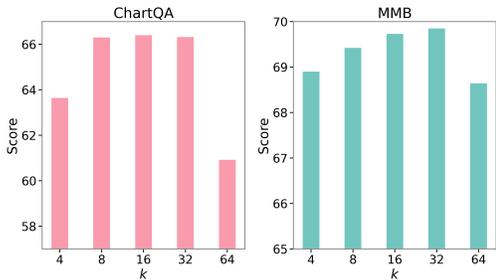


Figure 6: Influence of the filtered instruction token number k on model performance, measured on ChartQA and MMB benchmarks.

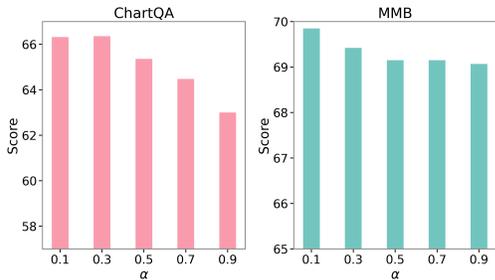


Figure 7: Influence of the cross-entropy loss weight α on model performance, measured on ChartQA and MMB benchmarks.

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.

5 CONCLUSION

In this work, we propose AVG-LLaVA, an LMM that can adaptively select appropriate visual granularity based on input image and instruction. AVG-LLaVA builds upon LLaVA-NeXT by introducing a visual granularity scaler and a visual granularity router, which are used to obtain multi-granularity visual features and select the appropriate visual granularity based on image and instruction, respectively. 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 based on image and instruction. 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. Particularly, on the AI2D benchmark, it reduces the number of visual tokens by 85.3% and speeds the inference by $2.53\times$. We hope our work can inspire more attention to the visual granularity in LMMs.

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. In the future, we plan to design a more suitable granularity scaling network to provide richer visual granularities.

REFERENCES

- 540
541
542 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson,
543 Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza
544 Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Mon-
545 teiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Shar-
546 ifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén
547 Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo,
548 S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neu-
549 ral Information Processing Systems 35: Annual Conference on Neural Information Process-
550 ing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,
551 2022*, 2022. URL [http://papers.nips.cc/paper_files/paper/2022/hash/
960a172bc7fbf0177cccbb411a7d800-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177cccbb411a7d800-Abstract-Conference.html).
- 552
553 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang
554 Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, local-
555 ization, text reading, and beyond. *CoRR*, 2023.
- 556
557 Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *CoRR*,
558 abs/2405.17430, 2024. doi: 10.48550/ARXIV.2405.17430. URL [https://doi.org/10.
48550/arXiv.2405.17430](https://doi.org/10.48550/arXiv.2405.17430).
- 559
560 Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhi-
561 hong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized
562 data for A lite vision-language model. *CoRR*, abs/2402.11684, 2024a. doi: 10.48550/ARXIV.
563 2402.11684. URL <https://doi.org/10.48550/arXiv.2402.11684>.
- 564
565 Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang.
566 An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-
567 language models. *CoRR*, abs/2403.06764, 2024b. doi: 10.48550/ARXIV.2403.06764. URL
<https://doi.org/10.48550/arXiv.2403.06764>.
- 568
569 Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong
570 Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl:
571 Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*,
572 abs/2312.14238, 2023. doi: 10.48550/ARXIV.2312.14238. URL [https://doi.org/10.
48550/arXiv.2312.14238](https://doi.org/10.48550/arXiv.2312.14238).
- 573
574 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,
575 Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An
576 open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL [https:
//lmsys.org/blog/2023-03-30-vicuna/](https://lmsys.org/blog/2023-03-30-vicuna/).
- 577
578 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng
579 Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-
580 purpose vision-language models with instruction tuning. In Alice Oh, Tristan Nau-
581 mann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances
582 in Neural Information Processing Systems 36: Annual Conference on Neural Informa-
583 tion Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,
584 2023*, 2023. URL [http://papers.nips.cc/paper_files/paper/2023/hash/
9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html).
- 585
586 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei
587 Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive eval-
588 uation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.
589 48550/ARXIV.2306.13394. URL <https://doi.org/10.48550/arXiv.2306.13394>.
- 590
591 Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman,
592 Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual ques-
593 tions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern
Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3608–3617.
Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.

- 594 2018.00380. URL [http://openaccess.thecvf.com/content_cvpr_2018/html/](http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html)
595 [Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html).
596
- 597 Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant
598 mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pp. 1735–1742. IEEE Computer
599 Society, 2006. doi: 10.1109/CVPR.2006.100. URL [https://doi.org/10.1109/CVPR.](https://doi.org/10.1109/CVPR.2006.100)
600 [2006.100](https://doi.org/10.1109/CVPR.2006.100).
601
- 602 Mark Hopkins and Jonathan May. Tuning as ranking. In *Proceedings of the 2011 Conference*
603 *on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John*
604 *McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of*
605 *the ACL*, pp. 1352–1362. ACL, 2011. URL <https://aclanthology.org/D11-1125/>.
- 606 Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang.
607 Matryoshka query transformer for large vision-language models. *CoRR*, abs/2405.19315, 2024.
608 doi: 10.48550/ARXIV.2405.19315. URL [https://doi.org/10.48550/arXiv.2405.](https://doi.org/10.48550/arXiv.2405.19315)
609 [19315](https://doi.org/10.48550/arXiv.2405.19315).
- 610 Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual
611 reasoning and compositional question answering. In *IEEE Conference on Computer Vision*
612 *and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–
613 6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL
614 [http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_](http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html)
615 [A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_](http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html)
616 [CVPR_2019_paper.html](http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html).
- 617 Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th*
618 *International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26,*
619 *2017, Conference Track Proceedings*. OpenReview.net, 2017. URL [https://openreview.](https://openreview.net/forum?id=rkE3y85ee)
620 [net/forum?id=rkE3y85ee](https://openreview.net/forum?id=rkE3y85ee).
621
- 622 Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali
623 Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and
624 Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam,*
625 *The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in*
626 *Computer Science*, pp. 235–251. Springer, 2016. doi: 10.1007/978-3-319-46493-0_15. URL
627 https://doi.org/10.1007/978-3-319-46493-0_15.
- 628 Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim,
629 Wonseok Hwang, Sangdoon Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document un-
630 derstanding transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria
631 Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference,*
632 *Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes*
633 *in Computer Science*, pp. 498–517. Springer, 2022. doi: 10.1007/978-3-031-19815-1_29. URL
634 https://doi.org/10.1007/978-3-031-19815-1_29.
- 635 Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa,
636 Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training
637 mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learn-*
638 *ing Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL
639 <https://openreview.net/forum?id=T5nUQDrM4u>.
- 640 Bo Li, Peiyuan Zhang, Jingkan Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-
641 resolution multi-modality model. *CoRR*, abs/2311.04219, 2023a. doi: 10.48550/ARXIV.2311.
642 [04219](https://doi.org/10.48550/ARXIV.2311.04219). URL <https://doi.org/10.48550/ARXIV.2311.04219>.
- 643 Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-
644 image pre-training with frozen image encoders and large language models. In Andreas Krause,
645 Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett
646 (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu,*
647 *Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742.
PMLR, 2023b. URL <https://proceedings.mlr.press/v202/li23q.html>.

- 648 Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng
649 Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models.
650 *CoRR*, abs/2403.18814, 2024a. doi: 10.48550/ARXIV.2403.18814. URL [https://doi.org/
651 10.48550/arXiv.2403.18814](https://doi.org/10.48550/arXiv.2403.18814).
- 652 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evalu-
653 ating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino,
654 and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural
655 Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Associa-
656 tion for Computational Linguistics, 2023c. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL
657 <https://doi.org/10.18653/v1/2023.emnlp-main.20>.
- 658 Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu,
659 and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-
660 modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
661 Recognition (CVPR)*, pp. 26763–26773, June 2024b.
- 662 Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and
663 Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *CoRR*, abs/2401.15947,
664 2024a. doi: 10.48550/ARXIV.2401.15947. URL [https://doi.org/10.48550/arXiv.
665 2401.15947](https://doi.org/10.48550/arXiv.2401.15947).
- 666 Chen Lin and Xing Long. Open-llava-next: An open-source implementation of llava-next se-
667 ries for facilitating the large multi-modal model community. [https://github.com/
668 xiaochen98/Open-LLaVA-NeXT](https://github.com/xiaochen98/Open-LLaVA-NeXT), 2024.
- 669 Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr
670 Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J.
671 Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014
672 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V,
673 volume 8693 of Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/
674 978-3-319-10602-1_48. URL [https://doi.org/10.1007/978-3-319-10602-1_
675 48](https://doi.org/10.1007/978-3-319-10602-1_48).
- 676 Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models
677 with visual tokens withdrawal for rapid inference. *CoRR*, abs/2405.05803, 2024b. doi: 10.48550/
678 ARXIV.2405.05803. URL <https://doi.org/10.48550/arXiv.2405.05803>.
- 679 Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi
680 Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and
681 Yu Qiao. SPHINX: the joint mixing of weights, tasks, and visual embeddings for multi-modal
682 large language models. *CoRR*, abs/2311.07575, 2023. doi: 10.48550/ARXIV.2311.07575. URL
683 <https://doi.org/10.48550/arXiv.2311.07575>.
- 684 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice
685 Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.),
686 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Infor-
687 mation Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16,
688 2023*, 2023a. URL [http://papers.nips.cc/paper_files/paper/2023/hash/
689 6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html).
- 690 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction
691 tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
692 (CVPR)*, pp. 26296–26306, June 2024a.
- 693 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee.
694 Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL [https://
695 llava-vl.github.io/blog/2024-01-30-llava-next/](https://llava-vl.github.io/blog/2024-01-30-llava-next/).
- 696 Yixin Liu, Pengfei Liu, Dragomir R. Radev, and Graham Neubig. BRIO: bringing order to ab-
697 stractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.),

- 702 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol-*
703 *ume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2890–2903. Association
704 for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.207. URL
705 <https://doi.org/10.18653/v1/2022.acl-long.207>.
706
- 707 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan,
708 Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal
709 model an all-around player? *CoRR*, abs/2307.06281, 2023b. doi: 10.48550/ARXIV.2307.06281.
710 URL <https://doi.org/10.48550/arXiv.2307.06281>.
- 711 Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey:
712 An ocr-free large multimodal model for understanding document. *CoRR*, abs/2403.04473, 2024c.
713 doi: 10.48550/ARXIV.2403.04473. URL [https://doi.org/10.48550/arXiv.2403.](https://doi.org/10.48550/arXiv.2403.04473)
714 [04473](https://doi.org/10.48550/arXiv.2403.04473).
- 715 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu,
716 Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reason-
717 ing via thought chains for science question answering. In Sanmi Koyejo, S. Mo-
718 hamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural*
719 *Information Processing Systems 35: Annual Conference on Neural Information Process-*
720 *ing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9,*
721 *2022*. URL [http://papers.nips.cc/paper_files/paper/2022/hash/](http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html)
722 [11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html).
- 723 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A
724 benchmark for question answering about charts with visual and logical reasoning. In Smaranda
725 Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Com-*
726 *putational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 2263–2279. Association
727 for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL
728 <https://doi.org/10.18653/v1/2022.findings-acl.177>.
729
- 730 Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on doc-
731 ument images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021,*
732 *Waikoloa, HI, USA, January 3-8, 2021*, pp. 2199–2208. IEEE, 2021. doi: 10.1109/WACV48630.
733 [2021.00225](https://doi.org/10.1109/WACV48630.2021.00225). URL <https://doi.org/10.1109/WACV48630.2021.00225>.
- 734 OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023a. doi: 10.48550/ARXIV.2303.
735 [08774](https://doi.org/10.48550/arXiv.2303.08774). URL <https://doi.org/10.48550/arXiv.2303.08774>.
- 736 OpenAI. Gpt-4v(ision) system card, September 2023b. URL [https://cdn.openai.com/](https://cdn.openai.com/papers/GPTV_System_Card.pdf)
737 [papers/GPTV_System_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf).
738
- 739 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
740 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya
741 Sutskever. Learning transferable visual models from natural language supervision. In Marina
742 Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Ma-*
743 *chine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Ma-*
744 *chine Learning Research*, pp. 8748–8763. PMLR, 2021. URL [http://proceedings.mlr.](http://proceedings.mlr.press/v139/radford21a.html)
745 [press/v139/radford21a.html](http://proceedings.mlr.press/v139/radford21a.html).
- 746 Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token
747 reduction for efficient large multimodal models. *CoRR*, abs/2403.15388, 2024. doi: 10.48550/
748 [ARXIV.2403.15388](https://doi.org/10.48550/arXiv.2403.15388). URL <https://doi.org/10.48550/arXiv.2403.15388>.
- 749 Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E.
750 Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-
751 of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017,*
752 *Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL
753 <https://openreview.net/forum?id=BlckMDqlg>.
754
- 755 Dachuan Shi, Chaofan Tao, Anyi Rao, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Cross-
get: Cross-guided ensemble of tokens for accelerating vision-language transformers. *CoRR*,

- 756 abs/2305.17455, 2023. doi: 10.48550/ARXIV.2305.17455. URL <https://doi.org/10.48550/arXiv.2305.17455>.
757
758
- 759 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi
760 Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference*
761 *on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-*
762 *20, 2019*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.
763 2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.
764
- 765 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,
766 Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly
767 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
768
- 769 IDEFICS Research Team. Introducing idefics: An open reproduction of state-of-the-art visual lan-
770 guage model, 2023. URL <https://huggingface.co/blog/idefics>.
- 771 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
772 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Ar-
773 mand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation
774 language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL
775 <https://doi.org/10.48550/arXiv.2302.13971>.
- 776 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez,
777 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von
778 Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman
779 Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on*
780 *Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp.
781 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
782
- 783 Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun,
784 Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-
785 language models. *CoRR*, abs/2312.06109, 2023. doi: 10.48550/ARXIV.2312.06109. URL
786 <https://doi.org/10.48550/arXiv.2312.06109>.
- 787 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei
788 Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with
789 modality collaboration. *CoRR*, abs/2311.04257, 2023. doi: 10.48550/ARXIV.2311.04257. URL
790 <https://doi.org/10.48550/arXiv.2311.04257>.
791
- 792 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
793 Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun,
794 Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and
795 Wenhui Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning
796 benchmark for expert AGI. *CoRR*, abs/2311.16502, 2023. doi: 10.48550/ARXIV.2311.16502.
797 URL <https://doi.org/10.48550/arXiv.2311.16502>.
- 798 Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng
799 Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init at-
800 tention. *CoRR*, abs/2303.16199, 2023. doi: 10.48550/ARXIV.2303.16199. URL <https://doi.org/10.48550/arXiv.2303.16199>.
801
- 802 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing
803 vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592,
804 2023. doi: 10.48550/ARXIV.2304.10592. URL <https://doi.org/10.48550/arXiv.2304.10592>.
805
806
807
808
809

810 A APPENDIX

811 A.1 VISUAL GRANULARITY SELECTION ALGORITHM

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

813 **Algorithm 1** Visual Granularity Selection Algorithm

- 814 **Require:** Multi-granularity visual tokens $\bar{\mathbf{X}}_v = [\mathbf{X}_v^1; \mathbf{X}_v^2; \dots; \mathbf{X}_v^N]$, Instruction tokens $\mathbf{X}_{instruct}$,
 815 Visual tokens of original granularity \mathbf{X}_v
 816 **Ensure:** Selected granularity visual tokens $\mathbf{X}_v^{selected}$.
- 817 1: Obtain the filtered instruction tokens $\bar{\mathbf{X}}_{instruct} = \text{Top-}k(\text{cosine_sim}(\mathbf{X}_{instruct}, \mathbf{X}_v))$
 - 818 2: Concatenate $\bar{\mathbf{X}}_v$ and $\bar{\mathbf{X}}_{instruct}$ to form the input for the router
 - 819 3: Apply a Transformer layer to facilitate token fusion $\mathbf{Z}_{fusion} = \text{Transformer}([\bar{\mathbf{X}}_v; \bar{\mathbf{X}}_{instruct}])$.
 - 820 4: Use an MLP to predict logits for each token $\mathbf{Z}_{out} = \text{MLP}(\mathbf{Z}_{fusion})$
 - 821 5: Aggregate the logits using a learnable weight matrix $\mathbf{Z}_{final} = \mathbf{W}\mathbf{Z}_{out}$.
 - 822 6: Compute the probability distribution using softmax $\mathbf{P} = \text{softmax}(\mathbf{Z}_{final})$.
 - 823 7: Identify the granularity with the highest probability $selected = \arg \max(\mathbf{P})$.
 - 824 8: **return** $\mathbf{X}_v^{selected}$

825 A.2 TRAINING DETAILS

826 We list the training hyperparameters for all stages in Table 5.

Hyperparameter	Stage 1	Stage 2	Stage 3	Stage 4
Data size	595K	1M	1M	1M
Batch size	256	128	128	128
lr	1e-3	2e-5	1e-5	1e-3
Vision encoder lr	-	2e-6	2e-5	-
lr schedule		cosine decay		
lr warmup ratio		0.03		
Weight decay		0		
Epoch		1		
Optimizer		AdamW		
DeepSpeed stage	2	3	3	3
Training cost (#Hours)	4	17	65	14

827 Table 5: Hyperparameters of AVG-LLaVA.

828 A.3 QUALITATIVE EVALUATION

829 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.

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

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917



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.

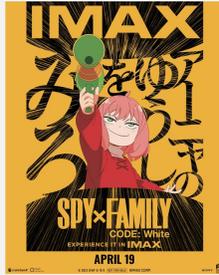
AVG-LLaVA (144): The image is unusual because it features a small white dog with wings, which is not a natural occurrence. Dogs do not have the ability to fly, and it is highly unlikely to see a dog with wings in a real-life scenario. The image is likely a digitally manipulated or artistically created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

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971



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.

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025



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.

A.4 ATTENTION MAP VISUALIZATION



Figure 11: 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 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 11, 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 primarily focused on 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 clearer attention map with fewer noise points in the background area, indicating more precise focus on the relevant regions, thereby improving model performance.

A.5 VISUALIZATION OF ROUTER SELECTION FOR DIFFERENT INSTRUCTIONS

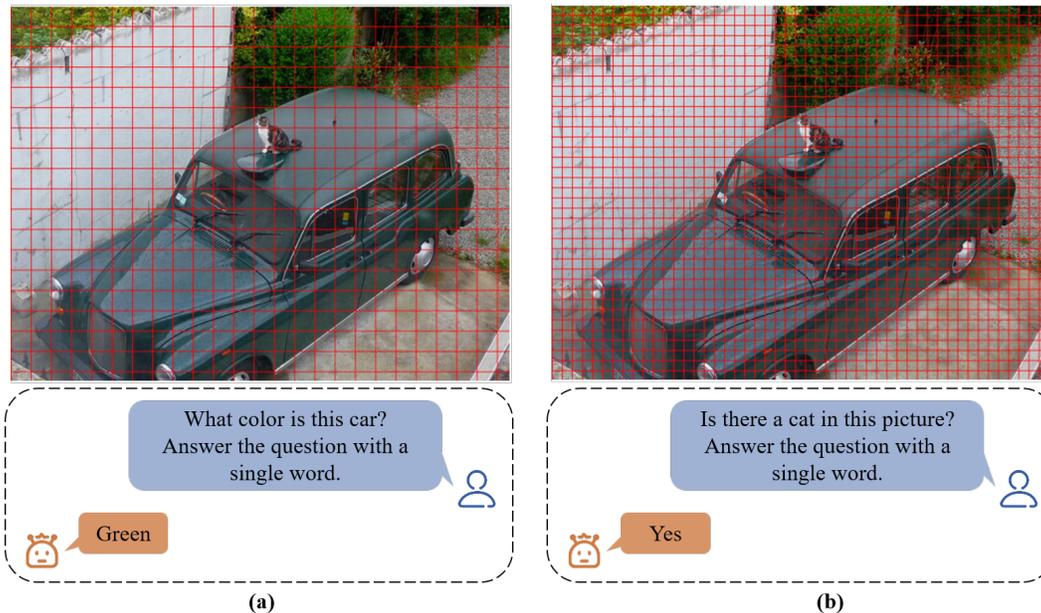


Figure 12: Visualization of granularity selection using different instructions.

As shown in Figure 12, 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

1080 about the color of the car, the model does not require such fine-grained visual information, whereas
1081 when asking whether there is a cat, the model requires finer-grained visual information.
1082

1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133