MATRYOSHKAFUSION: MULTI-GRANULARITY VISUAL FEATURE FUSION WITHIN PATCH REGIONS FOR MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-granularity understanding is essential for visual comprehension, as the human visual system processes visual signals at multiple granularities. However, multi-granularity understanding mechanisms remain scarce in MLLMs, a mainstream paradigm in vision-language understanding. A limited number of studies have introduced multi-granularity understanding to accelerate inference at the cost of some performance degradation, leaving the potential for enhancing visual understanding largely unexplored. In this paper, we propose MatryoshkaFusion, a novel Matryoshka multi-granularity method that fuses multi-granularity features within patch regions, thus enhancing the model's visual understanding ability without increasing the number of LLM input tokens or the volume of training data. Furthermore, we visually demonstrate and analyze the weight distribution of different granularities, offering a new interpretable perspective on the mechanism underlying multi-granularity visual features' enhancement of MLLMs' visual understanding ability. To validate the effectiveness of our method, two sets of experiments are conducted: (1) training a specialized model on the ScienceQA dataset to rapidly assess the feasibility of our method; (2) training a multimodal chatbot and evaluating its zero-shot performance across multiple benchmarks. In both experiments, our model achieves better results than existing advanced multi-granularity methods used in MQT-LLAVA and LLaVA-M^3.

1 INTRODUCTION

Multi-granularity understanding constitutes a critical research direction in visual comprehension, and its rationale is not only grounded in intuition but also supported by biological mechanisms (Bracci & Op de Beeck, 2023; Sigulinsky et al., 2024). From an intuitive perspective, when observing objects in real-world scenes, humans tend to interpret visual information at varying granularities depending on their focus of attention. Thus, multi-granularity understanding has been introduced into traditional computer vision tasks. The feature pyramid network (FPN) (Lin et al., 2017) was proposed for multi-granularity object detection in vision-only tasks, and some studies (Shi et al., 2024; Xiang et al., 2025) have also introduced multi-granularity understanding into image segmentation and visual grounding tasks. In image classification, Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) was proposed, which constructs visual representation vectors via multi-granularity understanding, enabling the model to maintain competitive performance with reduced vector dimensions. In recent years, as multimodal large language models (MLLMs) have become a mainstream paradigm owing to their powerful capabilities in visual understanding, reasoning, and instruction following, incorporating multi-granularity understanding into the MLLM framework has emerged as a critical direction that warrants in-depth exploration.

Current explorations of multi-granularity understanding methods within MLLMs remain insufficient. Mainstream MLLMs (Liu et al., 2024a; Zhu et al., 2025b; Bai et al., 2025) focus on adopting more advanced pre-trained models and larger-scale training datasets to enhance model performance. Only a limited number of studies (Hu et al., 2024; Cai et al., 2025) on multi-granularity MLLMs exist, and these studies focus on improving model inference efficiency. LLaVA-1.5-M^3 (Cai et al., 2025), inspired by MRL (Kusupati et al., 2022), employs average pooling to construct multi-granularity visual features and pairs them separately with text to form multiple image-text

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

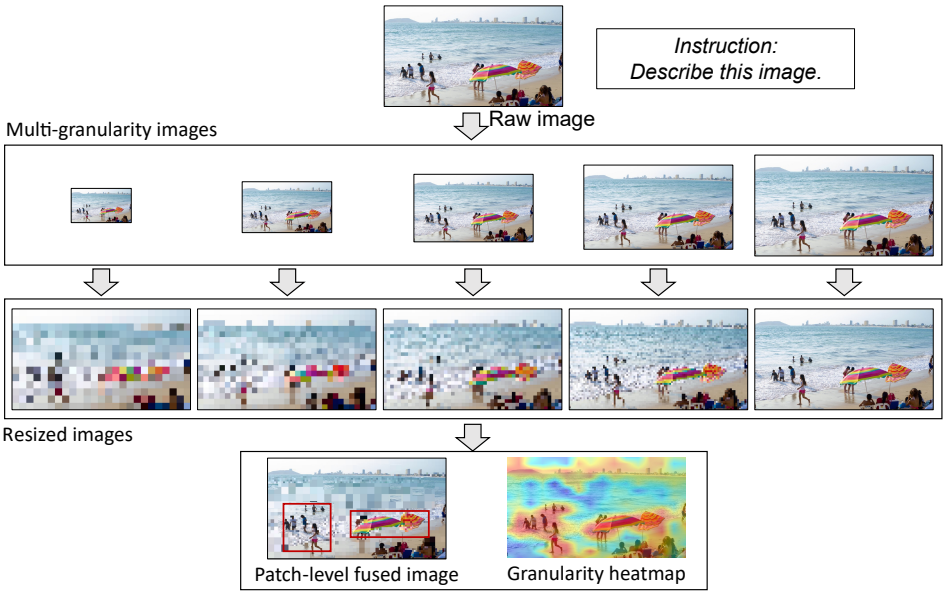

Figure 1: Matryoshka multi-granularity visual feature fusion method. We generate and fuse multi-granularity features within patch regions, thus enhancing the model's visual understanding ability without increasing the number of LLM input tokens or the volume of training data.

pairs for training. It allows coarse-grained features, which comprise fewer tokens, to replace fine-grained ones during inference, thereby accelerating inference at the cost of some performance degradation. Because this design focuses on inference acceleration, it overlooks the potential to enhance the model's visual comprehension capability and leaves the underlying structural mechanisms unexplored. Moreover, this design increases the number of image-text training pairs to approximately five times the original, substantially raising training costs. Another method, MQT-LLaVA (Hu et al., 2024), employs a set of query vectors with different numbers of tokens to extract visual information through attention mechanisms and feeds a randomly selected query vector into the LLM, thus enabling the model to utilize fewer tokens and improve inference speed. This approach abandons the original encoded visual features, causing information loss and degrading the model's visual understanding capability. Meanwhile, the entire learning process is implicit, making it difficult to determine the correspondence between each query token and the original image, which poses challenges for further analysis.

In this paper, we propose MatryoshkaFusion, a novel Matryoshka multi-granularity method to enhance the visual understanding ability of MLLMs. As shown in Fig. 1, it constructs multi-layered visual features at coarse-to-fine granularities and fuses these visual features within patch regions, without increasing the number of visual tokens or the amount of training data. Visualization of the learned granularity weights reveals that our method dynamically attends to different visual granularities at the patch level. This observation not only highlights the necessity of exploring multi-granularity understanding but also corroborates the effectiveness of our proposed mechanism. Then, two sets of experiments are conducted. First, we train a specialized model on ScienceQA, a large-scale vision-language scientific question answering dataset, to rapidly validate the effectiveness of MatryoshkaFusion. Second, following the same training pipeline and dataset of the advanced MLLM LLaVA-1.5, we evaluate the zero-shot performance on multiple benchmarks. Experimental results demonstrate the effectiveness of our method in enhancing visual understanding. In summary, the main contributions of this paper are as follows:

1. We propose a novel Matryoshka multi-granularity fusion method named MatryoshkaFusion to enhance the visual understanding ability of MLLMs. It constructs multi-layered visual features at coarse-to-fine granularities and fuses these visual features within patch regions to enhance the model's visual understanding without increasing the number of LLM input tokens or the amount of training data.

2. Furthermore, visualization and analysis of the granularity weight distributions in MatryoshkaFusion validate the effectiveness of our approach and provide a new interpretable perspective on how multi-granularity visual features enhance the visual understanding capability of MLLMs.

3. We conduct extensive experiments on both the ScienceQA and the Multimodal Chatbot. In the ScienceQA experiments, our model achieves competitive performance among advanced specialized MLLMs. In the multimodal chatbot experiments, our method yields improvements across multiple benchmarks when compared to LLaVA-1.5 and advanced multi-granularity MLLMs, LLaVA-1.5-M^3 and MQT-LLaVA, under the same training pipeline and dataset.

2 Related Work

2.1 Multimodal large language models

Large language models (LLMs) such as ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023c), LLaMA (Llama, 2023), and DeepSeek (Guo et al., 2025) have demonstrated remarkable capabilities in text understanding and reasoning. GPT-4V (OpenAI, 2023b) further incorporates visual information, pushing the boundaries beyond text-only tasks. Inspired by this progress, a series of works built upon open-source LLMs have emerged, among which LLaVA (Liu et al., 2024a) and InstructBLIP (Dai et al., 2023) are representative, since they combine the linguistic power of LLMs with CLIP's (Radford et al., 2021) image encoding ability, establishing a tripartite framework of "image understanding, cross-modal alignment, and text generation" for multimodal LLMs. More recently, multimodal LLMs targeting additional modalities, such as 3D-LMMs (Zhu et al., 2025a; Deng et al., 2025) and audio LLMs (Fu et al., 2026; Fang et al., 2025), have also been developed. Nevertheless, most existing methods directly extract patch-level features from a single layer of CLIP as input to subsequent modules, which is relatively simplistic and leaves room for deeper exploration of image representations. To address this, we propose to construct and fuse nested multi-granularity visual features, aiming to enrich visual representations and thereby enhance the model's visual understanding capability.

2.2 Matryoshka Representation Learning

In MLLMs, research on multi-granularity visual understanding remains limited. Only a few studies try to introduce visual understanding into MLLMs, and these works focus on using multi-granularity understanding to improve model inference speed. Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) constructs visual representations at varying granularities and applies them to classification and retrieval tasks, revealing that coarser representations can trade a moderate performance loss for improved efficiency. Inspired by MRL, MQT-LLAVA (Hu et al., 2024) and LLaVA-1.5-M^3 (Cai et al., 2025) introduce nested multi-granularity features into MLLMs, following the same principle of leveraging coarser features to accelerate inference. Technically, MQT-LLAVA incorporates a Q-Former (Li et al., 2023) to reduce the number of visual tokens, compressing representations for efficiency gains, though the Q-Former itself is computationally less efficient than the mainstream MLP projection. LLaVA-1.5-M^3, on the other hand, constructs multi-granularity visual features and pairs each with text separately; through data augmentation, it adapts the model to coarser visual inputs, thereby improving inference speed, though at the cost of substantially increased training overhead. Compared with the original LLaVA (Liu et al., 2024a), both methods incur performance degradation to varying degrees, with LLaVA-1.5-M^3 maintaining relatively better performance—a direct consequence of its design choices. In this paper, we revisit multi-granularity visual understanding in MLLMs and argue that it is a widespread phenomenon that inherently benefits visual understanding, and thus its role should not be confined to improving inference efficiency. Motivated by this perspective, we propose a novel approach for constructing and fusing multi-granularity features, and conduct extensive evaluations across multiple tasks to comprehensively validate its effectiveness.

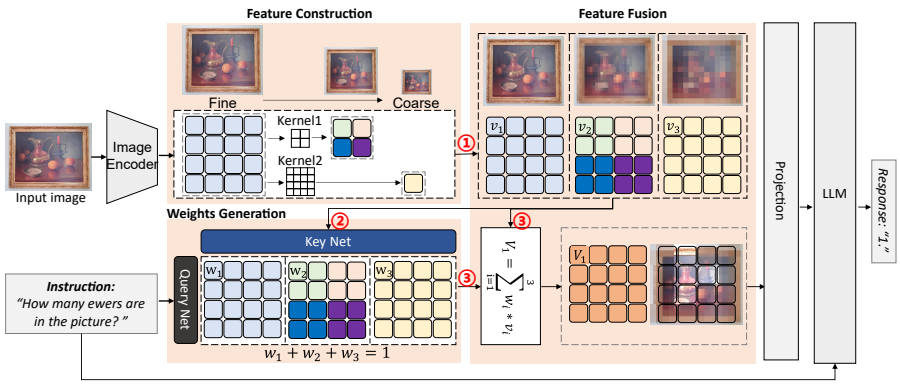

Figure 2: Architecture of our method. Feature Fusion constructs patch-level visual features at multiple granularities. Weights Generation determines the importance of features at different granularities. Feature Fusion establishes associations among multi-granularity features across image patch regions and performing weighted fusion based on their importance. Coarser granularity leads to less detailed visual information.

3 METHOD

3.1 OVERVIEW

In this paper, we aim to introduce multi-granularity understanding to enhance the visual understanding capability of the MLLM without imposing additional token overhead on the LLM. As illustrated in Fig 2, the image is first processed by an encoder to extract raw visual features, which are then fed into the Feature Construction module to generate multi-granularity visual features. Subsequently, the Weights Generation module determines the importance of each token based on the multi-granularity features and textual instructions. Finally, the Feature Fusion module integrates the features according to their importance, while ensuring that the total number of features remains the same as that of the raw visual features.

3.2 FEATURE CONSTRUCTION

First, we arrange the patch visual features extracted from the image encoder according to the 2D spatial layout of the image to obtain $V_P \in \mathbb{R}^{n \times n \times d}$, where n^2 is the number of tokens in V_P, and d is the feature dimension. Subsequently, 2D convolutional kernels $\{Kernel_1, \ldots, Kernel_M\}$ are employed to produce visual features at multiple granularities. The multi-granularity visual feature V_i is defined in Eq. 1.

$$V_i = Kernel_i(V_P). \tag{1}$$

Here, $i \in \{1, \cdots, M\}$ and $V_i \in \mathbb{R}^{n_i \times n_i \times d}$. The kernel size of $Kernel_i$ is 2^i, the stride is 2^i, and $n_i = n/2^i$. As i increases from 1 to M, n_i decreases from n to 1, and the granularity of the visual features changes from fine to coarse. Consequently, we obtain $M + 1$ groups of visual features $\{V_P, V_1, \cdots, V_M\}$ with different granularities. Because these features, like a set of Matryoshka dolls, retain a similar appearance while gradually decreasing in size, we name them Matryoshka multi-granularity visual features.

3.3 FEATURE FUSION

Due to the varying token counts of features at different granularities in $\{V_P, V_1, \cdots, V_M\}$, replication padding according to their spatial locations is adopted to ensure that all features have the same number of tokens. Coarser-grained features lose more details. The image features become more blurred in detail compared to the original image, as if they had been pixelated. Considering that features themselves are high-dimensional vectors, in order to minimize the introduction of erroneous information, the feature construction process does not involve interpolating to construct new features. Instead, it fills in the missing parts using the corresponding coarse-grained features that belong to the same spatial location. Subsequently, the padded features $\{V_P, V_1{}', \cdots, V_M{}'\}$ are fused

Table 1: Comparison with existing multi-granularity methods on the ScienceQA dataset under the same baseline model. We report the results on IMG, a task of ScienceQA designed to evaluate the visual understanding capabilities of models. MQT and M^3 refer to the methods used by MQT-LLAVA and LLaVA-1.5-M^3, respectively. Bold numbers denote the best results.

Setting	Baseline	+ MQT (Hu et al., 2024)	+ M^3 (Cai et al., 2025)	+ Ours
Acc (%)	91.22	85.97	89.34	**94.35**

Table 2: Abaltion studies on the ScienceQA dataset.

Feature Construction	Feature Fusion	Weights Generation	Acc (%)
√	√	√	94.35
√	√	×	93.60
√	×	×	92.81
×	×	×	89.89

on a per-patch basis using the weights $\{W_P, W_1, \cdots, W_M\}$ generated by the weights generation network. The fused visual features V are defined in Eq. 2.

$$V = V_P \cdot W_P + \sum_{i=1}^{M} V_i^{'} \cdot W_i. \tag{2}$$

Here, $W_P \in \mathbb{R}^{n\times n}$ and $W_i \in \mathbb{R}^{n\times n}$. The number of fused visual features $V \in \mathbb{R}^{n\times n}$ remains the same as that of the raw features V_P, imposing no additional overhead on LLM. The weights are required to satisfy certain constraints $W_P + \sum_{i=1}^{M} W_i = \mathbf{1}_{n\times n}$.

3.4 Weights Generation

The weight generation network assigns weights to $V_i^{'}$ corresponding to the regions of the image based on the text instruction and the visual features. The weights are defined in Eq. 3.

$$W = softmax(KQ). \tag{3}$$

Here, $W = \{W_p, W_1, \cdots, W_M\}$. The query vector $Q \in \mathbb{R}^d$ is obtained by averaging the instruction token embeddings. The key vector K is $\{V_P, V_1^{'}, \cdots, V_M^{'}\}$. For text vector acquisition, the averaging operation over word embeddings introduces no additional trainable parameters and achieves satisfactory performance.

3.5 Model Training

Training pipeline The training pipelines of the two experiments are different. In the ScienceQA experiments, the training pipeline contains only the instruction fine-tuning stage to reduce training cost and rapidly validate the effectiveness of our method. In the multimodal chatbot experiments, the training pipeline typically consists of two sequential stages: pre-training and instruction fine-tuning. In the pre-training stage, only the projection network is updated to map visual features into the LLM embedding space, thereby fulfilling the fundamental cross-modal alignment objective. In the instruction fine-tuning stage, the model is trained to generate coherent and contextually useful responses conditioned on a given image and user-provided instructions. This process effectively integrates visual understanding with robust language reasoning and conversational abilities, enabling the model to perform competently across a diverse range of tasks.

Training objective In all experiments, the autoregressive objective is employed to optimize the model parameters θ. Given the fused visual features V, text instruction features T, and the target text features $R = \{R_1, \cdots, R_L\}$, the loss function is defined as:

$$loss = -\sum_{j=1}^{L} log(R_j | V, T, R_1, \cdots, R_{j-1}; \theta) \tag{4}$$

Table 3: Efficiency of Feature Fusion module.

Settings	+w/ Fusion	+w/o Fusion	Overhead
Visual token number	256	341	+ 33.20%
Traing time/epoch	64min	72min	+ 12.50%
Inference time	60min	78min	+ 30.00%
GPU memory	21.31G	22.84G	+ 7.19%

Table 4: The general applicability of our method across different image encoders and LLMs (%).

Settings	Image Encoder	LLM	Acc (%)
Ours	CLIP	Mistral-7B	**94.35**
+ encoder	OpenCLIP	Mistral-7B	91.87
	DFN-CLIP	Mistral-7B	93.60
	EVA-CLIP	Mistral-7B	90.43
+ LLMs	CLIP	DeepSeek-7B	91.97
	CLIP	Qwen2.5-7B	93.31
	CLIP	Qwen3-4B	91.87

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Two sets of experiments are conducted: (1) training a specialized model on ScienceQA to evaluate the effectiveness of our method quickly; and (2) training a multimodal chatbot and evaluating its zero-shot performance on multiple benchmarks. The adapter (Bai & Bai, 2025) is inserted into the pre-trained models to reduce training costs. The projection is an MLP. In the ScienceQA experiments, we evaluate the model on ScienceQA (Lu et al., 2022) using the accuracy on the IMG task. the training epochs, batch size, learning rate, and adapter bottleneck dimension are 20, 32, 6e-3, and 24, respectively. The image encoder is CLIP ViT-L/14, and the LLM is Mistral-v0.2. The model is trained on 2 RTX4090D (24GB). In the chatbot experiments, there are two training stages: pre-training the projection for modality alignment on LLaVA-558K (Liu et al., 2024a) and instruction fine-tuning on LLaVA-665K (Liu et al., 2024a). The image encoder is ViT-L/14, and the LLM is Qwen2.5-7B. In the first stage, the number of training epochs, batch size, and learning rate are 1, 32, and 5e-3, respectively. In the second stage, the number of training epochs, batch size, and learning rate are 1, 32, and 5e-4, respectively. The adapter bottleneck dimension is 128. The model is trained on 8 RTX4090D (24GB) for 75 hours and evaluated on multiple benchmarks: AI2D (Kembhavi et al., 2016), VQAv2 (Goyal et al., 2017), $\text{MMB}_{en/cn}$ (Liu et al., 2024b), SEED_I (Li et al., 2024), VizWiz (Gurari et al., 2018), POPE (Guan et al., 2024), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019), MMMU (Yue et al., 2024), CMMMU (Zhang et al., 2024a).

4.2 RESULTS ON SCIENCEQA

Comparison with existing multi-granularity methods Table 1 presents the comparison between our method and advanced multi-granularity methods under the same baseline model. We report results on IMG, a ScienceQA task for evaluating the visual understanding capabilities of models. The baseline model is configured as follows: the image encoder ViT-L/14, the MLP projection network, and the LLM Mistral-v0.2. Our method achieves improvements over existing methods, confirming its effectiveness.

Ablation studies Table 2 presents ablation studies of the three modules in our method. When the Weights Generation module is ablated, all tokens share the same weight value. The accuracy drops from 94.35% to 93.60%. When the Feature Fusion component is ablated, the weights generation network also becomes ineffective. The constructed features are concatenated, which increases the token count and introduces considerable overhead; the details are shown in Table 3. The accuracy on the IMG task drops from 94.35% to 92.81%. When the Feature Construction module is ablated,

Table 5: Comparison with representative MLLMs on the ScienceQA dataset. P-D, T-P, and LLM-P denote the numbers of pre-training samples, tuning parameters, and LLM parameters, respectively. Bold and underlined numbers denote the best and second-best results, respectively.

Models	P-D	T-P	LLM-P	Acc (%)
Fully fine-tuned MLLMs.				
EfficientLLaVA (Liang et al., 2025)	558k	7B	7B	86.47
LLaVA (Liu et al., 2023)	558k	13B	13B	88.00
LLaVA+GPT-4 judge (Liu et al., 2023)	558k	13B	-	88.99
Honeybee (Cha et al., 2024)	200M	13B	13B	<u>93.75</u>
Parameter-efficient MLLMs.				
LLaMA-Adapter (Zhang et al., 2024b)	0k	1.8M	7B	80.32
LaVIN (Luo et al., 2023)	0k	5.4M	13B	87.65
PILL (Yin et al., 2024)	595k	45M	7B	88.65
Ours	0k	28.9M	7B	**94.35**

Table 6: Comparison with existing multi-granularity methods across multiple benchmarks using the same baseline model and 10% of the instruction fine-tuning data.

Methods	Baseline	+ MQT (Hu et al., 2024)	+ M^3 (Cai et al., 2025)	+ Ours
General VQA Benchmarks.				
MMB_{en}	60.7	51.2	60.2	**64.7**
MMB_{cn}	60.8	51.6	59.9	**64.0**
$SEED_I$	56.7	49.6	57.2	**65.4**
VizWiz	53.0	52.3	49.3	**57.4**
VQAv2	64.7	58.6	64.8	**70.8**
Reasoning Benchmarks.				
GQA	51.6	49.6	51.9	**57.9**
A-OKVQA	73.0	70.7	74.1	**76.4**
AI2D	62.7	62.7	63.2	**63.2**

our method is ineffective. The accuracy drops from 94.35% to 89.89%. In summary, the Feature Construction module is the cornerstone of our method; the Weights Generation module determines the effectiveness of feature fusion; and the Feature Fusion module ensures that our method does not increase the number of LLM input tokens or training samples, thereby effectively controlling the introduced overhead. Table 4 presents the general applicability of our method across different image encoders and LLMs. Our method achieves over 90% accuracy across all image encoder and LLM configurations, demonstrating broad applicability.

Comparison with representative MLLMs Table 5 presents a comparison between our model and representative MLLMs. Our model achieves the best performance among these MLLMs. The state-of-the-art MLLM, Honeybee, achieves the state-of-the-art result of 93.75% using 200M pre-training samples and fully fine-tuning the 13B LLM. Compared with Honeybee, our model achieves a 0.6% absolute improvement with 28.9M trainable parameters and a 7B LLM, which confirms the effectiveness of our model. Compared with parameter-efficient MLLMs, such as LaVIN, LLaMA-Adapter, and PILL, our model achieves significant improvements.

4.3 Results on Multimodal Chatbot

Comparison with existing multi-granularity methods Table 6 presents a comparison between our method and existing multi-granularity methods across multiple benchmarks using the same baseline model and 10% of the instruction fine-tuning data. The baseline model is configured as follows: the image encoder is ViT-L/14, the projection network is an MLP, and the LLM is Qwen2.5-7B. In general, our method achieves consistent performance improvements across multiple general visual understanding benchmarks, confirming the effectiveness of our method. MTQ-LLAVA and LLAVA-1.5-M^3 exhibit performance degradation on most benchmarks. Finally, Table 8 shows the results on the hallucination benchmark POPE. The results indicate that our method effectively avoids the risk

Table 7: Zero-shot performance on multiple benchmarks. P-D, T-D, T-P, and LLM-P denote the numbers of pre-training samples, instruction fine-tuning samples, tuning parameters, and LLM parameters, respectively. "–" means the result is not available.

Methods	BLIP-2	InstructBLIP	LLaVA-1.5	MQT-LLAVA	LLaVA-1.5-M^3	Ours
Settings.						
Res	224	224	336	336	336	224
P-D	129M	129M	558k	558k	558k	558k
T-D	–	1.2M	665k	665k	665k	665k
T-P	1.2B	188M	7B	7B	7B	131M
LLM-P	13B	13B	7B	7B	7B	7B
VQA Benchmarks.						
MMB$_{en}$	–	44.0	64.3	64.3	<u>65.9</u>	**70.3**
MMB$_{cn}$	–	–	58.3	55.4	<u>59.9</u>	**68.8**
SEED$_I$	–	58.8	<u>66.1</u>	62.5	64.9	**69.7**
VizWiz	19.6	33.4	50.0	53.1	<u>54.9</u>	**60.3**
VQAv2	65.0	–	**78.5**	76.8	<u>77.7</u>	**78.5**
Reasoning Benchmarks.						
AI2D	–	–	54.8	54.3	<u>55.7</u>	**65.3**
MMMU	35.4	<u>35.7</u>	35.3	34.8	33.2	**43.3**
CMMMU	–	–	21.8	22.4	<u>23.9</u>	**39.6**
GQA	41.0	49.2	<u>62.0</u>	61.6	61.9	**64.0**
A-OKVQA	72.1	75.7	**81.1**	77.2	<u>77.5</u>	**81.1**

Table 8: Comparison with existing multi-granularity methods across multiple benchmarks using the same baseline model and 10% of the instruction fine-tuning data.

Setting	Baseline	+ MQT Hu et al. (2024)	+ M^3 Cai et al. (2025)	+ Ours
POPE	81.2	77.1	79.8	**82.8**

of hallucination amplification that may arise from constructing multi-granularity features, while simultaneously achieving performance improvement.

Comparison with representative MLLMs Table 7 compares the performance among representative MLLMs, BLIP-2 (Li et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA-1.5 (Liu et al., 2024a), and advanced multi-granularity MLLMs, MQT-LLAVA (Hu et al., 2024) and LLaVA-1.5-M^3 (Cai et al., 2025). It is commonly believed that full fine-tuning, larger data scales, and higher resolutions lead to better performance. For instance, BLIP-2 and InstructBLIP introduce large-scale training data and achieve promising results. LLaVA-1.5 increases the resolution of LLaVA from 224×224 to 336×336. Our model adopts the same training pipeline and training data as LLaVA-1.5, MQT-LLAVA, and LLaVA-1.5-M^3. Due to GPU memory limitations of RTX4090 (24G), the CLIP with a resolution of 224×224 is applied instead of the 336×336. In general, our model achieves competitive results on all benchmarks, demonstrating its performance advantage.

4.4 Analysis

As shown in Figure 3, the granularity distributions on each benchmark are present for visually analyze. First, all granularity levels are indispensable. Layer 0 weights range from 34% to 40%, while Layers 1–4 contribute about 60% in total, each exceeding 12%. Second, the distributions vary notably across different benchmarks. Benchmarks of similar task types exhibit comparable distribution patterns: CMMMU and MMMU share highly similar weights; OKVQA, A-OKVQA, and GQA show closely related distributions, all with a preference for the coarse-grained Layer 4 features; VizWiz, characterized by low-quality images, tends to rely more on the fine-grained Layer 0 features; and general VQA tasks such as MMB$_{en/cn}$, SEED$_I$, VQAv2, and POPE present a distinct pattern where Layer 2 and Layer 3 have relatively low and similar weights, while Layer 4 is assigned higher importance. In summary, the weight variations across benchmarks indicate that different tasks

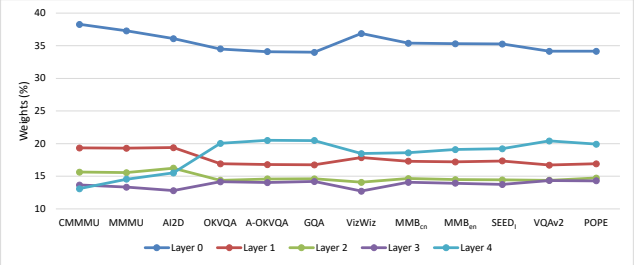

Figure 3: Comparison of granularity distribution in different tasks. There are five hierarchical levels of visual features (denoted as Layer 0 through Layer 4), progressing from fine-grained to coarse-grained representations.

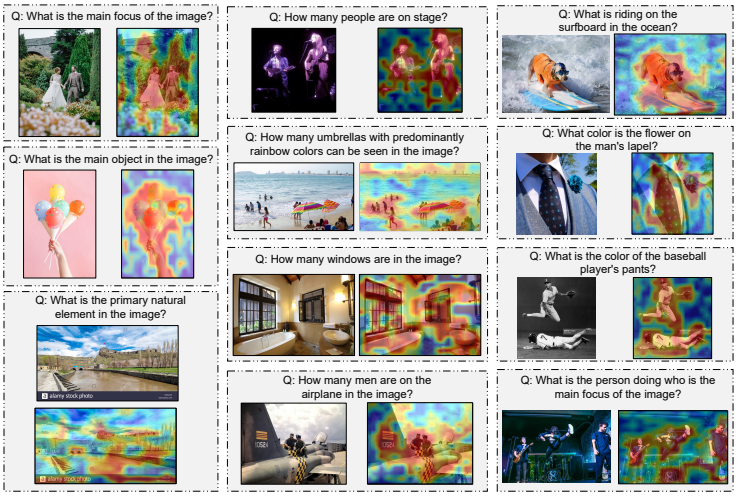

Figure 4: Granularity heatmaps. The higher intensity indicates a greater proportion of fine-grained features.

demand different granularities of visual features, underscoring the role of multi-granularity visual features for MLLMs in downstream applications.

Figure 4 presents heatmaps for different image-text instruction pairs. First, granularity heatmaps vary across samples, showing that our method adaptively generates diverse multi-granularity features. Second, regions relevant to the text or question show higher intensity (finer granularity), while irrelevant regions show lower intensity (coarser granularity), matching our expectation that task-related regions need fine-grained observation and unrelated regions need only coarse scanning. Third, our method shows some reasoning ability: in the upper-left sample, although the question does not mention "dog", it still identifies this region as the key target and assigns finer granularity to it. Note that our method operates before the LLM input, providing preliminarily selected multi-granularity features that emphasize richness and completeness rather than exact alignment with the final answer.

5 Conclusion

In this paper, we propose a novel patch-level Matryoshka multi-granularity visual feature generation and fusion method for MLLMs. Unlike previous studies that focus on improving inference efficiency, we explore and demonstrate that multi-granularity understanding contributes to enhancing visual comprehension capabilities. Extensive experiments on ScienceQA and multimodal chatbot tasks validate the effectiveness of our method. A promising direction for future work is to extend our patch-level method to the pixel level for finer visual understanding.

AI USE STATEMENT

We have not used generative AI tools for generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zixing Bai and Yuting Bai. Exploring the role of clip global visual features in multimodal large language models. In *2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

Stefania Bracci and Hans P Op de Beeck. Understanding human object vision: a picture is worth a thousand representations. *Annual review of psychology*, 74(1):113–135, 2023.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 46254–46272, 2025.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13817–13827, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, et al. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3d-llava: Towards generalist 3d lmms with omni superpoint transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782. IEEE, 2025.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 57607–57624, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, et al. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, June 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Danna Gurari, Qing Li, Abigale J Stangl, et al. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Wenbo Hu, Zi-Yi Dou, Liunian H Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. *Advances in Neural Information Processing Systems*, 37:50168–50188, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition*, 2019.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, et al. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Bohao Li, Yuying Ge, Yixiao Ge, et al. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yinan Liang, Ziwei Wang, Xiuwei Xu, Jie Zhou, and Jiwen Lu. Efficientllava: Generalizable auto-pruning for large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9445–9454. IEEE, 2025.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Pan Lu, Swaroop Mishra, Tanglin Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022.

Gen Luo, Yiyi Zhou, Tianhe Ren, et al. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36:29615–29627, 2023.

OpenAI. Chatgpt. `https://openai.com/blog/chatgpt/`, 2023a.

OpenAI. GPT-4V(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023b.

OpenAI. Gpt-4 technical report, 2023c.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, et al. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Bowen Shi, Peisen Zhao, Zichen Wang, et al. Umg-clip: A unified multi-granularity vision generalist for open-world understanding. In *European Conference on Computer Vision*, pp. 259–277. Springer, 2024.

Crystal L Sigulinsky, Rebecca L Pfeiffer, and Bryan William Jones. Retinal connectomics: a review. *Annual Review of Vision Science*, 10(1):263–291, 2024.

Wentao Xiang, Haoxian Tan, et al. Advancing visual large language model for multi-granular versatile perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22153–22164, 2025.

Yuyu Yin, Fangyuan Zhang, et al. Pill: Plug into llm with adapter expert and attention gate. *Applied Soft Computing*, 165:112115, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Ge Zhang, Xinrun Du, Bei Chen, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*, 2024a.

Renrui Zhang, Jiaming Han, Chris Liu, et al. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024b.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305. IEEE, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.