Leveraging Image as Compressed Visual Prompt and Hierarchical Visual Knowledge for Effective Image Utilization in MLLMs

Shezheng Song[1], Kangcheng Ding[2], Shan Zhao[2], Shasha Li[*1], Xiaopeng Li[1]
Chengyu Wang[3], Qian Wan[*4], Bin Ji[1], Jie Yu[1]
[1]National University of Defense Technology
[2]Hefei University of Technology
[3]Hunan University
[4]Central China Normal University
betterszsong@gmail.com, zhaoshan@hfut.edu.cn, lishasha@nudt.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) integrate text and images for complex reasoning tasks, but efficiently utilizing image remains a challenge due to redundancy and noise. Traditional methods take the entire image features as visual prompt into the MLLMs, leading to excessive visual tokens that disrupt textual information expression. Thus, recent studies treat image features as visual knowledge, storing them in the feed-forward network for retrieval when needed. These methods, completely removing images from the input, may hinder the activation of image-related knowledge. Besides, current visual knowledge focuses on fine-grained details but overlooks the hierarchical process of visual perception. As described in feature integration theory, global structure is first processed before details are integrated. Ignoring this process may lead to a fragmented visual understanding, making it difficult to capture high-level semantic relationships. To overcome these issues, we propose a novel image utilization mechanism in MLLMs. We leverage a compression-based attention mechanism to generate the compressed visual prompt, which not only mitigates the interference of excessively long visual prompts but also preserves crucial visual information necessary for activating knowledge in the MLLM. Furthermore, we extract hierarchical visual features as visual knowledge using wavelet transforms, allowing the model to capture both global structures and fine-grained details. Experiments show that our approach achieves state-of-the-art performance.

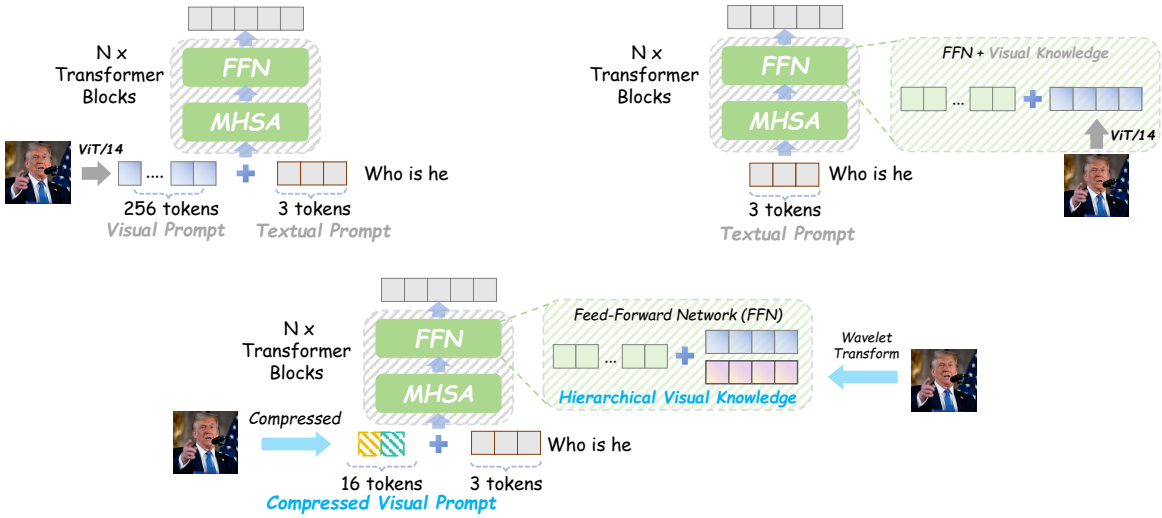
1 Introduction

With the rapid development of large language models (LLMs), researchers have increasingly explored their applications in the multimodal domain, leading to the emergence of multimodal large language models (MLLMs). These models can process multiple modalities, including text, images, audio, and video. Images serve as a crucial source of information, which is helpful for the understanding of complex scenarios such as visual question answering (Marino et al. 2019; Song et al. 2024) and image caption (Lin et al. 2014; Song et al. 2025).

Although images are crucial, the inherent noise and the modality gap between images and text require careful consideration of their usage. Currently, there are two major approaches to incorporating images in MLLMs. *(a) Input Visual Features (IVF)* : As shown in Figure 1a, traditional MLLMs (Liu et al. 2023b; Dai et al. 2023) typically encode raw images using vision transformer (ViT) (Radford et al. 2021) and then incorporate the extracted visual features as visual prompt for LLM. For example, in ViT/14, images are encoded with a 14×14 patch scheme, producing a sequence of up to 196 tokens. In contrast, the average text length is 12.84 tokens in the public dataset GQA (Hudson and Manning 2019) and 6.3 tokens in VQAv2 (Goyal et al. 2017). In fact, text serves as the primary carrier of human semantic expression (Jie et al. 2024). The long visual tokens cause image features to dominate the input sequence, potentially diminishing the model attention on textual information. *(b) Visual Knowledge (VK)*: To alleviate the impact of long visual tokens on textual information expression, researchers (Gao et al. 2023; Jie et al. 2024) remove image from input stage and instead treat it as an external visual knowledge in feed-forward network (FFN) for retrieval. As shown in Figure 1b, this approach allows the model to retrieve visual information only when necessary. However, completely ignoring images at the input stage may lead to insufficient knowledge activation. Images contain crucial information that activates knowledge within the MLLM. Relying solely on text for semantic modeling may lead to misunderstandings, especially in tasks that require integration between text and images. Therefore, it is necessary to utilize image as a compressed visual prompt to activate knowledge of LLM while avoiding excessively long visual prompts that may disrupt textual expression.

Although using compressed visual prompt could mitigate the interference introduced by long image inputs, it may also lead to the loss of visual information. Therefore, it is necessary to incorporate complete image information as visual knowledge into the LLM for retrieval. However, current visual knowledge is typically the feature encoded from raw image. The feature contains a large amount of fine-grained details, which may cause the model to focus excessively on local details while neglecting the overall structural understanding of the image. As revealed by feature integration theory (Treisman and Gelade 1980), there are two stages of attentional cognition. The first is the preattentive stage,

*Corresponding Author



(c) Hierarchical visual knowledge and compressed visual prompt (Ours).

Figure 1: Framework comparisons between previous methods and our method. MHSA is Multi-Head Self-Attention.

where basic visual features such as shapes and contours are distinguished, forming an initial perceptual representation of the image. The second is the focused attention stage, in which details are selectively integrated to form a complete perception of the image. For instance, in the process of analyzing a building, the overall structure is typically perceived first, followed by attention to specific rooms or architectural details. If the model lacks awareness of global structure, its understanding of text-image relationships may be limited.

To address these issues, as shown in Figure 1c, we propose a novel method, named IPK, that leverages Image as compressed visual Prompt and hierarchical visual Knowledge for effective image utilization in MLLMs. Our method consists of two key components: **(1) Compressed Visual Prompt:** At the input stage, retaining important image information is crucial. Omitting visual prompt may hinder knowledge activation in the MLLM while excessive visual prompt could disrupt attention balance. Thus we propose a compression-based mechanism for compact representations that preserve key visual semantics while reducing redundancy. Specifically, we first extract more compact abstract image features through pooling. Then, based on the abstract image features, we learn important information from the original image representations, ensuring that key information is effectively compressed into the abstract features. This approach enables the model to retain the most relevant visual information in a compressed form, thereby ensuring activation of relevant knowledge while reducing computational complexity and minimizing noise. **(2) Hierarchical Visual Knowledge:** We introduce the contour-aware and detailed image features as visual knowledge in FFN. The hierarchical visual knowledge ensures that the model perceives both the global structure and essential details of an image. Specifically, we employ wavelet transform to process images and extract key contour features. These features, together with the original image features (detail features), serve as visual knowledge to provide hierarchical information about the image. The contour features provide fundamental shape and contour information to the model during the first stage of feature integration theory (Treisman and Gelade 1980). In the second stage, the detail features contribute finer details, enabling the model to form a complete perception of the image. By combining compressed visual prompt in the input stage with hierarchical visual knowledge in FFN, our method could alleviate the redundancy and noise while keeping the visual information in MLLM. Our contributions are summarized as follows:

- We leverage a compression-based attention mechanism to generate the compressed visual prompt, which not only mitigates the interference of excessively long visual prompts on text but also preserves crucial visual information for activating knowledge in the MLLM.
- We introduce a wavelet-based hierarchical representation that extracts global contours and fine-grained details, addressing the lack of structural awareness in previous methods. The method aligns with human visual perception, enabling to capture hierarchical visual information.
- We achieve state-of-the-art results on multiple multimodal benchmarks, demonstrating the effectiveness of our approach in optimizing MLLM image utilization.

2 Related Work

Traditional methods for utilizing images typically involve directly using images as inputs to MLLMs. Specifically, (1) models such as LLaVA (Liu et al. 2023b), CogVLM (Wang et al. 2025), Qwen-VL (Bai et al. 2023) encode images through encoders like CLIP (Radford et al. 2021), then con-

catenate the encoded image features with text features as input to the MLLM. However, this approach introduces excessively long visual features. The long image token sequences can shift attention away from text, weakening textual reasoning. Moreover, since LLMs are primarily pretrained on textual data, their ability to interpret visual features is limited. Directly inputting visual features can therefore easily lead to hallucinations in the model. (2) models like BLIP-2 (Li et al. 2023), InstructBLIP (Dai et al. 2023), and Flamingo (Alayrac et al. 2022) design image-aware modules, such as VAE and QFormer, to more effectively leverage image information by inputting additional image features that are relevant to the task. This approach alleviates the issue of excessively long visual features to some extent. However, ensuring that the perceiver effectively captures meaningful visual information remains challenging.

A different method focuses on more efficient ways to use images, avoiding direct input of the entire image. LAVIN (Luo et al. 2023) builds an adapter to dynamically control the expression of visual and textual information across different samples, flexibly adjusting their importance, thereby effectively controlling the input length and selective expression of features. LLaMA-Adapter (Zhang et al. 2023; Gao et al. 2023) insert image information into the transformer layer of the LLM through zero-init attention, enabling closer interaction between visual and textual information. MemVP (Jie et al. 2024) adopts a different strategy by removing image information from the input and instead integrating images as additional knowledge into the FFN of MLLM. When the LLM requires image information, it can retrieve the relevant content from the FFN. However, these methods still overlook the impact of image information on the activation of internal knowledge in the MLLM and neglect the hierarchical structure of the image information itself.

3 Preliminary

3.1 Reformulation of FFN

The core of an LLM consists of multiple layers of MHSA (Multi-Head Self-Attention) and FFN (Feed-Forward Network), typically trained with Layer Normalization and Residual Connections. Specifically, the FFN consists of two fully connected layers. The process of passing $x \in \mathbb{R}^d$ through the FFN is as follows:

$$\text{FFN}(x) = \phi(xW_1)W_2. \tag{1}$$

in which ϕ is activation function, $W_1 \in \mathbb{R}^{d\times D}$ and $W_2 \in \mathbb{R}^{D\times d}$ are the weight matrices. In fact, W_1 and W_2 can be rewritten as:

$$W_1 = (k_1, k_2, \ldots, k_D), \quad W_2 = (v_1, v_2, \ldots, v_D)^T. \tag{2}$$

in which $k_i \in \mathbb{R}^d$ and $v_i \in \mathbb{R}^d$ are key and value, respectively. Then, the formulation of FFN can be reformulated as:

$$\text{FFN}(x) = \sum_{i=1}^{D} \phi(\langle x, k_i\rangle) \cdot v_i. \tag{3}$$

Therefore, the FFN can be interpreted as using input x as the query to calculate its similarity with keys, and gathering values based on the similarity (Jie et al. 2024). Previous work has found that FFN acts as a key-value memory storing factual knowledge (Geva et al. 2020).

3.2 Visual Knowledge Retrieval

FFN performs retrieval from its key-value memory. The retrieval process for visual features could be formulated as:

$$\text{Retrieval}(x) = \sum_{i=1}^{n} \phi(\langle x, K(z_i)\rangle) \cdot V(z_i) \tag{4}$$

where $K(z_i) = f(z_i), V(z_i) = f(z_i) \in \mathbb{R}^d$ are the key and value corresponding to visual feature z_i. f is a projector, which projects the visual features to the dimension of the textual token. Note that the formulation Equation (4) shares a similar form and performs a similar process as with Equation (3). From the perspective of FFN, $(K(z_i), V(z_i))$ are regarded as new memory entries to complement vision-related knowledge that language models used to lack. The new visual key-value entries are inserted into memory,

$$\text{FFN}(x) = \sum_{i=1}^{D} \phi(\langle x, k_i\rangle) \cdot v_i + \sum_{i=1}^{n} \phi(\langle x, K(z_i)\rangle) \cdot V(z_i). \tag{5}$$

Since the size of FFN's key-value memory D is usually much larger than the number of visual features n ($D = 11008$ in LLaMA-7B and $n = 196$ for ViT-L/14), the computation of retrieving visual features is insignificant (Jie et al. 2024).

4 Method

As shown in Figure 2, (a) we firstly extend the utilization of image information by introducing hierarchical visual knowledge. Wavelet transform extracts contour features, which are combined with detail features from the encoded raw image. Both are fed into the FFN for visual knowledge retrieval. Contour features capture the global structure, while detail features preserve high-resolution local details. (b) Furthermore, we optimize the injection of visual prompt into textual prompt. Given that overly long visual representations could disrupt textual information expression, we propose a visual representation compression method: The complete visual representation is compressed to generate a more compact feature for LLM knowledge activation. This strategy ensures that text remains the dominant modality in multimodal integration while retaining the essential visual information. In summary, we propose an effective image utilization strategy that uses a compressed visual prompt at the input stage to activate relevant knowledge within the MLLM, while the complete hierarchical visual information is placed in the FFN layer for retrieval.

4.1 Hierarchical Visual Knowledge

Given an image with three color channels (Red, Green, and Blue), we first separate the image into its individual RGB channels. Each channel is then processed independently using a Haar wavelet transform at level ρ. To extract the approximate contour information of an image, we perform the

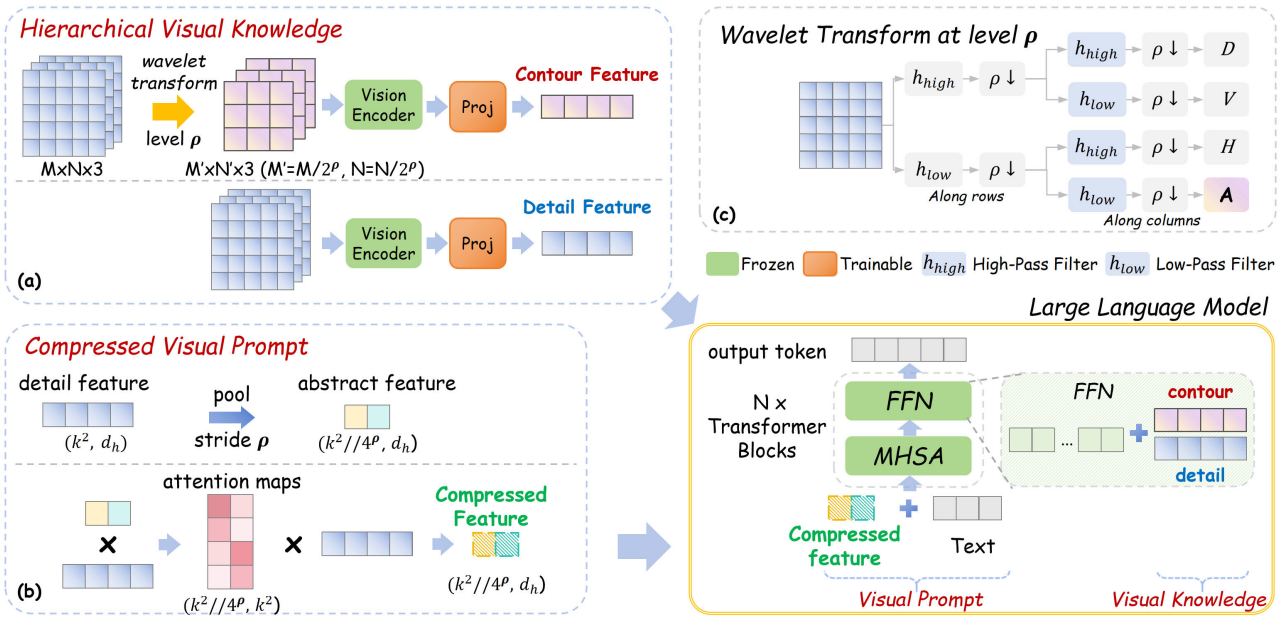


Figure 2: Method overview. Our method utilizes (a) hierarchical visual knowledge, incorporating contour features that depict shape outlines and detail features that capture fine-grained details, as comprehensive visual information. We also employ compressed features as (b) visual prompt to activate relevant knowledge in MLLM. (c) The wavelet transform decomposes an image into low-frequency and high-frequency components, where a high-pass filter preserves high-frequency details, and a low-pass filter retains low-frequency information.

wavelet transform on the input image I and utilize the resulting sub-band LL as the approximation. LL is the low-frequency content that captures the overall structure. Let I be a monochrome image of size $M \times N$, where i and j denote the row and column indices, respectively. The single-level decomposition process of the Haar wavelet transform, which aims to extract the low-frequency sub-band LL, consists of both horizontal and vertical transformations. In the horizontal transformation, for each row, the mean value is computed as:

$$LL'_{i,j} = \frac{I_{i,2j} + I_{i,2j+1}}{2}. \tag{6}$$

where $0 \le i < M, 0 \le j < \frac{N}{2}$. Then LL' is transformed vertically to generate the low-frequency sub-band LL:

$$LL_{i,j} = \frac{LL'_{2i,j} + LL'_{2i+1,j}}{2}. \tag{7}$$

where $0 \le i < \frac{M}{2}, 0 \le j < \frac{N}{2}$. To obtain approximate information at different scales, a multi-level wavelet transform is applied iteratively to LL. Given a decomposition level ρ, the low-frequency sub-band at level r, denoted as $LL^{(r)}$, is recursively computed as:

$$LL^{(r+1)}_{i,j} = \frac{LL^{(r)}_{2i,j} + LL^{(r)}_{2i+1,j}}{2}. \tag{8}$$

where $0 \le i < \frac{M}{2^{r+1}}, 0 \le j < \frac{N}{2^{r+1}}, r = 0, 1, \ldots, \rho - 1$. $LL^{(0)} = I$ represents the original image. After performing ρ levels of transformation, the final $I_{\text{LL}} = LL^{(\rho)}$ sub-band serves as the approximate contour information of the image, preserving the primary low-frequency structure.

Then I_{LL} from three channels is passed through a vision encoder $\mathcal{E}_{\text{v}}$, which processes these features to extract visual information $\mathbf{v}_c$:

$$\mathbf{v}_c = \mathcal{E}_{\text{v}}(I_{\text{LL}}). \quad \in \mathbb{R}^{(k^2, d_h)} \tag{9}$$

where k represents the patch size of visual encoder (such as 14 for ViT/14), and d_h is the feature dimension. The visual information is then passed through a projector to obtain the contour feature v_c. The projector aims to project the visual information to the dimension of the textual token d. $W \in \mathbb{R}^{d_h \times d}$ are the weight matrices.

$$v_c = W\mathbf{v}_c. \quad \in \mathbb{R}^{(k^2, d)} \tag{10}$$

The contour feature v_c represents the object contours that were emphasized by the wavelet transform, facilitating the learning of shape and structure of objects.

Similarly, the original image is encoded to get detailed visual information $\mathbf{v}_d$. And then $\mathbf{v}_c$ is passed to projector to get the detail feature v_d.

$$\mathbf{v}_d = \mathcal{E}_{\text{v}}(I) \in \mathbb{R}^{(k^2, d_h)} \quad v_d = W\mathbf{v}_d \in \mathbb{R}^{(k^2, d_h)} \tag{11}$$

4.2 Visual Prompt for Knowledge Activation

Given the detail feature v_d with shape (k^2, d_h), we apply average pooling to compress the spatial dimensions. The pooling operation is performed with a stride of ρ, which matches the level of the wavelet transform.

$$v_p = \text{AVGPOOL}(v_d). \quad \in \mathbb{R}^{(k^2//2^\rho, d_h)} \tag{12}$$

where k is the patch size, and in ViT-14, k is set to 14. Then we apply an attention mechanism to fuse the information from the original detail feature v_d and the pooled feature v_p. This attention operation ensures that the useful information in the original detail feature is compressed into a smaller representation, forming the compressed feature v_z:

The compressed feature v_z is obtained through an attention mechanism applied to the detail feature v_d and the pooled feature v_p.

$$v_z = \text{softmax}\left(\frac{(W_Q v_d)(W_K v_p)^T}{\sqrt{d_q}}\right) W_V v_p. \tag{13}$$

The learnable weight matrices $W_Q \in \mathbb{R}^{d_h \times d_q}$, $W_K \in \mathbb{R}^{d_h \times d_q}$, and $W_V \in \mathbb{R}^{d_h \times d_v}$ project the detail feature v_d and pooled feature v_p into the query, key, and value spaces, respectively. d_q is the dimension of the query and key representations, used for scaling in the attention.

The attention mechanism reduces the length of the feature from k^2 to $k^2//2^\rho$. The compressed feature v_z has a shorter length, containing only the most relevant information extracted from the image. This compressed feature is used to activate knowledge in the MLLM, while simultaneously reducing interference with the representation of textual information, allowing the model to focus on the most important visual content.

4.3 Flexible Image Utilization

We flexibly utilize image information at multiple positions and granularities. Specifically, on one hand, we concatenate the compressed feature v_z with the textual feature t. The combined features are used for activating relevant knowledge in MLLM.

$$H_{\text{out}} = \text{Transformers}([v_z, t]). \tag{14}$$

On the other hand, we extract image features at two granularities, obtaining the detail feature v_d and the contour feature v_c, respectively. These two features are then added to the FFN (Equation (3)) as hierarchical visual knowledge to provide hierarchical image information.

$$\text{FFN}(x) = \sum_{i=1}^{D} \phi(\langle x, k_i \rangle) \cdot v_i + \sum_{i=1}^{k^2} \phi(\langle x, K(v_c) \rangle) \cdot V(v_c) + \sum_{i=1}^{k^2} \phi(\langle x, K(v_d) \rangle) \cdot V(v_d). \tag{15}$$

5 Experiments

5.1 Benchmarks and Details

Following the previous work (Luo et al. 2024; Liu et al. 2023b), we evaluate the accuracy of our method on GQA (Goyal et al. 2017), VQAv2 (Goyal et al. 2017), OKVQA (Marino et al. 2019), VizWiz (Gurari et al. 2018) and SQA (ScienceQA) (Lu et al. 2022). In addition, we evaluate our method on MMBench and SEED to assess its reasoning capability, and on POPE and MME-Person to evaluate its dialogue ability. Our method is based on LLaMA-7B (Touvron et al. 2023), while the image encoder is implemented using ViT/14 (Radford et al. 2021), in line with previous works for consistency and comparison. The Pytorch version is 2.0 and all experiments are conducted on RTXA800. We reproduce the result of LLaMA for comparison where the image features encoded by ViT are concatenated with the textual information at input stage. The wavelet transform level is set to 2, with detailed experiments provided in Section 5.8.

5.2 Reference Methods

The reference methods include: (1) MLLMs: *BLIP-2* (Li et al. 2023) employs a two-stage pre-training strategy using a lightweight querying transformer. *Flamingo* (Alayrac et al. 2022) leverages learning on a variety of vision-language tasks. *InstructBLIP* (Dai et al. 2023) proposes a vision-language tuning framework that leverages instruction-aware Query Transformer. *IDEFICS* (Laurençon et al. 2023) an open-source replication of the Flamingo with 80-billion-parameter. *LLaVA and LLaVA-1.5* (Liu et al. 2023b,a) are multimodal models that leverage visual instruction tuning for various tasks. (2) Parameter-efficient Methods: *LLaMA* (Touvron et al. 2023) is one of the most influential open-source LLM. *LLaMA-Adapter-v2* (Gao et al. 2023) incorporates visual features through early fusion and image-text joint training. *LAVIN* (Luo et al. 2023) introduces mixture-of-modality adaptation, using lightweight adapters and a routing algorithm. *MemVP* (Jie et al. 2024) is a novel method that injects image into the FFN of LLM.

5.3 Result Analysis

As shown in Section 4.3,our approach achieves state-of-the-art performance on question-answering benchmarks, with an average score of 71.32, surpassing all existing methods, including MemVP (68.62) and LAVIN (63.78) among parameter-efficient models, as well as LLaVA-v1.5 (64.50) and Qwen2-VL (67.30). On reasoning datasets, we also attains the best score on MMBench and SEED. For dialogue tasks, our model demonstrates competitive performance, outperforming MemVP across the evaluated datasets. The superior performance could be attributed to hierarchical visual knowledge, which efficiently integrates both detail and contour features. Besides, the compressed feature effectively activates relevant knowledge while reducing redundancy.

5.4 Ablation Study

As shown in Section 5.3, to validate the effectiveness of each component, we conduct an ablation study on the GQA, VQAv2, and Vizwiz. Method 0 (No.0) is our proposed approach, which incorporates the following key features: *Hierarchical visual knowledge*, consisting of contour-level and detail-level features. *Compressed visual prompt*: we employ an attention mechanism to compress essential information from complete image into compressed feature v_z, ensuring that critical visual features are retained. Taking the results on GQA as an example, our analysis is as follows: *(1) Effect of different visual feature selection in the input stage:* We explore the choice of visual features: (No.0) using the compressed visual feature, (No.1) using abstract visual feature and (No.2) using the complete visual feature. The result

Method	Question Answer						Reasoning		Dialogue	
	GQA	VQAv2	OKVQA	VizWiz	SQA	Average	MMBench	SEED	POPE	MME-Per
BLIP-2	41.0	41.0	45.9	19.6	74.2	44.33	-	46.4	81.3	1293.8
InstructBILP-8B	49.2	51.9	49.9	34.5	73.3	51.77	36.0	25.8	78.9	1212.8
IDEFICS	38.4	50.9	38.4	35.5	77.4	48.12	48.2	54.5	-	-
LLaVA	49.5	63.6	45.2	30.7	56.5	49.08	36.2	32.5	79.9	1305.1
LLaVA-v1.5	62.0	74.0	45.6	50.0	90.9	64.50	64.3	56.6	82.4	1510.7
InternVL-MLP	62.9	79.3	42.9	52.5	90.1	65.55	76.7	50.0	83.1	1525.1
InternVL-QLLaMA	57.7	72.3	51.0	44.5	88.6	62.82	75.4	49.8	85.2	1298.5
Qwen-VL	59.3	78.8	46.9	35.2	67.1	57.46	61.8	56.3	-	1487.0
Qwen2-VL	63.7	**79.8**	48.5	65.4	79.1	67.30	73.8	58.1	82.4	**1939.0**
LLaMA	34.0	40.1	27.1	50.3	36.2	37.54	-	26.8	-	-
LLaMA-Adapterv2	33.8	41.6	37.8	42.2	85.2	48.11	39.5	34.7	-	1328.4
LAVIN	55.2	68.6	46.1	59.6	89.4	63.78	72.0	50.5	70.1	-
MemVP	58.1	73.2	49.9	70.0	91.9	68.62	83.3	57.4	78.8	1678.3
Ours	**66.2**	75.6	**51.1**	**71.4**	**92.3**	**71.32**	**85.4**	**58.3**	**85.7**	1784.6

Table 1: Comparison with existing methods on five vision-language tasks.

	K	C	A	CV	GQA		VQAv2		Vizwiz	
					Acc	Time	Acc	Time	Acc	Time
0	✓	×	×	✓	66.2	0.27	75.6	0.38	71.4	0.26
1	✓	×	✓	×	60.9	0.26	74.9	0.37	69.3	0.26
2	✓	✓	×	×	59.8	0.74	74.6	0.82	71.0	0.76
3	✓	×	×	×	60.5	0.25	73.9	0.23	70.5	0.26
4	×	×	×	×	58.1	0.23	72.6	0.25	70.0	0.23

Table 2: Ablation results. "K" is knowledge. "A" and "C" represent abstract features and compressed features in the visual prompt. "CV" is complete visual features encoded from raw image. Acc stands for accuracy (%), and time is the average training time (s/it).

	GQA	VQAv2	Vizwiz	Avg.Time
knowledge (Ours)	66.2	75.6	71.4	0.26
abstract knowledge	62.6	69.5	64.7	0.27
compressed knowledge	61.1	70.7	59.0	0.27

Table 3: Ablation on knowledge compression method.

using the compressed feature v_z (No.0) reaches 66.20, suggesting that the compressed attention mechanism effectively captures essential information from the raw image within v_z. Besides, the abstract feature (No.1) could achieve an accuracy of 60.90, outperforming the complete image feature (No.2), which only reached 59.75. This suggests that an overly long visual prompt may disrupt textual expression, while a shorter compressed visual prompt helps the model focus on text. *(2) Effect of visual prompts:* We remove visual prompts from the input stage (No.3), where images are only used as visual knowledge. The accuracy improves from 59.75 to 60.50, demonstrating the drawbacks of an excessively long visual prompt. *(3) Effect of visual knowledge:* When we additionally remove visual information from the knowledge of FFN (No.4), accuracy dropped from 60.50 to 58.10. This highlights the effectiveness of hierarchical visual knowledge, demonstrating that multi-level visual information provides valuable support.

Figure 3: Visualization of attention on image patches.

Besides, we have conducted ablation studies on the compression approaches used for knowledge representation in Section 5.3. As indicated by the results in the table, leveraging the complete knowledge yields the highest performance, while incurring minimal additional time cost.

5.5 Attention Visualization

Figure 3 illustrates the final layer attention distribution of LLM on visual tokens when answering multimodal questions of GQA. The left column displays the original images overlaid with a 7×7 patches. The right is the heatmap reflecting attention to image regions when processing text queries. In the first example, when answering the first question, IPK focuses on the region containing the rightmost player, aligning with the term "right boy" in text. The heatmap reveals a concentration of attention in this area, indicating that IPK effectively localizes relevant visual content. In the second example, the second question guides the attention towards the head and upper body of the bird, as these features are crucial for species identification. The heatmap confirms that IPK focuses on bird while minimizing attention to background.

5.6 Case Study

We conduct a case study under four challenging scenarios: occlusion, overlapping, blur, and multi-object scenes, as shown in Figure 4. Our model consistently produces correct

Figure 4: Case study and comparision between our method and previous representative methods.

answers, while both LLaVA (Liu et al. 2023b) and MemVP (Jie et al. 2024) fail. In the occlusion case, our wavelet contour representation accurately localizes the occluded surfboard by preserving global boundaries. In the overlapping case, the visual knowledge filtering module separates objects such as plates and forks by suppressing irrelevant textures and emphasizing semantic distinctions. Under blur, the model remains robust by retaining structural cues for reliable spatial reasoning. In multi-object scenes, it correctly counts humans where others fail.

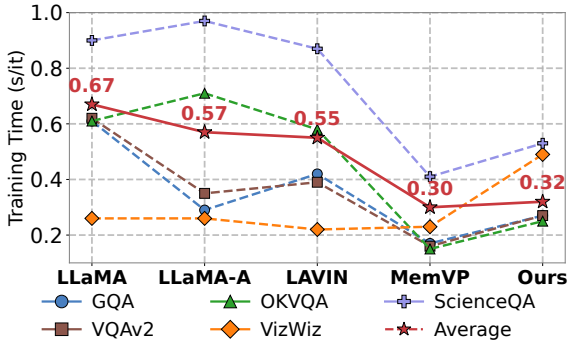


Figure 5: Average training time (s/it) across tasks.

5.7 Training Time Comparision

Figure 5 presents the average training time and the training time on different tasks, while Section 4.3 presents their accuracy. Based on the results in Figure 5 and Section 4.3, the proposed method IPK outperforms previous methods in accuracy while maintaining a competitive training speed. Specifically, IPK achieves significantly higher efficiency than LLaMA, LLaMA-Adapter-v2, and LAVIN, and is comparable to MemVP in speed. This demonstrates that IPK effectively balances accuracy and computational efficiency.

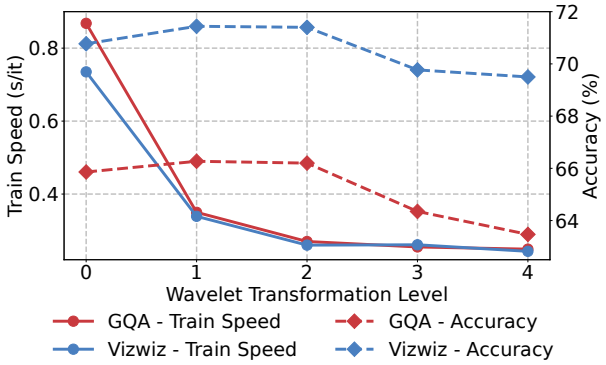


Figure 6: The effect of wavelet transformation level ρ.

5.8 Effect of Wavelet Transformation Level

As shown in Figure 6, we investigate the impact of wavelet transformation level ρ on model performance and evaluate its effect on the GQA and Vizwiz datasets. $\rho = 0$ indicates no transformation is applied to the image. The results show that when $\rho = 0$, the model exhibits lower accuracy and requires longer training time. This may be due to the model focusing excessively on fine-grained details, thereby overlooking the overall structure of the image. As ρ increases, accuracy initially improves and then declines. This trend suggests that higher ρ levels enhance the abstraction of the image, allowing the model to capture global structures more effectively. However, excessively abstracted images may lose critical information, leading to performance degradation. Meanwhile, training time consistently decreases as ρ increases, indicating that more abstract representations reduce computational complexity. To balance accuracy and training efficiency, we set $\rho = 2$, which achieves relatively high accuracy while maintaining a lower training cost.

6 Conclusion

In this paper, we propose a novel approach for enhancing image utilization in MLLMs by addressing key challenges related to visual redundancy, knowledge activation, and hierarchical perception. Our method leverages compressed visual prompt and hierarchical visual knowledge, ensuring an effective and structured incorporation of image into MLLMs. To alleviate the interference caused by excessively long visual prompts while preserving essential image semantics, we leverage a compression-based attention that generates a compact yet informative visual prompt at the input stage. This mechanism effectively balances the need for knowledge activation while reducing textual disruption. Furthermore, we incorporate wavelet-based hierarchical visual knowledge into the FFN, capturing both global structure and fine-grained details. By aligning with the hierarchical perception process, the proposed method enhances the ability to comprehend complex visual-textual relationships.

Acknowledgements

This work was supported by the National Key Research and Development Program under Grant 2024YFB4506200, the Science and Technology Innovation Program of Hunan Province under Grant 2024RC1048, and the National Key Laboratory Foundation Project under Grant 2024-KJWPDL-14. This work was supported by the National Natural Science Foundation of China under Grant 62302144, Grant 62401244, and Grant 72188101. This work was also supported by National Key Research and Development Program of China (2024YFC3308200) and National Natural Science Foundation of China (62307015, 62437002).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; Li, H.; and Qiao, Y. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jie, S.; Tang, Y.; Ding, N.; Deng, Z.-H.; Han, K.; and Wang, Y. 2024. Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning. *arXiv preprint arXiv:2405.05615*.

Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; et al. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36: 71683–71702.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Luo, G.; Zhou, Y.; Ren, T.; Chen, S.; Sun, X.; and Ji, R. 2023. Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models. *Advances in neural information processing systems (NeurIPS)*.

Luo, G.; Zhou, Y.; Zhang, Y.; Zheng, X.; Sun, X.; and Ji, R. 2024. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Song, S.; Li, X.; Li, S.; Zhao, S.; Yu, J.; Ma, J.; Mao, X.; Zhang, W.; and Wang, M. 2025. How to Bridge the Gap Between Modalities: Survey on Multimodal Large Language Model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9): 5311–5329.

Song, S.; Zhao, S.; Wang, C.; Yan, T.; Li, S.; Mao, X.; and Wang, M. 2024. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 19008–19016.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Treisman, A. M.; and Gelade, G. 1980. A feature-integration theory of attention. *Cognitive Psychology*, 12(1): 97–136.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2025. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37: 121475–121499.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.