LLaVA-UHD v2: Exploiting Hierarchical Vision Granularity in MLLMs via Inverse Semantic Pyramid

Yipeng Zhang[1*], Yifan Liu[1*], Zonghao Guo[1†], Yidan Zhang[5], Xuesong Yang[5], Xiaoying Zhang[6], Chi Chen[1], Jun Song[4], Yuan Yao[2,3†], Tat-Seng Chua[3], Maosong Sun[1]

[1]Tsinghua University
[2]Shanghai Qi Zhi Institute
[3]National University of Singapore
[4]Alibaba Group
[5]University of Chinese Academy of Sciences
[6]The Chinese University of Hong Kong
yipengzhang97@gmail.com, guozonghao96@outlook.com

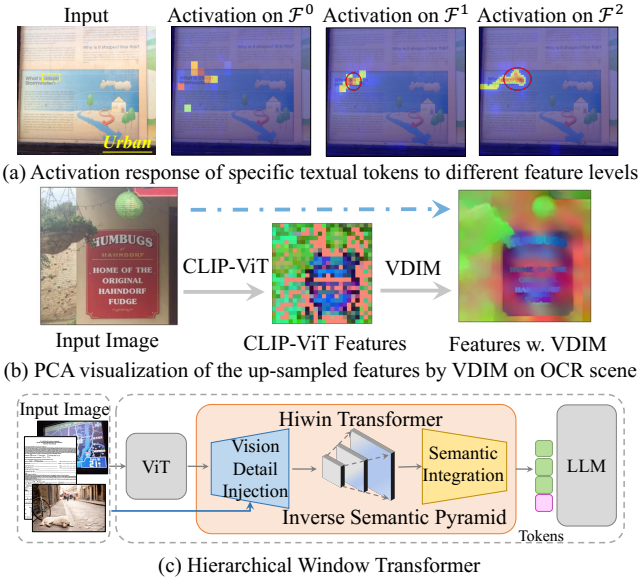
Abstract

Vision transformers (ViTs) are widely employed in multi-modal large language models (MLLMs) for visual encoding. However, they exhibit inferior performance on tasks regarding fine-grained visual perception. We attribute this to the inner limitations of ViTs in capturing diverse visual semantic levels. To address this, we present **Hi**erarchical **win**dow (**Hiwin**) transformer as a plug-and-play solution for MLLMs, centered around our inverse semantic pyramid (ISP). **Hiwin** transformer comprises two key modules: (i) a visual detail injection module, which progressively injects low-level visual details into high-level language-aligned semantics features, thereby constructing an ISP, and (ii) a hierarchical window attention module, which leverages cross-scale windows to condense multi-level semantics from the ISP. Notably, our design achieves an average boost of 3.7% across 14 benchmarks compared with the baseline method, 9.3% on DocVQA for instance.

Code — https://github.com/thunlp/LLaVA-UHD
Extended version — https://arxiv.org/abs/2412.13871

1 Introduction

Embedding visual information into large language models (LLMs) has significantly enhanced their multimodal abilities, such as visual question answering (Antol et al. 2015), document analysis (Mathew, Karatzas, and Jawahar 2021), and visual interaction (Chen et al. 2024c). Among these advancements, the CLIP (Radford et al. 2021) series, built upon the vision transformer (ViT) architecture (Dosovitskiy et al. 2021), has emerged as a paradigm for visual encoding in contemporary multi-modal large language models (MLLMs) (Dong, Zhang et al. 2024; Bai et al. 2023; Wang et al. 2024a; Liu et al. 2024b,a; Li et al. 2023b; Achiam et al. 2023; Anil et al. 2023; Yao et al. 2024).



(a) Activation response of specific textual tokens to different feature levels

(b) PCA visualization of the up-sampled features by VDIM on OCR scene

(c) Hierarchical Window Transformer

Figure 1: (a) OCR-like textual tokens yield finer-grained and more accurate activations at higher feature levels. (b) With VDIM, the high-resolution features could clearly depict object boundaries and text appearance. (c)Our Hiwin transformer to build an inverse semantic pyramid and compress it into visual tokens, providing various semantic granularity for language generation.

However, due to the low resolution and text-aligned nature, CLIP-ViT-based MLLMs often underperform in tasks requiring extensive low-level visual details, such as visual grounding (Yu et al. 2016; Chen et al. 2024a; Zhan, Zhu et al. 2024; You et al. 2023) and optical character recognition (Kim et al. 2022; Lee et al. 2023).

To remedy this, an intuitive idea is to inject fine-grained, low-level features from the input image to complement CLIP's high-level semantics. This approach, however, presents two primary challenges:

1. **Computational Overhead:** Extracting low-level details

*These authors contributed equally.
†Corresponding author.

by directly processing high-resolution images introduces prohibitive computational costs, as incorporating inputs at just $4\times$ resolution (leading to $16\times$ more tokens) can increase the Vision Transformer's (ViT) computational burden by $256\times$.

2. **Feature Fusion:** Merging high-level, language-aligned semantics from CLIP with low-level visual features is non-trivial. Naive fusion or resizing schemes often corrupt the well-aligned multimodal representations or lose critical spatial details.

To address these, we present **Hiwin** Transformer, a plug-and-play vision-enhancing solution for MLLMs, providing them with a rich and hierarchical visual granularity. As shown in Figure 1(c), **Hiwin** transformer captures diverse multi-modal granularity by constructing and compressing an inverse feature pyramid (**ISP**) from the output of ViT.

Specifically, the Hiwin transformer consists of two key modules as follows: **(i) a visual detail injection module (VDIM) for obtaining hierarchical vision granularity**. We propose VDIM to inject low-level details ($e.g.$, edges, textures) from images into text-aligned features from a CLIP-pretrained ViT, progressively building an up-sampled level upon the previous level, resulting in an inverse semantic pyramid. During the training stage, a reconstruction loss between fused and original CLIP features explicitly maintains vision-language alignment while enhancing visual granularity. This strategy can be extended to any ViT for building a feature pyramid while inheriting its powerful multi-modal representations. **(ii) a hierarchical window attention module for effective and efficient compression**. We propose utilizing a set of hierarchical windows to capture semantics from local regions across different semantic levels. A set of learnable queries is restricted to attending solely to the part of the pyramid within their respective windows. This attention mechanism performs effective compression on local dense features at the native resolution of each pyramid level, thereby enabling visual tokens to effectively capture both fine-grained visual details and high-level language-aligned semantics.

Figure 1 provides a visualization of our proposed method. As depicted in Figure 1(b), by injecting fine-grained visual details, VDIM produces up-sampled features that exhibit sharp, clearly-defined edges while preserving the original semantic integrity. Concurrently, Figure 1(a) shows that this multi-level feature representation allows OCR-like textual tokens to yield finer-grained and more accurate activations at higher feature levels, a critical factor in facilitating precise scene text recognition.

Extensive experiments demonstrate that Hiwin provides consistent growth for academic. Under our Hiwin-ISP framework, the base model (LLaVA-UHD (Guo et al. 2024)) enjoys a significant boost on 14 popular benchmarks by 3.7% in average, including document-centric visual question answering ($e.g.$, +9.3% on DocVQA), visual grounding ($e.g.$, average +5.7% on RefCOCOs (Yu et al. 2016)), and high-resolution image perception ($e.g.$, +3.4% on HR-Bench (Wang et al. 2024b)). Furthermore, our experimental findings demonstrate that ISP consistently enhances the visual perception capabilities of MLLMs across different construction methods (e.g., bilinear interpolation), providing insights for future research.

In summary, our contributions are threefold:

- We present Hiwin, an effective and efficient solution providing hierarchical visual semantics for MLLMs.
- We propose the Hiwin transformer, a novel vision-language projector that comprises a visual detail injection module and a hierarchical window attention module for capturing diverse multimodal visual granularities.
- Hiwin boasts performance growth for the base model LLaVA-UHD. Trained on merely academic-scale data, Hiwin-LLaVA-UHD achieves substantial improvements over the baseline method across 14 benchmarks.

2 Related Work

Hierarchical Representation. In deep learning, CNNs like ResNet and VGG inherently extract hierarchical features (He et al. 2016; Simonyan and Zisserman 2014), while innovations such as FPN and U-Net enhance semantic hierarchy for detection and segmentation (Lin, Dollár et al. 2017; Ronneberger, Fischer, and Brox 2015). Recent transformer-based models (Liu et al. 2021; Zhu et al. 2020) have further advanced multi-scale feature construction. However, MLLMs using CLIP-based ViTs underutilize such multi-level information, suggesting a research gap for integrating advanced hierarchical features.

Visual Encoding in MLLMs. CLIP-ViT, favored for its effective visual-linguistic alignment through contrastive pre-training, is widely adopted in MLLMs. Emerging research explores alternative visual representations in three categories: (1) Fusing CLIP-based CNNs and ViTs (Luo et al. 2024). (2) Fusing features from visual experts with different pre-training tasks (Lu et al. 2024). (3) Direct LLM encoding (Chen et al. 2024d). For more details, please refer to supplementary.

Token Projection and Compression. MLPs, perceiver resamplers (Yao et al. 2024) and Q-Formers (Dai et al. 2023) are basic projectors in modern MLLMs. Recently, various new designs have emerged: (1) Spatial-preserving compression, using linear or convolution layers for merging (Chen et al. 2023). (2) Cross-layer compression with cross attention (Li et al. 2024). (3) Merge semantically similar features (Shang et al. 2024).

3 Method

3.1 Overview

Our proposed Hiwin transformer's architecture is presented in Figure 1. We extend the LLaVA-UHD base model by integrating Hiwin, forming Hiwin-LLaVA-UHD, as detailed in Figure 2. The initial step involves processing the input image using LLaVA-UHD's adaptive slicing strategy, which produces CLIP (Radford et al. 2021) features of arbitrary size and shape. The features are passed to the Hiwin transformer for vision-language projection, which is carried out with two stages: (i) constructing an inverse semantic pyramid (ISP) and (ii) integrating the ISP by hierarchical window attention, which will be detailed in Sec.3.2 and Sec.3.3.

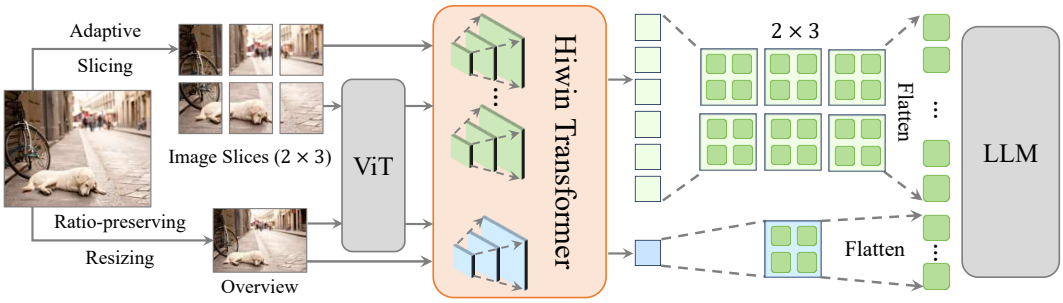


Figure 2: The overall architecture of proposed Hiwin-LLaVA-UHD, consisting of a ViT, our hierarchical window transformer (Hiwin transformer), and an LLM. The Hiwin transformer first injects high-frequency visual details from the image into the high-level semantics of ViT features, forming inverse semantic pyramids (ISP). Then it compresses the ISPs into spatially consistent tokens via cross-scale windows, for a better vision-language alignment. Details are illustrated in Figure 3 and 4.

The core of the Hiwin transformer lies in enhancing each ViT-encoded feature into a high-resolution semantic pyramid encoding, thereby achieving enriched semantic granularity for each slice or image. Followed by concatenating with the overview tokens, the visual tokens provide both high-level language-aligned semantics and high-resolution visual details for language decoding.

3.2 Inverse Semantic Pyramid

Preliminaries. Traditional convolutional neural networks (CNNs) naturally produce a pyramid of hierarchical bottom-up features $\{\mathcal{F}^l \in \mathbb{R}^{\frac{H}{p\cdot 2^l}\times\frac{W}{p\cdot 2^l}\times C}\}$. Note that l is the level index, (H, W) is the image resolution, C is the feature dimension, and p is a down-sampling ratio ($e.g.$, $p = 8$ in ResNet-50 (He et al. 2016)). In these pyramids, higher-resolution (lower-level) feature maps are rich in visual details, and lower-resolution (higher-level) ones contain abstract semantic information. However, ViTs produce single-scale feature maps ($i.e.$, $\mathcal{F}^0 \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times C}$, where $e.g., p = 14$). Lacking such feature pyramids, like CNNs, hinders their performance on MLLM tasks requiring both fine-grained and high-level visual information. Therefore, how to construct a ViT-based feature pyramid with varying semantic granularity remains a problem.

Visual detail injection module (VDIM). Due to the lack of high-resolution features, up-sampling ViT features becomes the necessary strategy to inversely construct the feature pyramid. Two simple approaches, (1) plain bilinear interpolation and (2) a deconvolution network, can be adopted. By doubling and quadrupling the last-layer feature maps, a ViT-based feature pyramid $\{\mathcal{F}^l \in \mathbb{R}^{\frac{H\cdot 2^l}{p}\times\frac{W\cdot 2^l}{p}\times C}, l = 0, 1, 2\}$ is then constructed. While effective, directly up-sampling language-aligned features from CLIP-ViT hardly introduces precise visual details, resulting in suboptimal performance, illustrated in Table 3. To address this, we design a VDIM to up-sample multi-modal semantic features guided by original image priors. Specifically, the objective of VDIM is to learn $(l-1)$ convolution layers on the image pyramid $\{\mathcal{I}^l \in \mathbb{R}^{\frac{H\cdot 2^l}{p}\times\frac{W\cdot 2^l}{p}\times 3}\}$ to capture high-frequency visual patterns of image texture for guiding the up-sampling process of semantic features, as shown in Figure 3. For each input image, its $(l+1)$-th level features is defined as

$$\mathcal{F}^{l+1} = \mathrm{Conv}\big(\mathrm{Up}(\mathcal{F}^l); \Theta^{l+1}(\mathcal{I}^{l+1})\big), \tag{1}$$

where $\mathrm{Up}(\cdot)$ denotes the up-sampling interpolation and $\mathrm{Conv}(\cdot)$ the convolutional operation on feature maps with customized kernel weights Θ^{l+1} learned on image $\mathcal{I}^{l+1}$.
Optimizing VDIM. We propose a multi-level reconstruction (MLR) loss between the higher-level feature maps $\{\mathcal{F}^1, \mathcal{F}^2\}$ and the lowest one $\mathcal{F}^0$ as

$$\mathcal{L} = \frac{1}{2}\sum_{l=1}^{2} \|\mathcal{F}^0 - \mathrm{Down}(\mathcal{F}^l; \Omega^l)\|_2^2, \tag{2}$$

where $\mathrm{Down}(\cdot)$ is a down-sampling operation with trainable weights Ω^l in each level. The proposed MLR loss drives the feature pathway to capture low-level textures while maintaining multi-modal semantics during the fusion procedure.

Construction of inverse semantic pyramid (ISP). As shown in Figure 3, VDIM acts as a progressive feature resolution expanding procedure conditioned on original image priors. During the inference, the resulting multi-level feature maps $\{\mathcal{F}^0, \mathcal{F}^1, \mathcal{F}^2\}$ form an ISP, which gathers a hierarchical multi-modal semantic representation with corresponding spatial resolutions supporting proper visual granularity.

3.3 Hierarchical Window Attention

The hierarchical nature of ISP necessitates an effective approach for compressing features at varying resolutions while maintaining cross-level spatial alignment.

Hierarchical window generation. To preserve 2D locality, we aggregate features with a set of hierarchical windows. Specifically, we first uniformly divide feature maps of each level into $N \times N$ windows, whose widths and heights are float-point values $(\frac{W}{N}, \frac{H}{N})$ rather than integers. Windows share the same "anchor" point form a set of hierarchical bounding boxes (coordinates of top-left and bottom-right) $\{\mathcal{R}^l_{i,j} \in \mathbb{R}^{1\times 4}, i, j \in 0, 1, 2...N-1, l = 0, 1, 2\}$, where l is the feature level and (i, j) the 2D index. Inside each window, we apply RoI-align (He et al. 2017) to sample key features.

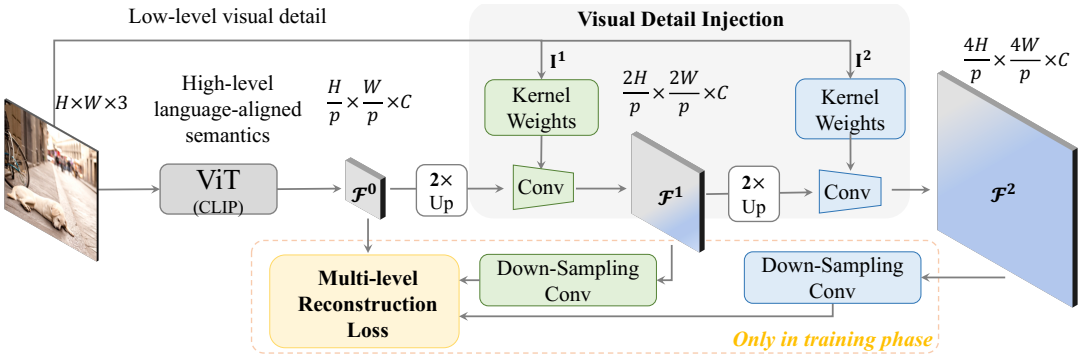


Figure 3: The flowchart illustrates the construction of the Inverse Semantic Pyramid (ISP). As the first level of ISP, $\mathcal{F}^0$ is the high-level language-aligned semantic features from CLIP-ViT. Subsequent levels, $\mathcal{F}^1$ and $\mathcal{F}^2$, are progressively built by injecting high-frequency visual details from the input image into upsampled features from the previous level, via the Visual Detail Injection Module (VDIM). A Multi-level Reconstruction (MLR) loss supervises in each scale, ensuring both text-aligned semantic coherence and fine-grained visual fidelity.

To mitigate the size distortion of feature maps caused by a difference between the aspect ratio of target RoI-aligned feature maps and the image, we define a pooling score to evaluate this difference:

$$S(W,H,r_w,r_h) = -\left|\log\frac{W}{H} - \log\frac{r_w}{r_h}\right|, \tag{3}$$

where (r_w, r_h) denotes the width and height of pooled features. By maximizing the score S, we select the optimal grid size (r_w^*, r_h^*) from pre-defined proposals $\{(3,3),(2,3),(3,2),(2,4),(4,2)\}$. Then, we carry out the RoI-align inside each window to form key feature maps for the following attention operation.

Cross-scale window querying. To compress the ISP $\{\mathcal{F}^l, l=0,1,2\}$ of one image or slice $\mathcal{I}$, we initialize a set of queries $\{\mathcal{Q}_{i,j}\in\mathbb{R}^{1\times C}, i,j\in 0,1,2...N-1\}$, each of which corresponds to the specific set of hierarchy windows $\{\mathcal{R}^l_{i,j}, l=0,1,2\}$ at its 2D position, in Figure 4 Regarding each query vector $\mathcal{Q}_{i,j}$, we prepare the key vector $\mathcal{K}^l_{i,j}\in\mathbb{R}^{(r_w^*\cdot r_h^*)\times C}$ of l-th level for window position as

$$\mathcal{K}^l_{i,j} = \text{RoI}(\mathcal{F}^l, \mathcal{R}^l_{i,j}) + \phi^l, \tag{4}$$

where ϕ^l is a level positional embedding. We then concatenate the $\mathcal{K}^l_{i,j}$ in length axis and form the final key vector $\mathcal{K}_{i,j}\in\mathbb{R}^{(3\cdot r_w^*\cdot r_h^*)\times C}$ for each window position. The corresponding value vector $\mathcal{V}_{i,j}$ is obtained in the same way yet without the level positional embedding. Thus, the cross-attention can be performed as

$$\mathcal{Q^*}_{i,j} = \text{CrossAttn}(\mathcal{Q}_{i,j}+\varphi_{i,j}, \mathcal{K}_{i,j}+\zeta_{i,j}, \mathcal{V}_{i,j}), \tag{5}$$

where $\mathcal{Q^*}_{i,j}$ denotes the updated query, and φ,ζ is the 2D spatial position embedding of query and key vector respectively. In the end, we concatenate all the $\mathcal{Q^*}_{i,j}$ into a feature map $\mathcal{P}\in\mathbb{R}^{N\times N\times C}$ according to the 2D position of its window to represent the visual token of $\mathcal{I}$.

4 Experiment

In this section, we conduct an empirical evaluation of Hiwin-LLaVA-UHD. We begin with a comprehensive outline of the implementation details of our model, followed by a comparative analysis of its performance across widely recognized benchmarks against competitive counterparts. Finally, we provide an in-depth ablation to further elucidate the capabilities and behaviors of Hiwin-LLaVA-UHD.

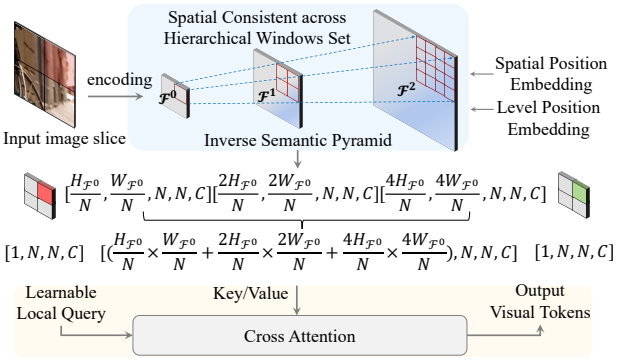


Figure 4: The flowchart of hierarchical window attention. We initialize a set of learnable queries to attend to local regions. Feature maps from the ISP are processed by a set of cross-scale windows, forming hierarchical and local-aware features at different levels. The features are then concatenated along the length axis, to serve as the key and value for the learnable queries. The output is condensed visual tokens rich in diverse and local-aware semantics.

4.1 Implementation Details

Model Setting. We adopt LLaVA-UHD (Guo et al. 2024) as the baseline method. Specifically, we employ CLIP-ViT-L/14-336 as the visual encoder, Vicuna-7B/13B (Chiang et al. 2023) or Qwen2-7B as the language model, and our proposed Hiwin transformer as the vision-language projector. We set the maximum slice number to 6 to cover a range of aspect ratios and image resolutions. The number of learnable local queries is set as 144 ($i.e.$,$N=12$). Before employing the VDIM in MLLM, we pre-train it with frozen CLIP-ViT on MS-COCO (Lin et al. 2014) with a global

Method	LLM	Data	Res.	FLOPs.	OCR & Chart				Knowledge				General					Vision Spatial	High Res.
					Doc	OCR	Chart	Text	AI2D	SQA	MMMU	Math.	GQA	SEED	MMB	MME	Star	RWQA	HR
mPLUG-Owl2(2023b)	Llama2-7B	401M	448	1.7T	-	-	-	58.2	-	68.7	-	25.5	56.1	57.8	64.5	72.5	34.8	-	-
UReader(2023a)	Llama2-7B	86M	1120	20.3T	65.4	-	59.3	57.6	-	-	-	-	-	-	-	-	-	-	-
VILA(2024)	Llama2-7B	51M	336	8.2T	-	-	-	64.4	-	68.2	-	-	62.3	61.1	68.9	76.7	-	-	-
SPHINX-2k(2023)	Llama2-7B	1.01B	762	42.2T	-	-	-	61.2	-	70.6	-	-	63.1	71.6	65.9	73.6	-	-	-
SPHINX-X(2024)	Llama2-7B	15.3M	448	21.3T	56.3	-	39.7	58.1	63.0	70.4	-	-	56.2	68.8	57.9	63.0	-	-	-
LLaVA-HR(2024)	Vicuna-7B	1.22M	1024	24.3T	-	-	-	67.1	-	65.1	-	-	64.2	64.2	-	77.7	-	-	-
Honey-bee(2024)	Vicuna-7B	52.5M	336	2.6T	-	-	-	-	-	-	35.3	-	-	64.5	70.1	77.2	-	-	-
Mini-Gemini(2024)	Vicuna-7B	3.0M	672	54.6T	61.9	47.7	47.4	65.2	68.2	69.6	36.8	-	64.5	66.9	65.8	77.3	-	51.1	50.1
Monkey(2023c)	Vicuna-7B	1.40B	1344	28.0T	66.5	51.4	65.1	67.6	62.6	69.4	38.9	33.5	60.7	64.3	59.8	73.6	37	51.6	38.0
LLaVA-1.5(2024b)	Vicuna-7B	1.22M	336	8.0T	21.8	31.8	17.8	45.5	55.5	66.8	37.0	25.5	62.0	65.8	66.5	75.3	33.1	54.8	36.1
LLaVA-Next(2024a)	Vicuna-7B	1.34M	672	44.4T	63.6	53.2	54.3	64.9	67.0	70.1	35.8	34.6	64.2	70.2	67.4	76.0	37.6	57.8	47.9
Token-Packer(2024)	Vicuna-7B	2.7M	1008	13.1T	60.2	45.2	-	68.0	-	-	35.4	-	-	67.4	74.5	-	-	-	
TextMonkey(2024c)	Vicuna-7B	1.45B	448	4.0T	66.7	-	59.9	64.3	-	-	-	-	-	-	-	-	-	-	-
LLaVA-1.5(2024b)	Vicuna-13B	1.22M	336	15.1T	-	-	-	61.3	-	71.6	-	-	63.3	61.6	67.7	76.5	-	-	-
LLaVA-Next(2024a)	Vicuna-13B	1.34M	672	67.0T	-	53.7	61.4	67.1	-	73.6	36.2	35.3	65.4	71.9	70.0	76.5	40.4	57.6	-
LLaVA-1.5(2024)	Qwen2-7B	1.22M	336	8.2T	-	-	-	-	64.9	-	40.7	33.6	62.7	69.4	72.0	76.0	-	-	-
Dense Connector(2025)	Llama3-8B	1.22M	384	11.6T	-	-	-	-	-	75.2	40.4	28.6	65.1	-	74.4	-	-	-	-
LLaVA-LLaMA3(2023)	Llama3-8B	1.22M	336	8.7T	-	-	-	-	-	73.3	36.8	-	63.5	-	68.9	-	-	-	-
PIIP-LLaVA(2025)	Llama3-8B	1.22M	1024	36.0T	-	-	-	67.1	-	68.3	-	-	63.9	69.4	67.0	-	-	-	-
MG-LLaVA(2024)	Llama3-8B	2.5M	768	33.0T	-	-	-	67.3	-	70.8	-	-	-	69.4	72.1	-	-	-	-
SliME(2024)	Llama3-8B	2.0M	2016	62.0T	-	-	-	64.8	-	-	41.2	-	63.9	-	75.0	-	-	-	-
DeepseekVL(2024)	Deepseek-7B	-	1024	-	-	45.6	-	-	-	-	36.6	36.1	-	70.4	73.2	-	37.1	-	-
Hiwin-LLaVA-UHD	Vicuna-7B	1.42M	1008	17.5T	68.1	53.9	64.5	67.6	70.5	71.3	38.2	34	65.4	70.0	68.2	74.7	40.2	58.2	51.5
Hiwin-LLaVA-UHD	Vicuna-13B	1.42M	1008	26.4T	68.2	55.6	67.4	70.0	72.4	73.3	37.7	35.2	**66.0**	71.1	70.3	73.1	42.0	59.6	55.3
Hiwin-LLaVA-UHD	Qwen2-7B	1.42M	1008	14.3T	**72.9**	**57.7**	**70.4**	**70.6**	**75.5**	**76.9**	**43.3**	**39.1**	65.1	**73.6**	**77.1**	**78.8**	**49.4**	**64.6**	**59.9**

Table 1: Main performance on popular benchmarks. Data denotes the volume of overall data during MLLM pre-training and supervised fine-tuning. "Res." is the maximum accessible resolution of MLLM. Note that our model uses a resolution of 1008×672. "Doc": DocVQA. "OCR": OCR-Bench. "Chart": ChartQA. "Text": TextVQA. "SQA": Science-QA. "MMMU": MMMU-val. "Math.": MathVista. "SEED": SEED-Image. "MME": perception sub-set of MME. "Star": MMStar. "RWQA": RealWorldQA. "HR": HR-Bench.

Method	Average	OCR & Chart				Knowledge			General				Vision Spatial		High Res.
		VQA^{D}	$Bench^{OCR}$	VQA^{C}	VQA^{T}	AI2D	SQA	$MMMU^{v}$	GQA	$SEED^{I}$	MMB	MME^{P}	RWQA	REC	$Bench^{HR}$
LLaVA-UHD (2024)	58.0	56.7	40.9	56.3	62.2	55.4	70.7	37.0	63.8	65.6	64.8	70.0	54.4	68.3	45.6
+ *VDIM(ISP)*	60.0	60.2	50.4	60.4	67.1	57.8	70.5	38.2	64.0	66.7	65.6	71.2	51.9	72.3	43.9
+ *Hiwin attention*	61.7	66.0	50.1	62.8	66.8	59.4	69.8	37.6	64.0	67.4	66.1	73.6	56.9	74.0	49.0
Δ	+3.7	+9.3	+9.2	+6.5	+4.6	+4.0	-0.9	+0.6	+0.2	+1.8	+1.3	+3.6	+2.5	+5.7	+3.4

Table 2: Ablation studies of modules in our proposed method. "Δ" denotes the overall improvement compared to the baseline. REC reports the average accuracy of RefCOCO/g/+.

batch of 16 on 8×A100. We leverage Adam optimizer with $1e^{-3}$ learning rate for 2000 steps. This is an independent phase for building a task-agnostic representation, and the weights of the ISP are always reused. Hiwin-LLaVA-UHD consists of a two-stage multi-modal training process as outlined below.

Stage 1: MLLM pre-training. In this stage, the parameters of the visual encoder, pre-trained VDIM, and LLM are frozen. We only fine-tune the parameters within the hierarchical window attention of the HiWin transformer using LLaVA-Pretrain (Liu et al. 2024b) for 1 epoch with a global batch size of 256. We employ the AdamW optimizer and a cosine learning rate scheduler. The learning rate is $1e^{-3}$ for Vicuna-7B, $2e^{-4}$ for Vicuna-13B, and Qwen2-7B. Note that, in this stage, we only encode the overview image without the slices for efficiency.

Stage 2: MLLM supervised fine-tuning. In this stage, we fine-tune all parameters except VDIM. The learning rate is $2e{-}5$ with a batch size of 128. To manage training costs, we use 825k data for analysis and ablation studies, including LLaVA-mix665k (Liu et al. 2024b) and 160k from Ureader (Ye et al. 2023a). For comparison with advanced MLLMs, we balance our data distribution and introduce an 858k-mixed dataset, detailed in the supplementary.

4.2 Experimental Setting

We present the experimental settings, detailing the benchmarks, evaluation metrics, and compared counterparts.

Benchmarks. Extensive benchmarks are used to analyze the effect of our modules. We categorize these benchmarks into the following folds: (1) General VQA benchmarks including MME (Fu et al. 2023), MMB (Liu et al.

2023a), SEED-Image (Li et al. 2023a), GQA (Hudson and Manning 2019), MMStar (Chen et al. 2024b) and HallusionBench: (Guan et al. 2024) ; (2) Knowledge-based VQA benchmarks including MMMU-val (Yue et al. 2024), Science-QA (Lu et al. 2022), AI2D (Kembhavi et al. 2016), MathVista (Lu et al. 2023); (3) OCR-based VQA benchmarks including ChartQA (Masry et al. 2022), OCR-Bench (Liu et al. 2023b), TextVQA (Singh et al. 2019) and DocVQA (Mathew, Karatzas, and Jawahar 2021); (4) Visual spatial understanding benchmarks such as Real-WorldQA (xAI 2024) and RefCOCOs (Yu et al. 2016); (5) High-resolution image perception benchmarks like HR-Bench(4K) (Wang et al. 2024b).

Evaluation Protocols. Beyond benchmark evaluations, we report additional metrics for comprehensive analysis: (1) overall volume of training data, (2) maximum supported image resolution for each method, and (3) computation cost of the entire MLLM at maximum resolution.

Counterparts. We compare our model with the advanced MLLM counterparts. (1) General MLLMs like Honeybee (Cha et al. 2024), Dense Connector (Yao et al. 2025), VILA (Lin et al. 2024) and LLaVA-1.5 (Liu et al. 2024b). (2) High-resolution MLLMs including Monkey (Li et al. 2023c), LLaVA-Next (Liu et al. 2024a), PIIP-LLaVA (Wang et al. 2025), SliME-Llama3-8B (Zhang et al. 2024), DeepseekVL-7B (Lu et al. 2024) and Token-Packer (Li et al. 2024). (3) Mixture of visual experts such as LLaVA-HR (Luo et al. 2024), SPHINX-series (Lin et al. 2023) , MG-LLaVA (Zhao et al. 2024) and Mini-Gemini (Anil et al. 2023). (4) OCR-centric MLLMs including UReader (Ye et al. 2023a) and TextMonkey (Liu et al. 2024c).

4.3 Main Performance

Table 1 showcases a comparative analysis of our proposed Hiwin-LLaVA-UHD against state-of-the-art MLLMs across 15 widely recognized benchmarks. **(1) Hiwin-LLaVA-UHD outperforms current counterparts**. Compared with general models (such as LLaVA-1.5, Dense Connector) and high-resolution MLLMs (like PIIP-LLaVA, SliME-Llama3-8 and DeepseekVL-7B), Hiwin-LLaVA-UHD demonstrates consistent improvements across various tasks, including general VQA (e.g., 77.1% on MMB and 49.4% on MM-Star), ultra-high-resolution image perception (e.g., 59.9% on HR-Bench). Notably, Hiwin-LLaVA-UHD surpasses OCR-centric models like TextMonkey on DocVQA (72.9% vs. 66.5%) and outperforms those with multiple experts (such as MG-LLaVA), achieving superior performance on general tasks like SEED (73.6%). These results underscore the value of rich semantics derived from multi-level multi-modal granularity, enhancing both the understanding and perception abilities of MLLMs. **(2) Hiwin-LLaVA-UHD indicates efficiency on data utilization and computation**. Compared to LLaVA-Next and Mini-Gemini, both operating at a 672×672 resolution, Hiwin-LLaVA-UHD supports 1.5 times the resolution (*i.e.*, 672×1008) and achieves superior performance with less than 40% of the computational cost. Furthermore, in contrast to Honey-bee and VILA, which utilize 52.5M and 51M data samples respectively, Hiwin-LLaVA-UHD attains comparable or superior performance using only ∼2.8% of the data, demonstrating the data efficiency of our model. As for the training duration, under the same model configuration and data volume, Hiwin-LLaVA-UHD requires ∼27 hours to train on 8×A100 GPUs, while LLaVA-Next needs ∼42 hours, which is well-suited for low-cost exploratory research in the academic community.

Method	Average	General		Knowledge	OCR & Chart			High Res.
		MME^P	GQA	AI2D	VQA^C	VQA^T	VQA^D	$Bench^{HR}$
LLaVA-UHD	59.9	70.0	63.8	55.4	56.3	62.2	56.7	45.6
w/ *ConvNext*	59.7	68.2	62.7	55.6	61.8	63.5	61.8	44.0
w/ *DeConv.*	61.7	71.2	64.2	57.4	61.8	67.8	63.4	46.3
w/ *Bilinear*	62.0	72.0	64.5	57.8	62.2	67.6	63.7	46.5
w/ *VDIM*	63.0	73.0	64.6	58.3	62.5	68.5	65.0	48.9

Table 3: Comparison of different methods for semantic pyramid construction. "*ConvNext*" means we replace the CLIP-ViT with CLIP-ConvNeXt (Liu et al. 2022) as visual encoder and directly use the feature maps from multiple stages as the final hierarchical feature pyramid.

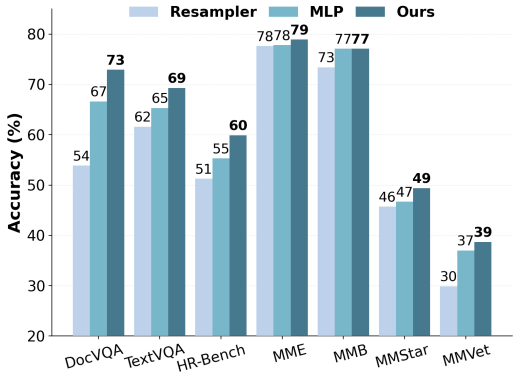


Figure 5: Performance comparison of using different projectors on compressing ISP. Hiwin attention exhibits a significant advantage.

4.4 Analytical Study

We conduct analytical experiments on the proposed modules to verify the effect of Hiwin-LLaVA-UHD. Without special instructions, we use Vicuna-7B as the base LLM.

Main module ablation. In Table 2, by replacing the low-resolution CLIP-ViT features with the highest-level ones of the inverse semantic pyramid (ISP) constructed by VDIM, 2.2% average improvement can be seen, especially on tasks depending on visual details like OCR-Bench (+9.5%) and RefCOCOs (+4.0%). Employing the Hiwin attention to integrate the ISP further increases 1.3% in average accuracy, especially 5.0% on ultra high-resolution perception (HR-Bench), demonstrating that rich visual granularity could facilitate precise language generation.

ISP demonstrates effectiveness on MLLM tasks compared to the traditional feature pyramid. In Table 3, it

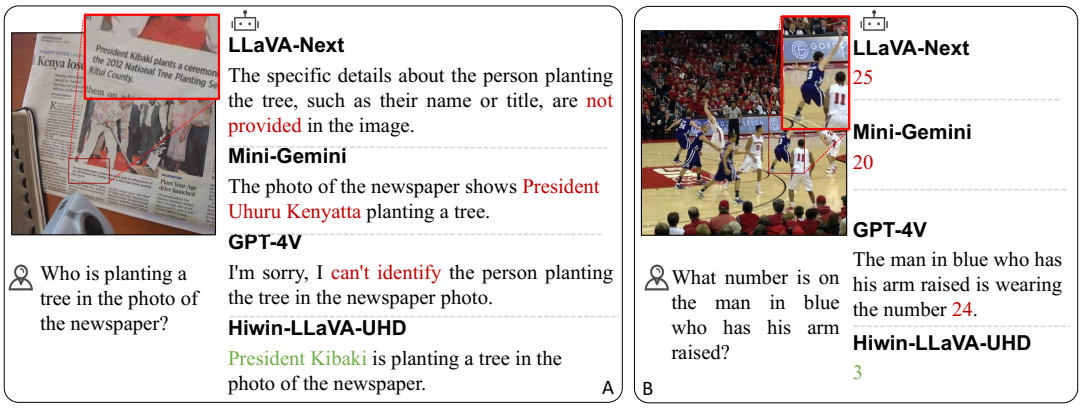


Figure 6: Qualitative comparison of proposed Hiwin-LLaVA-UHD and advanced MLLMs, including LLaVA-Next, Mini-Gemini, and GPT-4V. Our method outperforms its counterparts by providing both fine-grained visual information and high-level semantic contexts for the high-resolution complex perception tasks

is evident that feature pyramids, regardless of their construction method, could enhance performance across various tasks. Nonetheless, the ISP constructed by VDIM achieves an average performance gain of 1.0% over bilinear interpolation, indicating that the ISP further enhances beneficial visual representations ($e.g.$, high-frequency visual details).

Hierarchical window attention works on compressing multi-scale feature pyramid. As shown in Figure 5, Hiwin attention condenses visual tokens by using learnable queries on feature pyramids with arbitrary native resolutions, resulting in superior performance on fine-grained MLLM tasks (e.g., DocVQA and TextVQA). In contrast, directly downsampling high-resolution features via interpolation before MLP projection (Liu et al. 2024b) loses critical visual details and thereby degrades performance. Moreover, the Perceive Resampler (Alayrac et al. 2022; Yao et al. 2024) performs notably worse due to its lack of spatial prior constraints on each query, which adversely affects training convergence, as evidenced by its significantly higher pre-training loss compared to Hiwin attention (0.6110 vs. 0.4368).

4.5 Visualization Analysis

Case study. In Figure 6, we visualize the performance of well-known MLLMs on complex perception tasks, which requires MLLMs to well fuse both visual details and high-level semantics to accurately identify fine-grained targets ($e.g.$, OCR, colors) during the procedure of complex semantic perception ($e.g.$, semantic relation and visual behavior). It is evident that Hiwin-LLaVA-UHD correctly recognizes the tree planter in the newspaper photo and associates it with the name within the dense image caption (Case A). We can also see that Hiwin-LLaVA-UHD captures the player who raises hands and reads the "number 3" on clothes (Case B). In contrast, LLaVA-Next overlooks the name information within dense texts (Case A) and hallucinates on the player number (Case B). Mini-Gemini fails to extract the true name (Case A) and also hallucinates (Case B). Additionally, GPT-4V shows limitations in referencing the information in the newspaper (Case A) and falsely recognizes "number 24" due to wrong fine-grained action perception (Case B).

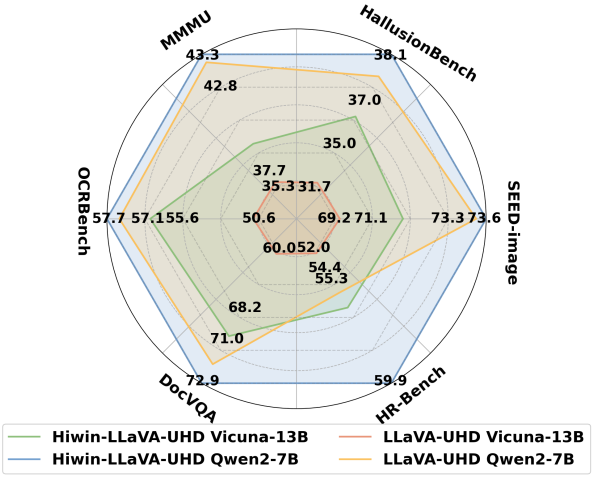


Figure 7: Comparison of performance when equipped with different LLMs. The proposed Hiwin transformer consistently improves performance across Vicuna-13B, and Qwen2-7B, demonstrating its strong generalization capability to different model scales and architectures. Especially in tasks that require visual details, such as DocVQA and HR-Bench, the improvement is even more pronounced.

5 Conclusion

Our proposed hierarchical window transformer, effectively addresses the limitations of conventional ViT-based MLLMs by capturing varying visual granularity essential for precise language generation. The Hiwin transformer adeptly constructs an inverse semantic pyramid for enriched multi-modal representation, which is then condensed into a compact set of visual tokens. This process enhances nuanced visual-linguistic alignment as well as facilitates efficient visual prompting for the LLM. Hiwin-LLaVA-UHD shows substantial gains over the baseline method across a range of MLLM benchmarks, demonstrating its capacity in tasks that demand both low-level details and high-level semantics. Furthermore, the Hiwin transformer offers versatility, presenting potential adaptability across diverse ViT-based MLLM architectures.

Acknowledgments

This work is also supported by the AI9Stars community.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Alayrac, J.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J. L.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*.

An, X.; Yang, K.; Dai, X.; Feng, Z.; and Deng, J. 2024. Multi-label cluster discrimination for visual representation learning. In *ECCV*, 428–444. Springer.

Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A family of highly capable multimodal models. arXiv:2312.11805.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *IEEE ICCV*, 2425–2433.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. arXiv:2308.12966.

Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *IEEE CVPR*, 13817–13827.

Chen, J.; Wei, F.; Zhao, J.; Song, S.; Wu, B.; Peng, Z.; Chan, S.-H. G.; and Zhang, H. 2024a. Revisiting Referring Expression Comprehension Evaluation in the Era of Large Multimodal Models. arXiv:2406.16866.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. arXiv:2310.09478.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024b. Are we on the right way for evaluating large vision-language models? arXiv:2403.20330.

Chen, P.; Bu, P.; Song, J.; Gao, Y.; and Zheng, B. 2024c. Can VLMs Play Action Role-Playing Games? Take Black Myth Wukong as a Study Case. arXiv:2409.12889.

Chen, Y.; Wang, X.; Peng, H.; and Ji, H. 2024d. A Single Transformer for Scalable Vision-Language Modeling. arXiv:2407.06438.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Contributors, X. 2023. XTuner: A Toolkit for Efficiently Fine-tuning LLM. https://github.com/InternLM/xtuner.

Dai, W.; Li, J.; Li, D.; et al. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dong, X.; Zhang, P.; et al. 2024. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD. arXiv:2404.06512.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394.

Gao, P.; Zhang, R.; Liu, C.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. arXiv:2402.05935.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 14375–14385.

Guo, Z.; Xu, R.; Yao, Y.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; and Huang, G. 2024. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. In *ECCV*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *IEEE ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE CVPR*, 770–778.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE CVPR*, 6700–6709.

Kembhavi, A.; Salvato, M.; Kolve, E.; et al. 2016. A diagram is worth a dozen images. In *ECCV*, 235–251.

Kim, G.; Hong, T.; Yim, M.; et al. 2022. Ocr-free document understanding transformer. In *ECCV*, 498–517.

Lee, K.; Joshi, M.; Turc, I. R.; et al. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *ICML*, 18893–18912.

Li, B.; Wang, R.; Wang, G.; et al. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv:2307.16125.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*.

Li, W.; Yuan, Y.; Liu, J.; et al. 2024. Tokenpacker: Efficient visual projector for multimodal llm. arXiv:2407.02392.

Li, Y.; Zhang, Y.; et al. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. arXiv:2403.18814.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; et al. 2023c. Monkey: Image resolution and text label are important things for large multi-modal models. arXiv:2311.06607.

Lin, J.; Yin, H.; Ping, W.; et al. 2024. Vila: On pre-training for visual language models. In *CVPR*, 26689–26699.

Lin, T.-Y.; Dollár, P.; et al. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; et al. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Lin, Z.; Liu, C.; Zhang, R.; et al. 2023. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. arXiv:2311.07575.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *NeurIPS*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; et al. 2023a. MMBench: Is your multi-modal model an all-around player? arXiv:2307.06281.

Liu, Y.; Li, Z.; Yang, B.; Li, C.; et al. 2023b. On the hidden mystery of ocr in large multimodal models. arXiv:2305.07895.

Liu, Y.; Yang, B.; Liu, Q.; et al. 2024c. Textmonkey: An ocr-free large multimodal model for understanding document. arXiv:2403.04473.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE ICCV*, 10012–10022.

Liu, Z.; Mao, H.; Wu, C.-Y.; et al. 2022. A convnet for the 2020s. In *CVPR*, 11976–11986.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. arXiv:2403.05525.

Lu, P.; Bansal, H.; Xia, T.; et al. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv:2310.02255.

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

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:2403.03003.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv:2203.10244.

Mathew, M.; Karatzas, D.; and Jawahar, C. V. 2021. DocVQA: A Dataset for VQA on Document Images. In *IEEE WACV*, 2199–2208.

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 *ICML*, 8748–8763.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. arXiv:2403.15388.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA models that can read. In *IEEE CVPR*, 8317–8326.

Wang, P.; Bai, S.; Tan, S.; et al. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191.

Wang, W.; Ding, L.; Zeng, M.; et al. 2024b. Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models. arXiv:2408.15556.

Wang, Z.; Zhu, X.; Yang, X.; et al. 2025. Parameter-Inverted Image Pyramid Networks for Visual Perception and Multimodal Understanding. arXiv:2501.07783.

xAI. 2024. RealWorldQA. https://huggingface.co/datasets/xai-org/RealworldQA.

Yao, H.; Wu, W.; Yang, T.; Song, Y.; Zhang, M.; Feng, H.; Sun, Y.; Li, Z.; Ouyang, W.; and Wang, J. 2025. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37: 33108–33140.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. arXiv:2408.01800.

Ye, J.; Hu, A.; Xu, H.; Ye, Q.; et al. 2023a. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. arXiv:2310.05126.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; et al. 2023b. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. arXiv:abs/2311.

You, H.; Zhang, H.; Gan, Z.; et al. 2023. Ferret: Refer and ground anything anywhere at any granularity. arXiv:2310.07704.

Yu, L.; Poirson, P.; Yang, S.; et al. 2016. Modeling context in referring expressions. In *ECCV*, 69–85.

Yue, X.; Ni, Y.; Zhang, K.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE CVPR*, 9556–9567.

Zhan, Y.; Zhu, Y.; et al. 2024. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, 405–422.

Zhang, Y.-F.; Wen, Q.; Fu, C.; Wang, X.; et al. 2024. Beyond llava-hd: Diving into high-resolution large multimodal models. arXiv:2406.08487.

Zhao, X.; Li, X.; Duan, H.; et al. 2024. Mg-llava: Towards multi-granularity visual instruction tuning. arXiv:2406.17770.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv:2010.04159.