LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training

LLaVA-OneVision Community Contributors

Abstract

We present LLaVA-OneVision-1.5, a novel family of Large Multimodal Models (LMMs) that achieve state-of-the-art performance with significantly reduced computational and financial costs. Different from the existing works, LLaVA-OneVision-1.5 provides an open, efficient, and reproducible framework for building high-quality vision-language models entirely from scratch. The LLaVA-OneVision-1.5 release comprises three primary components: ***(1) Large-Scale Curated Datasets:*** We construct an 85M concept-balanced pretraining dataset LLaVA-OneVision-1.5-Mid-Traning and a meticulously curated 22M instruction dataset LLaVA-OneVision-1.5-Instruct. ***(2) Efficient Training Framework:*** We develop a complete end-to-end efficient training framework leveraging an offline parallel data packing strategy to facilitate the training of LLaVA-OneVision-1.5 within a *$16,000* budget. ***(3) State-of-the-art Performance:*** Experimental results demonstrate that LLaVA-OneVision-1.5 yields exceptionally competitive performance across a broad range of downstream tasks. Specifically, LLaVA-OneVision-1.5-8B outperforms Qwen2.5-VL-7B on 18 of 27 benchmarks, and LLaVA-OneVision-1.5-4B surpasses Qwen2.5-VL-3B on all 27 benchmarks. (4) ***RL-based Post-training:*** We unlock the model's latent potential through a lightweight RL stage, effectively eliciting robust chain-of-thought reasoning to significantly boost performance on complex multimodal reasoning tasks.

	Code	`https://github.com/EvolvingLMMs-Lab/LLaVA-OneVision-1.5`
	RL Code	`https://github.com/EvolvingLMMs-Lab/LLaVA-OneVision-1.5-RL`
🤗	**Models**	`https://huggingface.co/lmms-lab/LLaVA-OneVision-1.5-8B-Instruct`
🤗	**RL Model**	`https://huggingface.co/mvp-lab/LLaVA-OV-1.5-8B-RL`
🤗	**Pretrain data**	`https://huggingface.co/datasets/mvp-lab/LLaVA-OneVision-1.5-Mid-Training-85M`
🤗	**Instruct data**	`https://huggingface.co/datasets/mvp-lab/LLaVA-OneVision-1.5-Instruct-Data`
🤗	**RL Data**	`https://huggingface.co/datasets/mvp-lab/LLaVA-OneVision-1.5-RL-Data`

1 Introduction

Recent advancements in Large Multimodal Models (LMMs) have demonstrated remarkable capabilities in multimodal understanding and reasoning (Comanici et al., 2025; Guo et al., 2025; Zhu et al., 2025). These developments enable artificial intelligence applications to effectively comprehend and analyze images, charts, and PDF documents. However, the most performant models remain proprietary, with neither their training data nor source code publicly available. Consequently, the broader research community lacks crucial insights into how such high-performing LMMs can be built from scratch.

To reduce barriers for community development, several research efforts have attempted to reproduce the capabilities of proprietary models using open architectures. Early efforts such as

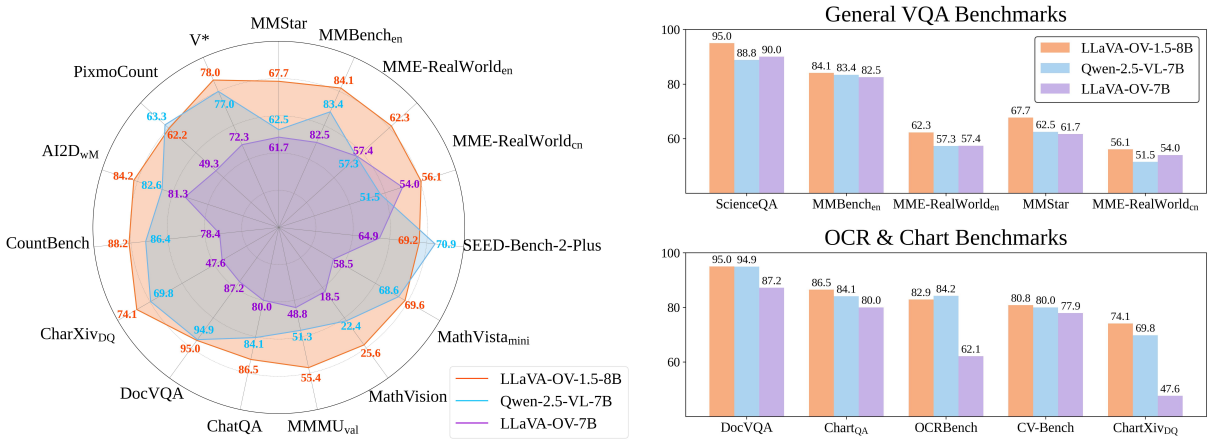


Figure 1 Performance of LLava-OV-1.5-8B across multiple benchmarks.

LLaVA (Liu et al., 2023), LLaVA-Next (Liu et al., 2024b), and LLaVA-OneVision (Li et al., 2025a) provided fully open training data and code, but their performance now falls substantially behind that of current state-of-the-art models (Bai et al., 2025; Zhu et al., 2025). More recent works have pushed the boundary: Molmo (Deitke et al., 2025) released model weights, datasets, and source code, enabling the community to train LMMs from scratch. Through careful architectural choices, a refined training pipeline, and high-quality data, Molmo achieves near-parity with GPT-4V on both academic benchmarks and user preference evaluations. Open-Qwen2VL (Wang et al., 2025b) introduces a 2B-parameter model pre-trained on only 0.36% of 1.4T multimodal tokens in Qwen2-VL, while outperforming Qwen2-VL-2B across various multimodal benchmarks. Despite these advances, the performance gap between open-source and proprietary models continues to widen as the field of LMMs rapidly evolves. Current open-source models are still constrained by substantial computational demands and suboptimal training efficiency.

To overcome the aforementioned limitations, we introduce LLaVA-OneVision-1.5, a fully open-source family of LMMs, extending the LLaVA series (Li et al., 2025a) to achieve superior performance with limited computational cost. Specifically, LLaVA-OneVision-1.5 adopts RICE-ViT (Xie et al., 2025) as the vision encoder, enabling native-resolution adaptation and fine-grained visual understanding based on stronger region-level semantic representation. Building upon LLaVA-OneVision (Li et al., 2025a), LLaVA-OneVision-1.5 adopts a three-stage training pipeline: (Stage-1) Language-Image Alignment, (Stage-1.5) High-Quality Knowledge Learning, and (Stage-2) Visual Instruction Tuning. Notably, we find that simply scaling data at the mid-training stage alone can produce state-of-the-art LMMs, eliminating the need for complex training paradigms. To foster open research, we release all assets to the community, including LLaVA-OneVision-1.5-Mid-Traning and LLaVA-OneVision-1.5-Instruct datasets, the training framework, and model checkpoints (LLaVA-OneVision-1.5-Base and LLaVA-OneVision-1.5-Instruct). In summary, our contributions are as follows:

- ***Large Multimodal Models.*** We propose LLaVA-OneVision-1.5 a family of fully open-source large multimodal models that achieve superior performance across multiple multimodal benchmarks compared to Qwen2.5-VL.
- ***Large-Scale Datasets.*** We construct an 85M concept-balanced pre-training dataset LLaVA-OneVision-1.5-Mid-Traning and a meticulously curated 22M instruction dataset LLaVA-OneVision-1.5-Instruct.
- ***Efficienct Training Framework.*** We develop a complete end-to-end training framework that

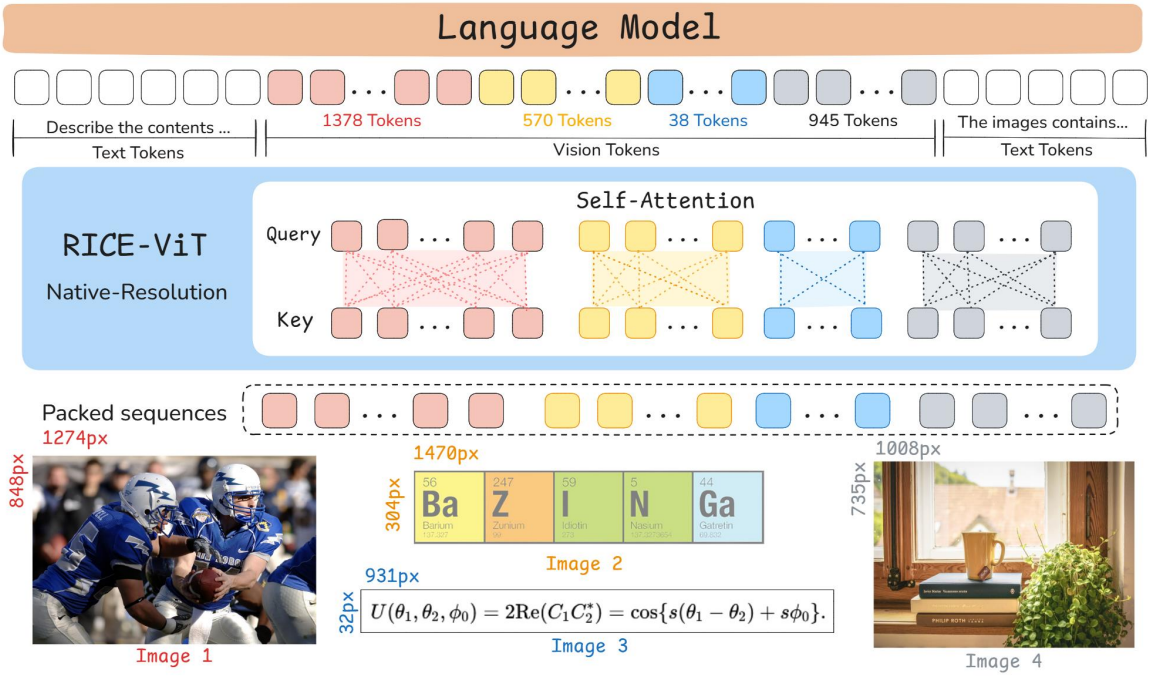


Figure 2 Overall architecture of LLaVA-OneVision-1.5. The framework integrates a pre-trained vision encoder with a language model decoder. The vision encoder adopts 2D RoPE for native-resolution processing and incorporates region-aware attention to enhance local semantic modeling. During pretraining, both object regions and OCR regions are jointly modeled to inject fine-grained text understanding capability. A lightweight projector maps visual features into the LLM embedding space, and the `[CLS]` token is preserved to retain global semantic capacity during multimodal alignment.

employs an offline parallel data packing strategy to optimize cost-effectiveness, enabling the training of LLaVA-OneVision-1.5 within a *$16,000* compute budget.

- ***Open-Source Release.*** We release all assets to the public including LLaVA-OneVision-1.5-Mid-Traning and LLaVA-OneVision-1.5-Instruct dataset, training framework, and the model checkpoints (LLaVA-OneVision-1.5-Base and LLaVA-OneVision-1.5-Instruct).

- ***RL-Enhanced Reasoning.*** We introduce a lightweight RL post-training stage using the asynchronous AReaL system. By adopting a discrepancy-driven data selection strategy and rigorous outcome-based verification, we effectively elicit latent reasoning capabilities, significantly boosting performance on complex tasks while maintaining robust general visual understanding.

2 Architecture

2.1 Overall Architecture

The overall architecture of LLaVA-OneVision-1.5 is illustrated in Fig. 2. LLaVA-OneVision-1.5 retains the "ViT–MLP–LLM" paradigm of the LLaVA series, comprising three core modules:

- ***Vision Encoder***: The vision encoder is responsible for extracting rich and semantically meaningful visual representations from input images, which serve as the foundation for multimodal alignment and downstream reasoning. Unlike previous works (Wang et al.,

2024b; Bai et al., 2025) that adopt SigLIP (Zhai et al., 2023) or DFN (Fang et al., 2023), LLaVA-OneVision-1.5 integrates our recently proposed cluster discrimination model RICE-ViT (Xie et al., 2025) to improve region-aware visual and OCR capabilities.

- ***Projector***: The projector bridges the modality gap between the vision encoder and the large language model by mapping visual embeddings into the text embedding space of the LLM. Following Qwen2.5-VL (Bai et al., 2025), we first group spatially adjacent sets of four patch features, which are then concatenated and passed through a two-layer multi-layer perceptron to map them into the text embedding space of the LLM.
- ***Large Language Model***: The large language model acts as the reasoning and generation core of the architecture. After receiving the projected multimodal embeddings, the LLM integrates visual information with linguistic context to perform complex reasoning, instruction following, and natural language generation. The LLaVA-OneVision-1.5 series utilize Qwen3 (Team, 2025) as the language backbone to significantly enhance performance on downstream tasks.

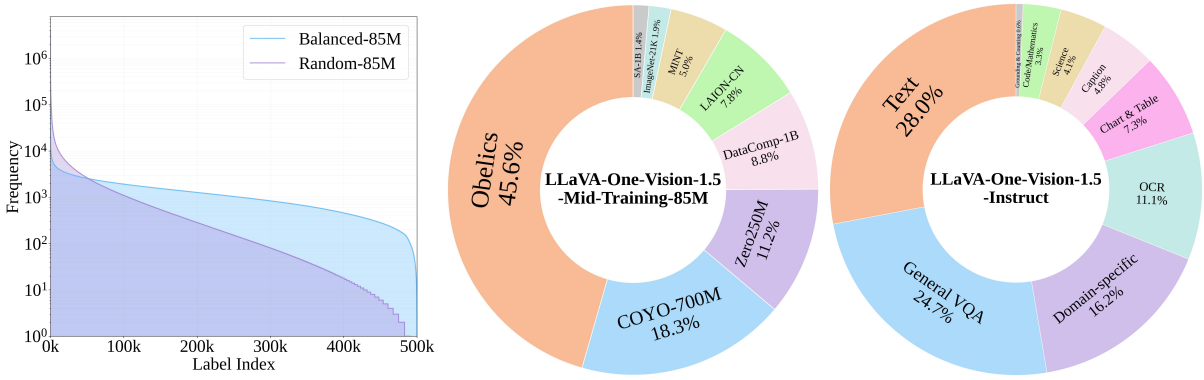
This modular design follows the LLaVA framework but incorporates more efficient training recipes and carefully selected encoders, enabling superior cost-effectiveness and scalability.

2.2 Vision Encoder via Region-Aware Cluster Discrimination

Fine-grained visual semantics are essential for dense prediction tasks such as grounding, OCR, and segmentation. Although large-scale vision–language contrastive models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) demonstrate strong performance through global vision-language alignment, they fail to capture the similarity structure of training data or the local region-level semantics within images. This shortcoming stems from instance-wise contrastive learning, which treats all instances as negatives regardless of their semantic similarity and represents each instance solely with a single global embedding.

To overcome these limitations, our LLaVA-OneVision-1.5 leverages the RICE-ViT (Xie et al., 2025) as its vision encoder, enabling precise multimodal alignment and enriched region-level representation. RICE-ViT enhances both object-centric and OCR capabilities by introducing a unified region cluster discrimination loss, trained on 450M images and 2.4B candidate regions. Its design combines a region-aware attention mechanism for local semantic modeling with 2D rotary positional encoding, which naturally supports variable input resolutions without requiring resolution-specific fine-tuning, unlike models such as Qwen2-VL (Wang et al., 2024b) and InternVL 2.5 (Chen et al., 2024c).

We integrate this pretrained encoder with a language model through joint training, yielding a streamlined multimodal pipeline. Compared to SigLIP2 (Tschannen et al., 2025), which depends on multiple specialized losses (SILC, TIPS, LocCa, and Sigmoid), our method adopts a single cluster discrimination loss that simultaneously strengthens general understanding, OCR, and localization. This unified formulation provides an elegant, computationally efficient solution that matches SigLIP2's performance while substantially reducing architectural complexity and training overhead.



(a) (b) (c)

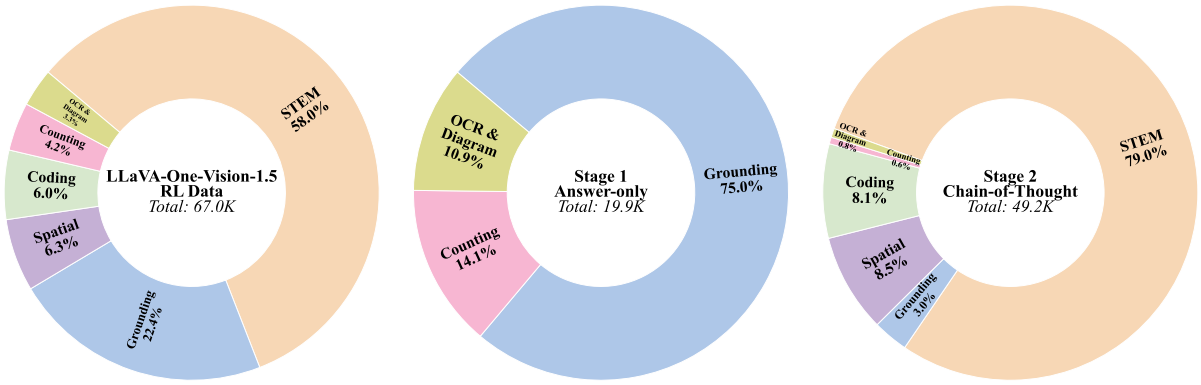
Figure 3 (a) The vocabulary coverage proportion in the LLaVA-OneVision-1.5-Mid-Traning dataset before and after concept balancing. (b) Distribution of data sources within the LLaVA-OneVision-1.5-Mid-Traning dataset. (c) Distribution of data sources within the LLaVA-OneVision-1.5-Instruct.

3 Data

3.1 Pre-Training Dataset

We use the LLaVA-1.5 558K (Liu et al., 2024a) to align the visual features into the word embedding space of LLMs. After that, LLaVA-OneVision-1.5 is underpinned by a large-scale multimodal dataset LLaVA-OneVision-1.5-Mid-Traning, which contains 85 million high-quality image-text pairs (20M in Chinese and 65M in English). The data of LLaVA-OneVision-1.5-Mid-Traning are from a wide range of sources: COYO-700M (Byeon et al., 2022), Obelics (Laurençon et al., 2023), DataComp-1B (Gadre et al., 2023), LAION-CN (Zhang et al., 2022), ImageNet-21K (Russakovsky et al., 2015), SAM-1B (Kirillov et al., 2023), MINT (Wang et al., 2024c), and Zero250M (Xie et al., 2023). To enrich the diversity of our pretraining data, we introduce a concept-balanced sampling strategy inspired by MetaCLIP (Xu et al., 2024). Unlike MetaCLIP, which depends on raw captions for concept matching and struggles with caption-free or interleaved datasets (*e.g.*, SAM-1B, ImageNet-21K, and Obelics), our method reduces reliance on caption quality, such as the brief and incomplete annotations common in COYO-700M. Instead, we adopt a feature-based matching approach that coarsely groups image sources. Specifically, using the pretrained MetaCLIP-H/14-Full-CC2.5B encoders (Xu et al., 2024), we project both images and MetaCLIP's 500K concept entries into a shared embedding space. Since MetaCLIP embeddings are already concept-balanced, this enables effective similarity-based concept induction: for each image, we retrieve its top-K nearest concept embeddings to construct refined pseudo-captions that enhance semantic alignment.

Top-K Concept Assignment and Balance Sampling. Given an image set $\mathcal{I} = \{i_0, i_1, ..., i_N\}$ and a concept vocabulary set $\mathcal{V} = \{v_0, v_1, ...v_M\}$, we first utilize the image encoder Φ_v and text encoder Φ_t to extract the image embeddings $\mathcal{E}_i = \{\Phi_v(i), i \in I\}$ and concept embeddings $\mathcal{E}_t = \{\Phi_t(v), v \in V\}$. Then we assign each image with the top-k nearest concepts based on the cosine similarity of the L2-normalized image and concept embeddings. After that, following MetaCLIP (Xu et al., 2024), we weight each image by the inverse frequencies of its concepts and then sample images based on the normalized image weights. This inverse-frequency methodology promotes a more balanced concept distribution, without relying on original captions, which may be noisy or missing. This process yields 85M images with balanced concepts. Subsequently, we apply a powerful captioner to produce English and Chinese captions for these images, followed by a validity filter to eliminate duplicates and excessively lengthy outputs. Ultimately, we establish



(a) (b) (c)

Figure 4 Distribution of task categories in the RL training data. (a) Total RL corpus (67K instances). (b) Stage 1: Answer-only training. (c) Stage 2: Chain-of-thought training.

an 85M concept-balanced mid-training dataset. The distribution of data sources of the LLaVA-OneVision-1.5-Mid-Traning dataset is illustrated in Fig. 3(b).

3.2 Instruction Dataset

Visual instruction tuning (Liu et al., 2023) is vital for enabling LMMs to understand and follow visual instructions, and its effectiveness hinges on the quality of the instruction datasets. To this end, we construct the LLaVA-OneVision-1.5-Instruct dataset by aggregating a wide range of instruction-tuning datasets from diverse sources. The data are carefully curated to ensure balanced coverage across seven categories: Caption, Chart & Table, Code & Math, Domain-specific, General VQA, Grounding & Counting, OCR, and Science. The resulting corpus comprises 22 million samples, with Fig. 3(c) showing the proportional distribution across categories.

4 Training Strategy

4.1 Training Pipeline

Following LLaVA-OneVision (Li et al., 2025a), LLaVA-OneVision-1.5 undergoes three distinct learning stages to enable LLM for multimodal capabilities:

- ***Stage-1: Language-Image Alignment.*** The stage aims to pretrain the projection layer with the LLaVA-1.5 558K to align the visual features into the word embedding space of LLMs.
- ***Stage-1.5: High-Quality Knowledge Learning.*** Building on the language-image alignment stage, we introduce the high-quality knowledge learning stage to strike a balance between computational efficiency and injecting new knowledge into LMMs. In this stage, we transition to full-parameter training of all modules using the LLaVA-OneVision-1.5-Mid-Traning dataset.
- ***Stage-2: Visual Instruction Tuning.*** To enable LMMs to handle a diverse range of visual tasks with desired responses, we continue full-parameter training with the proposed LLaVA-OneVision-1.5-Instruct as well as the FineVision (Wiedmann et al., 2025) dataset.

Table 1 Performance comparison across vision-language models on various benchmarks grouped by task type. All scores are reported as accuracy percentages unless otherwise specified.

Task	Benchmark	LLaVA-OV-1.5	LLaVA-OV-1.5 RL		Qwen2.5-VL	LLaVA-OV-1.5	Qwen2.5-VL	LLaVA-OV
Size		8B	8B		7B	4B	3B	7B
Mode		-	thinking	fast	-	-	-	-
General VQA	MMStar	67.7	68.2$^{\uparrow 0.5}$	**68.3**$^{\uparrow 0.6}$	62.5	64.9	55.9	61.7
	MMBench$_{en}$	84.1	**85.7**$^{\uparrow 1.6}$	85.7$^{\uparrow 1.6}$	83.4	84.2	78.0	82.5
	MMBench$_{cn}$	81.0	**84.2**$^{\uparrow 3.2}$	81.5$^{\uparrow 0.5}$	81.6	76.9	74.6	81.4
	MME-RealWorld$_{en}$	61.7	**63.4**$^{\uparrow 1.7}$	63.3$^{\uparrow 1.6}$	57.3	61.6	51.6	57.4
	MME-RealWorld$_{cn}$	56.1	56.1$^{\uparrow 0.0}$	**56.3**$^{\uparrow 0.2}$	51.5	49.6	45.4	54.0
	SeedBench$_{image}$	77.3	76.7	**77.6**$^{\uparrow 0.3}$	**77.5**	76.6	74.8	75.4
	CV-Bench	80.7	**82.9**$^{\uparrow 2.2}$	81.1$^{\uparrow 0.4}$	80.0	77.2	71.5	77.9
	SEED-Bench-2-Plus	69.2	69.5$^{\uparrow 0.3}$	69.2$^{\uparrow 0.0}$	**70.9**	68.9	68.6	64.9
	RealWorldQA	68.1	68.4$^{\uparrow 0.3}$	**70.6**$^{\uparrow 2.5}$	68.5	67.8	60.0	66.3
	Avg.	71.8	**72.8**$^{\uparrow 1.0}$	72.6$^{\uparrow 0.8}$	72.2	72.1	66.4	71.1
Reasoning	MathVista$_{mini}$	69.6	**72.3**$^{\uparrow 2.7}$	71.8$^{\uparrow 2.2}$	68.6	67.9	60.2	58.5
	WeMath	61.5	**69.4**$^{\uparrow 7.9}$	60.8	61.3	62.0	45.1	44.1
	MathVision	25.6	**34.4**$^{\uparrow 8.8}$	26.2$^{\uparrow 0.6}$	22.4	24.2	21.3	18.5
	MMMU$_{val}$	55.4	**58.8**$^{\uparrow 3.4}$	54.9	51.3	52.7	46.4	48.8
	MMMU-Pro$_{standard}$	37.4	**39.9**$^{\uparrow 2.5}$	38.0$^{\uparrow 0.6}$	36.3	35.3	31.1	28.0
	MMMU-Pro$_{vision}$	25.2	**35.7**$^{\uparrow 10.5}$	29.0$^{\uparrow 3.8}$	32.8	25.4	21.3	14.3
	Avg.	45.8	**51.8**$^{\uparrow 6.0}$	46.8$^{\uparrow 1.0}$	45.5	44.6	37.6	35.4
OCR & Chart	ChartQA	86.5	**87.4**$^{\uparrow 0.9}$	87.0$^{\uparrow 0.5}$	84.1	87.1	83.4	80.0
	CharXiv$_{DQ}$	70.9	68.4	**71.2**$^{\uparrow 0.3}$	69.8	63.8	58.2	47.6
	DocVQA	95.0	91.9	**95.0**$^{\uparrow 0.0}$	94.9	94.4	92.7	87.2
	OCRBench	82.9	81.7	82.3	**84.2**	80.0	79.2	62.1
	AI2D$_{w\ M}$	84.2	83.7	**84.3**$^{\uparrow 0.1}$	82.6	83.6	78.6	81.4
	AI2D$_{w/o\ M}$	**94.1**	93.7	93.9	93.4	93.3	90.7	90.8
	InfoVQA	78.4	76.6	78.7$^{\uparrow 0.3}$	**81.7**	76.1	75.6	68.8
	Avg.	84.6	83.3	**84.6**$^{\uparrow 0.0}$	84.4	82.6	79.8	74.0
Others	PixmoCount	62.2	65.7$^{\uparrow 3.5}$	**71.1**$^{\uparrow 8.9}$	**63.3**	52.2	50.9	49.3
	CountBench	88.2	86.8	**88.6**$^{\uparrow 0.4}$	86.4	79.8	72.5	78.4
	VL-RewardBench	47.7	44.0	**49.7**$^{\uparrow 2.0}$	49.7	48.2	42.1	44.5
	V*	78.0	**79.1**$^{\uparrow 1.1}$	78.0$^{\uparrow 0.0}$	77.0	74.9	69.6	72.3
	Avg.	69.0	66.0	**71.6**$^{\uparrow 2.6}$	69.1	63.8	58.8	61.1

4.2 Infrastructure

Load Balancing via Data Packing. A major source of training inefficiency arises from padding, where batch samples are standardized by adding padding tokens. This results in significant computational overhead and poor GPU utilization, particularly with heterogeneous multimodal data. To mitigate this, we propose an offline parallel data packing method that consolidates multiple shorter samples into packed sequences during preprocessing. Our approach employs hash buckets to handle large-scale data efficiently and leverages multi-threaded, strategy-aware batching to control packing success rate, sample count, and batch composition. Unlike online packing, which operates dynamically at runtime, our method processes entire datasets or large contiguous chunks offline, ensuring uniform output lengths. This yields up to an $11\times$ compression ratio on 85 million pretraining samples, substantially improving efficiency.

Hybrid Parallelism Framework. We adopt AIAK-Training-LLM[1] built upon Megatron-LM (Shoeybi et al., 2019) as our training framework. Its transformer engine and specialized optimizations enable efficient mid-training of LLaVA-OneVision-1.5-8B with a context length of 8K. By leveraging distributed optimizer parallelism and uniform recomputation, the mid-training process is conducted at native resolution on 85 million captions using 128 $\times$ A800 GPUs over 3.7 days.

[1]AIAK-Training-LLM: Baidu Cloud's optimized Megatron-LM

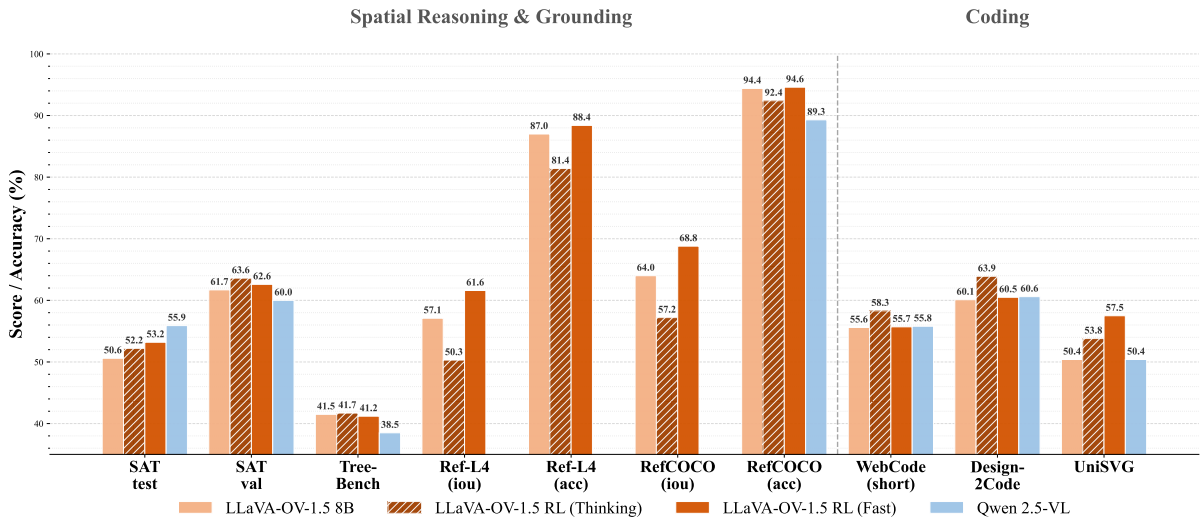


Figure 5 Performance comparison of LLaVA-OV-1.5 and corresponding RL version on Spatial Reasoning & Grounding and Coding tasks.

5 Post-training

To further enhance LLaVA-OneVision-1.5's performance on multimodal reasoning tasks, we perform a reinforcement learning-based post-training stage on top of the supervised LLaVA-OneVision-1.5-Instruct model.

5.1 RL Training Data

Discrepancy-Driven Data Selection. We curate the training data by measuring the divergence between *Pass@N* and *Pass@1* performance on diverse benchmarks. A significant gap indicates that the model possesses the *latent capability* to solve the task, as correct solutions do appear within its sampling distribution, yet its policy distribution fails to reliably assign high probability to the correct reasoning path. Under this lens, RL serves as an **elicitation** mechanism rather than knowledge injection, redirecting probability mass toward solutions the model can already generate but does not consistently prioritize. This selection paradigm ensures high training efficiency by targeting the model's effective learnable boundary, avoiding trivial tasks it has already mastered or unsolvable ones beyond its current grasp.

We construct the RL training corpus by aggregating diverse public data sources, including ViRL (Wang et al., 2025a), WebCode2M (Yun et al., 2024), UniSVG (Li et al., 2025b), Ref-L4 (Chen et al., 2024a), VigoRL-SA (Sarch et al., 2025), VigoRL-SAT (Sarch et al., 2025), Pixmo-Count (Deitke et al., 2025), AI2D (Kembhavi et al., 2016b), and InfoVQA (Mathew et al., 2022a). These sources cover a wide range of capabilities such as STEM reasoning, coding, grounding, counting, spatial reasoning, diagram understanding, and OCR.

Reward-Based Sampling. To further filter the high-quality training instances, we employ a **reward-based sampling** strategy. Specifically, we generate multiple candidate responses for each sample using the base model and compute their automatic rewards. We then retain only those examples where the average reward across candidates falls within a specified range. This filtering process effectively discards both trivial and unsolvable cases, biasing the corpus toward medium-difficulty instances that provide the most valuable learning signal.

Finally, we obtain a unified RL corpus of about 67K instances. The detailed composition is shown in Figure 4 (a). Specifically, **STEM** data (38.9K) comes from ViRL39K; **Grounding** (15K) aggregates Ref-L4 and VigoRL-SA; **Spatial** (4.2K) and **Counting** (2.8K) tasks are sourced from VigoRL-SAT and PixmoCount; **Coding** (4K) combines WebCode2M and UniSVG; while **OCR** (2K) and **Diagram** (0.2K) samples are selected from InfoVQA and AI2D, respectively. Each instance additionally records whether it is prompted in a short answer-only style or a longer chain-of-thought style, which we later exploit when designing different RL curricula over the same underlying pool.

5.2 Reward System

Our RL setup employs a rule-based reward paradigm, where rewards are derived directly from task outcomes rather than learned preference models. Since different answer types require distinct verification strategies, we design **answer-type-specific scoring rules**.

For **STEM** questions from ViRL39K, we face a key challenge: the model may express the same correct answer in vastly different formats. To address this, we implement a multi-stage verification pipeline. First, we extract answers via flexible parsing (prioritizing structured tags like `<answer>` but falling back to heuristics such as "Final Answer:" when necessary). Second, we normalize LaTeX artifacts (e.g., unifying `\frac12` and `\frac{1}{2}`). Finally, for numerical reasoning, we perform symbolic equivalence checking rather than string matching—thus $\frac{1}{2}$, 0.5, and even $\frac{2}{4}$ are all recognized as correct if the ground truth is any one of them. For multiple-choice problems, we match extracted option labels (A/B/C/D) against the reference, tolerating common formatting variations like `(A)`, `**A**`, or `A.`. In **Coding** benchmarks, WebCode2M samples are rewarded based on token- and tag-level overlap with the reference code, while UniSVG further incorporates an SVG rendering similarity score in $0, 1$ to encourage perceptually matched graphics.

Grounding data from Ref-L4 and VigoRL-SA are evaluated by the intersection-over-union (IoU) between predicted and reference bounding boxes, combined with standard accuracy for associated multiple-choice queries. **Spatial-reasoning** problems from VigoRL-SAT are scored purely by answer accuracy. For **Counting** tasks such as PixmoCount, we extract the final numeric token and require exact equality with the gold count. **OCR** instances from InfoVQA use text-similarity based rewards between the predicted and reference strings, and **Diagram** questions from AI2D are judged by multiple-choice accuracy. These answer-type-aware rules are collapsed into a single scalar reward per candidate response.

5.3 Training Procedure

We employ Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our core reinforcement learning algorithm. To maximize training efficiency and throughput, we adopt the GRPO implementation in AReaL (Fu et al., 2025), a state-of-the-art asynchronous RL framework. AReaL decouples generation from training, allowing rollout workers to continuously generate data while trainer workers update the model in parallel, significantly improving GPU utilization compared to synchronous implementations.

Regarding the optimization objective, we simplify the standard GRPO formulation by **omitting the KL divergence penalty**, relying instead on PPO-style clipping to maintain training stability. Furthermore, we discard the explicit format reward commonly used to enforce structural constraints (e.g., XML tagging). Instead, we rely solely on outcome-based correctness rewards. To better exploit the structure of our RL corpus, we employ a two-stage curriculum, as shown in Figure 4(b) and (c):

1. **Stage 1: Answer-only RL on normal data.** In the first stage, we train exclusively on the normal split, where instructions ask the model to output only the final answer. For these tasks we use the prompt

   ```
   Put ONLY your final answer within <answer></answer>.
   ```

   This warm-up stage solidifies basic perceptual skills like counting, serving as a critical foundation for subsequent reasoning tasks. This curriculum ensures the model retains precision on simple problems and avoids "over-thinking" when later advancing to complex reasoning chains.

2. **Stage 2: Chain-of-thought RL on long-reasoning data.** In the second stage, we switch to the long-reasoning split and encourage the model to produce explicit reasoning traces. The instruction for these tasks is

   ```
   Think and solve the following question step by step.  Please put your thinking
   and analysis procedure within <think></think>.  Put ONLY your final answer within
   <answer></answer>.
   ```

   The reward is still computed only from the content within `<answer></answer>`, ensuring that the optimization target remains answer correctness while the reasoning tokens inside `<think></think>` serve as auxiliary guidance.

A naive second stage that uses only long-reasoning tasks can cause the model to forget short, perception-heavy skills. To mitigate this, we interleave a small proportion of normal-set examples into Stage 2 mini-batches. These samples continue to use the answer-only prompt and reward, acting as an anchor that preserves the model's competence on concise tasks while RL emphasizes deeper reasoning. Overall, this two-stage, mixed-prompt curriculum allows LLaVA-OneVision-1.5-RL to simultaneously strengthen long-horizon reasoning and maintain strong performance on standard vision-language benchmarks, details in Section 6.6.

6 Experiments

6.1 Overall Performance

We use LMMs-Eval Zhang et al. (2025a) with the default prompt to evaluate the performance of LLaVA-OneVision-1.5 across multiple benchmarks in four categories of downstream tasks:(1) General Visual Question Answering (VQA): MMStar (Chen et al., 2024b), MMEBench series (Fu et al., 2023), MME-RealWorld series (Zhang et al., 2025b), SeedBench (Li et al., 2024b), Seed-Bench-2-Plus (Li et al., 2024a), CV-Bench (Tong et al., 2024), and RealWorldQA (Corp., 2024). (2) Multimodal Reasoning: MathVista (Lu et al., 2024), WeMath (Qiao et al., 2025), MathVision (Wang et al., 2024a), MMMU (Yue et al., 2024), and MMMU-Pro series (Yue et al., 2025).(3) OCR & Chart Understanding: ChartQA (Masry et al., 2022), CharXiv (Wang et al., 2024d), DocVQA (Mathew et al., 2021), OCRBench (Liu et al., 2024c), AI2D (Kembhavi et al., 2016a), and InfoVQA (Mathew et al., 2022b).(4) Others: PixmoCount (Deitke et al., 2025), CountBench (Paiss et al., 2023), VL-RewardBench (Li et al., 2025c), and V* (Wu and Xie, 2024). As shown in Tab. 1, LLaVA-OneVision-1.5-8B surpasses Qwen2.5-VL-7B on 18 of 27 benchmarks and LLaVA-OneVision-1.5-4B surpasses Qwen2.5-VL-3B on 27 of 27 benchmarks.

6.2 General Visual Question Answering

As detailed in Tab. 1, we evaluate the general visual question answering capability of LLaVA-OneVision-1.5 across multiple benchmarks, and LLaVA-OneVision-1.5-8B demonstrates superior

Table 2 Comparison of RICE-ViT with other vision encoders using the LLaVA-NeXT framework. All models are evaluated using identical configurations: Qwen2.5-7B as the language model, LLaVA-NeXT training data, and the same training pipeline. To ensure fair comparison, we adopt LLaVA-NeXT's tiling strategy (up to $2\times2+1$ tiles) for handling high-resolution images, as many vision encoders do not support native resolution processing.

Model Configuration		OCR & Document Understanding								General Vision Understanding							
Method	**Vision Tower**	InfoVQA	DocVQA	ChartQA	TextVQA	OCRBench	OCRBenchV2	LiveXivVQA	**OCR Avg**	AI2D	MMB^{EN}	MME^{Cog}	MME^{Per}	POPE	RealworldQA	MMStar	**Other Avg**
CLIP	ViT-L-14-336px	38.9	75.2	66.5	62.5	52.5	23.0	47.4	52.3	73.2	74.6	48.0	75.6	**88.8**	**63.7**	49.0	67.6
MLCD	ViT-L-14-336px	43.5	76.5	67.8	61.7	53.1	24.0	48.4	53.6	77.0	76.4	54.1	79.9	88.7	61.1	51.0	69.7
AIMv2	ViT-L-14-336px	35.4	77.2	**72.7**	**65.9**	57.2	23.9	47.3	54.2	75.4	**78.6**	48.3	75.0	88.4	62.2	50.2	68.3
RICE-ViT	ViT-L-14-336px	**45.2**	**79.2**	72.3	**65.9**	**57.5**	**24.1**	**48.9**	**56.2**	**77.9**	76.6	**54.6**	**80.7**	88.5	63.1	**51.8**	**70.5**
DFN5B	ViT-H-14-378px	38.6	70.9	64.4	59.4	47.3	21.9	46.2	49.8	73.5	73.4	45.8	76.9	88.6	59.9	49.1	66.7
SigLIP	ViT-SO400M-14-384px	41.4	76.7	69.3	64.7	55.4	24.0	48.4	54.3	76.2	77.0	46.1	79.9	88.8	**63.7**	47.3	68.4
SigLIPv2	ViT-SO400M-14-384px	43.7	79.1	70.2	66.2	58.7	25.4	48.6	56.0	**77.0**	77.1	46.6	**80.4**	**89.3**	63.4	**52.8**	69.5
RICE-ViT	ViT-L-14-378px	**48.1**	**82.6**	**75.1**	**66.2**	**58.8**	**25.8**	**49.5**	**58.0**	76.5	**77.6**	**54.1**	79.0	89.1	62.9	51.2	**70.1**
SigLIPv2	ViT-SO400M-16-560px	50.2	86.2	77.4	**70.2**	**62.7**	**26.5**	52.9	60.9	**77.0**	76.5	53.5	**79.9**	**89.3**	**68.2**	**53.1**	**71.1**
RICE-ViT	ViT-L-14-560px	**53.2**	**87.4**	**78.1**	69.0	60.7	26.1	**53.0**	**61.1**	76.9	**78.6**	**56.3**	79.3	88.9	65.1	50.5	70.8
Qwen-ViT from Qwen2.5-VL 7B	ViT-H-14-560px	**55.9**	85.8	78.8	73.7	66.2	26.8	53.4	62.9	78.8	78.4	**62.0**	80.8	88.6	64.2	55.0	72.5
RICE-ViT from OV-1.5 3B	ViT-L-14-560px	53.7	**87.1**	**81.9**	**73.8**	**73.3**	**30.4**	**53.6**	**64.8**	**80.3**	**79.6**	58.6	**82.2**	**89.0**	**67.3**	**56.6**	**73.4**

performance on MMStar (67.7), $\text{MMBench}_{\text{en}}$ (84.1), $\text{MME-RealWorld}_{\text{en}}$ (62.3), $\text{MME-RealWorld}_{\text{cn}}$ (56.1), CV-Bench (80.8), and ScienceQA (95.0). Besides, LLaVA-OneVision-1.5 also presents comparable performance on $\text{MMBench}_{\text{cn}}$ (81.0), $\text{SeedBench}_{\text{image}}$ (77.3), SEED-Bench-2-Plus (69.2), and RealWorldQA (68.1).

6.3 Multimodal Reasoning

LLaVA-OneVision-1.5 exhibits superior multimodal reasoning capabilities compared to Qwen2.5-VL. Specifically, LLaVA-OneVision-1.5-4B outperforms Qwen2.5-VL-3B on all evaluated benchmarks, leading in $\text{MathVista}_{\text{mini}}$ (67.9), WeMath (24.9), MathVision (24.2), MMMU_{val} (52.7), $\text{MMMU-Pro}_{\text{standard}}$ (35.3), and $\text{MMMU-Pro}_{\text{vision}}$ (25.4). Notably, LLaVA-OneVision-1.5-4B surpasses LLaVA-OneVision-7B across all benchmarks. Compared with Qwen2.5-VL-7B, LLaVA-OneVision-1.5-8B also demonstrates gains of 1.0%, 0.3%, 3.2%, 4.1%, and 1.1% on $\text{MatchVista}_{\text{mini}}$, WeMath, MathVision, MMMU_{val}, and $\text{MMMU-Pro}_{\text{standard}}$.

6.4 OCR & Chart Understanding

The interpretation of visual data, including documents and charts, requires a sophisticated array of skills from multimodal large language models, ranging from low-level Optical Character Recognition (OCR) to high-level semantic reasoning. To thoroughly evaluate these capabilities, we assess LLaVA-OneVision-1.5 across seven challenging benchmarks. LLaVA-OneVision-1.5-8B demonstrates robust outcomes on ChartQA (86.5), $\text{CharXiv}_{\text{DQ}}$ (74.1), DocVQA (95.0), $\text{AI2D}_{\text{w M}}$ (84.2), and $\text{AI2D}_{\text{w/o M}}$ (94.1). Notably, LLaVA-OneVision-1.5-4B outperforms Qwen2.5-VL-3B on all seven benchmarks.

6.5 Others

To further elucidate the capabilities of LLaVA-OneVision-1.5, we extend our evaluation to include PixmoCount, CountBench, VL-RewardBench, and V*. LLaVA-OneVision-1.5-8B records scores of 62.2 on PixmoCount, 88.2 on CountBench, 46.7 on VL-RewardBench, and 78.0 on

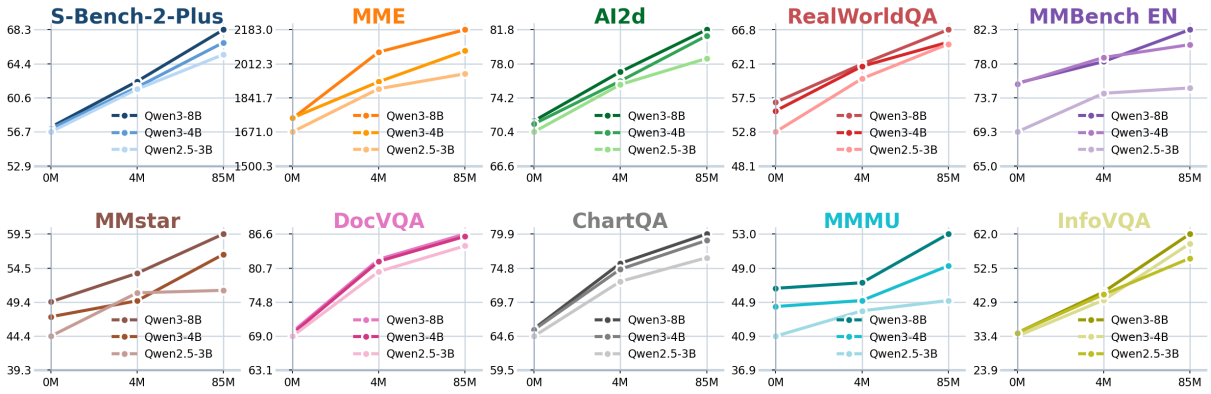


Figure 6 Performance comparison across different mid-training data scales on various benchmarks. Models initially undergo pre-training on LLaVA-558K and are then subjected to mid-training at different data scales (4M, 85M), followed by fine-tuning using the LLaVA-NeXT (Liu et al., 2024b) SFT framework. 0M denotes native pre-training without the mid-training stage.

V*, demonstrating proficiency in counting, visual perception, and visual grounding, on par with Qwen2.5-VL-7B.

6.6 Performance of RL Post-training

To validate the effectiveness of our lightweight RL post-training, we compare the RL-enhanced models (LLaVA-OV-1.5 RL-8B) against their supervised baselines and the strong competitor Qwen2.5-VL-7B. We report results for both standard "fast" inference and "thinking" mode (where the model generates reasoning chains) as shown in Tab. 1.

Core Capability Enhancement. As presented in Tab. 1, RL post-training yields consistent gains across major benchmarks. (1) **Multimodal Reasoning**: The most substantial improvements are observed in multimodal reasoning tasks. In "thinking" mode, our model achieves dramatic gains on WeMath (+7.9), MathVision (+8.8), and MMMU-$\text{Pro}_{\text{vision}}$ (+10.5), demonstrating that the RL-induced chain-of-thought capability effectively unlocks deeper problem-solving skills. (2) **General VQA & OCR**: On standard benchmarks like MMBench and DocVQA, RL maintains or slightly improves the already strong SFT performance, ensuring no regression in general capabilities while specializing in hard reasoning.

Extended Capability Analysis. Beyond the core benchmarks, we utilize Figure 5 to analyze the impact of RL on specific vertical capabilities not fully covered in Tab. 1. (1) **Spatial Reasoning & Grounding**: As shown in Figure 5, RL significantly enhances fine-grained perception. The RL model (Fast mode) consistently outperforms the SFT baseline on spatial tasks like SAT and Ref-L4. Interestingly, while "thinking" mode aids reasoning, it sometimes yields lower scores than "fast" mode on strictly perceptual metrics (e.g., Ref-L4 IoU), suggesting that verbose generation may occasionally interfere with precise coordinate regression. (2) **Coding**: In the coding domain (WebCode, UniSVG), Figure 5 shows that our RL models achieve consistent gains. The "thinking" mode proves particularly effective here, achieving the highest scores on Design2Code and UniSVG, indicating that chain-of-thought reasoning is beneficial for structural code generation.

These results confirm that our rule-based RL framework effectively enhances both the reasoning depth (via thinking chains) and perceptual precision (via direct feedback) of LLaVA-OneVision-1.5, positioning it as a state-of-the-art open model in the 8B parameter class.

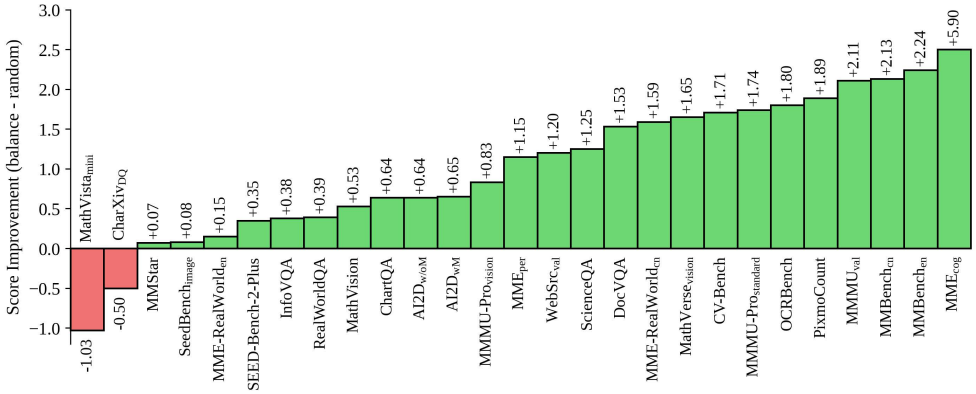


Figure 7 Experimental results using 2M blanced and unbalanced mid-training samples (LLaVA-NeXT-780k as the SFT data) show that using a balanced mid-training dataset yields consistent improvements over a random sampling strategy.

6.7 Ablation Study

6.7.1 Comparison of Different Vision Encoders

In Tab. 2, we evaluate various vision encoders, CLIP (Radford et al., 2021), MLCD (An et al., 2024), AIMv2 (Fini et al., 2025), DFN (Fang et al., 2023), SigLIP (Zhai et al., 2023), SigLIPv2 (Tschannen et al., 2025), and RICE-ViT (Xie et al., 2025), within the LLaVA-NeXT framework. At a resolution of 336 pixels, RICE-ViT surpasses CLIP across all benchmarks, achieving significant gains in InfoVQA (+6.3%) and OCRBench (+5.0%). It also demonstrates significant improvements in document understanding compared to AIMv2. At 378 pixels, RICE-ViT outperforms computationally intensive models such as SigLIPv2 in 9 of 14 benchmarks, notably in InfoVQA (+4.4%), DocVQA (+3.5%), and ChartQA (+4.9%). These results position RICE-ViT as a leading vision encoder, enhancing OCR capabilities and providing robust visual understanding, crucial for applications requiring advanced document analysis and visual reasoning. In addition, we further compare the performance of RICE-ViT with that of Qwen-ViT after incorporating LMM training, where RICE-ViT is derived from LLaVA-OneVision-1.5-3B and Qwen-ViT is derived from Qwen2.5-VL-7B. In the areas of OCR & Document Understanding and General Vision Understanding, RICE-ViT demonstrates average performance improvements of 1.9% and 0.9% compared to Qwen-ViT.

6.7.2 Mid-Training Data Scaling

As depicted in Fig. 6, we present the performance of three different LMMs trained with various scales of mid-training data across ten distinct benchmarks. Employing LLaVA-558K for language-image alignment and standard LLaVA-Next instruction tuning, our findings indicate that scaling the data volume during the high-quality knowledge learning phase consistently enhances model performance across all benchmarks. These results not only underscore the high quality and scalability of the proposed LLaVA-OneVision-1.5-Mid-Traning dataset but also confirm the efficacy of data scaling in improving the performance of LMMs.

6.7.3 Effectiveness of Concept Balance

Fig. 3(a) illustrates that after implementing concept balancing, LLaVA-OneVision-1.5-Mid-Traning exhibits a smoother distribution, thereby enhancing the model's capability to assimilate a more comprehensive set of knowledge. To further validate the effect of concept balance, we conducted a comparative analysis of models trained on 2M concept-balanced data versus those trained on

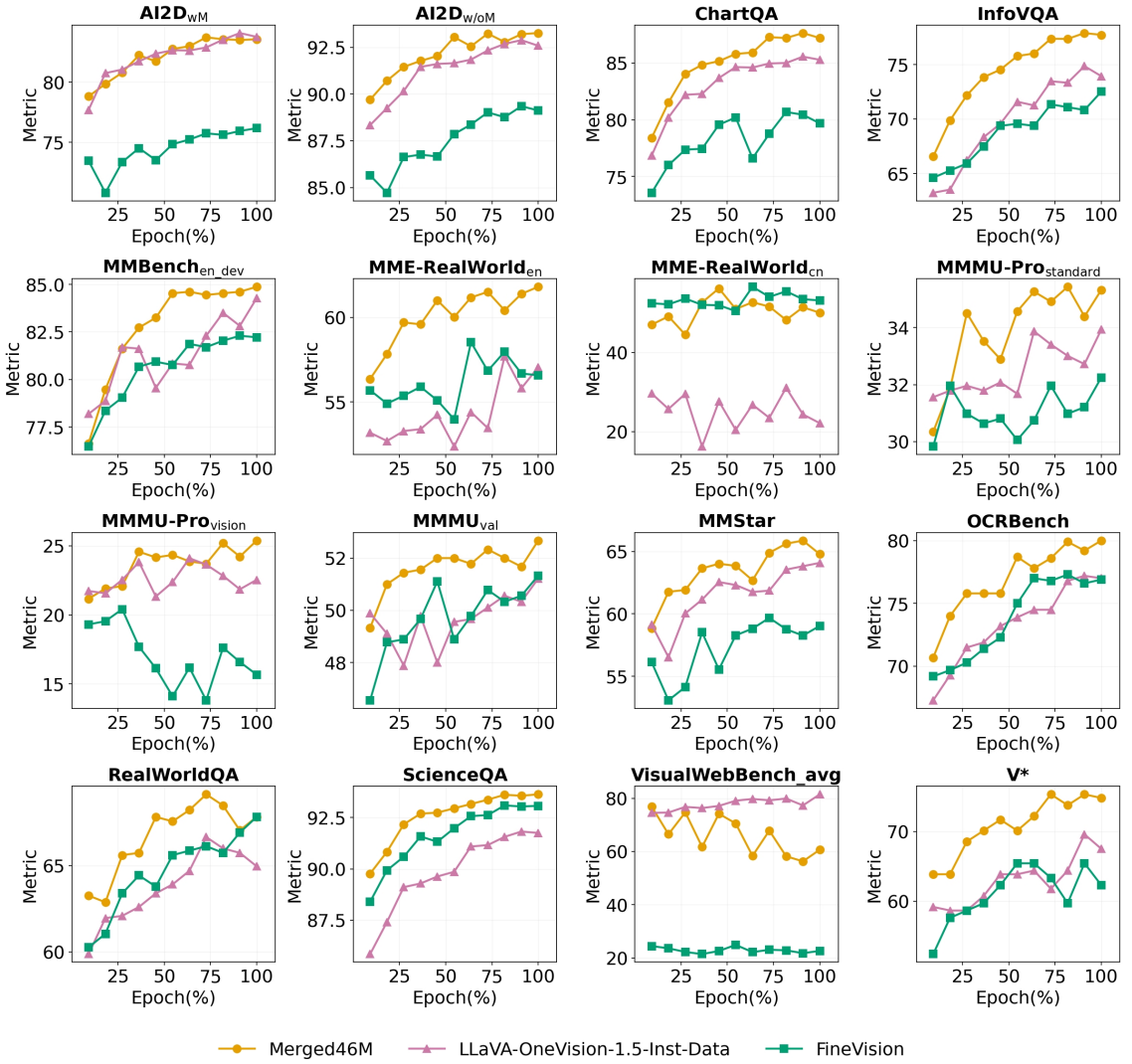


Figure 8 Performance comparison of three datasets (Merge46M, FineVision, and LLaVA-OneVision-1.5-Inst-Data) across 16 benchmarks during the SFT phase, demonstrating the superiority of Merge46M on most benchmarks.

2M data obtained through random sampling. As indicated in Fig. 7, the concept-balanced 2M data set demonstrates superior performance in 25 of 27 evaluated downstream benchmarks.

6.7.4 Instruction Data Quality and Scaling

To enhance performance across diverse VQA tasks, we compile 124 types of instruction data (LLaVA-OneVision-1.5-Inst-Data) for SFT training. We further scale the model capabilities by deduplicating and merging the recently proposed FineVision dataset (Wiedmann et al., 2025), resulting in the Merged46M SFT dataset. To maintain consistent training steps, we double the batch size for Merged46M due to its larger scale. Fig. 8 shows performance comparisons on 16 benchmarks during SFT using three datasets: LLaVA-OneVision-1.5-Inst-Data, FineVision, and Merged46M. LLaVA-OneVision-1.5-Inst-Data achieves performance comparable to FineVision, while the Merged46M dataset delivers the best results across nearly all benchmarks.

7 Conclusions

In this work, we introduce LLaVA-OneVision-1.5, a family of large multimodal models that establishes a new paradigm for constructing high-performance vision-language systems with improved efficiency and reproducibility. We demonstrate the feasibility of training competitive multimodal models from scratch under strict constraints. Our contributions are threefold: a large-scale, curated multimodal dataset; an efficient end-to-end training framework operable under a limited budget; and extensive empirical results demonstrating state-of-the-art performance across diverse benchmarks. The model excels particularly in resource-constrained settings, surpassing strong baselines such as Qwen2.5-VL-7B. This study underscores how open and efficient frameworks can drive progress in multimodal AI, democratizing access to state-of-the-art performance. We envision LLaVA-OneVision-1.5 as a foundational resource that empowers the community to build specialized applications and develop more powerful LMMs across diverse tasks through continued scaling.

References

Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, and Jiankang Deng. Multi-label cluster discrimination for visual representation learning. In ECCV, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. arXiv:2502.13923, 2025.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, and S.-H. Gary Chan. Revisiting referring expression comprehension evaluation in the era of large multimodal models. arXiv:2406.16866, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In NeurIPS, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv:2412.05271, 2024c.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261, 2025.

X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. `https://x.ai/blog/grok-1.5v`, 2024.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In CVPR, 2025.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In ICLR, 2023.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In CVPR, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394, 2023.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In NeurIPS, 2025.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In NeurIPS, 2023.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. arXiv:2505.07062, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In ECCV, 2016a.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In ECCV, 2016b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In ICCV, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In NeurIPS, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. TMLR, 2025a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. In ICLR, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. In CVPR, 2024b.

Jinke Li, Jiarui Yu, Chenxing Wei, Hande Dong, Qiang Lin, Liangjing Yang, Zhicai Wang, and Yanbin Hao. Unisvg: A unified dataset for vector graphic understanding and generation with multimodal large language models. In ACMMM, 2025b.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. In CVPR, 2025c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In CVPR, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 2024c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In ICLR, 2024.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In ACL, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In WACV, 2021.

Minesh Mathew, Viraj Bagal, Dimosthenis Karatzas, and C. V. Jawahar. Infographicvqa. In WACV, 2022a.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In WACV, 2022b.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In ICCV, 2023.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In ACL, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. IJCV, 2015.

Gabriel Herbert Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In NeurIPS, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv:1909.08053, 2019.

Qwen Team. Qwen3 technical report, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In NeurIPS, 2024.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. In arXiv:2502.14786, 2025.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In NeurIPS, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In NeurIPS, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191, 2024b.

Weizhi Wang, Yu Tian, Linjie Yang, Heng Wang, and Xifeng Yan. Open-qwen2vl: Compute-efficient pre-training of fully-open multimodal llms on academic resources. In COLM, 2025b.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In ICLR, 2024c.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In NeurIPS, 2024d.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. In arXiv:2510.17269, 2025.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In CVPR, 2024.

Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, et al. Ccmb: A large-scale chinese cross-modal benchmark. In ACMMM, 2023.

Yin Xie, Kaicheng Yang, Xiang An, Kun Wu, Yongle Zhao, Weimo Deng, Zimin Ran, Yumeng Wang, Ziyong Feng, Roy Miles, et al. Region-based cluster discrimination for visual representation learning. In ICCV, 2025.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In ICLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In CVPR, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In ACL, 2025.

Sukmin Yun, Haokun Lin, Rusiru Thushara, Mohammad Qazim Bhat, Yongxin Wang, Zutao Jiang, Mingkai Deng, Jinhong Wang, Tianhua Tao, Junbo Li, Haonan Li, Preslav Nakov, Timothy Baldwin, Zhengzhong Liu, Eric P. Xing, Xiaodan Liang, and Zhiqiang Shen. Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal LLMs. In NeurIPS, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In ICCV, 2023.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. arXiv:2209.02970, 2022.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In ACL, 2025a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. MME-Realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In ICLR, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479, 2025.

A LLaVA-OV-1.5 vs. Qwen2.5-VL with Same LLM

To enable a fair comparison with Qwen2.5-VL, we train LLaVA-Onevision-1.5-3B based on Qwen2.5-3B-Instruct. As shown in Fig. 9, LLaVA-Onevision-1.5-3B also demonstrates superior performance, achieving better results on 17 out of 27 downstream benchmarks.

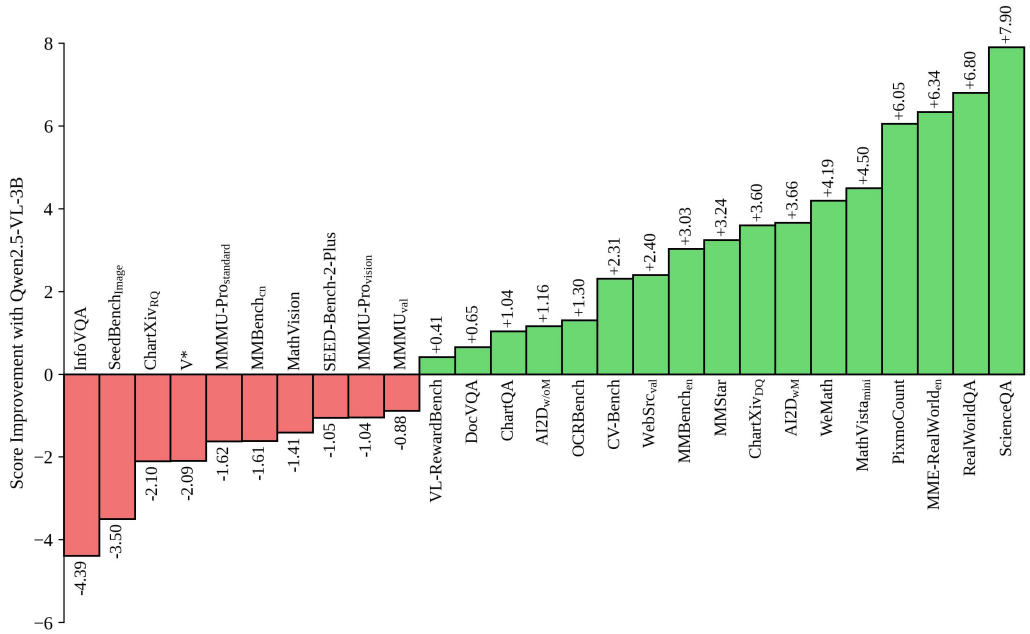


Figure 9 Comparison between LLaVA-OV-1.5-3B and Qwen2.5-VL-3B model on public datasets using the same LLM for both evaluations.

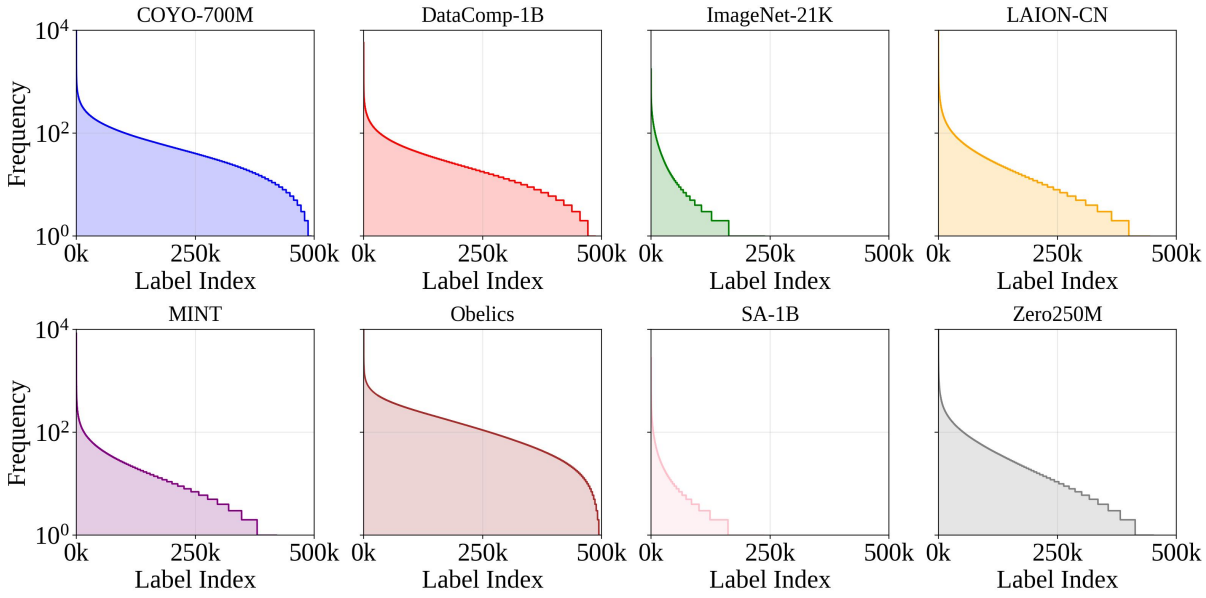


Figure 10 Original concept distributions across eight common vision datasets used in the LLaVA-OneVision-1.5-Mid-Traning dataset.

B Mid-Training Data: Concept Distribution and Top 50 Topics

Fig. 10 presents the raw concept distributions across eight common vision datasets used in the LLaVA-OneVision-1.5-Mid-Traning dataset. All sources exhibit a pronounced long-tail bias, indicating that the original data are far from comprehensive. Obelics displays the broadest and most uniform distribution (slowest tail decay), while others, such as ImageNet-21K and SA-1B, cover fewer concepts with higher frequency concentration. To characterize the semantic space of the mid-training mixture, we apply topic modeling to associated texts and extract the 50 most salient topics (Tab. 3). These span diverse domains, including wildlife, apparel, cuisine, interiors, engineering, electronics, healthcare, WebUI, and cultural activities, offering an interpretable summary of data coverage. This analysis confirms Obelics as the most comprehensive source and highlights coverage gaps in other datasets, informing subsequent data curation.

Table 3 Topic Modeling Results (50 Topics)

Concept	Related words
Wildlife	feathers, wildlife, beak, branch, birds, habitat, scales, plumage, claws, feather, behavior, spots, watchers, snout, species, creature, paws, limbs, flight, drawing
Jewelry and Aviation	jewelry, beads, airplane, aircraft, aviation, plane, sparkle, bracelet, flight, stones, elegance, pendant, accessory, loop, diamonds, gemstones, gemstone, pearl, landing, shine
Internet Humor	humor, phrase, eco, baby, parents, sustainability, frustration, references, meme, sentiment, surprise, mother, farm, disney, twist, internet, cartoonish, novelty, reaction, drawing
Education and Learning	slide, educators, learners, bullet, education, learning, classroom, study, math, guide, student, resource, writing, textbook, skills, problem, publisher, question, list, questions
Watches and Accessories	case, strap, watch, wristwatch, hour, brim, crown, markers, dial, barrel, numerals, grip, accessory, clock, luxury, gun, minute, buckle, rifle, stitching
Gaming and Sports	games, gamers, baseball, armor, jerseys, fi, franchise, sword, basketball, horror, competition, sport, court, gaming, anime, athleticism, readiness, arena, athletes, flames
Certification and Professionalism	stamp, clients, certification, seal, colleagues, stamps, standards, calligraphy, headshot, friendliness, compliance, certifications, profiles, mark, envelope, creases, impression, identification, trustworthiness, german
Vehicles and Racing	seat, wheel, tires, grille, truck, license, headlights, speed, driver, track, rims, quarter, mirrors, race, seats, racing, train, transportation, three-quarter, windshield
Architectural and Mechanical Plans	grooves, architects, drawings, plan, drawing, tire, hub, roller, wheel, tread, plans, coaster, distances, blueprint, 2023, rubber, planners, amusement, skateboard, traction

Continued on next page

Table 3 – continued from previous page

Concept	Related words
Ceremonies and Achievement	achievement, pride, ceremony, award, trophy, university, campus, awards, skeleton, victory, graduation, bones, success, achievements, alumni, casino, medal, certificate, dinosaur, tuxedo
Portraiture and Makeup	finger, eyebrows, jawline, cheek, makeup, cheeks, portraiture, index, nails, individual, strands, thumb, beer, cheekbones, brow, bangs, ear, shot, stubble, creases
Software and Development	software, computer, developers, options, system, user, code, management, file, version, application, menu, fields, input, process, arrows, programming, files, 3d, interfaces
Pets and Footwear	dog, footwear, owners, cat, shoe, laces, toe, sole, heel, paws, tongue, sneakers, soles, mesh, ankle, rubber, foot, dogs, toes, straps
Typography and Design	sans, gradients, letter, variations, symbolism, forms, spacing, thickness, variation, meaning, bold, alignment, curves, designers, representation, emotions, rectangle, curve, knowledge, trademark
Cuisine and Cooking	cooks, meal, ingredients, dish, freshness, vegetables, cooking, sauce, fruit, slices, pieces, herbs, spoon, meat, cuisine, crispy, slice, rice, fruits, chocolate
Home Interiors	coffee, lamp, shelf, living, vase, bedroom, cabinet, pillows, comfort, rug, countertop, cup, homeowners, sofa, pillow, cabinets, counter, flooring, couch, shade
Skincare and Packaging	skincare, liquid, lid, wine, container, ingredients, tube, supplements, screw, premium, luxury, benefits, transparency, labeling, jar, solutions, supplement, powder, oil, spray
Childhood and Holidays	parents, toy, holiday, baby, gift, toys, polka, kids, christmas, decorations, ribbon, bear, caregivers, fun, charm, decoration, greeting, rabbit, birthday, dot
Social Interactions	pen, smiles, postures, mid-20th-century, discussion, celebrity, hairstyles, collage, interactions, collaboration, jackets, notebook, -century, fourth, gestures, cinema, teamwork, interview, ties, plaid
Literature and Romance	love, spine, intimacy, works, couple, publisher, affection, novel, romance, proximity, moments, bond, hardcover, 19th-century, pages, edition, poetry, drama, collection, covers
Inspirational Quotes	quote, self, positivity, inspiration, resilience, philosophy, quotes, hope, help, introspection, journey, phrase, attribution, things, weight, motivation, freedom, contemplation, love, vulnerability

Continued on next page

Table 3 – continued from previous page

Concept	Related words
Waterside Leisure	beach, shore, ripples, relaxation, sand, roofs, boat, sea, pool, bridge, shoreline, turquoise, landscapes, boats, river, lake, deck, skyline, trunks, leisure
Engineering and Mechanics	component, engineers, holes, mechanics, diy, engineering, technicians, screws, hardware, hole, specifications, wires, electronics, manufacturers, steel, manufacturing, manufacturer, circuit, grip, ends
Retail and Pricing	price, package, shoppers, 10, barcode, promotion, pack, store, 20, sale, 100, tag, info, shipping, 50, medication, 30, sales, discount, quantity
Technology and Innovation	innovation, network, lightning, connectivity, bolt, globe, connections, nodes, ideas, cloud, gears, communication, networks, bulb, intelligence, integration, lightbulb, concept, telecommunications, networking
Industrial and Manufacturing	workers, machine, machinery, manufacturing, storage, safety, fish, facility, factory, workshop, pipes, warehouse, maintenance, logistics, task, site, fins, warning, steel, production
Trust and Publication	trust, shield, reliability, studies, china, stakeholders, clients, partners, publication, translation, institution, publisher, public, tradition, scholars, formality, stability, organizations, strength, academics
Healthcare and Fitness	healthcare, fitness, patients, training, wellness, exercise, hospital, practice, gym, strength, clinic, muscles, shirtless, support, lab, treatment, stethoscope, muscle, weight, doctor
Museum Artifacts	motifs, sculpture, bronze, museum, statue, coin, folds, engravings, artifacts, skull, scroll, item, carvings, coins, gallery, pedestal, scratches, artifact, elegance, artwork
Dining and Hospitality	tables, restaurant, café, experience, spot, counter, customers, seating, locals, diners, drinks, shop, establishment, patrons, menu, casual, fixtures, hats, market, drink
Real Estate	estate, tiles, homeowners, shrubs, bathroom, lawn, railings, bushes, porch, property, driveway, staircase, bricks, houses, panes, yard, homebuyers, neighborhood, chimney, concrete
Music and Geography	guitar, musicians, concert, instrument, instruments, country, artist, strings, regions, musician, roads, singer, drum, song, locations, geography, land, performers, performances, maps
Entomology and Containers	lid, antennae, container, insect, butterfly, spots, shell, spines, wing, ridges, candle, storage, rim, insects, hairs, abdomen, entomology, bee, iridescent, observation

Continued on next page

Table 3 – continued from previous page

Concept	Related words
Apparel Design	apparel, pocket, neckline, pockets, waist, garment, fit, cuffs, zipper, hip, belt, tag, straps, torso, cotton, crew, seams, hem, hood, knee
Crafting and DIY	crafters, craft, supplies, crafts, diy, artists, projects, creativity, lab, textiles, textile, laboratory, project, pieces, motifs, stationery, pencil, thread, tactile, workspace
Data and Chemistry	graph, comparison, values, axis, trends, analysis, chart, x-, analysts, market, chemistry, metrics, chemical, visualization, graphs, 2d, years, investors, axes, groups
Electronic Devices	devices, smartphone, computer, keyboard, laptop, lens, monitor, cable, indicator, case, audio, keys, speaker, tablet, usb, ports, security, workspace, electronics, solutions
Recreation and Law	police, law, sheet, superhero, notes, golf, enforcement, club, course, officer, officers, cape, notation, copyright, costume, security, piano, personnel, swing, musicians
Web and UI	navigation, app, user, header, links, webpage, search, smartphone, screenshot, web, options, link, yuan, tabs, battery, url, post, site, menu, email
Space and Music Media	moon, earth, exploration, record, wonder, vinyl, planet, purples, astronomy, records, sphere, thirds, crescent, two-thirds, landmasses, herbarium, rocket, oceans, cloud, magic
Business and Finance Documents	bullet, info, finance, header, management, money, address, documents, topics, question, headers, description, list, headings, job, email, fields, slide, dollar, newspaper
Weddings and Traditions	wedding, lace, embroidery, dance, festival, dresses, indian, traditions, bouquet, occasion, gown, couple, bride, costumes, couples, weddings, outfits, tradition, headscarf, veil
Transportation and Fiction	horse, ship, bike, mystery, bicycle, soldiers, handlebars, thriller, hull, motorcycle, crime, smoke, genre, rider, genres, cyclists, drama, horses, war, dramas
Botany and Gardening	stems, veins, cluster, stem, blooms, bloom, gardeners, soil, gardening, tips, clusters, bark, buds, trunk, stamens, botany, canopy, blossoms, fishing, droplets
Vast Landscapes	terrain, peaks, landscapes, desert, expanse, slopes, range, formations, vastness, isolation, grandeur, peak, hikers, valley, stones, mist, appeals, earth, slope, awe
Biological Diagrams	diagram, arrows, biology, research, documentation, anatomy, cell, diagrams, purposes, cells, blood, flow, study, emergency, specimen, scientists, tissue, understanding, measurements, response

Continued on next page

Table 3 – continued from previous page

Concept	Related words
Religious Architecture	carvings, grandeur, church, robe, dome, reverence, christian, robes, temple, site, statue, landmark, spires, arches, european, towers, statues, arch, castle, spire
Time and Health Data	dates, days, calendar, pm, masks, covid-19, times, info, week, month, ray, pandemic, schedule, daily, 1st, sunday, flyer, dna, 2020, helix
Public Speaking	speech, bubble, press, speaker, podium, politics, pin, government, formality, communication, mark, bubbles, microphones, announcement, lapel, u.s., gesture, exclamation, gravity, gestures
Urban Streetscape	signage, mid-morning, -morning, sidewalk, pavement, parking, pole, traffic, hotel, poles, store, locals, awning, lot, streetlights, pedestrians, station, residents, storefront, japan

C Contributors

Contributors

- Xiang An
- Yin Xie
- Kaicheng Yang
- Wenkang Zhang
- Xiuwei Zhao
- Zheng Cheng
- Yirui Wang
- Songcen Xu
- Changrui Chen
- Didi Zhu
- Chunsheng Wu
- Huajie Tan
- Chunyuan Li
- Jing Yang
- Jie Yu
- Xiyao Wang
- Bin Qin
- Yumeng Wang
- Zizhen Yan

Project Leaders

- Ziyong Feng
- Ziwei Liu
- Bo Li
- Jiankang Deng