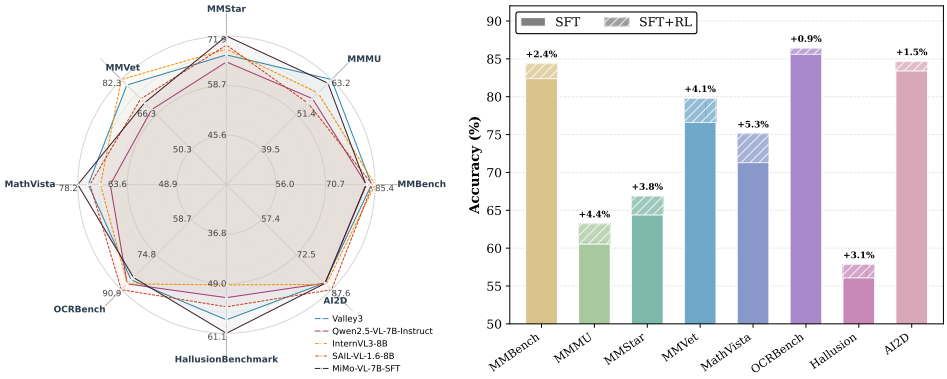
Valley3 Technical Report

Valley Team, ByteDance Group

Abstract

The rapid advancement of Multimodal Large Language Models (MLLMs) is often hindered by the opacity of leading architectures and training methodologies, limiting reproducibility and community progress. In this paper, we introduce **Valley3**, a high-performance 8-billion-parameter MLLM, and provide a fully transparent blueprint for its construction. We propose a quantitative framework for optimal backbone selection—integrating Qwen3-8B-Base with Qwen2-VL-ViT—and demonstrate that rigorous, model-based data curation is the primary driver of performance scaling. Beyond the typical two-stage training pipeline, Valley3 employs a novel post-training paradigm that combines Mixed Preference Optimization (MPO) for general alignment with the GThinker framework. This approach utilizes supervised cold-start, text-only reinforcement warm-up, and multimodal reinforcement fine-tuning to significantly enhance complex reasoning capabilities. Extensive evaluations on OpenCompass and other benchmarks demonstrate that Valley3 achieves competitive performance among similarly sized open-source models. We publicly release our complete codebase, data recipes, and model checkpoints to facilitate future research.



(a) Comparison with baselines. (b) Reinforcement Learning Gains.

Figure 1: (a) Valley3 achieves consistently strong and well-balanced performance across a diverse set of benchmarks, demonstrating competitive reasoning, perception, and robustness abilities compared to state-of-the-art multimodal LLMs. (b) Introducing reinforcement learning yields further improvements across all OpenCompass tasks, validating its role in enhancing both factual reliability and complex reasoning, with accuracy gains of up to +5.3% on mathematical reasoning benchmarks.

1 Introduction

Multimodal large language models (MLLMs) have advanced rapidly and competitively [2–4, 39, 54, 56, 57, 66, 70, 75], with open-source models continuously pushing the boundary of multimodal understanding and reasoning. These developments empower AI systems to comprehend and analyze

Last updated: November 27, 2025

images and videos, bringing them closer to natural human perception and interaction with the world. The recent remarkable progress of MLLMs can be attributed to the continuous evolution of foundation language and vision models, advances in training strategies, and the increasing quality of open-source data. In particular, with the rise of reinforcement learning (RL) [1, 59, 73], these MLLMs have evolved from handling simple perception and cognition tasks to performing complex, open-ended reasoning and decision-making tasks.

However, most leading models keep their architectural designs, training configurations, and data compositions closed-source. This opacity limits the community's ability to gain practical insights into effective model design and optimization, making it difficult to reproduce or build powerful multimodal models from scratch, slowing progress in this field. To bridge the gap, we introduce **Valley3**, a high-performance, open-source Multimodal Large Language Model built at the 8-billion-parameter scale. The core philosophy of **Valley3** is centered on radical transparency: we provide comprehensive details of every component from foundation model selection and data curation to pre-training and post-training strategies, providing a fully reproducible blueprint for building next-generation MLLMs.

We retain the modular design philosophy of Valley2 [70] while updating the pre-trained backbones to *Qwen3-8B-Base*[72] as the language model and *Qwen2-VL-ViT*[62] as the vision encoder. To quantitatively analyze the composition of the backbone models, we introduce an evaluation framework that scores LLM–ViT pairings by their cross-modal alignment ability. Transparently, the primary gains in **Valley3** come from the high-quality data construction: we scale a two-stage curation pipeline that 1) performs rigorous heuristic filtering, including domain balancing, safety filtering, and deduplication; 2) applies model-based quality scoring to assess the quality of data and perform filtering to retain high-quality samples. During training, we observe predictable scaling behavior with respect to data volume, demonstrating the effectiveness of our data quality and cleaning pipeline.

Benefiting from the high-quality data, we further undertake a deeper exploration of the training paradigm, instilling strong reasoning capabilities and yielding unexpectedly substantial gains. While the pre-training stage largely follows the same three classic phases as Valley2 with improvements primarily in data scale and quality, our core innovations are concentrated in the post-training paradigm. The post-training phase consists of two main parts, spanning four distinct stages. Firstly, Mixed Preference Optimization (MPO)[63] on curated pairwise preferences remains central for alignment, for which we further optimize the MPO algorithm and conduct several comprehensive ablation studies. Secondly, we further enhance the model's reasoning capabilities under the GThinker framework [80], which includes (1) a supervised cold start to stabilize instruction following; (2) a text-only RL warm-up to strengthen reasoning capabilities; and (3) multimodal RL to achieve multimodal reasoning generalization. This integrated pipeline, guided by lightweight visual cues for reflection, yields more coherent and verifiable multimodal reasoning than the Valley2 recipe.

Comprehensive evaluations across eight public OpenCompass benchmarks and other widely adopted benchmarks demonstrate Valley3's highly competitive performance, achieving state-of-the-art multimodal reasoning among similarly sized open-source models. Furthermore, we evaluate the model's performance at each stage, noting that the post-training phase under the GThinker framework yields consistent performance improvements across all benchmarks.

In summary, our contributions are as follows:

- **A High-Performance and Transparent MLLM.** We introduce Valley3, an 8B open-source MLLM, and provide a fully reproducible blueprint for its creation, detailing its architectural design, data curation, and comprehensive training pipeline to advance community research.

- **A Principled Blueprint for Architecture and Data.** We establish a systematic framework for model design, including a novel evaluation method for selecting optimal vision and language backbones. We also construct a two-stage data curation pipeline that combines heuristic filtering with model-based scoring to ensure high-quality data at a billion-sample scale.

- **A Comprehensive Training Recipe.** Building on established pre-training principles, we build a well-designed and advanced post-training paradigm, which first employs MPO for preference alignment to enhance the model's general ability and then applies the GThinker framework to specifically boost complex reasoning with visual cues for reflection.

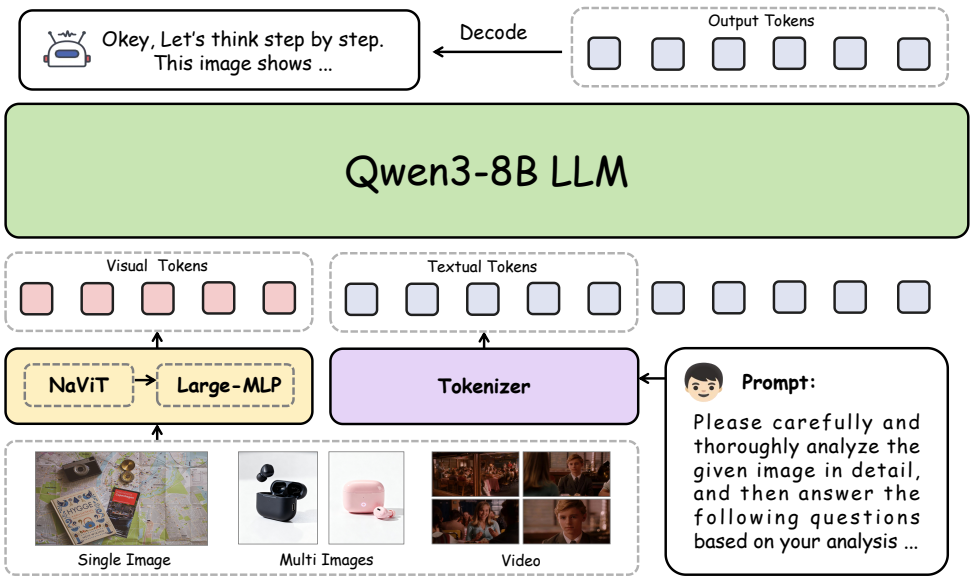


Figure 2: **The architectural overview of Valley3.** The model integrates *Qwen2-VL-ViT* as the vision encoder to handle dynamic resolutions and *Qwen3-8B-Base* as the strong language backbone. These components are bridged by a two-layer large MLP with a hidden dimension of 64k to ensure robust Visual-Text alignment. Valley3 excels at handling tasks across various modalities, including single-image, multi-image, and video.

2 Model Architecture

2.1 Overview

For the LLM, we adopt Qwen3-8B-Base due to its strong reasoning and language comprehension capabilities. The vision encoder utilizes Qwen2-VL-ViT, which supports dynamic-resolution inputs—a more robust alternative to the commonly used tiling approach when processing images with extreme aspect ratios. The projector applies a 2×2 pixel-shuffle downsampling on visual tokens, followed by a two-layer MLP with a 64k hidden dimension, offering strong alignment capacity between modalities. This architectural design achieves a balanced trade-off among representational power, computational efficiency, and multimodal adaptability.

2.2 Language Backbone Analysis

We adopt the Qwen3-8B family as the language model backbone, which is available in two official variants: Qwen3-8B-Base and Qwen3-8B. We paired both variants with Qwen2-VL as the vision encoder and trained them on large-scale captioning datasets. Results show that Qwen3-8B-Base achieves approximately **0.01** lower average training loss compared to Qwen3-8B, while also exhibiting superior performance on the held-out caption validation dataset. Consequently, we adopt Qwen3-8B-Base as the foundation LLM for Valley3.

2.3 Vision Encoder Evaluation

Due to resource and data constraints, we opted not to further pre-train the ViT with a SigLip-style loss [79]. To select the most suitable model from a wide range of publicly available ViTs, we designed a comprehensive evaluation framework and a series of systematic experiments to compare various combinations of models and architectures.

Leveraging Qwen3-1.7B-Base as a lightweight proxy, we evaluated a broad spectrum of ViTs classified into two distinct categories: **Dynamic-Resolution ViTs** (e.g., *Qwen2.5VL-ViT*[5], *Qwen2VL-ViT*[61], *MoonViT*[55]), which adaptively adjust the number of visual tokens based on input resolution, and **Fixed-Resolution ViTs** (e.g., *Aimv2*[17], *InternViT-v2.5*[11], *Oryx-ViT*[34], *SigLip*[79]), where we employ a tiling strategy to accommodate variable-resolution inputs.

Based on the hypothesis that a robust ViT must demonstrate superior alignment capabilities across diverse contexts, we established a systematic evaluation framework to guide the selection. To evaluate cross-modal alignment robustness, we established a systematic framework using 2.5 million image-text pairs. To ensure fairness, we froze both ViT and LLM parameters and exclusively trained the projectorto bridge the modality gap. We further constructed a comprehensive benchmark covering Captioning, Grounding, and OCR. Crucially, we stratified Image Captioning by resolution (Low: $< 500^2$, Medium: 500^2–800^2, High: $> 800^2$ pixels) and Video Captioning by duration (Short: 0–30s, Medium: 30s–60s, Long: 1–3 min) to assess multi-scale adaptability. As illustrated in Table 1, Qwen2VL-ViT delivers optimal performance with the Qwen3-1.7B-Base proxy, justifying its selection as our foundational vision encoder.

Table 1: **Quantitative evaluation of vision encoder backbones**. We compare different ViT variants under variable and fixed resolution settings with a *Qwen3-1.7B-Base* proxy LLM. The results demonstrate that *Qwen2VL-ViT* consistently outperforms other architectures across diverse tasks—including Captioning, Grounding, and OCR—justifying its selection as the optimal visual encoder for Valley3.

Metric	**Qwen2.5VL-ViT**	**Qwen2VL-ViT**	**MoonViT**	**Aimv2**	**InternViT-v2.5**	**Oryx-ViT**	**SigLip**
Resolution	Variable	Variable	Variable	Fixed (Tiling)	Fixed (Tiling)	Fixed (Tiling)	Fixed (Tiling)
Parameters	600M	600M	400M	600M	300M	400M	400M
Resolution	NaViT	NaViT	NaViT	448	448	384	384
Final Score	33.26	**34.42**	33.39	33.76	33.88	30.41	32.48
E-com Product Caption	47.26	51.99	50.54	49.47	45.77	44.79	50.69
E-com Live Caption	25.07	25.99	26.87	25.86	25.19	24.48	25.48
E-com Video Caption	26.29	27.39	27.18	27.82	26.75	25.59	25.58
Image Caption (Low)	6.38	4.26	7.37	16.81	19.54	13.95	21.19
Image Caption (Medium)	42.41	45.80	45.76	41.00	43.23	34.99	37.19
Image Caption (High)	58.16	58.00	57.16	56.87	61.76	49.46	56.50
Video Caption (Short)	15.42	15.67	14.44	12.40	12.21	8.84	9.49
Video Caption (Medium)	11.93	12.71	10.45	8.71	8.36	6.61	7.42
Video Caption (Long)	8.68	8.79	5.00	4.21	4.76	4.36	1.76
Grounding Caption	41.56	43.71	43.36	42.64	44.93	39.69	41.96
OCR Caption	82.70	84.30	79.20	85.60	80.20	81.80	80.10

3 Pre-Train

3.1 Data Construction

Prior work has shown that high-quality pre-training data is foundational to building powerful models [6, 8, 23, 24, 26, 28, 30, 31, 37, 38, 41, 44, 49, 50, 52, 53, 58, 68, 69, 81, 83, 85]. However, most releases provide only data sources or partial processing pipelines, offering limited end-to-end reproducibility and quality control. To address this gap, we construct and open-source an end-to-end, high-quality data curation pipeline for pre-training, detailed below.

For Valley3, we curated a dataset of approximately 200 million raw samples, primarily sourced from open-source platforms such as HuggingFace, ArXiv, and GitHub. To address the suboptimal quality inherent in such large-scale raw data, we implemented a rigorous pipeline for data cleaning and type management, ensuring high quality and diversity throughout all training stages. This chapter first presents the overall data cleaning pipeline for Valley3's pre-training data, followed by a detailed explanation of the collection and management of three critical data types. The resulting datasets will be publicly released to support further research.

3.1.1 Data Curation

To ensure data quality and reduce noise, all samples pass through a rigorous two-stage curation pipeline that combines heuristic filtering with model-based scoring to maximize quality while preserving diversity, as shwon in Figure 3.

The initial stage involves a heuristic-based approach to filter outliers: **Image Quality Check** is ensured by excluding images and videos with extreme aspect ratios(exceeding $5:1$) or resolutions outside a standard range (shorter side below 28px or longer side exceeding 4096px) to target corrupted files and unrepresentative formats. **Text Quality Inspection** is maintained by removing samples with excessively long text (over 8k tokens) or high repetition rates (detected via n-gram frequency) to mitigate noise. **Image-Text Deduplication** removes exact and near-duplicates using a

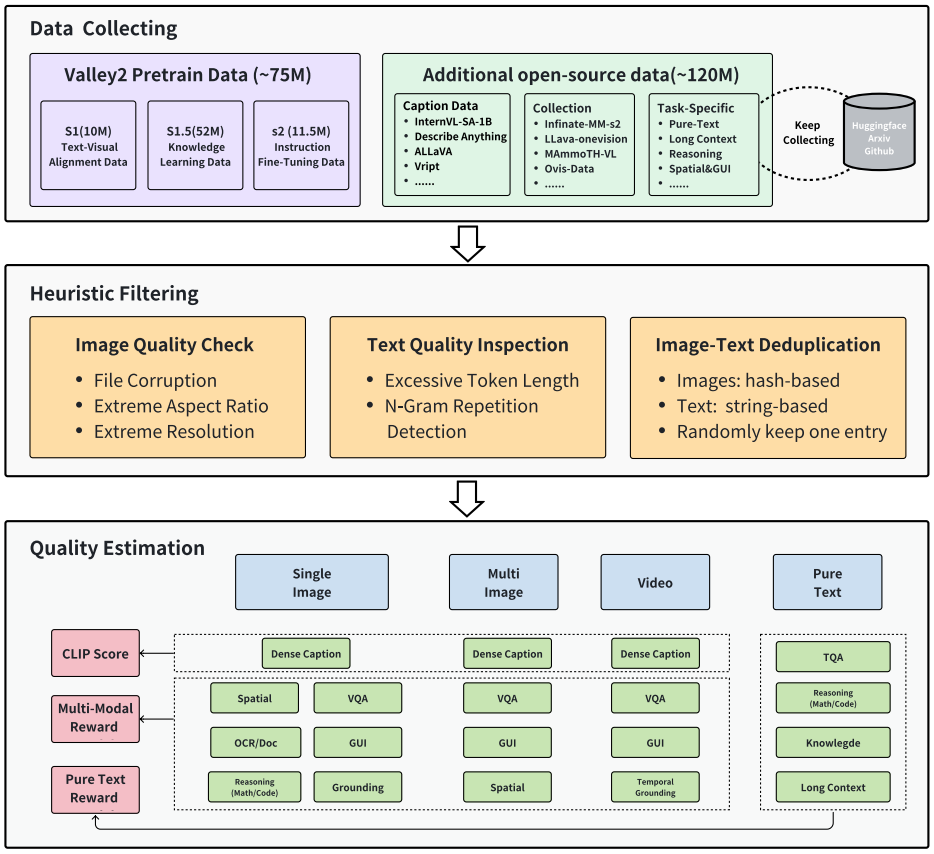


Figure 3: **The data curation pipeline for pre-training.** Building upon the Valley2 data foundation, we aggregated extensive open-source datasets to form a raw corpus of approximately 200M samples. To ensure high quality at this scale, the data first undergoes rigorous heuristic filtering (e.g., resolution checks, deduplication) to remove noise. This is followed by model-based scoring (e.g., CLIP Score, Reward Model) to retain only the high-quality samples, balancing diversity with quality.

multimodal procedure where images are indexed with perceptual hashes (pHash)[27] and text fields undergo Unicode NFKC normalization, lowercasing, punctuation/URL stripping, and whitespace canonicalization, retaining the highest-scoring representative to resolve conflicts.

The second stage employs a domain-specific model-based scoring system to further refine the data. This includes **Caption Data** scoring using CLIP-based [47] similarity as a proxy for semantic coherence to filter samples below empirically determined thresholds; **General VQA** scoring via a dedicated reward model [67] trained on prior model outputs and human-verified labels to predict answer correctness; and **Pure-text Data** evaluation leveraging an LLM-based reward model [35] that conducts multi-dimensional scoring across correctness, completeness, and helpfulness metrics while incorporating self-consistency checks to mitigate systematic evaluation biases.

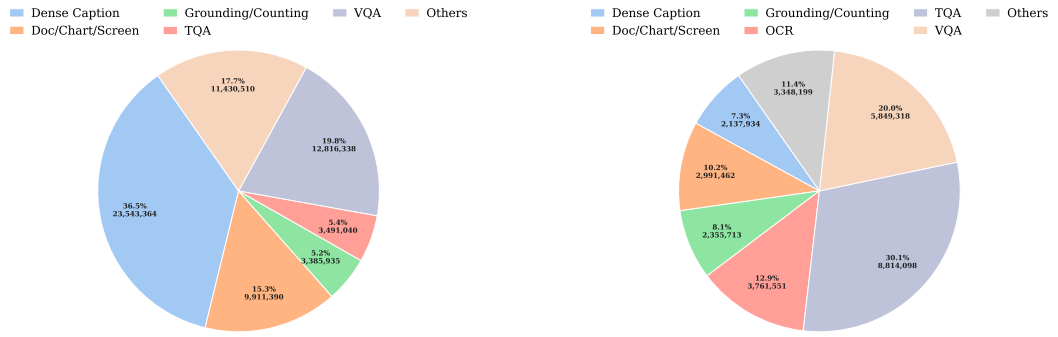
3.1.2 Data Composition

Visual-Text Alignment Data Visual-text alignment data, primarily consisting of caption data, forms the foundation of the model's visual understanding capabilities. To build a robust dataset, we curate a diverse collection of images and videos from prominent open-source datasets[7, 10, 12, 16, 20, 45, 84]. These data undergo the visual quality, text quality, and duplicate detection components of the aforementioned cleaning pipeline to ensure initial data quality. Subsequently, Visual-Text alignment scoring is particularly crucial, guaranteeing semantic consistency between image/video content and text descriptions.

Knowledge Learning Data The pre-training dataset for Valley3 is constructed to broaden multimodal capabilities across image, video, and text tasks. Building upon previous training corpora,

we expanded the dataset by aggregating numerous open-source datasets spanning the three modalities. The resulting dataset constitutes a unified, large-scale collection designed to support Dense Captioning, Image–Text Interleave, Grounding, OCR, Document–Chart–Screen, and general VQA tasks. To ensure consistency and high quality, the entire data curation pipeline described above is applied end-to-end throughout dataset construction. The task distribution of the constructed dataset is illustrated in Figure 4a, and the detailed data sources are provided in Appendix C .

Instruction Fine-Tuning Data Instruction fine-tuning data aims to build the highest quality dataset by using a reward model to score the samples from the knowledge learning data, retaining high-quality data for instruction fine-tuning. Concurrently, we supplement new types of knowledge to enhance new model capabilities, such as GUI understanding, Spatial Reasoning, and complex Reasoning data. This portion of the data is also processed using the complete data curation pipeline described above. The task distribution of the constructed dataset is illustrated in Figure 4b, and the detailed data sources are provided in Appendix D



(a) Distribution of Knowledge Learning Data. (b) Distribution of Instruction Fine-Tuning Data.

Figure 4: Task distribution comparison.

3.2 Training Recipe

3.2.1 Training Phase

Our pretraining pipeline, as illustrated in Figure 5, consists of three main stages, each designed to gradually enhance the model's multimodal capabilities: first establishing a cross-modal foundation, then expanding the model's knowledge base, and finally refining its instruction-following capabilities. The detailed training configurations and data statistics for each stage are summarized in **Table 2**.

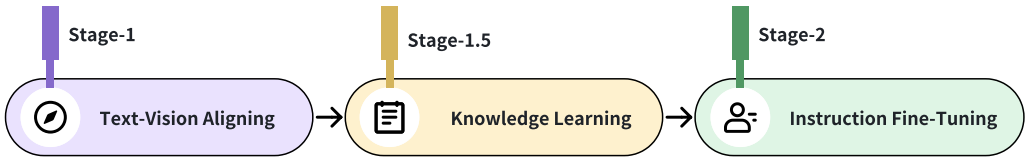


Figure 5: The comprehensive pre-training roadmap of Valley3

Text-Vision Alignment (Stage 1). This initial stage focuses on bridging the modality gap between the pre-trained vision encoder and the Large Language Model (LLM). Using a dataset of 10 million image-caption pairs, we train only the projector. The goal is to map visual features into the LLM's embedding space, establishing a fundamental cross-modal understanding. This stage is conducted with a sequence length of 4k and a learning rate of 1e-4.

Knowledge Learning (Stage 1.5). The core objective of this stage is to infuse the model with a vast amount of world knowledge. We expanded the training to unfreeze the LLM weights and utilize a massive, diverse dataset of 62 million samples (85B tokens). This broad exposure enables the model to learn rich, cross-modal concepts. The learning rate is reduced to 1e-5, and the sequence length is increased to 8k.

Instruction Fine-Tuning (Stage 2). The final stage hones the model's ability to follow complex instructions and perform nuanced reasoning. The focus shifts from raw knowledge acquisition to quality and diversity of tasks. We use a curated, high-quality dataset of 30 million samples (43B tokens) that includes all data types from the previous stage, plus more challenging instruction data focused on Spatial, GUI, Long-context, and Reasoning. This stage refines the model's conversational and problem-solving skills, preparing it for real-world applications. The learning rate and sequence length remain consistent with Stage 1.5, at 1e-5 and 8192, respectively.

Table 2: Detailed training configurations and data statistics for pre-training stages.

	Stage 1	**Stage 1.5**	**Stage 2**
Datasets	Image-Caption Pairs	Pure Text, Dense Caption, OCR/Doc/Chart, QA, Grounding	Pure Text, Dense Caption, OCR/Doc/Chart, QA, Grounding, GUI, Spatial, Long Reasoning
Learning Rate	1.0×10^{-4}	1.0×10^{-5}	1.0×10^{-5}
Samples	10M	62M	30M
Tokens	15B	85B	43B
Optimizer	AdamW	AdamW	AdamW
Sequence Length	4096	8192	8192
Compute Resources	48× H100 GPUs	48× H100 GPUs	48× H100 GPUs

3.2.2 Scaling Law

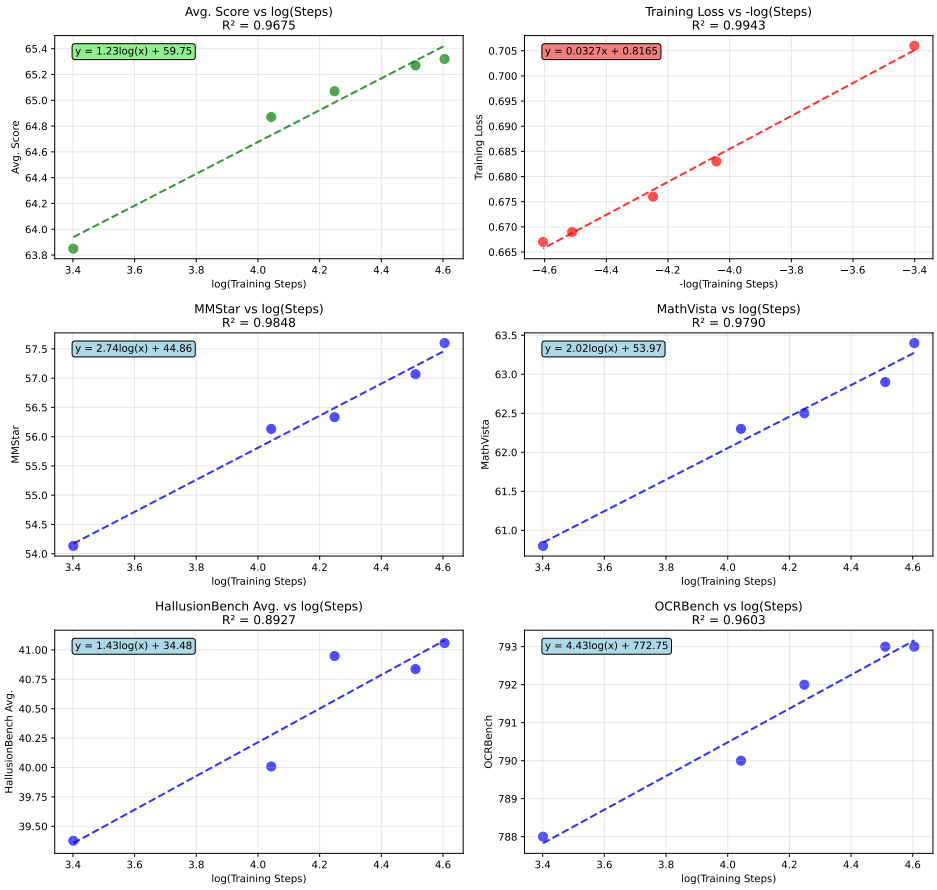


Figure 6: **Scaling behavior during the Knowledge Learning stage (Stage 1.5).** Periodic evaluations on OpenCompass demonstrate a consistent log-linear scaling law. This trend holds true across individual domains (e.g., Math, OCR), validating the effectiveness of our data curation pipeline.

To validate the scalability and quality of our pretraining data pipeline, we analyzed the model's learning trajectory during Stage 1.5 across eight OpenCompass benchmarks. As illustrated in Figure 6, the model strictly adheres to scaling laws, evidenced by a strong log-linear correlation between training steps and average scores ($R^2 = 0.9675$) and a consistent decline in training loss ($R^2 = 0.9943$). This predictable behavior confirms that our high-quality data effectively translates compute into capability, avoiding the noise or stagnation often seen in less robust datasets.

Furthermore, domain-specific analysis reveals that while performance improves universally, the rates of gain vary. We observed the steepest growth in OCR and multi-modal reasoning (e.g., MMStar), highlighting the pipeline's effectiveness in unlocking text-centric and logical capabilities, while mathematical reasoning and hallucination mitigation showed steady, continuous progress. These trends underscore the breadth and efficacy of our data construction, ensuring holistic model improvement throughout extended training steps.

In conclusion, the predictable scaling behavior validates the efficacy of our overall pretraining data construction, suggesting that extended training will yield further improvements.

4 Post-Train

Current leading multimodal large models [4, 22, 74, 86] often rely on carefully designed post-training phases to reach their full potential and ensure alignment with human preferences. However, the details of these training steps are often not fully disclosed. Based on our early exploration of DPO in Valley2 [70] and insights from GThinker [80], we propose a novel four-stage post-training pipeline for Valley3. This pipeline starts with **Preference Optimization** to further align the model with human preferences and improve its general capabilities. Next, a **Cold Start** stage helps the model learn the cue–reflection reasoning pattern. Finally, we strengthen the model's reasoning abilities through two stages: **Text-Only Warm-up** and **Multimodal Reinforcement Learning**. To support this process, we systematically construct high-quality datasets and design a comprehensive reward system, as detailed in the following subsections.

4.1 Data Construction

To support post-training, we collect data from public sources and produce high-quality annotations. Preference Optimization Data is sampled from open-source datasets, while Cold Start and RL Data follow the unified processing pipeline in Figure 7.

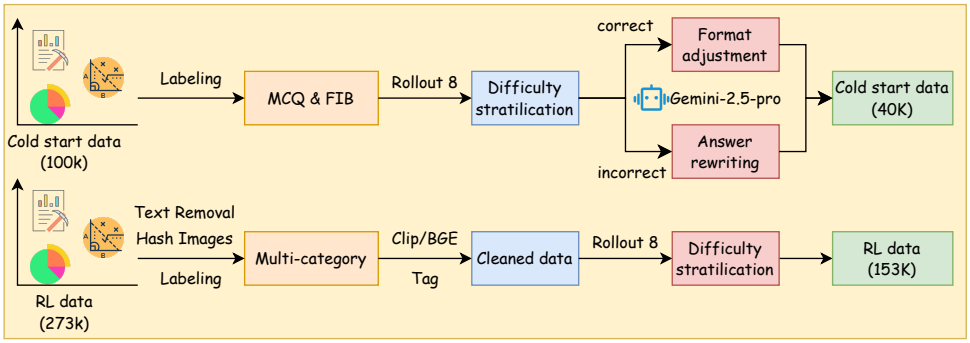


Figure 7: **The data processing pipeline for Cold Start and RL Data.** The Cold Start branch (top) refines 100K raw samples into 40K high-quality instances using rollout-based difficulty stratification and Gemini-2.5-pro for response refinement. The RL Data branch (bottom) filters 273K samples through multi-stage deduplication and difficulty assessment to yield 153K verifiable samples.

4.1.1 Preference Optimization Data.

To enable effective alignment with human preferences, we curate a preference optimization dataset that provides high-quality pairwise supervision signals. Specifically, we sample approximately 100K preference pairs from the MMPR [65, 66] dataset, covering diverse domains such as General-VQA,

Science, Chart, Mathematics, OCR, and Document understanding. Each pair consists of a chosen and a rejected response, reflecting human preference judgments in terms of helpfulness, accuracy, and reasoning quality.

4.1.2 Cold Start Data.

The cold-start stage serves as the first step of the reasoning training, with the primary objective of enabling the model to learn the reasoning pattern. Therefore, high-quality data is essential for stable and effective RL training. In this section, we describe the construction of our cold-start dataset from three perspectives: Data Composition, Think Format, and Response Curation.

Data Composition. During the cold-start data construction stage, we collected 100K samples and converted them into question–answer pairs spanning multiple domains, including general knowledge, mathematics, video understanding, and graphical user interface (GUI) tasks. To ensure both data reliability and task relevance, we prioritized verifiability in the selection process—each task is designed such that correctness can be determined through explicit and objective criteria. This design constraint enables the cold-start dataset to provide a stable and high-quality supervisory signal before reinforcement learning begins.

Think Format. To enable effective incorporation of visual information during multimodal reasoning, we adopt the think formatting strategy of the GThinker [80] framework. The objective is to preserve the natural reasoning format to the greatest extent possible while introducing visual cues to facilitate reflection.

Response Format

<think> <vcues_*> visual cue description </vcues_*> </think>
<answer> final answer </answer>

Response Curation. To ensure that the cold-start dataset contains responses that are both accurate and diverse in reasoning behavior, we run eight rollouts per sample and categorize sample difficulty based on the corresponding pass rate. Additionally, we leverage Gemini-2.5-Pro [13] as an external evaluator to normalize the output structure of correct responses and correct erroneous ones. This process enforces unified formatting and enhances the reliability of the curated responses, ultimately providing a more stable foundation for multimodal reinforcement learning.

4.1.3 Reinforcement Learning Data

The objective of reinforcement learning data selection is to maximize the proportion of verifiable data in each sub-domain, thereby enabling more effective policy improvement through reinforcement learning. In the following, we provide a detailed description of the reinforcement learning data used in our experiments from two perspectives: Data Distribution and Data Selection.

Data Distribution. For the reinforcement learning (RL) stage, we integrated a dataset encompassing a broad spectrum of tasks, including mathematical reasoning, chart interpretation, and visual question answering. Consistent with the cold-start phase, our data selection strategy prioritizes verifiable samples. This design enables reliable reward computation, thereby ensuring stable optimization dynamics and promoting more consistent policy improvement throughout training.

Data Selection. To ensure dataset quality and diversity, we construct a unified data processing pipeline that merges data from multiple sources into a consistent multimodal format. The resulting dataset contains 273K samples, along with an additional 32K pure-text mathematics problems. The cleaning procedure involved three key steps: (1) applying image hashing with a Hamming distance threshold to remove duplicate text and near-duplicate images, retaining only samples with distinct questions; (2) performing CLIP+BGE joint embedding with a similarity threshold to filter out redundant data; and (3) assigning task-specific labels—including OCR, Math, GEN, and General (mathematics and science)—to the remaining samples. This process established a high-quality data foundation for subsequent RL.

To assess problem difficulty and support curriculum learning, we use the cold-start model to perform eight rollouts for each sample. Problems that are consistently answered correctly are labeled easy, those consistently answered incorrectly are labeled hard, and the remaining are labeled medium. This difficulty stratification provides a reliable basis for reinforcement learning by offering structured guidance during policy optimization.

4.2 Reward System

For Reinforcement Learning with Verifiable Rewards, it's essential to reward the rollout completion for each sample accurately and consistently. Instead of focusing on a single type, we build a comprehensive and unified reward system to support nearly all common tasks related to visual perception, understanding, reasoning, and interaction.

Table 3: **Reward design specifications across different domains**. Our evaluation framework for reward design is built to ensure verifiable feedback across a diverse set of tasks. It is organized into six primary categories (STEM, Chart & OCR, Grounding, Spatial, GUI, and Video), which are composed of ten specific sub-domains in total.

Category	Domain	Reward Design Details
STEM	Math	Numeric matching via SymPy with tolerance.
	Physics	Numeric: similar to Math; String: $R_{\text{string}}(s) = \textbf{EM}$ if $\lvert s\rvert < r,\ \textbf{ED}$ otherwise.
	Chemistry	Similar to Physics.
Chart & OCR	OCR	Numeric: **EM**; Equation: similar to Math; String: $R_{\text{string}}(s) = \textbf{EM}$ if $\lvert s\rvert < r,\ \textbf{ED}$ otherwise.
	Chart	Similar to OCR.
Grounding	Grounding	IoU reward $= \frac{\#\{\text{boxes s.t. IoU}>\tau\}}{\#\text{boxes}}$; String: $R_{\text{string}}(s) = \textbf{EM}$ if $\lvert s\rvert < r,\ \textbf{ED}$ otherwise.
Spatial	Counting	Similar to Math.
	Geo Guess	Numeric: similar to Math; String: $R_{\text{string}}(s) = \textbf{EM}$ if $\lvert s\rvert < r,\ \textbf{ED}$ otherwise.
GUI	GUI	Grounding: **IoU**; String: $R_{\text{string}}(s) = \textbf{EM}$ if $\lvert s\rvert < r,\ \textbf{ED}$ otherwise; Action: **EM**.
Video	Video	Grounding: **IoU**.

Overall. We follow the classical implementation and include both the format reward and the accuracy reward. The format reward assigns a binary (0/1) reward score to each completion by judging whether the completion strictly follows the *<think></think><answer></answer>* structure. The accuracy reward is calculated based on the task type and also the completion, following our accuracy reward design. The overall reward for a completion is then computed as:

$$\text{Reward} = \text{Reward}_{\text{Format}} + \lambda \times \text{Reward}_{\text{Acc}} \tag{1}$$

Accuracy Reward Function. Though several tasks can be verified directly by exact matching (EM) and math verification tools, open-ended questions are not fully supported. Instead of employing an online reward model, we design two metrics for the main perception and understanding tasks. For grounding tasks, the reward is defined based on the Intersection over Union (IoU) between predictions and ground truths. For understanding tasks, where the outputs consist of a few words or sentences, we compute similarity using the edit distance (ED) between predictions and ground truths. The detailed formulations of all three reward types are provided in Appendix A.

Task Coverage. In our system, we mainly support 6 types of tasks, including STEM, Chart & Doc, Grounding, Spatial Reasoning, GUI, and Video. For each task, we select appropriate reward strategies based on both the task type and the specific nature of the expected answers, enabling tailored rewards that capture the unique challenges and requirements of each domain. The detailed reward design for all 6 task types is summarized in Table 3.

Independent Evaluation of Reward Before global mixed training, we evaluate each domain independently to ensure the reward signals are stable and learnable. As illustrated in Figure 8, the model successfully optimizes these rewards in isolation, exhibiting stable upward trends across all domains. This empirical validation confirms the reliability of our reward design and supports the feasibility of subsequent joint training.

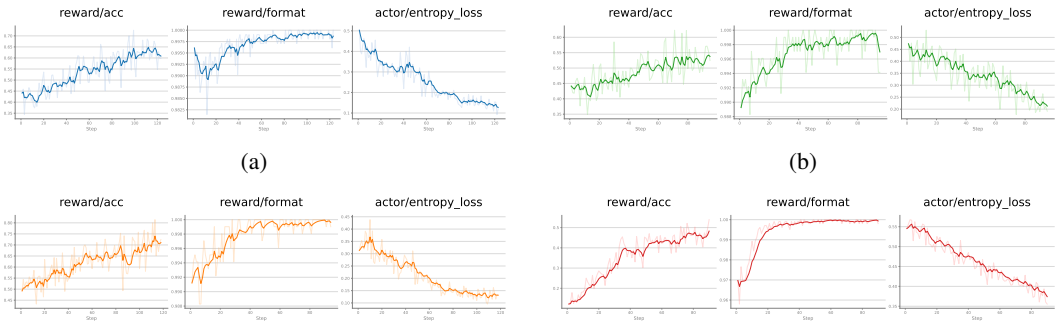


Figure 8: **Detailed training curves across four different domains.** Each sub-figure illustrates the progression of the accuracy score, format score, and entropy loss. The consistent upward trends in accuracy and format scores, paired with decreasing entropy loss, demonstrate that the model successfully optimizes the designed rewards in isolation for all domains. This validates the reward metrics as stable supervision signals for joint training.

4.3 Training Recipe

To comprehensively enhance the model's multimodal reasoning ability, we adopt a multi-stage training framework, including preference optimization, cold start, text-only warm-up, and finally, large-scale multimodal reinforcement learning. As shown in Figure 9, the complete experimental framework is organized into four principal stages, and their interdependencies and core evaluation metrics are summarized in Table 4. The full training curves are provided in Appendix B.

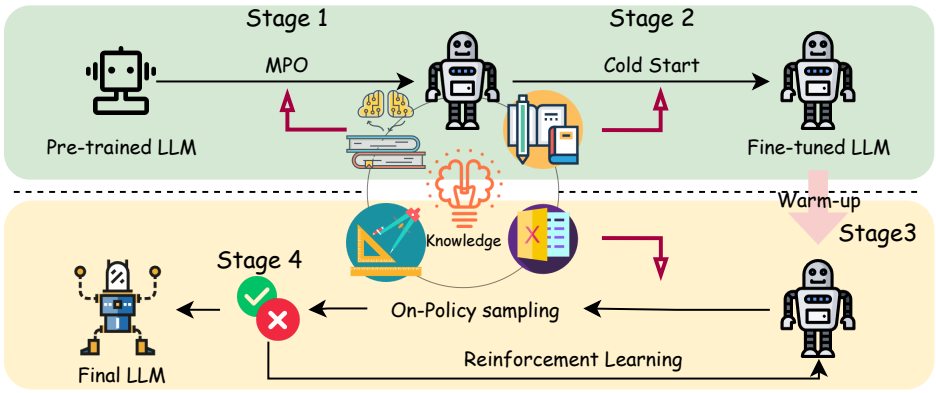


Figure 9: **Overview of the post-training pipeline.** The model undergoes four sequential stages: it is first optimized through preference optimization, then supervised fine-tuned during the cold-start phase, followed by text-only warm-up, and finally optimized through multimodal reinforcement learning with on-policy sampling.

Preference Optimization In the initial phase of post-training, we introduce a dedicated Preference Optimization phase to explicitly align the model's outputs with human preference signals. Specifically, we adopt the MPO [65] (Mixed Preference Optimization) algorithm and utilize approximately 100K preference pairs sampled from the MMPR dataset. This large-scale human preference data enables the model to better capture nuanced human judgments, thereby enhancing its general reasoning and alignment capabilities.

Cold Start. During the cold start phase, the preference-aligned model undergoes supervised fine-tuning on a curated dataset of 40K high-quality reasoning paths, each iteratively annotated. This process aims to familiarize the model with the cue–reflection reasoning pattern while ensuring that

its responses adhere to a strictly structured format—a prerequisite for the effective design of reward signals in subsequent reinforcement learning.

Text-Only Warm-up. In the text-only warm-up phase, we apply the GRPO [48] algorithm to further optimize the cold-started model on a 38K text-only reasoning dataset via reinforcement learning. This stage is designed to strengthen the model's internal "thinking" behavior by leveraging purely textual supervision. Although the cold-start phase enables basic task-oriented reasoning, the model's long CoT generation remains unstable and susceptible to degradation. Initiating multimodal RL training prematurely may exacerbate this issue, as incomplete vision–text alignment could disrupt the emerging reasoning patterns.

Multimodal Reinforcement Learning. In this phase, we introduce multimodal reinforcement learning, further training the model from the text-only warm-up stage using the GRPO [48] algorithm to enhance its generalization and reasoning abilities on multimodal tasks. This stage effectively transitions the model from text-only reasoning to multimodal decision-making, enabling it to leverage visual and interactive cues for more comprehensive inference.

Table 4: Detailed training configurations for post-training stages.

	Stage 1: Preference Optimization	**Stage 2: Cold Start**	**Stage 3: Text-Only Warm-up**	**Stage 4: Multimodal RL**
Base Model	Pre-trained Model	Output from Stage 1	Output from Stage 2	Output from Stage 3
Steps	6250	942	245	400
Learning Rate	2.0×10^{-6}	1.0×10^{-5}	1.0×10^{-6}	1.0×10^{-6}
Global Batch Size	16	32	64	64
Optimizer	xx AdamW	AdamW	AdamW	AdamW
Response Length	4096	4096	4096	4096
Compute Resources	8× H100 GPUs	8× H100 GPUs	8× H100 GPUs	8× H100 GPUs

5 Experiments

5.1 Overall Performance

To thoroughly assess the performance of Valley3, we adopted eight widely used benchmarks from OpenCompass [14], covering diverse domains and task types for a rigorous performance assessment. These benchmarks include MMBench [32], MMStar [9], MMMU [77], MathVista [36], HallusionBench [19], AI2D [25], OCRBench [33], and MMVet [76]. As shown in Table 5, Valley3 demonstrates competitive performance against other MLLMs on these benchmarks.

Table 5: Performance comparison of Valley3 with other powerful MLLMs on the OpenCompass leaderboard. The best score among open-source models is marked in **bold**, and the second-best result is indicated by underlining.

Models	**AVG**	**MMBench**	**MMStar**	**MMMU**	**MathVista**	**HallusionBench**	**AI2D**	**OCRBench**	**MMVet**
Closed-source MLLMs									
SenseNova-V6-5-Pro	82.20	87.3	76.1	77.0	82.8	66.7	90.2	885	89.4
Gemini-2.5-Pro	80.10	88.3	73.6	74.7	80.9	64.1	89.5	862	83.3
GPT-5-20250807	79.90	86.6	75.7	81.8	81.9	65.2	89.5	807	77.6
Open-source MLLMs									
R-4B	**75.50**	82.8	72.6	64.7	78.0	60.0	86.2	837	76.2
SAIL-VL2-8B	74.00	84.5	70.7	57.6	76.2	53.7	87.8	912	70.7
InternVL3-8B	73.60	82.1	68.7	62.2	70.5	49.0	85.1	884	82.8
SAIL-VL-1.6-8B	73.60	84.0	69.5	55.4	74.2	54.4	87.5	905	73.3
Ola-7B	72.60	84.3	70.8	57.0	68.4	53.5	86.1	822	78.6
WeThink-7B	72.50	81.8	64.2	61.0	71.6	55.3	84.0	890	73.2
Ovis2-8B	71.80	83.6	64.6	57.4	71.8	56.3	86.6	891	65.1
Qwen2.5-VL-7B	70.90	82.2	64.1	58.0	68.1	51.9	84.3	888	69.7
Valley3	<u>74.30</u>	84.7	67.3	62.1	74.4	56.3	84.4	870	77.9

While the OpenCompass leaderboard provides a high-level overview of model capabilities, a granular analysis is essential to understand specific strengths and weaknesses across different modalities. To this end, we conducted a rigorous assessment on a diverse set of benchmarks categorized into four core domains: General Multimodal Understanding, Math & Reasoning, OCR & Chart Interpretation, and Hallucination. Table 6 presents a detailed comparison of **Valley3** against its predecessor, Valley2, and other powerful open-source MLLMs.

Table 6: Detailed performance comparison of Valley3 against Valley2 and other competitive open-source MLLMs across four core domains. **Bold** denotes the best performance, and underlined denotes the second-best.

Benchmark	Valley3	Valley2	SAIL-VL2-8B	Qwen2.5-VL-7B	Ovis2-8B	InternVL3.5-8B	LLava-OV-1.5
Multimodal Understanding							
$\text{MMBench}_{\text{EN}}$ [32]	**85.45**	83.29	84.50	83.28	84.52	82.51	84.10
$\text{MMBench}_{\text{CN}}$ [32]	83.98	81.66	**86.84**	81.97	83.90	79.88	81.00
MME [18]	85.09	79.76	84.76	82.02	83.14	**85.15**	-
MMStar [9]	67.27	62.53	**70.70**	64.10	64.60	68.40	67.70
RealWorldQA [15]	70.46	67.45	**76.73**	67.97	73.20	68.24	68.10
AI2D [25]	84.36	81.47	**87.80**	84.30	86.60	84.26	84.20
MMVet [76]	77.90	62.94	70.70	69.70	65.10	**83.39**	-
MMMU_{val} [77]	**62.11**	57.66	57.60	58.00	57.40	58.11	55.40
Multimodal Reasoning							
MathVision [60]	33.65	24.93	27.63	25.36	25.43	**34.70**	25.60
$\text{MathVerse}_{\text{mini}}$ [82]	**54.85**	44.16	46.19	44.72	47.41	39.49	-
$\text{MathVista}_{\text{mini}}$ [36]	74.40	69.20	**76.20**	68.10	71.80	73.20	69.60
LogicVista [71]	**53.24**	41.39	44.97	47.65	40.27	52.35	-
$\text{OlympiadBench}_{\text{mini}}$ [21]	12.64	8.97	14.10	8.36	8.44	**16.07**	-
WeMath [46]	**38.19**	30.19	36.00	35.14	27.52	32.38	33.60
DynaMath [87]	**32.73**	16.26	17.16	21.75	21.15	18.56	-
OCR & Chart							
DocVQA [42]	94.68	91.33	**95.51**	94.93	94.17	91.41	95.00
TextVQA [51]	83.75	78.90	85.07	**85.34**	83.23	77.64	-
$\text{ChartQA}_{\text{test}}$ [40]	**88.00**	82.20	86.96	86.12	84.64	86.56	86.50
$\text{InfoVQA}_{\text{val}}$ [43]	81.53	69.43	81.02	**82.52**	80.30	78.91	78.40
OCRBench [33]	87.00	86.10	**91.20**	88.80	89.10	84.00	82.90
Hallucination							
HallusionBench [19]	**56.34**	52.84	53.70	51.90	56.30	54.39	-
$\text{CRPE}_{\text{relation}}$ [64]	76.09	73.95	72.02	75.91	**76.99**	75.10	-
POPE [29]	**88.95**	88.24	88.70	86.39	88.56	88.31	-

Multimodal Understanding Strong multimodal understanding forms the basis of MLLM performance. In this core area, Valley3 shows excellent ability, achieving top scores on $\text{MMBench}_{\text{EN}}$ and MMMU (Table 6). Compared with Valley2, Valley3 delivers clear gains across general benchmarks, reflecting a marked improvement in overall perception. Its leading results on multi-domain tasks further demonstrate robust performance across a wide range of practical scenarios.

Multimodal Reasoning Advanced reasoning is essential for turning perception into effective problem solving. Beyond basic understanding, Valley3 excels in complex multi-step reasoning, outperforming all baselines on MathVerse, LogicVista, WeMath, and DynaMath. Notably, its score on the challenging DynaMath benchmark nearly doubles that of Valley2 (16.26% $\rightarrow$ 32.73%), showing a clear transition from a perception-oriented model to one capable of strong logical and mathematical reasoning.

OCR & Chart Accurately interpreting text and charts is crucial for detailed document analysis and data-centered tasks. Valley3 achieves top performance on ChartQA (88.00%) while maintaining strong results on other benchmarks. It also shows steady improvement over Valley2, especially on InfoVQA (+12.1%), highlighting its enhanced ability to process high-resolution documents and extract accurate information from dense visual data.

Hallucination Reducing hallucinations remains a key challenge for MLLMs. In our reliability assessments, Valley3 achieves the highest scores on POPE (88.95%) and HallusionBench (56.34%). These results show that Valley3 adheres more closely to the visual input than both Valley2 and current competitors, effectively reducing incorrect or fabricated content.

5.2 Ablation Study

5.2.1 Ablation on MPO

MPO is an effective optimization framework for aligning model behaviors with preference data. In this section, we systematically investigate the impact of three critical factors within MPO: loss function composition, training data scale, and data contamination. Our results demonstrate that a simplified loss objective, coupled with a curated dataset of moderate size, yields significant performance gains. Furthermore, a rigorous decontamination analysis confirms that these improvements stem from genuine capability enhancement rather than test set leakage.

Simplicity Prevails: Removing SFT Loss Enhances Performance. The standard MPO algorithm integrates three loss terms: SFT, BCO, and DPO. We initially adopted the weighting strategy from the original paper ($W_{\text{DPO}} = 0.8, W_{\text{BCO}} = 0.2, W_{\text{SFT}} = 1.0$) using 100K training samples. However, this configuration underperforms compared to a DPO-only baseline. Hypothesizing that the SFT term might be redundant or detrimental in this context, we experimented with removing it while retaining the DPO and BCO losses. As shown in Table 7, this simplified configuration outperforms the DPO-only approach. Consequently, we adopted this simplified MPO variant as the default method for subsequent experiments.

Table 7: Ablation of loss composition on the OpenCompass benchmark. All experiments are conducted based on an intermediate checkpoint (Valley3-Pretrain-V1) from the pretraining stage.

Model	Algorithm	OpenCompass Score
Valley3-Pretrain-V1	-	66.87
	DPO	70.75
	DPO+BCO+SFT (MPO)	70.25
	DPO+BCO	**71.06**

Scaling Limits: Saturation at High Data Volumes. To determine the optimal data volume for MPO, we evaluated model performance across scales ranging from 10k to 200k samples. Table 8 illustrates a consistent performance improvement as data increases from 10k to 100k, aligning with established scaling laws. However, expanding the dataset to 200k resulted in performance degradation. Attempts to mitigate this by adjusting the learning rate (ranging from 5×10^{-7} to 5×10^{-6}) failed to surpass the 100k baseline. These findings suggest that for the MMPR source, Valley3 reaches performance saturation at approximately 100k samples, beyond which additional data may introduce noise or distribution shifts that hamper training.

Table 8: Ablation of data scale and learning rate during the preference optimization stage. All experiments are conducted based on an intermediate checkpoint (Valley3-Pretrain-V2) from the pretraining stage.

Model	Algorithm	Data Scale	Learning Rate	OpenCompass Score
Valley3-Pretrain-V2	-	-	-	67.48
	DPO + BCO	10k	1×10^{-6}	68.71
	DPO + BCO	50k	1×10^{-6}	70.37
	DPO + BCO	100k	1×10^{-6}	**71.48**
	DPO + BCO	200k	1×10^{-6}	70.12
	DPO + BCO	200k	5×10^{-7}	70.48
	DPO + BCO	200k	2×10^{-6}	70.92

Validating Genuine Gains: Robustness to Data Leakage. To verify that the performance gains during the MPO phase were not artifacts of test set leakage, we conducted a rigorous contamination check. We computed the cosine similarity of image embeddings (extracted via `CLIP-large-P14`) between the MPO training set and the OpenCompass benchmarks. As detailed in Table 9, removing samples with high similarity (0.80–1.00) resulted in a negligible performance drop (< 0.5 PP), with the model still significantly outperforming the pretraining baseline. This confirms that the gains are driven by learned patterns rather than memorization. Conversely, when we artificially introduced contaminated samples (increasing from 3% to 10%), performance declined. This indicates that

excessive duplication distorts the training distribution, causing the model to overfit specific examples and reducing its generalization capability.

Table 9: Ablation of data composition strategies during the preference optimization stage. All experiments are conducted based on an intermediate checkpoint (Valley3-Pretrain-V2) from the pretraining stage.

Model	Algorithm	Data Details	OpenCompass
Valley3-Pretrain-V2	-	-	67.48
	DPO + BCO	100k data	**71.48**
	DPO + BCO	97k samples (after removing 0.80–1.00 similarity)	71.18
	DPO + BCO	97k samples (after removing 0.80–1.00 similarity) + 3k uncontaminated data	70.95
	DPO + BCO	95k uncontaminated samples + 5k contaminated data	71.23
	DPO + BCO	90k uncontaminated samples + 10k contaminated data	69.97

5.2.2 Ablation on MCQ & FIB

To investigate how reward signal structure influences policy optimization, we compare models trained on mixed Multiple-Choice Questions (MCQ) versus those where MCQs are reformulated into Fill-in-the-Blank (FIB) tasks. MCQs provide binary, coarse-grained rewards, whereas FIB requires open-ended generation, offering fine-grained supervision. As illustrated in Figure 10, the FIB-trained model exhibits higher entropy (Fig. 10-a) and lower training rewards (Fig. 10-b) but achieves superior generalization across 11 Mini-Benchmarks (Fig. 10-c). Note that these Mini-Benchmarks (approx. 200 samples each) were rigorously validated to ensure their rankings align consistently with full authoritative datasets. The discrepancy between high training rewards and lower test performance in the MCQ setting suggests the occurrence of reward hacking, whereas the FIB format promotes more robust policy exploration and generalization.

Figure 10: Performance comparison between MCQ & FIB mixed training versus pure FIB training.

5.2.3 Ablation on RL Data Ratio

We systematically evaluate the impact of data difficulty distribution during the RL stage. Samples are stratified into Easy, Medium, and Hard tiers (Figure 11). We propose four sampling strategies:

Strategy 1: Exclusive Use of Medium-Difficulty Data. Focusing on the most informative signals, we utilize 76K medium-difficulty samples (including 10K Video and 3K GUI data) to balance learning efficiency by avoiding both trivial and excessively noisy instances.

Strategy 2: Mixed Difficulty with Fixed Proportion. To test the benefit of diversity, we construct a 73K dataset with an Easy:Medium:Hard ratio of 1:3:1 (12K Easy, 36K Medium, 12K Hard), supplemented by 10K Video and 3K GUI samples.

Strategy 3: Fixed Sample Size with Medium-Difficulty Domain Balance. We subsample 25K instances from the Strategy 1 dataset (76K Medium) to ensure a uniform distribution across domains, allowing us to isolate the effect of domain balancing.

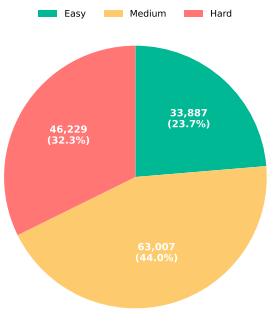


Figure 11: Difficulty distribution of the RL data.

Strategy 4: Fixed Sample Size with Mixed Difficulty and Domain Balance. Similarly, we subsample 25K instances from the Strategy 2 dataset (73K Mixed) to control sample size while maintaining both difficulty diversity and domain balance.performance.

For the four aforementioned data sampling strategies, we conduct reinforcement learning on the SFT-initialized model and evaluate the resulting policies on the Mini-Benchmarks. The evaluation results, summarized in Table 10, indicate that the strategy leveraging mixed-difficulty multimodal data yields the strongest overall performance among all sampling configurations.

5.2.4 Ablation on Text-Only Warm-Up

After completing the text-only warm-up, our next objective is to determine the most effective data configuration for the subsequent Multimodal RL stage. Since the warm-up has already equipped the model with a preliminary reasoning prior, we aim to understand how this prior influences the model's ability to leverage different multimodal data distributions.

To this end, we conduct a targeted ablation comparing two candidate data settings for Multimodal RL: (1) a focused dataset containing only medium-difficulty samples (76K), and (2) a mixed-difficulty dataset that blends easy, medium, and hard examples (73K). This comparison allows us to assess whether the warm-up stabilizes the model sufficiently to benefit from diverse multimodal signals, or whether a more concentrated difficulty scope remains advantageous at this stage.

The results, shown in the final columns of Table 10, indicate that **Warm-up + 73K** consistently achieves superior performance. This suggests that once the text-only warm-up establishes a solid reasoning foundation, exposing the model to a broader span of difficulty levels during Multimodal RL becomes more effective. The increased diversity supports better reasoning transfer and yields greater robustness across complex multimodal tasks.

Table 10: **Performance comparison of data configurations for the Multimodal RL stage.** We evaluate four RL data sampling strategies (Strategies 1–4) and two data configurations used *after* the text-only warm-up (76K medium-difficulty vs. 73K mixed-difficulty). Results across multiple mini-benchmarks show that the mixed-difficulty 73K setting yields the best overall performance.

Benchmarks/Data ratio	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Warm-up + 76K	Warm-up + 73K
Mini_MMBench	0.88	0.86	0.875	0.855	0.875	0.87
Mini_MMMU	0.64	0.70	0.62	0.57	0.59	0.634
Mini_MMStar	0.67	0.67	0.62	0.62	0.66	0.65
Mini_MMVet	0.83	0.83	0.84	0.82	0.85	0.857
Mini_MathVista	0.73	0.77	0.74	0.72	0.68	0.79
Mini_OCRBench	0.85	0.84	0.86	0.85	0.85	0.82
Mini_HallusionBenchmark	0.63	0.64	0.65	0.65	0.65	0.64
Mini_InfoVQA	0.29	0.30	0.33	0.31	0.33	0.31
Mini_CountBench	0.88	0.85	0.88	0.89	0.85	0.90
Avg	0.7111	0.7178	0.7128	0.6983	0.7039	**0.7190**

6 Conclusion

In this work, we present Valley3, an 8-billion-parameter Multimodal Large Language Model that balances high performance with full reproducibility. By implementing a quantitative architecture evaluation and a model-based data curation pipeline, we demonstrate that rigorous methodological choices are essential for ensuring predictable scaling and training efficiency. Crucially, our post-training strategy—integrating our optimized Mixed Preference Optimization (MPO) with the advanced GThinker framework—validates the effectiveness of preference optimization and verifiable reinforcement learning in stabilizing complex reasoning. Extensive evaluations on OpenCompass and domain-specific benchmarks confirm that Valley3 achieves competitive performance among similarly sized open-source models. By publicly releasing our complete codebase, data recipes, and model checkpoints, we aim to bridge the transparency gap and provide a solid blueprint for future research in the multimodal community.

References

[1] Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, et al. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *arXiv preprint arXiv:2507.08306*, 2025.

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[3] Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025.

[4] 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 preprint arXiv:2502.13923*, 2025.

[5] 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 preprint arXiv:2502.13923*, 2025.

[6] Ali Furkan Biten, Ruben Tito, Marcal Rusiñol, et al. Scene text visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context, 2018. URL `https://arxiv.org/abs/1612.03716`.

[8] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024.

[9] 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? *arXiv preprint arXiv:2403.20330*, 2024.

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[11] 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 preprint arXiv:2412.05271*, 2024.

[12] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[13] 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 preprint arXiv:2507.06261*, 2025.

[14] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

[15] 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.

[16] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale, 2020. URL `https://arxiv.org/abs/1909.04422`.

[17] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9641–9654, 2025.

[18] 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[19] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, June 2024.

[20] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. URL `https://arxiv.org/abs/1908.03195`.

[21] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, 2024.

[22] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507, 2025.

[23] Kamil Iskakov et al. Can ai assistants know what they don't know? In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[24] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[25] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[26] Douwe Kiela et al. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

[27] Neal Krawetz. Looks like it. `https://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html`, 2011. URL `https://www.hackerfactor.com/blog/index.php?/archives/432-Looks-Like-It.html`.

[28] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556, 2024.

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[30] Haotian Liu et al. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[31] Jiali Liu et al. G-llava: Solving geometric problems with multi-modal large language models. *arXiv preprint arXiv:2406.06578*, 2024.

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2025.

[33] 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*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL `http://dx.doi.org/10.1007/s11432-024-4235-6`.

[34] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024.

[35] Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024.

[36] 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 *International Conference on Learning Representations (ICLR)*, 2024.

[37] Pan Lu et al. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

[38] Pan Lu et al. CLEVR-Math: A dataset for solving multi-modal math word problems. *arXiv preprint arXiv:2206.08644*, 2022.

[39] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

[40] 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 *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[41] Ahmed Masry et al. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[42] Minesh Mathew, Dimosthenis Karatzas, et al. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[43] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[44] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536, 2020.

[45] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation, 2025. URL `https://arxiv.org/abs/2407.02371`.

[46] Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070, 2025.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[48] 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 preprint arXiv:2402.03300*, 2024.

[49] Mohit Shridhar, Shubham Gupta, Jason Lee, et al. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

[50] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020.

[51] Amanpreet Singh, Viraj Prabhu Singh, Jarad Mitrovic, et al. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[52] Alane Suhr et al. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[53] Benny J. Tang, Angie Boggust, Arvind Satyanarayan, et al. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023.

[54] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. URL `https://arxiv.org/abs/2506.03569`.

[55] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[56] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[57] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL `https://arxiv.org/abs/2507.01006`.

[58] Andreas Veit, Tomas Matera, Lukas Neumann, et al. Coco-text: A large scale dataset for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[59] 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. *arXiv preprint arXiv:2504.08837*, 2025.

[60] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[61] 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 preprint arXiv:2409.12191*, 2024.

[62] 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 preprint arXiv:2409.12191*, 2024.

[63] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[64] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024.

[65] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2025. URL `https://arxiv.org/abs/2411.10442`.

[66] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[67] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

[68] Zhi Wang et al. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

[69] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[70] Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu. Valley2: Exploring multimodal models with scalable vision-language design, 2025. URL `https://arxiv.org/abs/2501.05901`.

[71] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[72] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[73] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[74] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025.

[75] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

[76] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

[77] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

[78] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[79] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[80] Yufei Zhan, Ziheng Wu, Yousong Zhu, Rongkun Xue, Ruipu Luo, Zhenghao Chen, Can Zhang, Yifan Li, Zhentao He, Zheming Yang, et al. Gthinker: Towards general multimodal reasoning via cue-guided rethinking. *arXiv preprint arXiv:2506.01078*, 2025.

[81] Jing Zhang et al. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[82] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[83] Shouta Zhang et al. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*, 2025.

[84] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025. URL `https://arxiv.org/abs/2410.02713`.

[85] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022.

[86] 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 preprint arXiv:2504.10479*, 2025.

[87] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

A Reward Formulations

We provide formal definitions of the reward mechanisms used across domains: Exact Match (EM), Edit Distance (ED), and Intersection-over-Union (IoU). These rewards are referenced in Table 3 for different task categories.

Exact Match (EM). EM is applied to tasks where the answer is categorical or symbolic and must exactly match the ground truth:

$$r_{\text{EM}}(\mathbf{o}_{\text{resp}}, \mathbf{o}_{\text{gt}}) = \begin{cases} 1, & \text{if } \mathbf{o}_{\text{resp}} = \mathbf{o}_{\text{gt}}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

IoU-based Reward. For grounding tasks involving spatial regions (e.g., bounding boxes), the reward depends on the Intersection-over-Union (IoU) between prediction and ground truth:

$$r_{\text{IoU}}(\mathbf{o}_{\text{resp}}, \mathbf{o}_{\text{gt}}) = \begin{cases} \frac{|I_{\text{pred}} \cap I_{\text{gt}}|}{|I_{\text{pred}} \cup I_{\text{gt}}|}, & \text{if } \frac{|I_{\text{pred}} \cap I_{\text{gt}}|}{|I_{\text{pred}} \cup I_{\text{gt}}|} > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where τ=0.6 denotes the acceptance threshold.

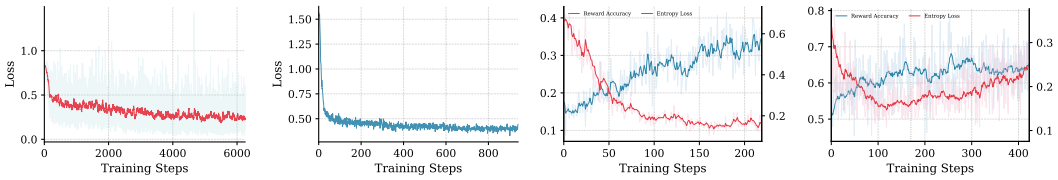
Edit Distance (ED). For open-form text outputs, we compute a normalized similarity score based on the Levenshtein edit distance:

$$r_{\text{ED}}(\mathbf{o}_{\text{resp}}, \mathbf{o}_{\text{gt}}) = \begin{cases} 1 - \dfrac{d_{\text{edit}}(\mathbf{o}_{\text{resp}}, \mathbf{o}_{\text{gt}})}{\max(|\mathbf{o}_{\text{resp}}|, |\mathbf{o}_{\text{gt}}|)}, & \text{if } 1 - \dfrac{d_{\text{edit}}(\mathbf{o}_{\text{resp}}, \mathbf{o}_{\text{gt}})}{\max(|\mathbf{o}_{\text{resp}}|, |\mathbf{o}_{\text{gt}}|)} > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $d_{\text{edit}}(\cdot, \cdot)$ is the Levenshtein distance [78] and τ=0.6.

We use EM, IoU, and ED across domains depending on the output modality: structured symbolic responses (EM), spatial grounding (IoU), and free-form textual reasoning (ED).

B Training Dynamics



(a) Stage 1 (b) Stage 2 (c) Stage 3 (d) Stage 4

Figure 12: Details of the four post-training stages.

We present the learning curves for the four-stage training pipeline outlined in Section **??**. As illustrated in Figure 12, the first stage (Stage 1) employs a pre-trained model with Model-based Policy Optimization (MPO) to initiate supervised fine-tuning. And the supervised fine-tuning stage (Stage 2) rapidly reduces the training loss, indicating efficient adaptation to the curated reasoning data. In Stage 3, the text-only warm-up introduces reinforcement learning on textual supervision, where the reward accuracy steadily improves while entropy gradually decreases, demonstrating that the model becomes more confident and policy-aligned. Finally, during Stage 4, multimodal reinforcement fine-tuning further enhances reward accuracy on visual-language tasks while stabilizing entropy, suggesting strengthened grounding ability without excessive policy collapse.

Overall, these results verify the effectiveness of our progressive curriculum: from controlled reasoning learning to text-based policy shaping, and finally to multimodal alignment.

C Knowledge Learning Data Sources

To support the Knowledge Learning stage(Pretrain-Stage1.5), we aggregated a massive and diverse corpus aimed at establishing a comprehensive multi-modal world model. As detailed in Table 11, this dataset spans single-image, multi-image, video, and pure-text modalities, prioritizing domain breadth to cover fundamental tasks such as dense captioning, OCR, and general VQA. This extensive exposure ensures the model acquires robust cross-modal alignment capabilities and rich world knowledge before entering the instruction-tuning stage

Table 11: Knowledge Learning Data

Modality	Task	Datasets
SI	**Caption**	ALLaVA-Caption-LAION-4V, ALLaVA-Caption-VFLAN-4V, LAION-GPT4V, CC3M, InternVL-SA-1B-Caption, LLaVA-ReCap, LLaVA-Wild, ShareGPT-4o, ShareGPT4V, LVIS-Instruct, TextCaps, UReader, DenseFusion, OVIS-Data, PixMo, WebSight
	OCR	SynthDog, Infinity-MM, TextOCR, SROIE, K12 Printing, IAM, ORAND, IIIT-5K, HME100K, UReader, ST-VQA, LaTeX OCR, Rects
	Math / Geometry	GeomVerse, Geo3K, MAVIS, VisualWebInstruct, GeoQA+, Geo170K, GEOS, MathV360K, MathVision, VisualGenome
	Knowledge Reasoning	MathV360K, TQA, AI2D, VQA-RAD, PMC-VQA, ScienceQA, PathVQA, WIT, Viquae
	Code / Document QA	design2code, Docmatix, FigureQA, Robut-WTQ, DVQA, Screen2Words, InfographicVQA, MapQA, UReader, TabMWP, IconQA, Robut-WikiSQL, VisText, VisualMRC, HiTab, Chart2Text, Robut-SQA, ChartQA, Geo170K, LRV-Chart, TQA, MultiHierTT, MMC, DocVQA, DocStruct4M, DocLIE, PlotQA, ArxivQA, TAT-QA, TinyChart, IDEFICS, PIXMO
	Grounding / Counting	GRIT, Flickr30k, TallyQA, CLEVR-Math, COCO-Text, Sherlock, PIXMO, Describe-Anything-Dataset, BLIP-3 Grounding
	General VQA	LVIS-Instruct4V, Vision-Flan, LLaVAR, Cambrian, ALLAVA, Raven, Visual7W, Super-CLEVR, ST-VQA, VizWiz, A-OKVQA, Hateful Memes, LRV, CLEVR, DocReason, ShareGPT4V, VisDial, AI2D, GQA, TextVQA, SVIT, FinQA, PIXMO
MI	**Caption**	FlintstonesSV, PororoSV, Twitter, VIST
	OCR / VQA	R-VQA, Interleave, Coinstruct, Contrastive Caption, MMC4Doc, IconQA, TQA, DocVQA
	Spatial Reasoning	3D-LLM, Alfred, DreamSim, HQ-Edit, HQ-Edit-Diff, IEdit, MagicBrush, MagicBrush-Diff, ScanQA, Spot-the-Diff
Video	**Dense Caption**	OpenVid-1M, Video-Caption-Pretrain-900K, Vript, WebVid
	GUI / Spatial Reasoning	GUI-World, SpaceR, GPT4Scene
	General VQA	Academic-QA, ActivityNet-QA, LLaVA-Next-Video-SFT, NextQA, Perception Test-QA, ShareGPTVideo, YouTube-QA
Pure-Text	**Math**	AIME, AoPS Forum, CN-K12 Math, GSM8K, Infinity-Instruct, Math Olympiads, MathInstruct, MathQA, Orca
	Code	Infinity-Instruct
	Generation	Infinity-Instruct
	General QA / Dialogue	Evol-Instruct-GPT4-Turbo, Infinity-Instruct, Infinity-Preference, Magpie-Pro, Orca, WizardLM, K-12 Vista

D Instruction Fine-Tuning Dataset

The Instruction Fine-Tuning stage(Pretrain-Stage2) focuses on refining the model's ability to follow complex constraints and perform multi-step reasoning. As listed in Table 12, this dataset is rigorously curated to prioritize quality and task diversity. We significantly enriched the mixture with high-difficulty samples—including Chain-of-Thought (CoT) reasoning, GUI navigation, and spatial understanding tasks—to effectively transition the model from general knowledge recall to practical, agentic problem-solving.

Table 12: Instruction Fine-Tuning Data

Modality	Task	Datasets
SI	Caption	DenseFusion, LVIS-Instruct, coco_colors, Face_Emotion, Google_landmarks, LAION_GPT4V, Localized Narratives, ShareGPT, ShareGPT-4V, TextCaps
	OCR	ArT, Captcha, ChromeWriting, COCO-Text, CTW, FUNSD, HME100K, HW-SQuAD, IAM Handwriting, IIIT5K, Imgur5K, K12 Printing, Im2LaTeX, Latex Handwritten, MapText, Math-Writing Google, Memotion OCR, ORAND-CAR A, Rendered Text, SROIE, SVRD, Synth-CodeNet, SynthDog, SynthFormulaNet, TAL OCR English, WordArt, olmOCR-Mix 0225 Documents, olmOCR-Mix 0225 Books, OCR-VQA, UReader
	Math	CLEVR, CLEVR-Math, CoSyn 400k Math, GEOS, Geo170K, Geo3K, GeoQA, GeoQA Plus, GeomVerse, InterGPS, Mavis Math Metagen, Mavis Math Rule Geo, Raven, Super-CLEVR, UniGeo, VisualWebInstruct
	Knowledge	AI2D, PMC-VQA, PathVQA, ScienceQA, ScienceQA NoAnnotation, TQA, VQA-RAD
	GUI	Android Control, RICO-ScreenQA, SeeClick, ScreenQA, Screen2Words, WebSight
	Doc	Chart2Text, ChartQA, Diagram Image-to-Text, Docmatix, DVQA, FigureQA, Geo170k, HiTab, IconQA, InfographicVQA, LRV-Chart, MapQA, MultiHiertt, Robust SQA, Robust WikiSQL, Robust WTQ, Screen2Words, TabMWP, TQA, UReader-IE, UReader-QA, VisText, VisualMRC, VQAonBD
	Grounding / Counting	AGuVIS, CLEVR-Math, GroundUI, Objects365-QA, OODVQA, TallyQA
	Spatial	VSR, SpaceR, GPT4Scene
	Reasoning	CLEVR-CoAgent-R1, CLEVR-Cogent-R1, MCOT-R1-VQA-66k, MM-R1-Combined, Open-R1-8k, OpenVLThinker, QVQ-R1, R1-90k Instruct, R1-Think-Med, R1-Vision AI2D, R1-Vision PixMo-Cap-QA, R1-Vision PixMo-Cap-QA-ZH, R1-Vision Rendered-Stratos-17k, R1-Vision ScienceQA, RAVEN, RedStar-Geo-R1, Vision-R1 (LLaVA CoT), Vision-R1 (Mulberry SFT), VLAA-Thinking, VQASynth-R1-12k, WeThink Multimodal Reasoning
	Spatial / Reasoning / General VQA	A-OKVQA, ALFWorldGPT, ALLaVA LAION, ALLaVA VFLAN, Cambrian-7M, Chinese Meme, COCO-QA, DaTikZ, DriveLM, Hateful Memes, IconQA, IDK VQAv2-IDK, Indoor QA, LLaVA-1.5-660k, LLaVA-Instruct-150K, LNQA, LRV-Instruction, LRV-Chart, LVIS-Instruct, MIMIC-CGD, MMEvol, NLVR2, MathV360K VQA AS, MM-K12, OneVision Sample 348k, OpenSource VQA Mix, SPARK, SketchyVQA, ST-VQA, Super-CLEVR, Vision-FLAN, Visual7W, VizWiz, WebSight, WildVision, YesBut, MS COCO Captions, DenseFusion-1M, Face Emotion Captions, Google Landmarks, Image Textualization, LAION-GPT4V, Localized Narratives, ShareGPT-4o, ShareGPT-4V, TextCaps, Chart2Text, ChartQA, CoSyn 400k Chart, CoSyn 400k Table, DVQA, FigureQA, FinQA, HiTab, MultiHiertt, PlotQA, Robust SQA, Robust WikiSQL, Robust WTQ, SynthChartNet, TabMWP, TAT-QA, Unichart, VisText, VQAonBD
MI	Generate Stories	In-house Data, ShareGPT
	General VQA	In-house Data
Video	Dense Caption	VideoChat-Flash
	GUI	GUI-World
	Reasoning	Video-R1
	Spatial	SpaceR, Scene-30K
	General VQA	VideoITG-40K, VideoChat-Flash
Pure-Text	Reasoning	Mixture-of-Thoughts, K12-Vista, OpenR1
	Long Context	LongAlign, LongReward, LongAlpaca, LongQLoRA
	General QA	Infinity Preference, STEM ZH Instruction, LongAlign, LongReward, LongAlpaca, LongQLoRA, Mixture of Thoughts, K12 Vista, OpenR1, Code Feedback, CodeFeedback Instruction, InfinityMath, MathInstruct, MathQA, MathStep DPO 10k, NuminaMath CoT, OpenHermes 2.5, OpenOrca, OrcaMath, PythonCode25k, PythonCode Alpaca, Ruozhiba, TheoremQA, WizardLM Evol-Instruct, OpenMathInstruct 2