MiMo-VL Technical Report

LLM-Core Xiaomi

Abstract

We open-source MiMo-VL-7B-SFT and MiMo-VL-7B-RL, two powerful vision-language models delivering state-of-the-art performance in both general visual understanding and multimodal reasoning. MiMo-VL-7B-RL outperforms Qwen2.5-VL-7B on 35 out of 40 evaluated tasks, and scores 59.4 on OlympiadBench, surpassing models with up to 78B parameters. For GUI grounding applications, it sets a new standard with 56.1 on OSWorld-G, even outperforming specialized models such as UI-TARS. Our training combines four-stage pre-training (2.4 trillion tokens) with Mixed On-policy Reinforcement Learning (MORL) integrating diverse reward signals. We identify the importance of incorporating high-quality reasoning data with long Chain-of-Thought into pre-training stages, and the benefits of mixed RL despite challenges in simultaneous multi-domain optimization. We also contribute a comprehensive evaluation suite covering 50+ tasks to promote reproducibility and advance the field. The model checkpoints and full evaluation suite are available at https://github.com/XiaomiMiMo/MiMo-VL.

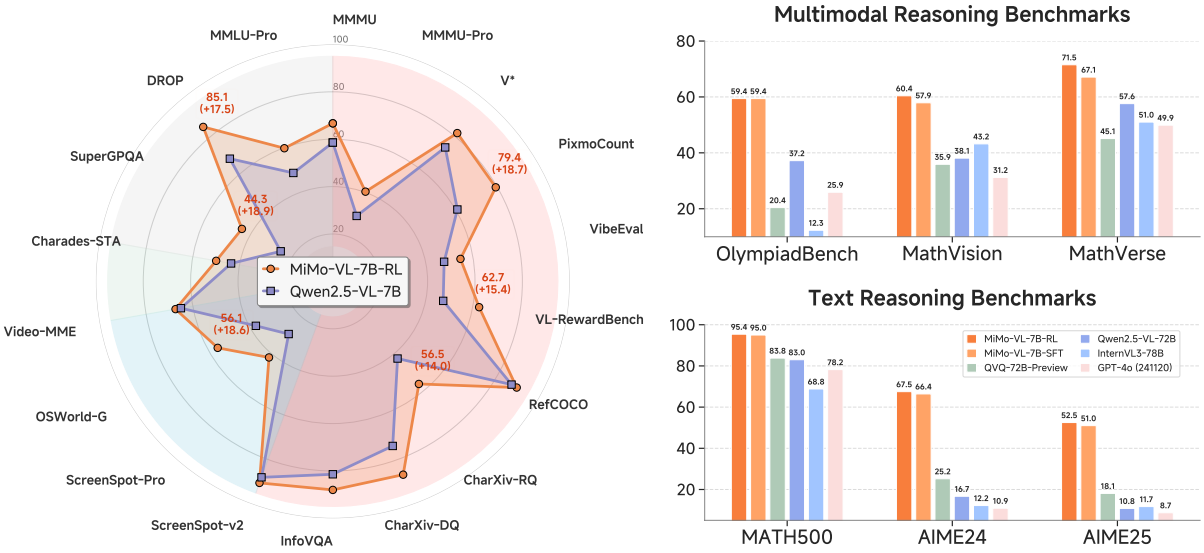

Figure 1 Benchmark performance of MiMo-VL-7B.

Contents

1 Introduction 4

2 Pre-Training 5

2.1 Architecture 5

2.2 Pre-training Data 6

2.2.1 Image Caption Data 6

2.2.2 Interleaved Data 7

2.2.3 OCR and Grounding Data 7

2.2.4 Video Data 7

2.2.5 Graphical User Interface Data 8

2.2.6 Synthetic Reasoning Data 8

2.3 Pre-training Stages 8

3 Post-Training 9

3.1 Reinforcement Learning with Verifiable Rewards 9

3.2 Reinforcement Learning from Human Feedback 10

3.3 Mixed On-Policy Reinforcement Learning 11

4 Evaluation 12

4.1 Evaluation Setting 12

4.2 General Capabilities 12

4.3 Reasoning Tasks 12

4.4 GUI Tasks 14

4.5 Elo Rating 16

5 Discussion 16

5.1 Boosting Reasoning Capability in Pre-training 16

5.2 On-Policy RL v.s. Vanilla GRPO 17

5.3 Interference Between RL Tasks 17

6 Case Study 17

7 Conclusions 20

A Contributions and Acknowledgments 26

B Model Configuration of MiMo-VL-7B 27

C Evaluation Benchmarks 27

D GUI Action Space 28

E More Qualitative Examples 28

1 Introduction

Vision-language models (VLMs) have emerged as the foundational backbone for multimodal AI systems, enabling autonomous agents to perceive the visual world, reason over multimodal content (Yue et al., 2024b), and interact with both digital (Xie et al., 2024; OpenAI, 2025) and physical environments (Zitkovich et al., 2023; Black et al., 2024). The signi cance of these capabilities has motivated extensive exploration across multiple dimensions, including novel architectural designs (Alayrac et al., 2022; Team, 2024; Ye et al., 2025) and innovative training methodologies with optimized data recipes (Karamcheti et al., 2024; Dai et al., 2024), leading to rapid progress in the eld (Liu et al., 2023; Tong et al., 2024; Bai et al., 2025a).

In this report, we share our e orts to build a compact yet powerful VLM, MiMo-VL-7B. MiMo-VL-7B comprises three components: (1) a native-resolution Vision Transformer (ViT) encoder that preserves ne-grained visual details, (2) a Multi-Layer Perceptron (MLP) projector for e cient cross-modal alignment, and (3) our MiMo-7B (Xiaomi, 2025) language model, speci cally optimized for complex reasoning tasks.

The development of MiMo-VL-7B involves two sequential training processes: (1) A four-stage pre-training phase, which includes projector warmup, vision-language alignment, general multi-modal pre-training, and long-context Supervised Fine-Tuning (SFT). Throughout these stages, we curate high-quality datasets by strategically combining open-source collections with synthetic data generation techniques, consuming 2.4 trillion tokens, adjusting the dataset distribution in di erent stages to facilitate training. This phase yields the MiMo-VL-7B-SFT model. (2) A subsequent post-training phase, where we introduce Mixed On-policy Reinforcement Learning (MORL), a novel framework that seamlessly integrates diverse reward signals spanning perception accuracy, visual grounding precision, logical reasoning capabilities, and human preferences. We adopt the idea of GRPO (Shao et al., 2024) and enhance training stability by exclusively performing on-policy gradient updates during this stage. This phase yields the MiMo-VL-7B-RL model.

During this journey, we nd:

(1) Incorporating high-quality, broad-coverage reasoning data from the pre-training stage is crucial for enhancing model performance. In the current era of thinking models, vast quantities of multimodal pre-training data are undergoing a signi cant re-evaluation. Traditional question-answering (QA) data, with its direct, short answers, often restricts models to super cial pattern matching and leads to over tting. In contrast, synthesized reasoning data with long Chain-of-Thought (CoT) enables learning of complex logical relationships and generalizable reasoning patterns, providing richer supervision signals that markedly improve both performance and training e ciency. To leverage this advantage, we curate high-quality reasoning data by identifying diverse queries, employing large reasoning models to regenerate responses with long CoT, and applying rejection sampling to ensure quality. Furthermore, rather than treating this as supplementary ne-tuning data, we incorporate substantial volumes of this synthetic reasoning data directly into the later pre-training stages, where extended training yields continued performance improvements without saturation.

(2) Mixed On-policy Reinforcement Learning further enhances model performance, while achieving stable simultaneous improvements remains challenging. We apply RL across diverse capabilities, including reasoning, perception, grounding, and human preference alignment, spanning modalities including text, images, and videos. While this hybrid training approach further unlocks the model's potential, interference across data domains remains challenging. Disparities in the growth trends of response length and task di culty levels hinder stable, simultaneous improvements across all capabilities.

MiMo-VL-7B-RL delivers exceptional results across the full spectrum of multimodal capabilities. (1) On fundamental visual perception tasks, it achieves state-of-the-art performance among open-source VLMs of comparable scale, scoring 66.7 on MMMU (Yue et al., 2024b) and outperforming Qwen2.5-VL-7B (Bai et al., 2025a) on 35 out of 40 evaluated tasks. (2) For complex multimodal reasoning, MiMo-VL-7B-RL exhibits outstanding performance, scoring 59.4 on OlympiadBench (He et al., 2024) and surpassing models up to 72B parameters.(3) In GUI grounding for agentic applications, our model sets a new standard by achieving a score of 54.7 on OSWorld-G (Xie et al., 2025), even exceeding specialized models like UI-TARS (Qin et al., 2025b). (4) In terms of user experience and preference, MiMo-VL-7B-RL achieves the highest Elo rating among all open-source VLMs on our in-house user preference evaluation, performing competitively with proprietary models such as Claude 3.7 Sonnet.

These results validate our approach: by combining strong perception, sophisticated reasoning, and precise grounding capabilities through our MORL framework, MiMo-VL-7B-SFT and MiMo-VL-7B-RL establish new standards for open-source vision-language models. To promote transparency and reproducibility, we also contribute a comprehensive evaluation suite covering 50+ tasks with complete prompts and protocols, enabling the community to build upon our work.

2 Pre-Training

In this section, we introduce the architecture design of our MiMo-VL-7B, followed by the data curation process and training recipes of pre-training stages.

2.1 Architecture

MiMo-VL-7B consists of three components: (1) a Vision Transformer (ViT) for encoding visual inputs such as images and videos; (2) a projector that maps the visual encodings into a latent space aligned with the Large Language Model (LLM); and (3) the LLM itself, which performs textual understanding and reasoning. To support native resolution inputs, we adopt Qwen2.5-ViT (Bai et al., 2025a) as our visual encoder. We initialize the LLM backbone with MiMo-7B-Base (Xiaomi, 2025) to leverage its strong reasoning capability, and utilize a randomly initialized Multi-Layer Perceptron (MLP) as the projector. The overall architecture is illustrated in Figure 2, and the model con guration can be found in Appendix B.

Stages	Stage 1	Stage 2	Stage 3	Stage 4
Purpose	Projector Warmup	Vision-Language Alignment	Multimodal Pre-training	Long-context SFT
Dataset	Image-Caption Pairs	+ Interleaved Data	+ Pure Text, OCR, Grounding, QA, Video, GUI, Instruction Data, Reasoning Data	+ Long Pure Text, Long Documents, High-resolution Images, Extended Videos, Long Reasoning Data
Learning Rate	1e-3	1e-4, 1e-5	1e-5	2.5e-5
Training Tokens	300B	167B	1.4T	550B
Sequence Length	8K	8K	8K	32K
Trainable Components	Projector	ViT, Projector	All	All

Table 1 Overview of MiMo-VL-7B training stages.

Figure 2 Model architecture of MiMo-VL-7B.

2.2 Pre-training Data

The MiMo-VL-7B pre-training dataset comprises 2.4 trillion tokens of high-quality, diverse multimodal data spanning images, videos, and text. This comprehensive dataset includes general image captions, interleaved data, Optical Character Recognition (OCR) data, grounding data, video content, GUI interactions, reasoning examples, and text-only sequences.

To ensure the quality of each data modality, we implement dedicated data curation pipelines tailored to the characteristics of each data type. Throughout the training process, we systematically adjust the proportions of di erent data modalities across various training stages to optimize both training e ciency and model stability. Additionally, we employ phash-based image deduplication to eliminate potential overlaps between our training datasets and evaluation benchmarks, thereby minimizing data contamination.

We detail the speci c processing procedures for each data type in the following sections.

2.2.1 Image Caption Data

The construction of our image caption dataset involves a multi-stage process designed to ensure high quality and distributional balance. We begin by aggregating a substantial volume of publicly available caption data from web sources. This initial corpus then undergoes a rigorous deduplication phase, employing image perceptual hashing (phash) in conjunction with text ltering, to yield a re ned set of unique raw captions.

Subsequently, leveraging both the images and their original textual descriptions as priors, we utilize a specialized captioning model to re-caption the entire raw caption dataset. Following re-