本文由 AINLP 公众号整理翻译，更多 LLM 资源请扫码关注!

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.

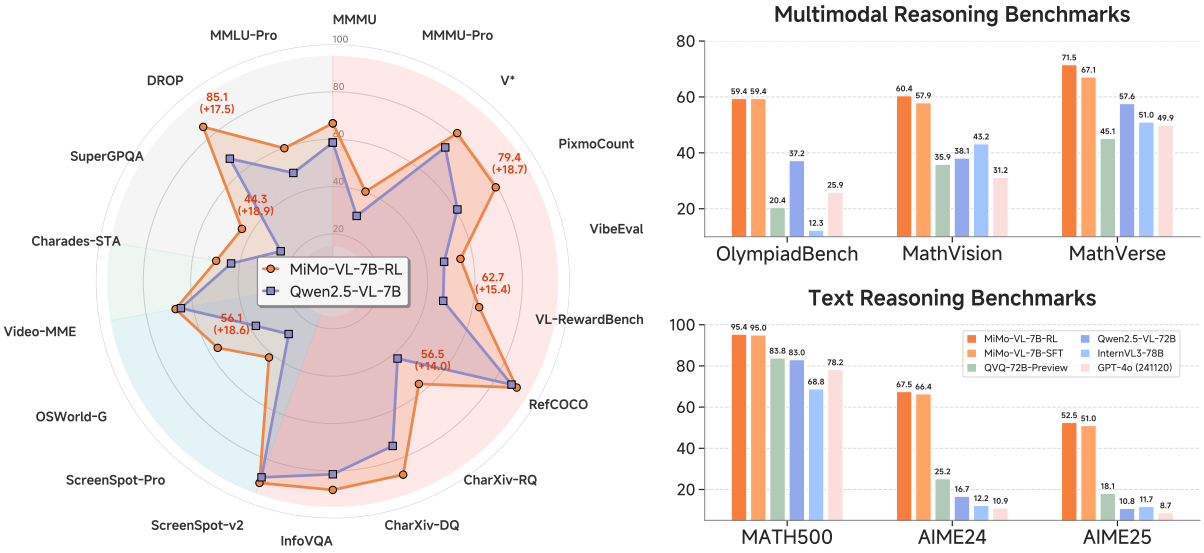


Figure 1 Benchmark performance of MiMo-VL-7B.

MiMo-VL 技术报告

LLM-Core 小米

摘要

我们开源了MiMo-VL-7B-SFT和MiMo-VL-7B-RL，这两款强大的视觉-语言模型在通用视觉理解和多模态推理方面都达到了最先进的性能。MiMo-VL-7B-RL在40个评估任务中有35个超越Qwen2.5-VL-7B，在OlympiadBench上的得分为59.4，超越了参数高达78B的模型。在GUI基础应用中，它以56.1的成绩在OSWorld-G中树立了新标准，甚至优于专门的模型如UI-TARS。我们的训练结合了四阶段预训练（2.4万亿个Token）与融合多样奖励信号的混合策略强化学习（MORL）。我们认识到在预训练阶段引入高质量的长链式思考（Chain-of-Thought）推理数据的重要性，以及尽管在多领域同时优化存在挑战，混合强化学习的优势。我们还贡献了一个涵盖50+任务的全面评估套件，以促进可复现性和推动该领域的发展。模型检查点和完整的评估套件可在https://github.com/XiaomiMiMo/MiMo-VL获取。

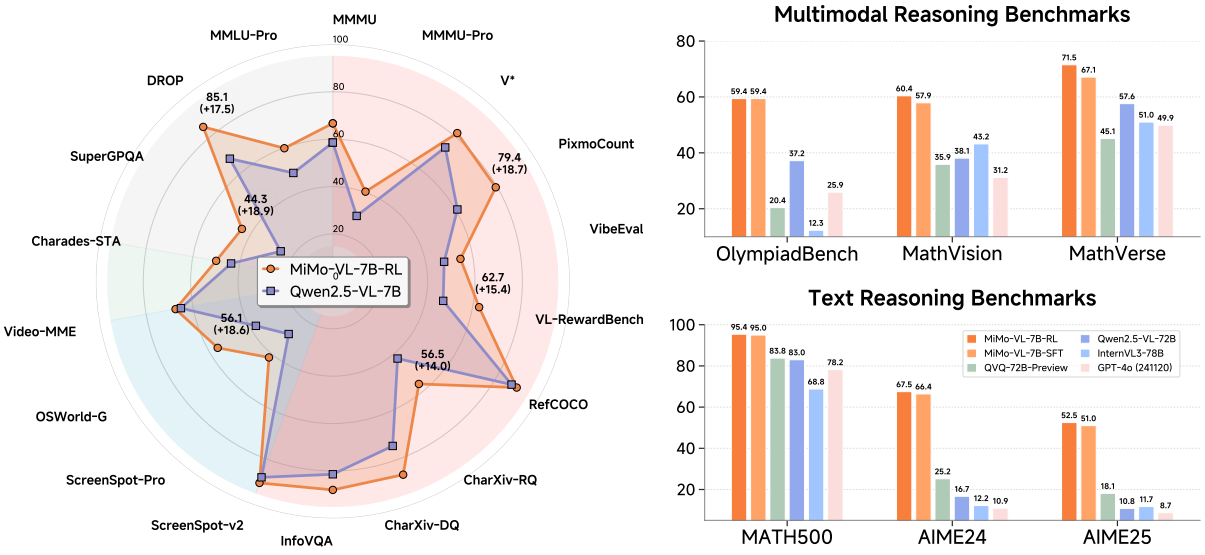


图 1 MiMo-VL-7B 的基准性能。

Contents

Contents 目录

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 significance 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 field (Liu et al., 2023; Tong et al., 2024; Bai et al., 2025a).

In this report, we share our efforts 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 fine-grained visual details, (2) a Multi-Layer Perceptron (MLP) projector for efficient cross-modal alignment, and (3) our MiMo-7B (Xiaomi, 2025) language model, specifically 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 different 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 find:

(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 significant re-evaluation. Traditional question-answering (QA) data, with its direct, short answers, often restricts models to superficial pattern matching and leads to overfitting. 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 efficiency. 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 fine-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 difficulty levels hinder stable, simultaneous improvements across all capabilities.

1 引言

视觉-语言模型（VLMs）已成为多模态人工智能系统的基础支柱，使自主代理能够感知视觉世界、推理多模态内容（Yue 等，2024b），并与数字（Xie 等，2024；OpenAI，2025）和物理环境（Zitkovich 等，2023；Black 等，2024）进行交互。这些能力的重要性促使在多个维度进行广泛探索，包括新颖的架构设计（Alayrac 等，2022；Team，2024；Ye 等，2025）和具有优化数据方案的创新训练方法（Karamcheti 等，2024；Dai 等，2024），推动该领域快速发展（Liu 等，2023；Tong 等，2024；Bai 等，2025a）。

在本报告中，我们分享了构建紧凑而强大的 VLM——MiMo-VL-7B 的努力。MiMo-VL-7B 由三个部分组成：(1) 保留细粒度视觉细节的原生分辨率视觉变换器（ViT）编码器，(2) 用于高效跨模态对齐的多层感知机（MLP）投影器，以及 (3) 我们的 MiMo-7B（小米，2025）语言模型，特别针对复杂推理任务进行了优化。

MiMo-VL-7B 的开发包括两个连续的训练过程：(1) 四阶段预训练阶段，包括投影器预热、视觉-语言对齐、通用多模态预训练和长上下文监督微调（SFT）。在这些阶段中，我们通过战略性地结合开源数据集与合成数据生成技术，策划高质量的数据集，消耗了2.4万亿个标记，并在不同阶段调整数据集分布以促进训练。此阶段产生了 MiMo-VL-7B-SFT 模型。(2) 后续的后训练阶段，我们引入了混合策略强化学习（MORL），这是一种无缝整合多样奖励信号的创新框架，涵盖感知准确性、视觉定位精度、逻辑推理能力和人类偏好。我们采用 GRPO（Shao 等，2024）的思想，并通过在此阶段仅进行策略梯度更新来增强训练的稳定性。此阶段产生了 MiMo-VL-7B-RL 模型。

在这段旅程中，我们发现：

(1) 在预训练阶段融入高质量、覆盖面广的推理数据对于提升模型性能至关重要。在当今“思考”模型的时代，大量多模态预训练数据正经历着一次重大重新评估。传统的问答（QA）数据，因其直接、简短的答案，常常限制模型仅进行表面模式匹配，导致过拟合。相比之下，采用长链式思维（CoT）合成的推理数据，能够学习复杂的逻辑关系和具有泛化能力的推理模式，提供更丰富的监督信号，显著提升性能和训练效率。为了充分利用这一优势，我们通过识别多样化的查询、利用大型推理模型生成带有长CoT的响应，并采用拒绝采样确保质量，来策划高质量的推理数据。此外，我们不将其视为补充的微调数据，而是将大量合成推理数据直接融入后续的预训练阶段，在该阶段延长训练时间可以持续带来性能提升而不出现饱和。

(2) 混合的在线策略强化学习进一步提升了模型性能，但实现稳定的同时多能力提升仍然具有挑战性。我们在推理、感知、基础和人类偏好对齐等多种能力上应用强化学习，涵盖文本、图像和视频等多模态。虽然这种混合训练方法进一步释放了模型的潜力，但不同数据领域之间的干扰仍然具有挑战性。响应长度和任务难度水平的增长趋势差异阻碍了所有能力的稳定、同步提升。

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 configuration 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.

MiMo-VL-7B-RL 在全方位的多模态能力方面都表现出色。 (1) 在基础视觉感知任务中，它在同等规模的开源VLM中达到了最先进的性能，在MMMU（岳等，2024b）中得分为66.7，并在40个评估任务中有35个优于Qwen2.5-VL-7B（白等，2025a）。 (2) 在复杂的多模态推理方面，MiMo-VL-7B-RL表现出色，在OlympiadBench（何等，2024）中得分为59.4，超越了参数高达72B的模型。 (3) 在面向智能体应用的GUI基础方面，我们的模型树立了新标准，在OSWorld-G（谢等，2025）中获得54.7的分数，甚至超过了专用模型如UI-TARS（秦等，2025b）。 (4) 在用户体验和偏好方面，MiMo-VL-7B-RL在我们内部的用户偏好评估中，在所有开源VLM中获得了最高的Elo评分，与专有模型如Claude 3.7 Sonnet表现相当。

这些结果验证了我们的方法：通过在我们的MORL框架下结合强大的感知能力、复杂的推理能力和精确的基础能力，MiMo-VL-7B-SFT和MiMo-VL-7B-RL为开源视觉-语言模型树立了新的标准。为了促进透明度和可复现性，我们还提供了一个涵盖50+任务的全面评估套件，配备完整的提示和协议，帮助社区在我们的基础上进行创新。

2 预训练

在本节中，我们介绍了我们的MiMo-VL-7B的架构设计，随后介绍了数据整理过程和预训练阶段的训练方案。

2.1 架构

MiMo-VL-7B 由三个部分组成：(1) 用于编码视觉输入（如图像和视频）的视觉变换器（ViT）；(2) 一个投影器，将视觉编码映射到与大型语言模型（LLM）对齐的潜在空间；以及 (3) 作为执行文本理解和推理的 LLM 本身。为了支持原生分辨率输入，我们采用 Qwen2.5-ViT（Bai 等，2025a）作为我们的视觉编码器。我们用 MiMo-7B-Base（小米，2025）初始化 LLM 的骨架，以利用其强大的推理能力，并使用随机初始化的多层感知机（MLP）作为投影器。整体架构如图 2 所示，模型配置详见附录 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

表1 MiMo-VL-7B 训练阶段概述。

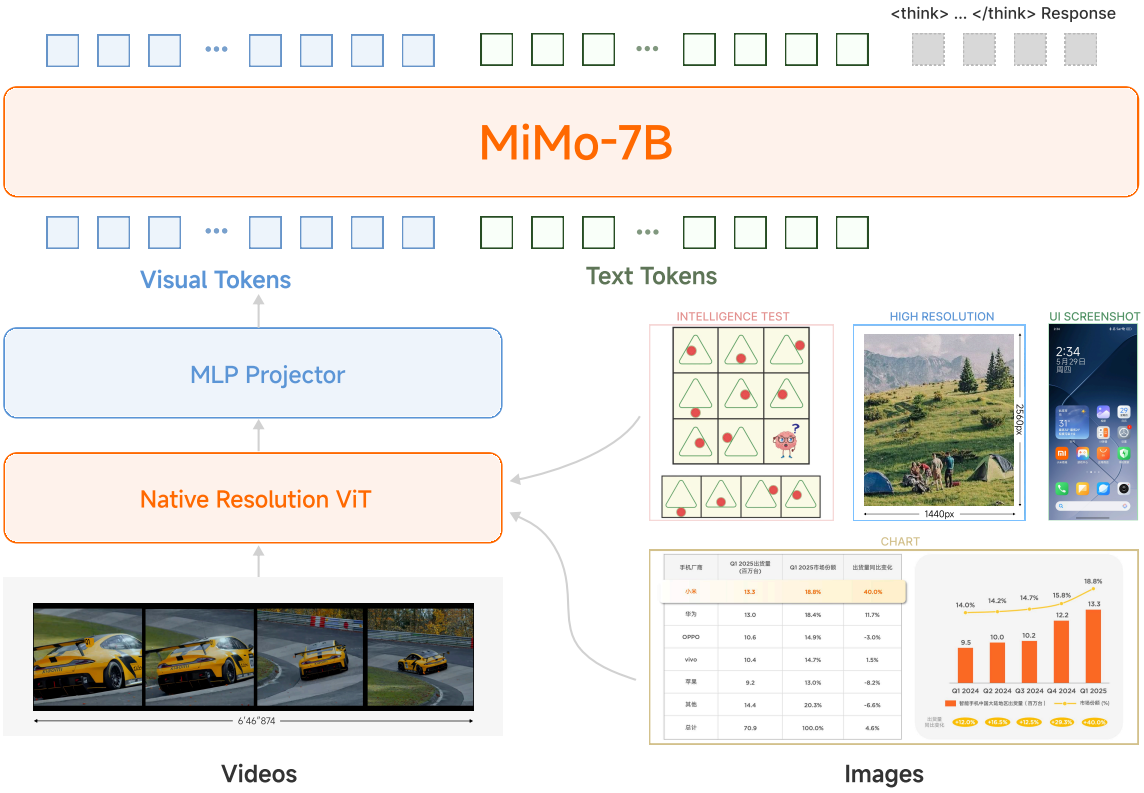


Figure 2 Model architecture of MiMo-VL-7B.

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 different data modalities across various training stages to optimize both training efficiency 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 specific 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 filtering, to yield a refined 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-

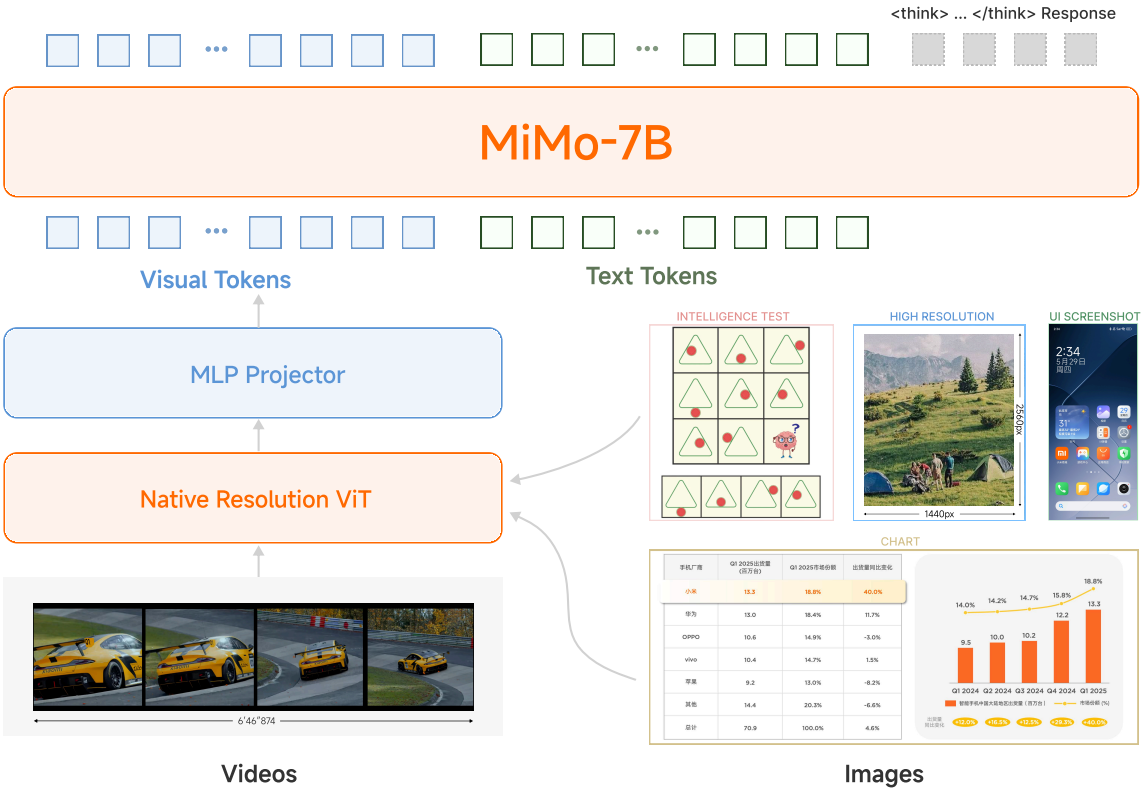


图 2 MiMo-VL-7B 的模型架构。

2.2 预训练数据

MiMo-VL-7B预训练数据集包含2.4万亿个高质量、多样化的多模态数据，涵盖图像、视频和文本。这一全面的数据集包括通用图像描述、交错数据、光学字符识别（OCR）数据、基础数据、视频内容、GUI交互、推理示例以及纯文本序列。

为了确保每种数据模态的质量，我们实施了专门的数据整理流程，针对每种数据类型的特性进行定制。在整个训练过程中，我们系统地调整不同数据模态在各个训练阶段的比例，以优化训练效率和模型稳定性。此外，我们采用基于phash的图像去重方法，消除训练数据集与评估基准之间的潜在重叠，从而最大程度地减少数据污染。

我们在以下各节中详细说明每种数据类型的具体处理流程。

2.2.1 图像标题数据

我们的图像标题数据集的构建涉及一个多阶段的过程，旨在确保高质量和分布的平衡。我们首先从网络资源中汇总大量公开可用的标题数据。然后对这个初始语料库进行严格的去重阶段，结合图像感知哈希（phash）和文本过滤，生成一组经过优化的唯一原始标题。

随后，利用图像及其原始文本描述作为先验，我们使用一种专门的描述模型对整个原始描述数据集进行重新描述。

captioning, we apply filtering mechanisms to the generated captions based on linguistic consistency and repetition patterns to ensure the quality of the re-captioned text. To address inherent data imbalances, we adopt the methodology of MetaCLIP (Xu et al., 2023), which involves constructing novel bilingual (Chinese and English) metadata. This critical step serves to refine the caption distribution, thereby mitigating the over-representation of high-frequency entries and reducing dataset noise.

The culmination of this meticulous process is a balanced, high-quality, and diverse caption dataset. Crucially, we observe that this rich dataset significantly enhances model generalization and qualitative performance, offering advantages not always fully reflected in evaluations on existing, specialized benchmarks.

2.2.2 Interleaved Data

We compile an extensive corpus of interleaved image-text data from diverse sources, including webpages, books, and academic papers. For content originating from books and papers, we employ advanced PDF parsing toolkits for content extraction and cleansing. The filtering process prioritizes and retains data types rich in world knowledge, such as textbooks, encyclopedias, manuals, guides, patents, and biographies. Within this interleaved dataset, textual segments are evaluated based on metrics including knowledge density and readability. For the visual components, we implement filters to exclude images of diminutive size, anomalous aspect ratios, unsafe content, and those with minimal visual information (e.g., decorative chapter headings and illustrations). Finally, image-text pairs are scored on relevance, complementarity, and the balance of information density, ensuring the retention of high-quality data. This curated dataset significantly augments the model's knowledge repository, thereby establishing a robust foundation for its subsequent reasoning capabilities.

2.2.3 OCR and Grounding Data

To enhance the model's capabilities in OCR and object grounding, we compile an extensive corpus of OCR and grounding data from open-source datasets for pre-training.

For the OCR data, the images encompass a diverse textual content from documents, tables, general scenes, product packaging, and mathematical formulas. To increase learning difficulty, in addition to standard printed text, we specifically incorporate images containing handwritten script, typographically deformed text, and blurry/occluded text, thereby improving the model's recognition performance and robustness. For a portion of this data, textual regions are annotated with bounding boxes, enabling the model to simultaneously predict these locations.

For the grounding data, our images feature scenarios with both single and multiple objects. We further improve the model's capacity to comprehend complex referential intentions by employing intricate object expressions within localization prompts. In all scenarios involving localization, we use absolute coordinates for representation.

2.2.4 Video Data

Our video dataset primarily draws from publicly available online videos, spanning a wide array of domains, genres, and durations. Based on the videos, we design a video re-captioning pipeline that produces dense, fine-grained event-level descriptions. Each caption is temporally grounded with precise start and end timestamps, enabling the model to develop general video perception with time-awareness. From the caption dataset, we further collect a subset with a balanced distribution of event durations for temporal grounding pretraining. We also curate video analysis data that

字幕，我们应用过滤机制对生成的字幕进行筛选，基于语言一致性和重复模式，以确保重新字幕的质量。为了解决固有的数据不平衡问题，我们采用MetaCLIP（Xu 等人，2023）的方法，该方法涉及构建新颖的中英文双语元数据。这一关键步骤旨在优化字幕分布，从而减轻高频条目的过度代表，并降低数据集噪声。

这一细致过程的高潮是一个平衡的、高质量的、多样化的字幕数据集。关键的是，我们观察到这个丰富的数据集显著提升了模型的泛化能力和定性表现，提供的优势并不总是在现有的、专门的基准测试中得到充分体现。

2.2.2 交错数据

我们编译了一个由来自多种来源的交错图像-文本数据组成的庞大语料库，包括网页、书籍和学术论文。对于来自书籍和论文的内容，我们采用先进的PDF解析工具包进行内容提取和清洗。过滤过程优先保留丰富世界知识的数据类型，如教科书、百科全书、手册、指南、专利和传记。在这个交错的数据集中，文本片段根据知识密度和可读性等指标进行评估。对于视觉部分，我们实施过滤，排除尺寸过小、比例异常、不安全内容以及信息量极少的图片（例如装饰性章节标题和插图）。最后，图像-文本对根据相关性、互补性和信息密度的平衡进行评分，确保保留高质量数据。这一经过策划的数据集大大丰富了模型的知识库，从而为其后续的推理能力奠定了坚实的基础。

2.2.3 OCR 与基础数据

为了增强模型在光学字符识别（OCR）和对象定位方面的能力，我们从开源数据集中整理了大量的OCR和定位数据，用于预训练。

对于OCR数据，图像涵盖了来自文档、表格、一般场景、产品包装和数学公式的多样文本内容。为了增加学习难度，除了标准印刷文本外，我们还特别加入了包含手写文字、排版变形文本以及模糊/遮挡文本的图像，从而提升模型的识别性能和鲁棒性。对于其中一部分数据，文本区域被标注了边界框，使模型能够同时预测这些位置。

关于基础数据，我们的图像展示了包含单个和多个对象的场景。我们通过在定位提示中使用复杂的对象表达，进一步提升模型理解复杂指称意图的能力。在所有涉及定位的场景中，我们采用绝对坐标进行表示。

2.2.4 视频数据

我们的视频数据集主要来自公开的在线视频，涵盖了各种领域、类型和时长。基于这些视频，我们设计了一个视频重新描述流程，生成密集的、细粒度的事件级描述。每个字幕都具有精确的开始和结束时间戳，确保时间上的定位，使模型能够发展具有时间感知的通用视频理解能力。在字幕数据集的基础上，我们进一步收集了一个具有平衡事件时长分布的子集，用于时间定位的预训练。我们还整理了视频分析数据，

summarize the global semantics of videos, such as narrative structures, stylistic elements, and implicit intentions. These analytical paragraphs enable the model to grasp in-depth comprehension of videos and extended world knowledge. To enhance the model's conversational coherence, we collect diverse and challenging questions about videos and synthesize corresponding responses. We also incorporate open-source video captioning and conversation datasets to further enrich our video pretraining data.

2.2.5 Graphical User Interface Data

To enhance the model's capabilities in navigating Graphical User Interfaces (GUIs), we collect open-source pre-training data covering all sorts of platforms such as mobile, web, and desktop. A synthetic data engine is also devised to compensate for the limitations of open-source data and to strengthen specific aspects of the model's capabilities. For example, we have constructed a vast amount of Chinese GUI data to enable the model to better handle Chinese GUI scenarios.

For GUI Grounding, we gather data for both element grounding and instruction grounding. Element grounding trains the model to precisely locate interface elements based on textual descriptions, establishing a robust perception of static user interfaces. Instruction grounding requires the model to identify target objects on screenshots according to user instructions, enhancing comprehension of GUI interaction logic. For this part, we additionally introduce a pre-training task that involves predicting intermediate actions based on before-and-after screenshots. Empirical evidence demonstrates that this approach significantly improves the model's dynamic perception of GUI interfaces.

For GUI Action, we collect a large scale of long GUI action trajectories. To ensure consistency across different platforms, we unified actions from mobile, web, and desktop environments into a standardized action space. A detailed specification of this action space is provided in Appendix D. This harmonization prevents action conflicts while maintaining platform-specific functionality.

2.2.6 Synthetic Reasoning Data

Our approach to generating synthetic reasoning data begins with the comprehensive curation of open-source questions. This diverse collection spans perceptual question answering, document question answering, video question answering, and visual reasoning tasks, supplemented by question-answer pairs derived from web content and literary works.

Following initial filtering of these source questions, we leverage a large reasoning model to produce answers that integrate explicit reasoning. A cornerstone of our methodology is rigorous, multi-stage quality control. Beyond verifying the factual correctness of answers, we apply strict filtering criteria to the reasoning processes themselves, evaluating clarity of thought, eliminating redundancy, and ensuring consistent formatting.

The resulting high-fidelity dataset plays a critical role in empowering our model. It facilitates the effective inheritance of strong reasoning abilities inherent in MiMo-7B-Base (Xiaomi, 2025), enabling their seamless transfer and adaptation to multimodal contexts. Consequently, this allows our model to exhibit powerful and versatile multimodal reasoning capabilities across a broad array of domains.

2.3 Pre-training Stages

Our model undergoes a four pre-training stages as illustrated in Table 1:

Stage 1: We freeze the ViT and LLM components, and warm up the randomly initialized projector

总结视频的全球语义，例如叙事结构、风格元素和隐含意图。这些分析段落使模型能够深入理解视频和扩展的世界知识。为了增强模型的对话连贯性，我们收集了关于视频的多样且具有挑战性的问题，并合成了相应的回答。我们还结合了开源的视频字幕和对话数据集，进一步丰富我们的视频预训练数据。

2.2.5 图形用户界面数据

为了增强模型在导航图形用户界面（GUIs）方面的能力，我们收集了涵盖各种平台（如移动端、网页和桌面端）的开源预训练数据。还设计了一个合成数据引擎，以弥补开源数据的局限性，并强化模型的特定能力。例如，我们构建了大量的中文GUI数据，以使模型更好地应对中文GUI场景。

对于GUI定位，我们收集了元素定位和指令定位的数据。元素定位训练模型根据文本描述准确定位界面元素，建立对静态用户界面的稳固感知。指令定位则要求模型根据用户指令在截图中识别目标对象，增强对GUI交互逻辑的理解。为此，我们还引入了一项预训练任务，涉及根据前后截图预测中间动作。实验证明，这种方法显著提升了模型对GUI界面动态感知的能力。

对于GUI操作，我们收集了大量的长GUI操作轨迹。为了确保不同平台之间的一致性，我们将来自移动端、网页和桌面环境的操作统一到一个标准化的操作空间中。该操作空间的详细规范在附录D中提供。这种统一避免了操作冲突，同时保持了平台特定的功能性。

2.2.6 合成推理数据

我们生成合成推理数据的方法始于对开源问题的全面整理。这一多样化的集合涵盖感知问答、文档问答、视频问答和视觉推理任务，并辅以来自网络内容和文学作品的问答对。

在对这些源问题进行初步筛选后，我们利用大型推理模型生成包含明确推理的答案。我们方法的基石是严格的多阶段质量控制。除了验证答案的事实正确性外，我们还对推理过程本身应用严格的筛选标准，评估思路的清晰度，消除冗余，并确保格式的一致性。

生成的高保真数据集在增强我们的模型方面发挥着关键作用。它促进了MiMo-7B-Base（小米，2025）固有的强大推理能力的有效继承，使其能够无缝转移和适应多模态环境。因此，这使我们的模型能够在广泛的领域中展现出强大且多样的多模态推理能力。

2.3 预训练阶段

我们的模型经历了如表1所示的四个预训练阶段：

阶段1：我们冻结ViT和LLM组件，并预热随机初始化的投影器

using image-caption pairs. This ensures the projector learns to map visual concepts to the language model's representation space effectively, providing informative gradient signals for subsequent training stages rather than noisy updates from a poorly aligned projector.

Stage 2: The ViT is then unfrozen, and interleaved data is introduced to further strengthen vision-language alignment. The inclusion of complex and diverse images within the interleaved data enhances the ViT's performance and robustness.

Stage 3: In this stage, all parameters are trainable. We introduce a more diverse array of data and tasks, including OCR, grounding, video, and GUI data—accumulating to 1.4 trillion tokens—to bolster the model's general multimodal capabilities. To ensure stable mid-stage evaluation monitoring, small quantities of QA, instruction-following, and reasoning data are incorporated. Furthermore, a limited amount of text-only data is utilized to preserve MiMo-7B-Base's textual capability.

Stage 4: This stage is dedicated to enhancing the model's adaptability to long-context inputs. The training sequence length is extended from 8K to 32K tokens. We introduce additional data types such as long pure text, high-resolution images, long documents, extended videos, and long reasoning data to augment its long-context processing capabilities. As long-context packing significantly increases the effective batch size, the learning rate is adjusted from 1e-5 to 2.5e-5. Relative to Stage 3, this stage features a markedly increased proportion of reasoning data, alongside the introduction of long-form reasoning patterns.

These four stages create a powerful model, MiMo-VL-7B-SFT. With particular emphasis on Stage 4, the model's reasoning capabilities are fully realized, enabling it to address highly intricate STEM problems. This advanced reasoning aptitude also generalizes effectively to common perception tasks. Consequently, our model demonstrates exceptionally high performance across various downstream benchmarks.

3 Post-Training

Building upon the visual perception capabilities and multimodal reasoning established during pre-training, we conduct post-training to further enhance MiMo-VL-7B. Our approach employs a novel Mixed On-policy Reinforcement Learning (MORL) framework that seamlessly integrates Reinforcement Learning with Verifiable Rewards (RLVR) (Shao et al., 2024; Lambert et al., 2025) with Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) to improve MiMo-VL-7B on challenging reasoning tasks and alignment with human preferences.

3.1 Reinforcement Learning with Verifiable Rewards

RLVR relies exclusively on rule-based reward functions, enabling continuous model self-improvement. In the post-training of MiMo-VL-7B, we design multiple verifiable reasoning and perception tasks where final solutions can be precisely validated using predefined rules.

Visual Reasoning Visual reasoning capabilities are fundamental for multimodal models to understand and solve complex problems that require both visual perception and logical thinking. To facilitate this capability, we compile diverse verifiable STEM questions from open-source communities and proprietary K-12 collections. An LLM is prompted to filter proof-based problems and rewrite multiple-choice questions with numerical or symbolic answers into free-answer formats, alleviating potential reward hacking. We further refine question quality through comprehensive model-based difficulty assessment, excluding questions that either cannot be solved by advanced

使用图像-标题对。这确保投影仪能够有效地将视觉概念映射到语言模型的表示空间，为后续的训练阶段提供有用的梯度信号，而不是来自对齐不良的投影仪的噪声更新。

阶段2：然后解冻ViT，并引入交错数据以进一步增强视觉-语言对齐。在交错数据中加入复杂多样的图像，提升了ViT的性能和鲁棒性。

第3阶段：在此阶段，所有参数均可训练。我们引入了更多样化的数据和任务，包括光学字符识别（OCR）、定位、视频和图形用户界面（GUI）数据——累计达到1.4万亿个标记——以增强模型的通用多模态能力。为了确保中期评估的稳定监控，加入了少量的问答（QA）、指令遵循和推理数据。此外，还使用了有限的纯文本数据，以保持MiMo-7B-Base的文本能力。

第4阶段：此阶段旨在增强模型对长上下文输入的适应能力。训练序列长度从8K扩展到32K个标记。我们引入了额外的数据类型，如长纯文本、高分辨率图像、长文档、扩展视频和长推理数据，以增强其长上下文处理能力。由于长上下文打包显著增加了有效批次大小，学习率从1e-5调整为2.5e-5。与第3阶段相比，此阶段的推理数据比例显著增加，同时引入了长形式的推理模式。

这四个阶段构建了一个强大的模型，MiMo-VL-7B-SFT。特别强调第4阶段，模型的推理能力得以充分发挥，使其能够解决高度复杂的STEM问题。这种先进的推理能力也能有效推广到常见的感知任务。因此，我们的模型在各种下游基准测试中表现出色。

3 训练后

在预训练期间建立的视觉感知能力和多模态推理基础上，我们进行了后续训练，以进一步增强MiMo-VL-7B。我们的方法采用了一种新颖的混合策略强化学习（MORL）框架，巧妙地将具有可验证奖励的强化学习（RLVR）（Shao 等，2024；Lambert 等，2025）与来自人类反馈的强化学习（RLHF）（Ouyang 等，2022）相结合，以提升MiMo-VL-7B在具有挑战性的推理任务中的表现以及与人类偏好的对齐。

3.1 可验证奖励的强化学习

RLVR 完全依赖基于规则的奖励函数，从而实现模型的持续自我提升。在 MiMo-VL-7B 的后训练阶段，我们设计了多个可验证的推理和感知任务，最终解决方案可以通过预定义的规则进行精确验证。

视觉推理 视觉推理能力对于多模态模型理解和解决需要视觉感知和逻辑思维的复杂问题至关重要。为了促进这一能力，我们从开源社区和专有的K-12资料库中整理了多样的可验证STEM问题。通过提示大型语言模型（LLM）筛选基于证明的问题，并将带有数值或符号答案的多项选择题改写为自由回答格式，以减轻潜在的奖励黑客行为。我们还通过全面的模型难度评估来优化问题质量，排除那些无法由高级模型解决的问题。

VLMs or are too easy, with a MiMo-VL-7B rollout pass rate exceeding 90%. Additionally, we remove questions solvable even without image inputs. After data cleaning and category balancing, we curate a visual reasoning dataset of 80K problems. For evaluation, we use the rule-based Math-Verify library to determine response correctness.[1]

Text Reasoning Since most visual reasoning data is limited to K-12 level questions, the reasoning performance of RL-trained models could be constrained. In contrast, text-only reasoning datasets include more challenging queries requiring college or competition-level intelligence. To unleash the full reasoning potential of our model, we incorporate mathematical reasoning data from Xiaomi (2025). Rewards are computed using the same rule-based Math-Verify library to ensure consistent evaluation across both visual and textual reasoning tasks.

Image Grounding Accurate spatial localization is essential for models to understand object relationships and spatial reasoning in images. We include both general and GUI grounding tasks in our RLVR framework to enhance MiMo-VL-7B's grounding capability. For bounding box predictions, rewards are calculated using the Generalized Intersection over Union (GIoU) (Rezatofighi et al., 2019) between predicted and ground-truth boxes. For point-style outputs, rewards are determined by whether the predicted point falls within the ground-truth bounding box.

Visual Counting Precise counting abilities are essential for quantitative visual understanding and mathematical reasoning in visual contexts (Chen et al., 2025a). We enhance visual counting capabilities through RL training, where rewards are defined by the accuracy of the model's counting predictions compared to ground-truth counts.

Temporal Video Grounding Beyond static image understanding and reasoning, we extend our RLVR framework to dynamic video content to capture temporal dependencies. We incorporate temporal video grounding tasks that require the model to localize video segments corresponding to natural language queries (Wang et al., 2025). The model outputs timestamps in `[mm:ss,mm:ss]` format to indicate the start and end times of the target video segments. Rewards are computed as the Intersection over Union (IoU) between predicted and ground-truth temporal segments.

3.2 Reinforcement Learning from Human Feedback

To align model outputs with human preferences and mitigate undesirable behaviors, we employ Reinforcement Learning from Human Feedback (RLHF) as a complementary approach to our verifiable reward framework.

Query Collection Query diversity is paramount to the success of RLHF. Our methodology commences with gathering multimodal and text-only queries from open-source instruction tuning datasets and in-house human-written sources. All collected queries, both text and multimodal, then undergo a dedicated screening process. To further enhance diversity, we employ techniques such as clustering queries based on their embeddings and analyzing the resultant patterns. Crucially, we balance the proportions of Chinese and English queries, as well as those targeting helpfulness and harmlessness, before curating the final query set. For each selected query, MiMo-VL-7B and multiple other top-performing VLMs are prompted to generate responses. These responses are subsequently pairwise ranked by an advanced VLM to construct the definitive dataset for reward

[1] https://github.com/huggingface/Math-Verify

VLMs 或者说太容易了，MiMo-VL-7B 的推出通过率超过 90%。此外，我们还去除了即使没有图像输入也能解决的问题。在数据清洗和类别平衡后，我们整理了一个包含 80K 个问题的视觉推理数据集。为了评估，我们使用基于规则的 Math-Verify 库来判断回答的正确性。[1]

文本推理 由于大多数视觉推理数据仅限于K-12水平的问题，RL训练模型的推理性能可能受到限制。相比之下，纯文本推理数据集包含更多具有挑战性的问题，要求具备大学或竞赛级别的智能。为了释放我们模型的全部推理潜力，我们引入了来自小米（2025）的数学推理数据。奖励使用相同的基于规则的Math-Verify库进行计算，以确保在视觉和文本推理任务中的评估一致性。

图像定位 准确的空间定位对于模型理解图像中的对象关系和空间推理至关重要。我们在RLVR框架中同时包含通用和GUI定位任务，以增强MiMo-VL-7B的定位能力。对于边界框预测，奖励通过预测框与真实框之间的广义交并比（GIoU）（Rezatofighi等，2019）进行计算。对于点式输出，奖励取决于预测点是否落在真实边界框内。

视觉计数 精确的计数能力对于视觉环境中的定量理解和数学推理至关重要（Chen 等，2025a）。我们通过强化学习训练提升视觉计数能力，其中奖励由模型的计数预测与真实计数的准确性决定。

超越静态图像理解与推理，时间视频定位扩展了我们的RLVR框架，以捕捉时间依赖性。我们引入需要模型定位与自然语言查询对应的视频片段的时间视频定位任务（Wang 等，2025）。模型输出的时间戳采用`[mm:ss,mm:ss]`格式，以指示目标视频片段的起始和结束时间。奖励计算为预测的和真实的时间片段之间的交并比（IoU）。

3.2 从人类反馈中进行强化学习

为了使模型输出与人类偏好保持一致并减轻不良行为，我们采用来自人类反馈的强化学习（RLHF）作为我们可验证奖励框架的补充方法。

查询收集 查询多样性对于RLHF的成功至关重要。我们的方法从收集多模态和纯文本查询开始，数据来源包括开源的指令调优数据集和内部人工编写的资料。所有收集到的查询，无论是文本还是多模态，随后都经过专门的筛选过程。为了进一步增强多样性，我们采用了基于嵌入的聚类等技术，并分析由此产生的模式。关键的是，在策划最终查询集之前，我们平衡了中文和英文查询的比例，以及针对有用性和无害性的查询比例。对于每个选中的查询，MiMo-VL-7B和其他多个表现优异的VLM会被提示生成响应。这些响应随后由先进的VLM进行两两排序，以构建最终的奖励数据集。

[1] `https://github.com/huggingface/Math-Verify`

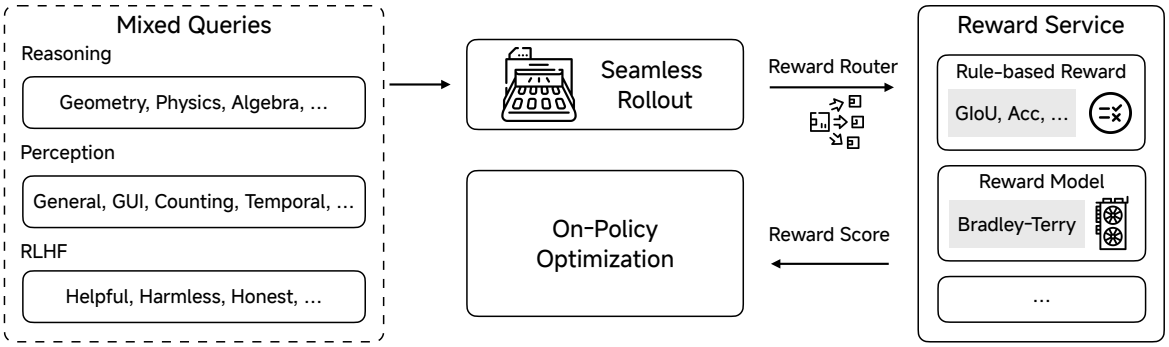


Figure 3 Mixed On-policy Reinforcement Learning in post-training phase.

model training. Notably, to mitigate potential reward hacking, this same query set is utilized for both reward model training and the RLHF process.

Reward Model We develop two specialized reward models tailored to different input modalities, training them using the Bradley-Terry reward modeling objective (Ouyang et al., 2022). The text-only reward model is initialized from MiMo-7B (Xiaomi, 2025) to leverage its strong language understanding capabilities, while the multimodal reward model builds upon MiMo-VL-7B to effectively process queries containing visual inputs. This dual-model approach ensures optimal performance across both textual and multimodal evaluation scenarios.

3.3 Mixed On-Policy Reinforcement Learning

In the post-training phase of MiMo-VL-7B, we implement Mixed On-policy Reinforcement Learning (MORL) to simultaneously optimize RLVR and RLHF objectives. As illustrated in Figure 3, we integrate rule-based and model-based rewards as unified services within the verl framework (Sheng et al., 2024), enhanced by the Seamless Rollout Engine (Xiaomi, 2025).

On-Policy RL Recipe We adopt a fully on-policy variant of GRPO (Shao et al., 2024) as the RL algorithm, which demonstrates robust training stability and effective exploration capabilities (Chen et al., 2025b). For each problem q, the algorithm samples a group of responses $\{o_1, o_2, ..., o_G\}$ from the policy π_θ, and updates the policy by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim D,\{o_i\}_{i=1}^{G}\sim\pi_\theta(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{j=1}^{|o_i|}A_{i,j}\right], \tag{1}$$

where $A_{i,j}$ is the advantage, which is computed by the rewards $\{r_1, r_2, ..., r_G\}$ of responses in the same group:

$$A_{i,j} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}. \tag{2}$$

Compared to vanilla GRPO, this on-policy variant performs single-step policy updates following response rollout, eliminating the need for a clipped surrogate training objective. Following Xiaomi (2025), we integrate several advancements, including removal of the KL loss, dynamic sampling, easy data filter, and re-sampling strategies, into our RL training recipe.

Reward-as-a-Service The MORL process integrates tasks across reasoning, perception, grounding, multimodal RLHF, and text-only RLHF, each requiring distinct reward functions or dedicated

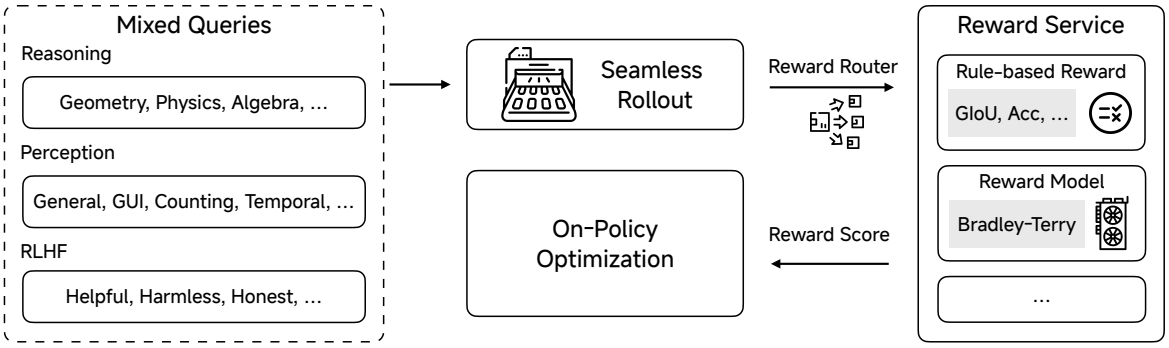


图3 训练后阶段的混合策略强化学习

模型训练。值得注意的是，为了减轻潜在的奖励黑客行为，采用相同的查询集用于奖励模型训练和RLHF过程。

奖励模型 我们开发了两种专门的奖励模型，针对不同的输入模态进行定制，并使用Bradley-Terry奖励建模目标（Ouyang等，2022）进行训练。纯文本奖励模型从MiMo-7B（小米，2025）初始化，以利用其强大的语言理解能力，而多模态奖励模型则基于MiMo-VL-7B，能够有效处理包含视觉输入的查询。这种双模型方法确保在文本和多模态评估场景中都能实现最佳性能。

3.3 混合在策略强化学习

在MiMo-VL-7B的后训练阶段，我们实现了混合策略强化学习（MORL），以同时优化RLVR和RLHF目标。如图3所示，我们将基于规则和基于模型的奖励整合为统一的服务，嵌入在verl框架中（Sheng等人，2024），并由无缝展开引擎（Xiaomi，2025）增强。

基于策略的强化学习方案我们采用了GRPO（Shao等人，2024）的完全基于策略的变体作为强化学习算法，该算法表现出稳健的训练稳定性和有效的探索能力（Chen等人，2025b）。对于每个问题q，算法从策略π_θ中采样一组响应$\{o_1, o_2, ..., o_G\}$，并通过最大化以下目标来更新策略：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim D,\{o_i\}_{i=1}^{G}\sim\pi_\theta(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{j=1}^{|o_i|}A_{i,j}\right], \tag{1}$$

其中 $A_{i,j}$ 是优势值，它通过同一组中响应的奖励 $\{r_1, r_2, ..., r_G\}$ 计算得出：

$$A_{i,j} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}. \tag{2}$$

与普通的GRPO相比，这个基于策略的变体在响应滚动后进行单步策略更新，省去了裁剪的代理目标训练。根据小米（2025年），我们将多项改进整合到我们的强化学习训练方案中，包括去除KL损失、动态采样、简易数据过滤和重采样策略。

奖励即服务（Reward-as-a-Service）MORL 过程整合了推理、感知、基础、多模态 RLHF 和纯文本 RLHF 等任务，每个任务都需要不同的奖励函数或专用的

reward models. To provide a unified interface and near-zero latency reward computation, we introduce Reward-as-a-Service (RaaS). A reward router dynamically selects the appropriate reward function based on the query's task type. To minimize latency, reward models are deployed as standalone services, ensuring scalable reward computation accessible via HTTP. All rewards are normalized to the range $[0, 1]$. No additional reward, such as format rewards, is incorporated in our training process.

4 Evaluation

We evaluate MiMo-VL-7B across 50 tasks to comprehensively assess its capabilities. Appendix C lists all the benchmarks adopted in our evaluation. We also assess model performance with an in-house evaluation set.

4.1 Evaluation Setting

For image understanding benchmarks, we set the max pixels for the input image to $4096 \times 28 \times 28$ and the maximum generation tokens to 32,768, employing greedy search for decoding. For video benchmarks, videos are sampled at 2 FPS, with a maximum of 256 frames and a total token limit of 16,384. For text evaluation benchmarks, we set the max new tokens for generation as 32,768, temperature as 0.6 and top-p as 0.95. We adapt the existing framework based on LMMs-Eval (Zhang et al., 2024a) to better accommodate long-CoT reasoning models. We further optimize the evaluation logic for specific tasks to ensure better evaluation consistency. To facilitate open research, we open-source our evaluation framework with all prompts used.[2]

4.2 General Capabilities

Table 2 presents benchmark results assessing the general capabilities of VLMs. Our MiMo-VL-7B models demonstrate consistently leading performance across a diverse range of vision-language and text benchmarks, establishing new state-of-the-art results among open-source models and even surpassing proprietary counterparts.

Specifically, our MiMo-VL-7B models achieve state-of-the-art results among open-source models. MiMo-VL-7B-SFT and **(i)** On general vision-language tasks, our models achieve exceptional performance that leads the open-source field. MiMo-VL-7B-SFT and MiMo-VL-7B-RL obtain 64.6% and 66.7% on MMMU_{val} respectively, outperforming much larger models such as Gemma 3 27B. For document and chart understanding, MiMo-VL-7B-RL excels with a top open-source score of 56.5% on $\text{CharXiv}_{\text{RQ}}$, significantly exceeding Qwen2.5-VL (42.5%) by 14.0 points and InternVL3 (37.6%) by 18.9 points. **(ii)** Our models demonstrate superior video understanding capabilities while maintaining strong textual performance. MiMo-VL-SFT achieves a leading 53.1% on Video-MMMU, and MiMo-VL-RL obtains an impressive 50.0% mIoU on Charades-STA. **(iii)** Compared with MiMo-7B, our models maintain decent performance on text-only benchmarks. **(iv)** Remarkably, our MoRL yields comprehensive improvements, with the most impressive gains observed on challenging benchmarks such as VibeEval and CountBench.

4.3 Reasoning Tasks

Table 3 presents evaluation results for multimodal and text reasoning benchmarks. In multimodal reasoning, both the SFT and RL models significantly outperform all compared open-source

[2] https://github.com/XiaomiMiMo/lmms-eval

奖励模型。为了提供统一的接口和几乎零延迟的奖励计算，我们引入了奖励即服务（Reward-as-a-Service，RaaS）。奖励路由器根据查询的任务类型动态选择合适的奖励函数。为了最小化延迟，奖励模型作为独立的服务部署，确保通过HTTP实现可扩展的奖励计算。所有奖励都被归一化到范围[0, 1]。在我们的训练过程中，没有加入额外的奖励，例如格式奖励。

4 评估

我们在50个任务中评估了MiMo-VL-7B，以全面评估其能力。附录C列出了我们评估中采用的所有基准。我们还使用内部评估集对模型性能进行了评估。

4.1 评估设置

对于图像理解基准测试，我们将输入图像的最大像素数设置为 $4096 \times 28 \times 28$，最大生成令牌数为 32,768，采用贪婪搜索进行解码。对于视频基准测试，视频采样频率为每秒2帧，最大帧数为256，总令牌限制为16,384。对于文本评估基准，我们将最大新令牌数设置为 32,768，温度为0.6，top-p为0.95。我们基于LMMs-Eval（Zhang等人，2024a）调整了现有框架，以更好地适应长链推理模型。我们还进一步优化了特定任务的评估逻辑，以确保更好的评估一致性。为了促进开放研究，我们开源了我们的评估框架以及所有使用的提示。[2]

4.2 一般能力

表2展示了评估VLMs通用能力的基准测试结果。我们的MiMo-VL-7B模型在各种视觉-语言和文本基准测试中表现出持续领先的性能，创造了新的开源模型的最先进水平，甚至超越了专有模型。

具体而言，我们的 MiMo-VL-7B 模型在开源模型中实现了最先进的结果。MiMo-VL-7B-SFT 和 (i) 在通用视觉-语言任务中，我们的模型表现出色，领先于开源领域。MiMo-VL-7B-SFT 和 MiMo-VL-7B-RL 在 MMMU_{val} 上分别获得了 64.6% 和 66.7%，超越了如 Gemma 3 27B 这样规模更大的模型。在文档和图表理解方面，MiMo-VL-7B-RL 以 56.5% 的开源最高分在 $\text{CharXiv}_{\text{RQ}}$ 上表现出色，显著超过 Qwen2.5-VL（42.5%）14.0 点和 InternVL3（37.6%）18.9 点。 (ii) 我们的模型在视频理解能力方面表现优越，同时保持强大的文本性能。MiMo-VL-SFT 在 Video-MMMU 上取得了 53.1% 的领先成绩，MiMo-VL-RL 在 Charades-STA 上获得了令人印象深刻的 50.0% mIoU。 (iii) 与 MiMo-7B 相比，我们的模型在纯文本基准测试中保持了不错的性能。 (iv) 值得注意的是，我们的 MoRL 实现了全面的提升，在 VibeEval 和 CountBench 等具有挑战性的基准测试中取得了最令人印象深刻的增益。

4.3 推理任务

表3展示了多模态和文本推理基准的评估结果。在多模态推理中，SFT和RL模型都显著优于所有对比的开源模型

[2] https://github.com/XiaomiMiMo/lmms-eval

Benchmark	Metrics	MiMo-VL 7B-SFT	MiMo-VL 7B-RL	Qwen2.5-VL 7B	InternVL3 8B	Gemma-3 27B-IT	GPT-4o	Claude 3.7 Sonnet
General								
$\text{MMMU}^{\dagger}_{\text{val}}$	Acc.	64.6	**66.7**	58.6	62.7	64.9	70.7	69.8*
$\text{MMMU-Pro}_{\text{standard}}$	Acc.	45.2	**46.2**	34.7*	45.6*	37.8*	42.5*	56.5*
$\text{MMMU-Pro}_{\text{vision}}$	Acc.	39.4	**40.3**	29.4*	37.8*	24.9*	36.1*	45.8*
$\text{MMBench-en}_{\text{test}}$	Acc.	**84.5**	84.4	83.5	83.4	81.6*	84.6*	84.8*
$\text{MMBench-cn}_{\text{test}}$	Acc.	81.9	82.0	**83.4**	82.2	82.4*	84.5*	83.7*
Mantis	Acc.	**78.8**	78.3	74.7*	72.8*	70.0*	75.6*	75.1*
$\text{MME-RealWorld}_{\text{en}}$	Acc.	57.4	**59.1**	57.4	56.1*	51.9*	57.5*	50.8*
$\text{MME-RealWorld}_{\text{cn}}$	Acc.	55.0	55.5	51.2*	**58.5***	47.9*	58.5*	40.6*
AI2D	Acc.	83.2	83.5	83.9	**85.2**	84.5	82.6*	81.4*
$\text{BLINK}_{\text{val}}$	Acc.	**62.5**	62.4	56.4	55.5	53.3*	60.0	62.3*
CV-Bench	Acc.	81.8	**82.3**	75.4*	81.0*	70.4*	76.0*	75.4*
$\text{VibeEval}^{\dagger}$	GPT-Score	47.2	**54.7**	47.7*	43.6*	44.0*	64.7	39.0*
$\text{VL-RewardBench}^{\dagger}$	Macro Acc.	61.9	**62.7**	47.3*	49.7*	51.9*	62.4	67.4*
V*	Acc.	80.6	**81.7**	73.8*	72.8*	50.8*	73.9	-
VLMs are Blind	Acc.	78.0	**79.4**	37.4*	36.8*	18.6*	49.8*	72.1*
PixmoCount	Acc.	**79.4**	**79.4**	60.7*	62.0	48.6*	54.4*	53.5*
CountBench	Acc.	87.0	**90.4**	74.1*	80.0*	77.2*	85.7*	90.2*
$\text{RefCOCO}^{\text{avg}}_{\text{val}}$	Acc.@0.5	85.7	89.6	87.1	**90.1**	-	-	-
Doc & OCR								
$\text{ChartQA}^{\dagger}$	Acc.	**92.9**	91.7	90.2*	89.6*	78.0	86.7	92.2*
$\text{CharXiv}^{\dagger}_{\text{RQ}}$	Acc.	54.4	**56.5**	42.5	37.6	29.2*	52.0	63.0*
$\text{CharXiv}^{\dagger}_{\text{DQ}}$	Acc.	**87.0**	86.8	73.9	73.6	63.8*	86.5	89.5*
$\text{DocVQA}^{\dagger}_{\text{val}}$	Acc.	95.2	**95.7**	95.5*	89.4*	86.6	93.0*	94.1*
$\text{InfoVQA}^{\dagger}_{\text{val}}$	Acc.	87.2	**88.0**	81.4*	70.7*	70.6	82.1*	65.5*
SEED-Bench-2-Plus	Acc.	71.9	**72.4**	70.7	69.7	66.3*	71.1*	72.9*
$\text{OCRBench}^{\dagger}$	Acc.	87.6	86.6	**89.7***	88.0	77.6*	84.3*	80.6*
GUI								
$\text{VisualWebBench}_{\text{avg}}$	Acc.	**80.2**	**80.2**	72.9*	62.7	49.7*	80.2	79.3*
$\text{WebSrc}_{\text{avg}}$	SQuAD F1	**96.5**	95.3	94.6*	91.1	89.0*	89.1*	91.1*
ScreenSpot	Center Acc.	**87.3**	87.2	84.7	79.5	-	18.3	-
ScreenSpot-v2	Center Acc.	89.5	**90.5**	88.0	81.4	-	18.5	-
$\text{ScreenSpot-Pro}_{avg}$	Center Acc.	39.9	**41.9**	29.0	-	-	-	-
$\text{OSWorld-G}_{\text{no\_refusal}}$	Center Acc.	54.7	**56.1**	37.5*	-	-	-	-
Video								
$\text{Video-MME}_{\text{w/o sub.}}$	Acc.	66.9	**67.4**	65.1	66.3	-	-	-
Video-MMMU	Acc.	**53.1**	43.3	47.4	48.9*	-	-	-
$\text{EgoSchema}_{\text{val}}$	Acc.	60.4	59.6	62.4*	**68.2***	-	-	-
Charades-STA	mIoU	38.5	**50.0**	43.6	25.4	-	-	-
Text								
GPQA Diamond	Pass@1	56.3	**58.3**	30.3*	33.5*	40.9	50.6*	66.0*
SuperGPQA	Pass@1	42.6	**44.3**	25.4*	29.4*	29.3*	31.6*	54.4*
DROP	3-shot F1	82.7	**85.1**	67.6*	77.5*	72.2	81.5	87.2*
MMLU-Pro	EM	59.8	**64.8**	48.7*	58.3*	45.3	69.1	81.0
IF-Eval	Prompt Strict	75.3	75.9	75.1*	87.2*	**88.9**	82.5	88.7*

Table 2 Comparison of MiMo-VL-7B models with other models on diverse visual-language and text benchmarks. Results marked with * are obtained using our evaluation framework. † indicates that evaluation is performed using GPT-4o. The best results among open-source models are **bolded**.

Benchmark	Metrics	MiMo-VL 7B-SFT	MiMo-VL 7B-RL	Qwen2.5-VL 7B	InternVL3 8B	Gemma-3 27B-IT	GPT-4o	Claude 3.7 Sonnet
General								
$\text{MMMU}^{\dagger}_{\text{val}}$	Acc.	64.6	**66.7**	58.6	62.7	64.9	70.7	69.8*
$\text{MMMU-Pro}_{\text{standard}}$	Acc.	45.2	**46.2**	34.7*	45.6*	37.8*	42.5*	56.5*
$\text{MMMU-Pro}_{\text{vision}}$	Acc.	39.4	**40.3**	29.4*	37.8*	24.9*	36.1*	45.8*
$\text{MMBench-en}_{\text{test}}$	Acc.	**84.5**	84.4	83.5	83.4	81.6*	84.6*	84.8*
$\text{MMBench-cn}_{\text{test}}$	Acc.	81.9	82.0	**83.4**	82.2	82.4*	84.5*	83.7*
Mantis	Acc.	**78.8**	78.3	74.7*	72.8*	70.0*	75.6*	75.1*
$\text{MME-RealWorld}_{\text{en}}$	Acc.	57.4	**59.1**	57.4	56.1*	51.9*	57.5*	50.8*
$\text{MME-RealWorld}_{\text{cn}}$	Acc.	55.0	55.5	51.2*	**58.5***	47.9*	58.5*	40.6*
AI2D	Acc.	83.2	83.5	83.9	**85.2**	84.5	82.6*	81.4*
$\text{BLINK}_{\text{val}}$	Acc.	**62.5**	62.4	56.4	55.5	53.3*	60.0	62.3*
CV-Bench	Acc.	81.8	**82.3**	75.4*	81.0*	70.4*	76.0*	75.4*
$\text{VibeEval}^{\dagger}$	GPT-Score	47.2	**54.7**	47.7*	43.6*	44.0*	64.7	39.0*
$\text{VL-RewardBench}^{\dagger}$	Macro Acc.	61.9	**62.7**	47.3*	49.7*	51.9*	62.4	67.4*
V*	Acc.	80.6	**81.7**	73.8*	72.8*	50.8*	73.9	-
VLMs are Blind	Acc.	78.0	**79.4**	37.4*	36.8*	18.6*	49.8*	72.1*
PixmoCount	Acc.	**79.4**	**79.4**	60.7*	62.0	48.6*	54.4*	53.5*
CountBench	Acc.	87.0	**90.4**	74.1*	80.0*	77.2*	85.7*	90.2*
$\text{RefCOCO}^{\text{avg}}_{\text{val}}$	Acc.@0.5	85.7	89.6	87.1	**90.1**	-	-	-
Doc & OCR								
$\text{ChartQA}^{\dagger}$	Acc.	**92.9**	91.7	90.2*	89.6*	78.0	86.7	92.2*
$\text{CharXiv}^{\dagger}_{\text{RQ}}$	Acc.	54.4	**56.5**	42.5	37.6	29.2*	52.0	63.0*
$\text{CharXiv}^{\dagger}_{\text{DQ}}$	Acc.	**87.0**	86.8	73.9	73.6	63.8*	86.5	89.5*
$\text{DocVQA}^{\dagger}_{\text{val}}$	Acc.	95.2	**95.7**	95.5*	89.4*	86.6	93.0*	94.1*
$\text{InfoVQA}^{\dagger}_{\text{val}}$	Acc.	87.2	**88.0**	81.4*	70.7*	70.6	82.1*	65.5*
SEED-Bench-2-Plus	Acc.	71.9	**72.4**	70.7	69.7	66.3*	71.1*	72.9*
$\text{OCRBench}^{\dagger}$	Acc.	87.6	86.6	**89.7***	88.0	77.6*	84.3*	80.6*
GUI								
$\text{VisualWebBench}_{\text{avg}}$	Acc.	**80.2**	**80.2**	72.9*	62.7	49.7*	80.2	79.3*
$\text{WebSrc}_{\text{avg}}$	SQuAD F1	**96.5**	95.3	94.6*	91.1	89.0*	89.1*	91.1*
ScreenSpot	Center Acc.	**87.3**	87.2	84.7	79.5	-	18.3	-
ScreenSpot-v2	Center Acc.	89.5	**90.5**	88.0	81.4	-	18.5	-
$\text{ScreenSpot-Pro}_{avg}$	Center Acc.	39.9	**41.9**	29.0	-	-	-	-
$\text{OSWorld-G}_{\text{no\_refusal}}$	Center Acc.	54.7	**56.1**	37.5*	-	-	-	-
Video								
$\text{Video-MME}_{\text{w/o sub.}}$	Acc.	66.9	**67.4**	65.1	66.3	-	-	-
Video-MMMU	Acc.	**53.1**	43.3	47.4	48.9*	-	-	-
$\text{EgoSchema}_{\text{val}}$	Acc.	60.4	59.6	62.4*	**68.2***	-	-	-
Charades-STA	mIoU	38.5	**50.0**	43.6	25.4	-	-	-
Text								
GPQA Diamond	Pass@1	56.3	**58.3**	30.3*	33.5*	40.9	50.6*	66.0*
SuperGPQA	Pass@1	42.6	**44.3**	25.4*	29.4*	29.3*	31.6*	54.4*
DROP	3-shot F1	82.7	**85.1**	67.6*	77.5*	72.2	81.5	87.2*
MMLU-Pro	EM	59.8	**64.8**	48.7*	58.3*	45.3	69.1	81.0
IF-Eval	Prompt Strict	75.3	75.9	75.1*	87.2*	**88.9**	82.5	88.7*

表2 不同视觉-语言和文本基准测试中MiMo-VL-7B模型与其他模型的比较。标有*的结果是使用我们的评估框架获得的。† 表示使用GPT-4o进行评估。在开源模型中，最佳结果以加粗显示。

Reasoning Benchmark	Metrics	MiMo-VL 7B-SFT	MiMo-VL 7B-RL	QVQ-72B Preview	Qwen2.5-VL 72B	Intern-VL3 78B	Qwen2.5 72B	GPT-4o	Gemini-2.5 Pro
Multi-modal									
OlympiadBench	Acc.	**59.4**	**59.4**	20.4	37.2*	12.3	-	25.9	69.8
MathVision	Acc.	57.9	**60.4**	35.9	38.1	43.2	-	31.2	69.1
MathVerse$^{\dagger}_{\text{vision_only}}$	Acc.	67.1	**71.5**	45.1*	57.6	51.0	-	49.9	76.7
DynaMath	Worst-case Acc.	**46.9**	45.9	30.7	38.1*	35.1	-	48.5	56.3
WeMath	Strict Score	65.1	**66.3**	37.7*	50.6*	46.1	-	50.6	78.0
LogicVista	Acc.	61.2	**61.4**	53.8*	57.1*	55.9	-	64.4	73.8
MathVista$^{\dagger}_{\text{mini}}$	Acc.	**81.8**	81.5	71.4	74.8	72.2	-	63.8	80.9
Text									
MATH500	Pass@1	95.0	**95.4**	83.8*	83.0	68.8*	82.8*	78.2*	95.2
AIME24	Pass@1	66.4	**67.5**	25.2*	16.7*	12.2*	19.4*	10.9*	92.0
AIME25	Pass@1	50.9	**52.5**	18.1*	10.8*	11.7*	13.3*	8.7*	86.7

Table 3 Comparison of MiMo-VL-7B with other models on reasoning benchmarks. Results marked with * are obtained using our evaluation framework. † indicates that evaluation is performed using GPT-4o. The best results among open-source models are **bolded**.

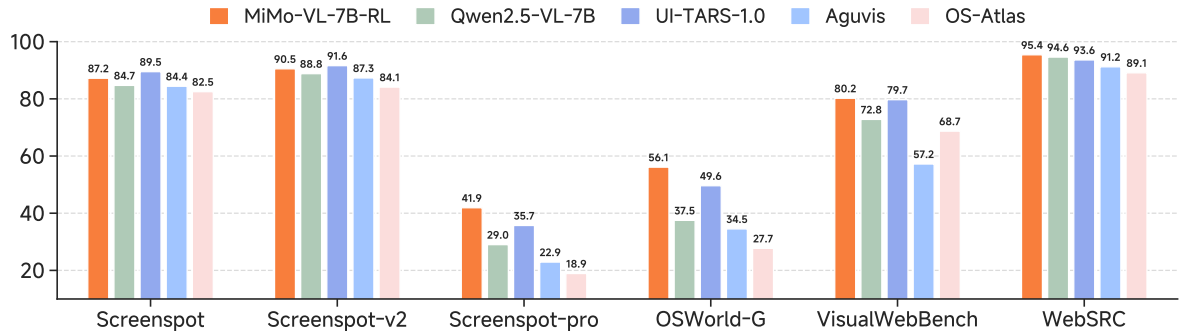


Figure 4 GUI understanding and grounding results. MiMo-VL-7B-RL achieves comparable results with GUI specialized models.

baselines across these benchmarks. Notably, MiMo-VL-7B-SFT surpasses much larger models, including Qwen2.5-VL-72B and QVQ-72B-Preview. The RL model further improves performance on most reasoning benchmarks. For example, MiMo-VL-7B-RL boosts accuracy on MathVision from 57.9% to 60.4%. MiMo-VL-7B models also exhibit impressive reasoning capabilities on pure text benchmarks, even outperforming Qwen2.5-72B. These results demonstrate that our multimodal pre-training and post-training recipes effectively endow the model with exceptional visual capabilities and strong text intelligence.

4.4 GUI Tasks

In addition, we demonstrate that MiMo-VL-7B models possess exceptional GUI understanding and grounding capabilities. In Table 2, MiMo-VL-7B-RL outperforms all other general VLMs compared. In Figure 4, we further compare MiMo-VL-7B-RL with GUI-specialized models (UI-TARS-1.0 (Qin et al., 2025a), Aguvis (Xu et al., 2024), OS-Atlas (Wu et al., 2024)) of similar size on GUI Understanding (WebSrc, VisualWebBench) and Grounding (Screenspot, Screenspot-v2, Screenspot-Pro, OSWorld-G) benchmarks. As a general-purpose VLM, MiMo-VL achieves comparable or even superior performance to GUI-specialized models, particularly on the more challenging Screenspot-Pro and OSWorld-G benchmarks.

Reasoning Benchmark	Metrics	MiMo-VL 7B-SFT	MiMo-VL 7B-RL	QVQ-72B Preview	Qwen2.5-VL 72B	Intern-VL3 78B	Qwen2.5 72B	GPT-4o	Gemini-2.5 Pro
Multi-modal									
OlympiadBench	Acc.	**59.4**	**59.4**	20.4	37.2*	12.3	-	25.9	69.8
MathVision	Acc.	57.9	**60.4**	35.9	38.1	43.2	-	31.2	69.1
MathVerse$^{\dagger}_{vision_only}$	Acc.	67.1	**71.5**	45.1*	57.6	51.0	-	49.9	76.7
DynaMath	Worst-case Acc.	**46.9**	45.9	30.7	38.1*	35.1	-	48.5	56.3
WeMath	Strict Score	65.1	**66.3**	37.7*	50.6*	46.1	-	50.6	78.0
LogicVista	Acc.	61.2	**61.4**	53.8*	57.1*	55.9	-	64.4	73.8
MathVista$^{\dagger}_{mini}$	Acc.	**81.8**	81.5	71.4	74.8	72.2	-	63.8	80.9
Text									
MATH500	Pass@1	95.0	**95.4**	83.8*	83.0	68.8*	82.8*	78.2*	95.2
AIME24	Pass@1	66.4	**67.5**	25.2*	16.7*	12.2*	19.4*	10.9*	92.0
AIME25	Pass@1	50.9	**52.5**	18.1*	10.8*	11.7*	13.3*	8.7*	86.7

表3 在推理基准测试中，MiMo-VL-7B 与其他模型的比较。带有*的结果是使用我们的评估框架获得的。† 表示使用 GPT-4o 进行评估。在开源模型中，最佳结果以加粗显示。

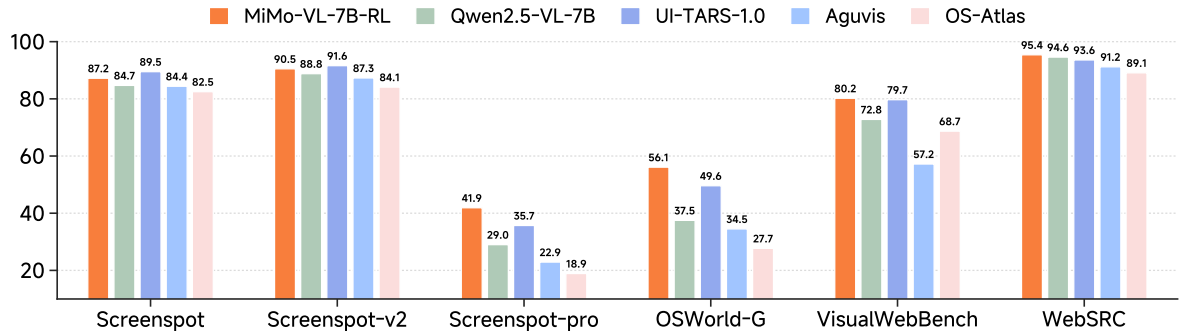


图4 GUI理解与定位结果。MiMo-VL-7B-RL在性能上与专门针对GUI的模型相当。

在这些基准测试中，基线表现均有提升。值得注意的是，MiMo-VL-7B-SFT 超越了许多更大规模的模型，包括 Qwen2.5-VL-72B 和 QVQ-72B-Preview。RL 模型进一步提升了大多数推理基准的性能。例如，MiMo-VL-7B-RL 将 MathVision 的准确率从 57.9% 提升到 60.4%。MiMo-VL-7B 模型在纯文本基准测试中也展现出令人印象深刻的推理能力，甚至优于 Qwen2.5-72B。这些结果表明，我们的多模态预训练和后训练方案有效赋予模型卓越的视觉能力和强大的文本智能。

4.4 图形界面任务

此外，我们证明了MiMo-VL-7B模型具有卓越的GUI理解和定位能力。在表2中，MiMo-VL-7B-RL优于所有其他通用VLM。在图4中，我们进一步比较了MiMo-VL-7B-RL与类似规模的专门针对GUI的模型（UI-TARS-1.0（Qin等，2025a）、Aguvis（Xu等，2024）、OS-Atlas（Wu等，2024））在GUI理解（WebSrc、VisualWebBench）和定位（Screenspot、Screenspot-v2、Screenspot-Pro、OSWorld-G）基准测试中的表现。作为一种通用VLM，MiMo-VL在性能上与专门针对GUI的模型相当，甚至在更具挑战性的Screenspot-Pro和OSWorld-G基准测试中表现出更优的性能。

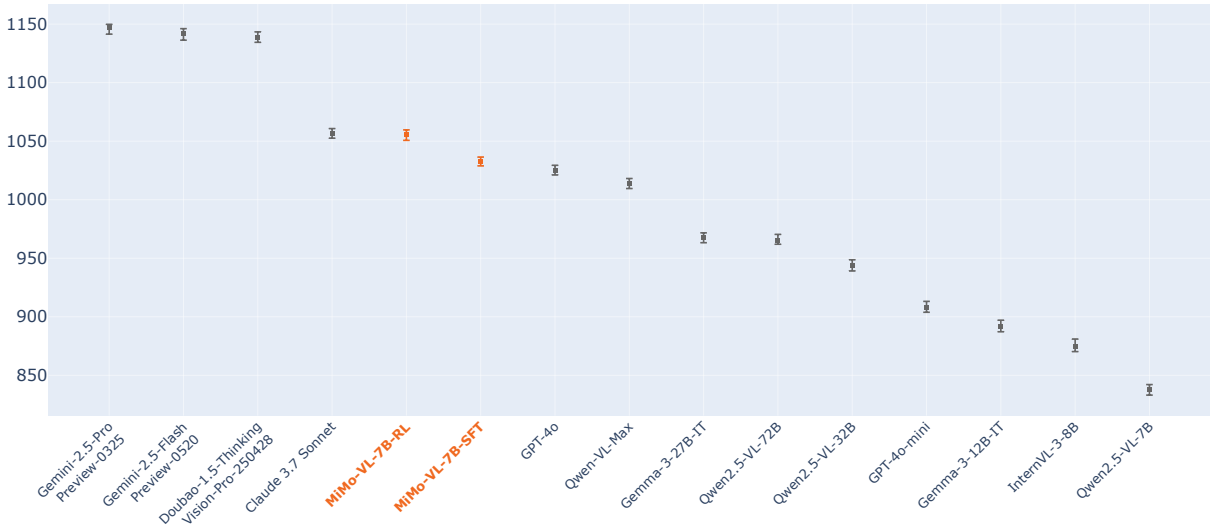


Figure 5 Elo ratings comparison across VLMs. MiMo-VL-7B-RL achieves the highest rating among open-source models, approaching the performance of proprietary alternatives such as Claude 3.7 Sonnet.

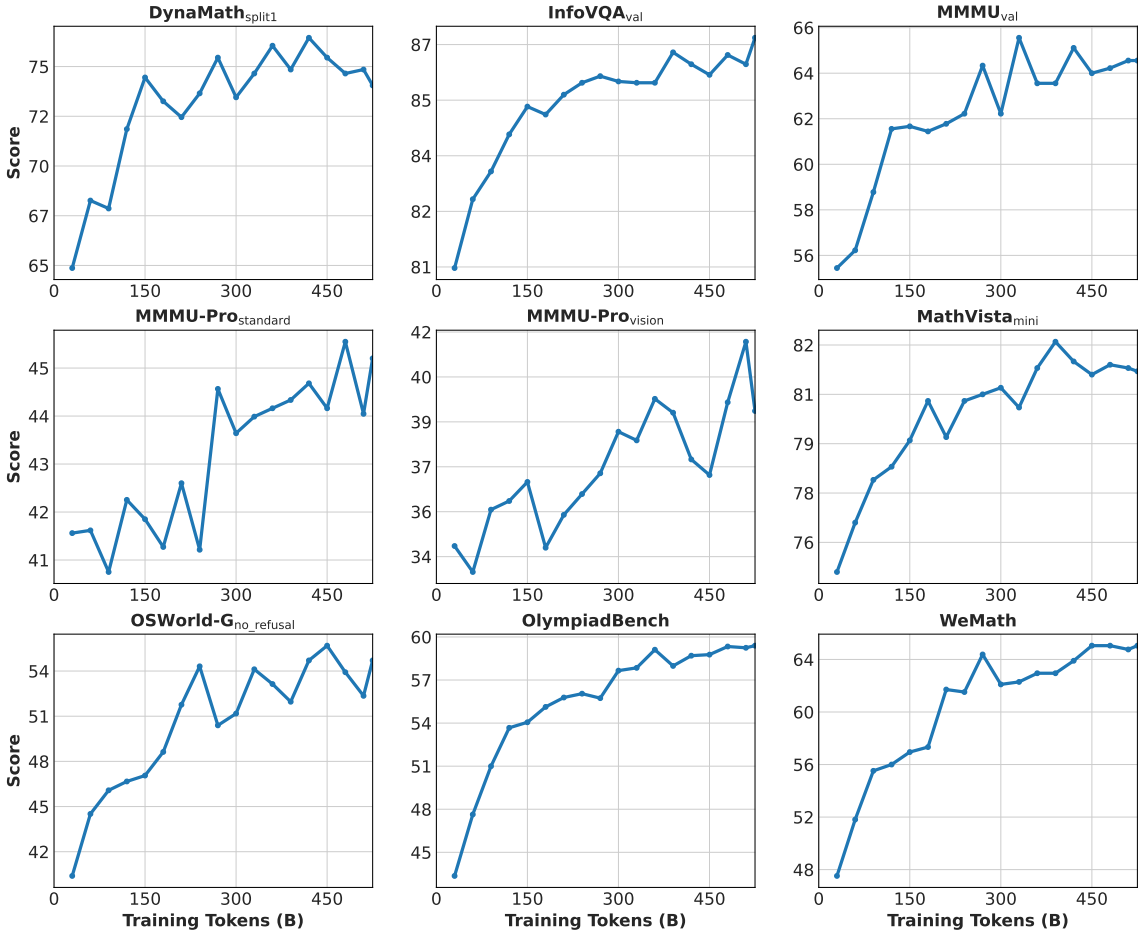


Figure 6 MiMo-VL-7B-SFT training curves in Stage 4.

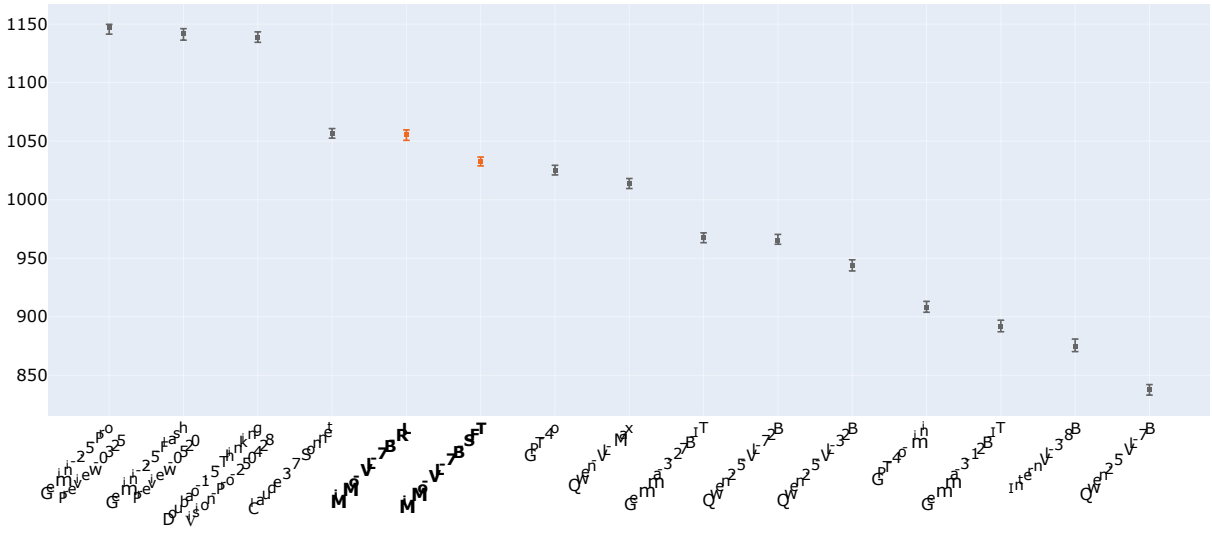


图5 不同VLM的Elo评分比较。MiMo-VL-7B-RL在开源模型中获得最高评分，接近专有模型如Claude 3.7 Sonnet的性能。

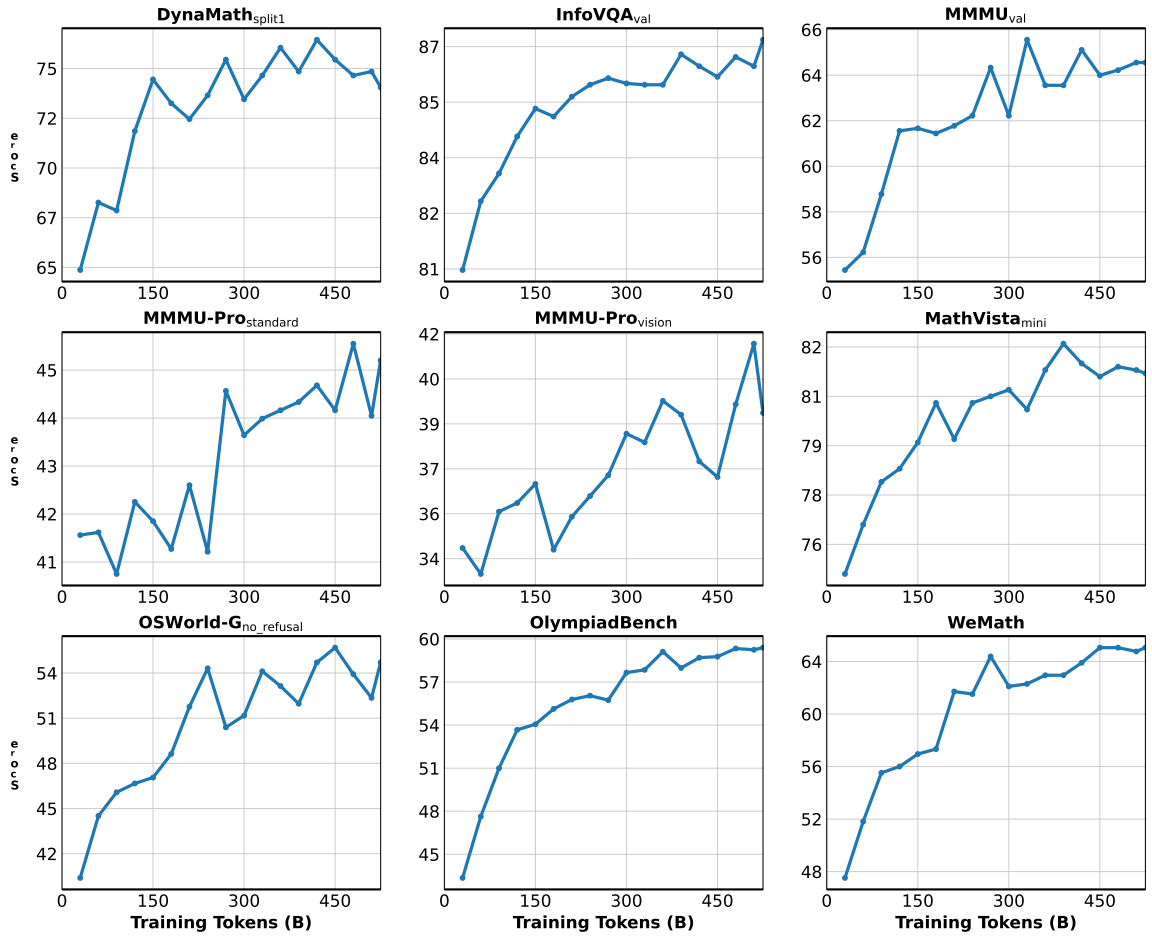


图6 MiMo-VL-7B-SFT在第4阶段的训练曲线。

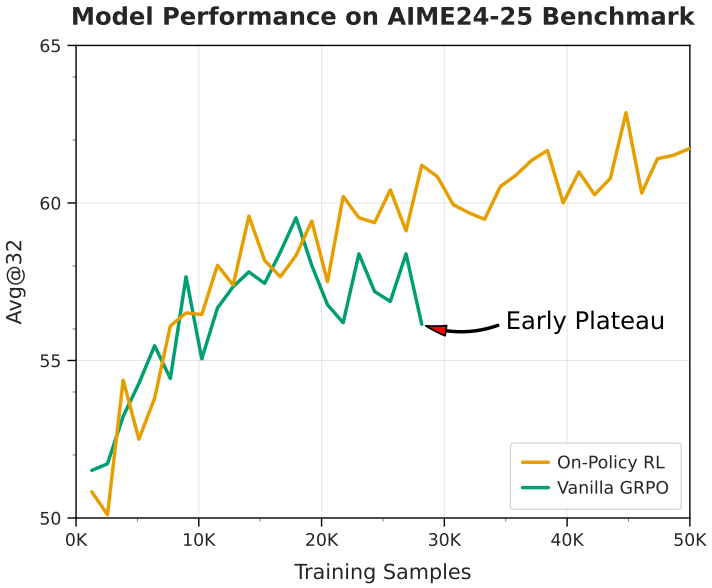


Figure 7 On-policy RL and vanilla GRPO shows contrasting scaling behavior: on-policy RL performance continuously improves with more data, while vanilla GRPO reaches a plateau around 20,000 samples.

4.5 Elo Rating

Inspired by ChatbotArena (Zheng et al., 2023), we construct a balanced bilingual (Chinese and English) in-house evaluation dataset comprising real user prompts. This approach assesses user preference beyond traditional benchmark scores, providing insights into practical model performance in real-world scenarios.

Following the methodology of (Chou et al., 2024), we conduct pairwise comparisons between MiMo-VL-7B and competing models, including leading proprietary models and open-source VLMs ranging from 7B to 72B parameters. We compute Elo ratings based on GPT-4o judgments with style-controlled evaluation protocols. Our evaluation covers a diverse range of visual-linguistic tasks, including multimodal reasoning, image understanding, and GUI interaction scenarios, therefore serving as a good proxy of user preference.

As illustrated in Figure 5, MiMo-VL-7B-RL achieves the highest Elo rating among all evaluated open-source VLMs, ranking first across models spanning from 7B to 72B parameters. This demonstrates superior user experience across the evaluation set, with our model's performance closely approaching that of proprietary models such as Claude 3.7 Sonnet. Moreover, MORL brings a boost of 22+ points for the MiMo-VL-7B-SFT. These results highlight the competitive capability of our models and validate the effectiveness of our training methodology.

5 Discussion

5.1 Boosting Reasoning Capability in Pre-training

Figure 6 shows the performance of MiMo-VL-7B-SFT during Stage 4, its final pre-training phase. In this stage, substantial volumes of synthetic long-form reasoning data are incorporated, and model performance increases sharply, e.g., +9 on MMMU, +14 on OSWorld-G, and +16 on OlympiadBench. Notably, model performance continuously improves without saturation. These

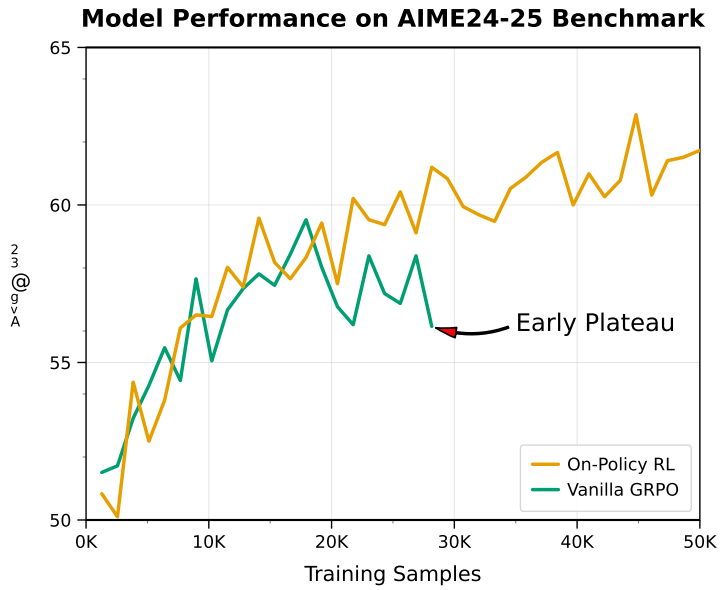


图7：策略内强化学习（On-policy RL）和普通的GRPO显示出不同的扩展行为：策略内RL的性能随着数据量的增加而持续提升，而普通的GRPO在大约20,000个样本时达到一个平台期。

4.5 Elo 评分

受到ChatbotArena（Zheng等，2023）的启发，我们构建了一个平衡的中英双语内部评估数据集，包含真实用户的提示。这种方法超越了传统基准分数，评估用户偏好，为实际场景中的模型性能提供了洞察。

遵循（Chou 等人，2024）的方法，我们对MiMo-VL-7B与竞争模型进行了成对比较，包括领先的专有模型和参数范围从7B到72B的开源VLM。我们基于GPT-4o的判断，采用风格控制的评估协议计算Elo评分。我们的评估涵盖了多样的视觉-语言任务，包括多模态推理、图像理解和GUI交互场景，因此可以作为用户偏好的良好代理。

如图5所示，MiMo-VL-7B-RL在所有评估的开源VLM中实现了最高的Elo评分，在从7B到72B参数的模型中排名第一。这显示了在整个评估集上优越的用户体验，我们的模型性能紧贴专有模型如Claude 3.7 Sonnet。此外，MORL为MiMo-VL-7B-SFT带来了22+点的提升。这些结果突显了我们模型的竞争能力，并验证了我们训练方法的有效性。

5 讨论

5.1 在预训练中提升推理能力

图6显示了MiMo-VL-7B-SFT在第4阶段，即其最终预训练阶段的表现。在此阶段，加入了大量合成的长形式推理数据，模型性能显著提升，例如在MMMU上达到+9，在OSWorld-G上达到+14，在OlympiadBench上达到+16。值得注意的是，模型性能持续提升，没有达到饱和状态。

improvements are attributed to an increased depth in the model's reasoning. For instance, on MMMU, the model's average number of response tokens grows from 680 to 2.5K per question after Stage 4, indicating a more detailed and profound level of reasoning when tackling problems.

5.2 On-Policy RL v.s. Vanilla GRPO

We explore the benefits of on-policy RL v.s. vanilla GRPO with text-only reasoning tasks. As illustrated in Figure 7, the on-policy algorithm demonstrates a consistent positive correlation between training data volume and performance score. Its learning curve shows no signs of saturation within the observed training window, suggesting potential for further enhancement with additional computational resources and data. Conversely, the vanilla GRPO algorithm initially exhibits higher sample efficiency, achieving robust performance early in training. This advantage, however, is transient. The algorithm's performance generally saturates around 20,000 training samples, beyond which further training yields negligible improvements.

5.3 Interference Between RL Tasks

While MORL training enhances performance on nearly all evaluated tasks, achieving stable and simultaneous improvements across diverse task domains remains a significant challenge. During training, we observe reasoning tasks exhibit disparities with visual perception and grounding tasks, such as visual grounding and counting, making it difficult to match the performance of standalone RL on individual tasks.

The potential cause lies in the opposing growth trends of response length: reasoning tasks encourage longer CoT during the RL process, whereas grounding and counting tasks lead to shrinking ones. Disparities in task difficulty and the risk of reward hacking may also contribute to this interference. We are actively investigating the underlying causes of this phenomenon and seeking solutions to achieve consistent and persistent growth across all tasks.

6 Case Study

We present qualitative results in Figure 8 and Appendix E. As depicted in the top example in Figure 8, our model showcases strong plot understanding capabilities, successfully converting an intricate plot into a well-structured markdown table. Additionally, we highlight the models' superior reasoning capabilities in STEM tasks. In the examples shown in Figure 10 and Figure 13, the model effectively addresses multiple STEM questions within a single response. Furthermore, our model exhibits strong agentic capabilities. As illustrated in Figure 9, MiMo-VL-7B successfully navigates a webpage to complete the task of purchasing a Xiaomi SU7 with customized paint and interior options.

7 Conclusions

In this report, we present our efforts in building MiMo-VL-7B models. Leveraging curated high-quality pre-training datasets and our MORL framework, MiMo-VL-7B-SFT and MiMo-VL-7B-RL demonstrate state-of-the-art performance across evaluated benchmarks. We share key observations from our development process: the consistent performance gains from incorporating reasoning data in later pre-training stages, the advantages of on-policy RL over vanilla GRPO, and the challenges of task interference when applying MORL across diverse capabilities. Alongside the released model checkpoints, we open-source our comprehensive evaluation suite to promote transparency

改进归因于模型推理深度的增加。例如，在MMMU上，模型每个问题的平均响应标记数从680增加到2.5K，显示出在解决问题时推理变得更加详细和深刻。

5.2 基于策略的强化学习与普通的GRPO

我们探讨了基于策略的强化学习（on-policy RL）与纯粹的GRPO在仅文本推理任务中的优势。如图7所示，基于策略的算法在训练数据量与性能得分之间表现出持续的正相关关系。其学习曲线在观察到的训练窗口内没有显示出饱和的迹象，表明通过增加计算资源和数据，仍有潜力进一步提升。相反，纯粹的GRPO算法在初期表现出更高的样本效率，训练早期就取得了稳健的性能。然而，这一优势是暂时的。该算法的性能通常在大约20,000个训练样本左右达到饱和，之后的进一步训练几乎没有明显的提升。

5.3 RL任务之间的干扰

虽然MORL训练在几乎所有评估任务上都能提升性能，但在不同任务领域实现稳定且同时的改进仍然是一个重大挑战。在训练过程中，我们观察到推理任务与视觉感知和基础任务（如视觉基础和计数）存在差异，这使得难以在单独任务上达到独立RL的性能。

潜在原因在于响应长度的相反增长趋势：推理任务在强化学习过程中倾向于产生更长的CoT，而基础和计数任务则导致长度缩减。任务难度的差异以及奖励操控的风险也可能加剧这种干扰。我们正在积极研究这一现象的根本原因，并寻求解决方案，以实现所有任务中一致且持续的增长。

6 案例研究

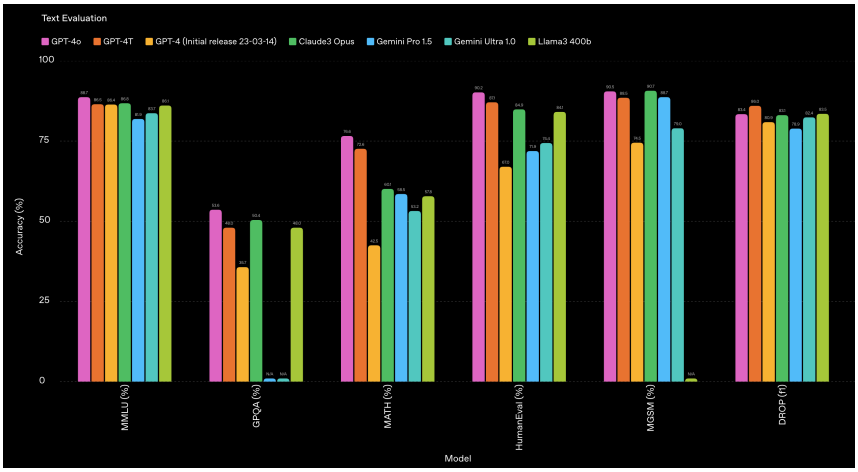
我们在图8和附录E中展示了定性结果。如图8顶部的示例所示，我们的模型展现出强大的情节理解能力，成功将复杂的情节转换为结构良好的markdown表格。此外，我们还强调模型在STEM任务中的优越推理能力。在图10和图13所示的示例中，模型在一次响应中有效地解决了多个STEM问题。此外，我们的模型还表现出强大的代理能力。如图9所示，MiMo-VL-7B成功导航网页，完成购买一台具有定制涂装和内饰选项的小米SU7的任务。

7 结论

在本报告中，我们介绍了构建MiMo-VL-7B模型的努力。利用经过精心筛选的高质量预训练数据集和我们的MORL框架，MiMo-VL-7B-SFT和MiMo-VL-7B-RL在所评估的基准测试中展示了最先进的性能。我们分享了开发过程中的关键观察结果：在后续预训练阶段引入推理数据带来的持续性能提升，基于策略的强化学习（on-policy RL）优于传统的GRPO，以及在应用MORL于多样能力时面临的任务干扰挑战。除了发布的模型检查点外，我们还开源了全面的评估套件，以促进透明度。

and reproducibility in multimodal research. We hope our work advances the development of capable open-source vision-language models and provides valuable insights for the community.

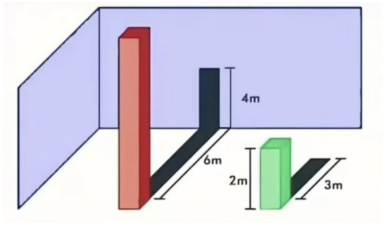
以及多模态研究中的可重复性。我们希望我们的工作推动具有能力的开源视觉-语言模型的发展，并为社区提供有价值的见解。



Instruction: Convert it to a table.

MiMo-VL-7B-RL:

Model	GPT-4.0	GPT-4-T	GPT-4 (Initial release 23-03-14)	Claude3 Opus	Gemini Pro 1.5	Gemini Ultra 1.0	Llama3 400b
MMLU (%)	88.7	86.5	86.4	86.8	81.9	83.7	86.1
GPQA (%)	53.6	48.0	35.7	50.4	N/A	N/A	48.0
MATH (%)	76.6	72.6	42.5	60.1	58.5	53.2	57.8
HumanEval (%)	90.2	87.1	67.0	84.9	71.9	74.4	84.1
MGSM (%)	90.5	88.5	74.5	90.7	88.7	79.0	N/A
DROP (f1)	83.4	86.0	80.9	83.1	78.9	82.4	83.5



Instruction: How tall is the red pillar?

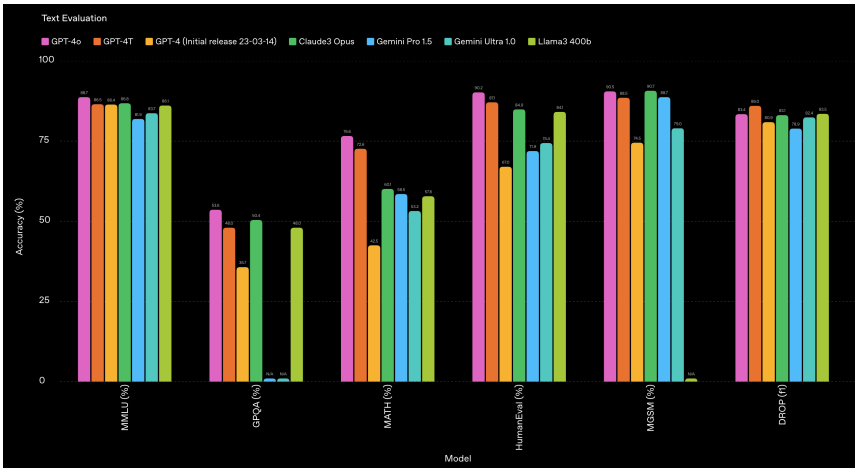
MiMo-VL-7B-RL: `<think>...</think>`

To determine the height of the red pillar, we analyze the **proportionality of distance and height** using the given elements in the image:

1. **Identify key elements**:
   - The dark vertical shape near the black pillar is 4m tall.
   - The green pillar is 2m tall and sits 3m from the dark shape.
   - The red pillar is 6m from the dark shape.
2. **Recognize the proportionality**:
   The distance from the dark shape to the green pillar (3m) and to the red pillar (6m) is a 1:2 ratio. Since the green pillar's height (2m) is half the dark shape's height (4m), this confirms a proportional relationship between distance and height.
3. **Calculate the red pillar's height**:
   If the dark shape is 4m tall (and the green pillar, at half the distance, is half its height: $4 \div 2 = 2m$ the red pillar—being twice as far from the dark shape—must be twice the dark shape's height: $4m \times 2 = 8m$.

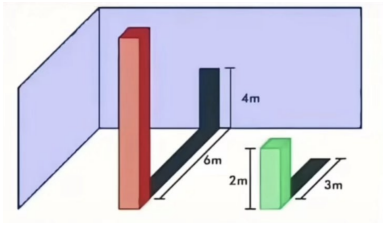
The height of the red pillar is 8 meters.

Figure 8 Examples of MiMo-VL-7B solving complicated problems.



| Instruction | Convert it to a table.
MiMo-VL-7B-RL: |

Model	GPT-4.0	GPT-4-T	GPT-4 (Initial release 23-03-14)	Claude3 Opus	Gemini Pro 1.5	Gemini Ultra 1.0	Llama3 400b
MMLU (%)	88.7	86.5	86.4	86.8	81.9	83.7	86.1
GPQA (%)	53.6	48.0	35.7	50.4	N/A	N/A	48.0
MATH (%)	76.6	72.6	42.5	60.1	58.5	53.2	57.8
HumanEval (%)	90.2	87.1	67.0	84.9	71.9	74.4	84.1
MGSM (%)	90.5	88.5	74.5	90.7	88.7	79.0	N/A
DROP (f1)	83.4	86.0	80.9	83.1	78.9	82.4	83.5



指令：红色柱子有多高?

MiMo-VL-7B-RL: `<think>...</think>`
为了确定红色柱子的高度，我们使用图像中的已给元素分析距离与高度的比例关系：

1. 识别关键元素：
   - 黑色柱子附近的深色垂直形状高4米。 - 绿色柱子高2米，
   距离深色形状3米。 - 红色柱子距离深色形状6米。

2. 识别成比例关系：
   从黑暗形状到绿色柱子（3米）和到红色柱子（6米）的距离比例为1:2。由于绿色柱子的高度（2米）是黑暗形状高度（4米）的一半，这确认了距离与高度之间的成比例关系。

3. 计算红色柱子的高度：
   如果黑暗的形状高4米（而绿色的柱子，距离的一半，只有它一半的高度：$4 \div 2 = 2m$，那么红色的柱子——距离黑暗的形状是它的两倍——必须是黑暗形状高度的两倍：$4m \times 2 = 8m$。红色柱子的高度是8米。

图8 MiMo-VL-7B 解决复杂问题的示例。

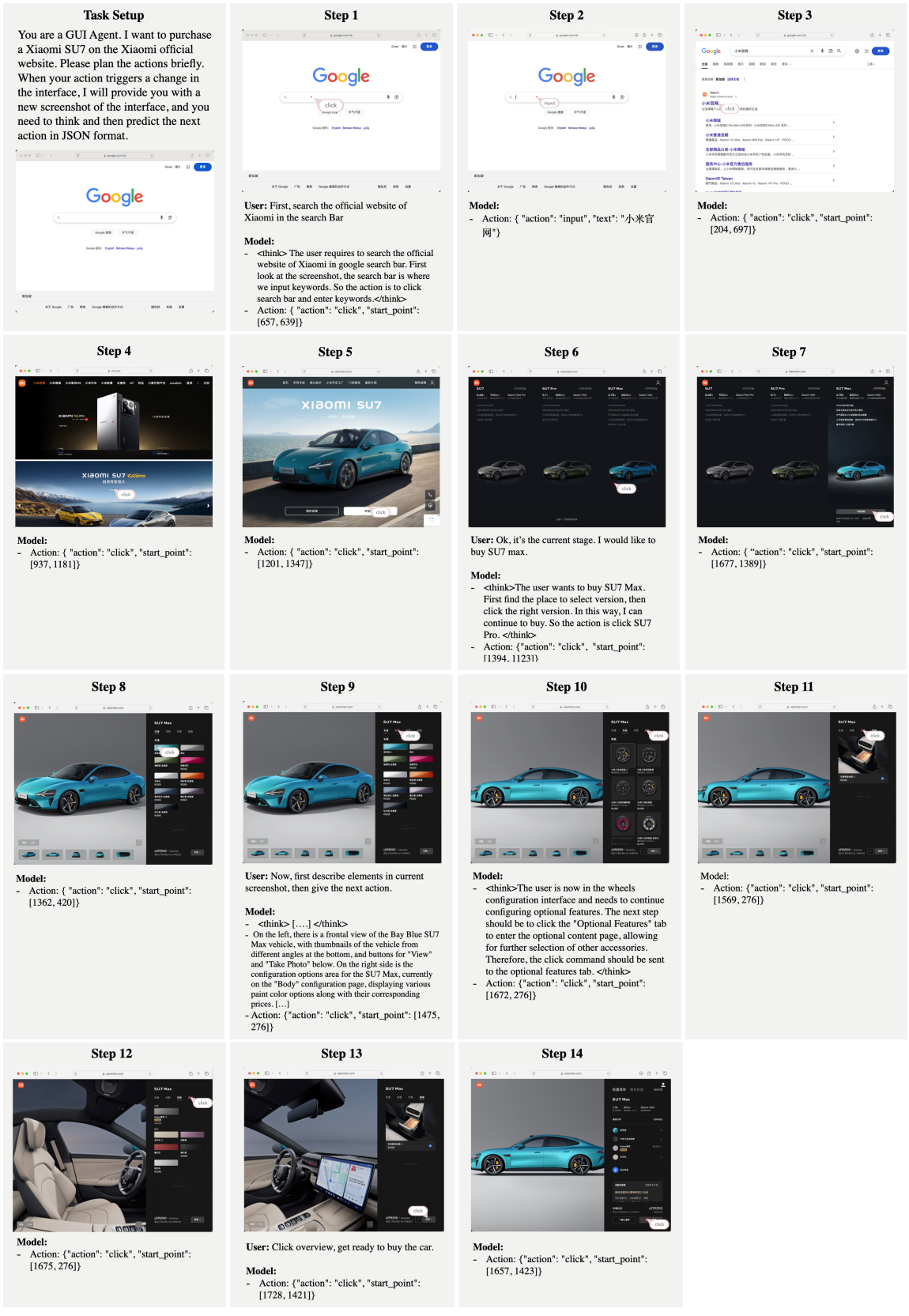


Figure 9 A case demonstrating the agentic capabilities of our model. MiMo-VL-7B successfully navigates a website to purchase Xiaomi SU7 with customized paint and interior options.

Task Setup

You are a GUI Agent. I want to purchase a Xiaomi SU7 on the Xiaomi official website. Please plan the actions briefly. When your action triggers a change in the interface, I will provide you with a new screenshot of the interface, and you need to think and then predict the next action in JSON format.

Step 1

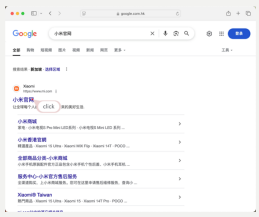
User: First, search the official website of Xiaomi in the search Bar

Model:
- <think> The user requires to search the official website of Xiaomi in google search bar. First look at the screenshot, the search bar is where we input keywords. So the action is to click search bar and enter keywords.</think>
- Action: { "action": "click", "start_point": [657, 639]}

Step 2

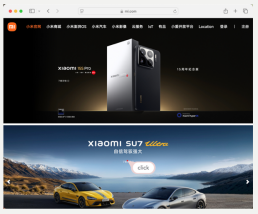
Model:
- Action: { "action": "input", "text": "小米官网"}

Step 3



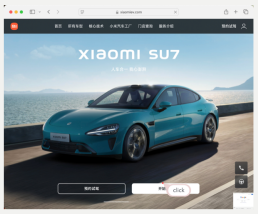
Model:
- Action: { "action": "click", "start_point": [204, 697]}

Step 4



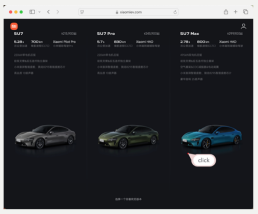
Model:
- Action: { "action": "click", "start_point": [937, 1181]}

Step 5



Model:
- Action: { "action": "click", "start_point": [1201, 1347]}

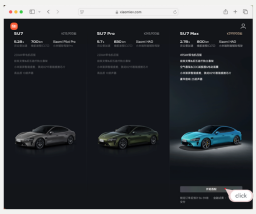
Step 6



User: Ok, it's the current stage. I would like to buy SU7 max.

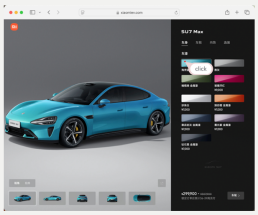
Model:
- <think>The user wants to buy SU7 Max. First find the place to select version, then click the right version. In this way, I can continue to buy. So the action is click SU7 Pro. </think>
- Action: {"action": "click", "start_point": [1394, 1123]}

Step 7



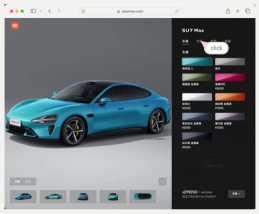
Model:
- Action: { "action": "click", "start_point": [1677, 1389]}

Step 8



Model:
- Action: { "action": "click", "start_point": [1362, 420]}

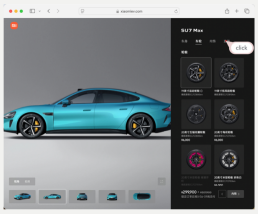
Step 9



User: Now, first describe elements in current screenshot, then give the next action.

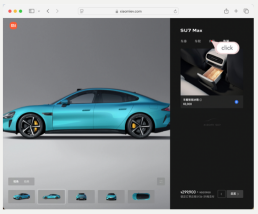
Model:
- <think> [….] </think>
- On the left, there is a frontal view of the Bay Blue SU7 Max vehicle, with thumbnails of the vehicle from different angles at the bottom, and buttons for "View" and "Take Photo" below. On the right side is the configuration options area for the SU7 Max, currently on the "Body" configuration page, displaying various paint color options along with their corresponding prices. […]
- Action: {"action": "click", "start_point": [1475, 276]}

Step 10



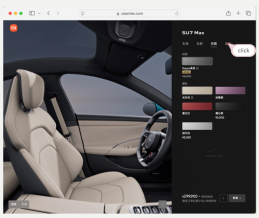
Model:
- <think>The user is now in the wheels configuration interface and needs to continue configuring optional features. The next step should be to click the "Optional Features" tab to enter the optional content page, allowing for further selection of other accessories. Therefore, the click command should be sent to the optional features tab. </think>
- Action: {"action": "click", "start_point": [1672, 276]}

Step 11



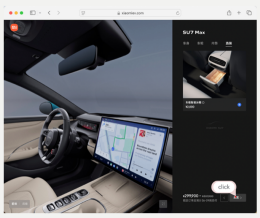
Model:
- Action: {"action": "click", "start_point": [1569, 276]}

Step 12



Model:
- Action: {"action": "click", "start_point": [1675, 276]}

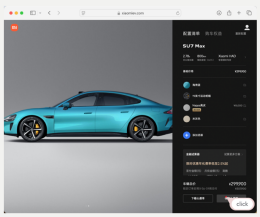
Step 13



User: Click overview, get ready to buy the car.

Model:
- Action: {"action": "click", "start_point": [1728, 1421]}

Step 14



Model:
- Action: {"action": "click", "start_point": [1657, 1423]}

图9 一个展示我们模型主动能力的案例。MiMo-VL-7B 成功浏览网站，购买了带有定制涂装和内饰选项的小米 SU7。

References

J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning. In NeurIPS, 2022.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025a.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025b.

K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. $\pi 0$: A vision-language-action flow model for general robot control. ArXiv, abs/2410.24164, 2024. URL https://api.semanticscholar.org/CorpusID:273811174.

L. Chen, L. Li, H. Zhao, Y. Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V, 2025a. Accessed: 2025-02-02.

X. Chen, Z. Zhao, L. Chen, D. Zhang, J. Ji, A. Luo, Y. Xiong, and K. Yu. Websrc: a dataset for web-based structural reading comprehension. arXiv preprint arXiv:2101.09465, 2021.

Y. Chen, Z. Yang, Z. Liu, C. Lee, P. Xu, M. Shoeybi, B. Catanzaro, and W. Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. arXiv preprint arXiv:2505.16400, 2025b.

K. Cheng, Q. Sun, Y. Chu, F. Xu, Y. Li, J. Zhang, and Z. Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. arXiv preprint arXiv:2401.10935, 2024.

C. Chou, L. Dunlap, K. Mashita, K. Mandal, T. Darrell, I. Stoica, J. Gonzalez, and W.-L. Chiang. Visionarena: 230k real world user-vlm conversations with preference labels. ArXiv, abs/2412.08687, 2024. URL https://api.semanticscholar.org/CorpusID:274655992.

W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping. Nvlm: Open frontier-class multimodal llms. arXiv preprint, 2024.

M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. ArXiv preprint, abs/2409.17146, 2024.

X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. ArXiv preprint, abs/2502.14739, 2025. URL https://arxiv.org/abs/2502.14739.

D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In J. Burstein, C. Doran, and T. Solorio,

参考文献

J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, 和 K. Simonyan。Flamingo：一种用于少样本学习的视觉语言模型。发表于 NeurIPS，2022。

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu 和 J. Lin。Qwen2.5-vl 技术报告。arXiv 预印本 arXiv:2502.13923, 2025a。

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang 等. Qwen2. 5-vl 技术报告. arXiv 预印本 arXiv:2502.13923, 2025b.

K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, 和 U. Zhilinsky。π 0：一种用于通用机器人控制的视觉-语言-动作流程模型。ArXiv，abs/2410.24164，2024。网址 https://api.semanticscholar.org/CorpusID:273811174。

L. Chen, L. Li, H. Zhao, Y. Song 和 Vinci. R1-v：在视觉-语言模型中用少于 {v*} 3 的方法增强超泛化能力，2025a。访问日期：2025-02-02。

X. Chen, Z. Zhao, L. Chen, D. Zhang, J. Ji, A. Luo, Y. Xiong, 和 K. Yu. Websrc：一个用于基于网页结构的阅读理解的数据集。arXiv 预印本 arXiv:2101.09465, 2021。

Y. Chen, Z. Yang, Z. Liu, C. Lee, P. Xu, M. Shoeybi, B. Catanzaro, 和 W. Ping. Acereason- nemotron：通过强化学习推进数学和代码推理。arXiv 预印本 arXiv:2505.16400, 2025b。

K. Cheng, Q. Sun, Y. Chu, F. Xu, Y. Li, J. Zhang, 和 Z. Wu. Seeclick：利用图形界面基础实现先进的视觉图形界面代理。arXiv 预印本 arXiv:2401.10935, 2024。

C. Chou, L. Dunlap, K. Mashita, K. Mandal, T. Darrell, I. Stoica, J. Gonzalez 和 W.-L. Chi- ang. Vision arena：23万真实世界用户-VLM对话及偏好标签。ArXiv，abs/2412.08687，2024。网址 https://api.semanticscholar.org/CorpusID：274655992。

W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, 和 W. Ping. Nvlm：开放前沿级多模态大模型。arXiv 预印本，2024。

M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini 等. Molmo 和 pixmo：用于最先进多模态模型的开源权重和开源数据。ArXiv 预印本，abs/2409.17146，2024。

X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, 等. Supergpqa：在285个研究生学科中扩展大模型评估。ArXiv预印本，abs/2502.14739，2025。网址 https://arxiv.org/abs/2502.14739。

D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh 和 M. Gardner. DROP：一个需要对段落进行离散推理的阅读理解基准。在 J. Burstein、C. Doran 和 T. Solorio,

editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2368–2378, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246.

C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024a.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pages 148–166. Springer, 2024b.

J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In Proceedings of the IEEE international conference on computer vision, pages 5267–5275, 2017.

C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. arXiv preprint arXiv:2402.14008, 2024.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. ArXiv preprint, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, and Z. Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826, 2025.

D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen. Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483, 2024.

S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In ICML, 2024.

S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 787–798, 2014.

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pages 235–251. Springer, 2016.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, and H. Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

B. Li, Y. Ge, Y. Chen, Y. Ge, R. Zhang, and Y. Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. arXiv preprint arXiv:2404.16790, 2024a.

K. Li, Z. Meng, H. Lin, Z. Luo, Y. Tian, J. Ma, Z. Huang, and T.-S. Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. arXiv preprint arXiv:2504.07981, 2025.

编辑，2019年北美计算语言学会人类语言技术会议论文集：第1卷（长篇和短篇论文），第2368–2378页，明尼阿波利斯，明尼苏达州，2019年。计算语言学协会。doi：10.18653/v1/N19-1246。网址 https://aclanthology.org/N19-1246。

C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, 等. Video-mme：首个多模态大模型在视频分析中的全面评估基准. arXiv 预印本 arXiv:2405.21075, 2024a.

X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, 和 R. Krishna. Blink：多模态大型语言模型可以“看见”但不能“感知”。发表于欧洲计算机视觉会议，页码 148–166。Springer，2024b。

J. Gao, C. Sun, Z. Yang, 和 R. Nevatia. Tall：通过语言查询进行时间活动定位。收录于IEEE国际计算机视觉会议论文集，第5267–5275页，2017年。

C. He, R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang 等。Olympiadbench：一个具有挑战性的基准，用于通过奥林匹克级别的双语多模态科学问题促进AGI。arXiv预印本 arXiv:2402.14008，2024。

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, 和 J. Steinhardt. 使用数学数据集衡量数学问题解决能力。ArXiv预印本，abs/2103.03874，2021年。网址 https://arxiv.org/abs/2103.03874。

K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, 和 Z. Liu. Video-mmmu：评估来自多学科专业视频的知识获取。arXiv预印本 arXiv:2501.13826, 2025。

D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, 和 W. Chen. Mantis：交错多图像指令调优。arXiv预印本 arXiv:2405.01483, 2024。

S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar 和 D. Sadigh. 棱柱形视觉条件语言模型：探索视觉条件语言模型的设计空间。发表于 ICML，2024。

S. Kazemzadeh, V. Ordonez, M. Matten, 和 T. Berg. Referitgame：指涉自然场景照片中的对象。收录于2014年自然语言处理经验方法会议（EMNLP）论文集，第787–798页，2014年。

A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi 和 A. Farhadi. 一张图表胜过十几张图片。在《计算机视觉–ECCV 2016：第十四届欧洲会议》，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第IV部分，第14卷，页码235–251。 Springer，2016。

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, 和 H. Hajishirzi。Tulu 3：推动开放语言模型后训练的前沿，2025。网址 https://arxiv.org/abs/2411.15124。

B. Li, Y. Ge, Y. Chen, Y. Ge, R. Zhang, 和 Y. Shan. Seed-bench-2-plus：基于丰富文本的视觉理解对多模态大型语言模型的基准测试。arXiv 预印本 arXiv:2404.16790, 2024a。

K. Li, Z. Meng, H. Lin, Z. Luo, Y. Tian, J. Ma, Z. Huang, 和 T.-S. Chua. Screenspot-pro：面向专业高分辨率计算机使用的GUI基础。arXiv预印本 arXiv:2504.07981, 2025。

L. Li, Y. Wei, Z. Xie, X. Yang, Y. Song, P. Wang, C. An, T. Liu, S. Li, B. Y. Lin, L. Kong, and Q. Liu. Vlrewardbench: A challenging benchmark for vision-language generative reward models. ArXiv, abs/2411.17451, 2024b. URL https://api.semanticscholar.org/CorpusID:274281459.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In NeurIPS, 2023.

J. Liu, Y. Song, B. Y. Lin, W. Lam, G. Neubig, Y. Li, and X. Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? arXiv preprint arXiv:2404.05955, 2024a.

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European conference on computer vision, pages 216–233. Springer, 2024b.

Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, and X. Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 67 (12):220102, 2024c.

P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.

MAA. American invitational mathematics examination - aime. In American Invitational Mathematics Examination - AIME, 2024. URL https://maa.org/math-competitions/american-invitational-mathematics-examination-aime.

MAA. American invitational mathematics examination - aime. In American Invitational Mathematics Examination - AIME, 2025. URL https://maa.org/math-competitions/american-invitational-mathematics-examination-aime.

K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems, 36: 46212–46244, 2023.

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

M. Mathew, D. Karatzas, and C. Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021.

M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022.

V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pages 792–807. Springer, 2016.

OpenAI. Computer-using agent: Introducing a universal interface for ai to interact with the digital world. 2025. URL https://openai.com/index/computer-using-agent.

L. Li, Y. Wei, Z. Xie, X. Yang, Y. Song, P. Wang, C. An, T. Liu, S. Li, B. Y. Lin, L. Kong, 和 Q. Liu. Vlrewardbench：一个具有挑战性的视觉-语言生成奖励模型基准。ArXiv，abs/2411.17451，2024b。网址 https://api.semanticscholar.org/CorpusID:274281459。

H. Liu, C. Li, Q. Wu, 和 Y. J. Lee. 视觉指令调优。发表于 NeurIPS，2023。

J. Liu, Y. Song, B. Y. Lin, W. Lam, G. Neubig, Y. Li, 和 X. Yue. Visualwebbench：多模态大模型在网页理解与定位方面的发展到何种程度？arXiv 预印本 arXiv:2404.05955, 2024a。

Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu 等. Mmbench：你的多模态模型是全能选手吗？发表于欧洲计算机视觉会议，页码 216–233。Springer，2024b。

Y. Liu, Z. Li, M. Huang, B. Yang, W. Yu, C. Li, X.-C. Yin, C.-L. Liu, L. Jin, 和 X. Bai. OCRBench：关于大型多模态模型中 OCR 的隐藏奥秘。中国科学：信息科学，67 (12):220102，2024c。

P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, 和 J. Gao. Mathvista：在视觉环境中评估基础模型的数学推理能力。arXiv预印本 arXiv:2310.02255, 2023。

MAA。美国邀请数学考试 - AIME。在2024年的美国邀请数学考试 - AIME中。网址 https://maa.org/math-competitions/american-invitational-mathematics-examination-aime。

MAA。美国邀请数学考试 - AIME。在2025年的美国邀请数学考试 - AIME中。网址 https://maa.org/math-competitions/american-invitational-mathematics-examination-aime。

K. Mangalam、R. Akshulakov 和 J. Malik。Egoschema：一种用于超长视频语言理解的诊断基准。神经信息处理系统进展，36：46212–46244，2023年。

A. Masry, D. X. Long, J. Q. Tan, S. Joty, 和 E. Hoque. Chartqa：一个关于图表的问答基准，结合了视觉和逻辑推理。arXiv 预印本 arXiv:2203.10244, 2022。

M. Mathew、D. Karatzas 和 C. Jawahar。Docvqa：用于文档图像的VQA数据集。发表于IEEE/CVF冬季计算机视觉应用会议论文集，页码2200–2209，2021年。

M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, 和 C. Jawahar. 信息图VQA。在IEEE/CVF计算机视觉应用冬季会议论文集，页码1697–1706，2022年。

V. K. Nagaraja、V. I. Morariu 和 L. S. Davis。用于指称表达理解的对象之间上下文建模。收录于《计算机视觉–ECCV 2016：第十四届欧洲会议》，2016年10月11日至14日，荷兰阿姆斯特丹，第十四部分，论文集，第792–807页。Springer，2016。

OpenAI。使用计算机的代理：引入一个用于人工智能与数字世界交互的通用接口。2025年。网址 https://openai.com/index/computer-using-agent。

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In NeurIPS, 2022.

P. Padlewski, M. Bain, M. Henderson, Z. Zhu, N. Relan, H. Pham, D. Ong, K. Aleksiev, A. Ormazabal, S. Phua, et al. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. ArXiv preprint, abs/2405.02287, 2024.

R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel. Teaching clip to count to ten. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3170–3180, 2023.

R. Qiao, Q. Tan, G. Dong, M. Wu, C. Sun, X. Song, Z. GongQue, S. Lei, Z. Wei, M. Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? arXiv preprint arXiv:2407.01284, 2024.

Y. Qin, Y. Ye, J. Fang, H. Wang, S. Liang, S. Tian, J. Zhang, J. Li, Y. Li, S. Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025a.

Y. Qin, Y. Ye, J. Fang, H. Wang, S. Liang, S. Tian, J. Zhang, J. Li, Y. Li, S. Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326, 2025b.

P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. URL https://arxiv.org/abs/2407.06581.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In First Conference on Language Modeling, 2024.

H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 658–666, 2019.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. Hybridflow: A flexible and efficient rlhf framework. ArXiv preprint, abs/2409.19256, 2024. URL https://arxiv.org/abs/2409.19256.

C. Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024.

P. Tong, E. Brown, P. Wu, S. Woo, A. J. V. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37:87310–87356, 2024.

K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169, 2024a.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, 和 R. Lowe. 使用人类反馈训练语言模型以遵循指令。发表于 NeurIPS，2022。

P. Padlewski, M. Bain, M. Henderson, Z. Zhu, N. Relan, H. Pham, D. Ong, K. Aleksiev, A. Ormazabal, S. Phua 等. Vibe-eval：一种用于衡量多模态语言模型进展的硬评估套件。ArXiv 预印本，abs/2405.02287，2024年。R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani 和 T. Dekel. 教学剪辑数到 {v*}。发表于 IEEE/CVF 国际计算机视觉会议论文集，第 3170–3180 页，2023年。

R. Qiao, Q. Tan, G. Dong, M. Wu, C. Sun, X. Song, Z. GongQue, S. Lei, Z. Wei, M. Zhang, 等. We-math：你的大型多模态模型是否实现了类人数学推理？arXiv预印本 arXiv:2407.01284, 2024。

Y. Qin, Y. Ye, J. Fang, H. Wang, S. Liang, S. Tian, J. Zhang, J. Li, Y. Li, S. Huang, 等. Ui-tars：开创性地使用本地代理实现自动化GUI交互。arXiv预印本 arXiv:2501.12326, 2025a.

Y. Qin, Y. Ye, J. Fang, H. Wang, S. Liang, S. Tian, J. Zhang, J. Li, Y. Li, S. Huang, 等. Ui-tars：开创性地使用本地代理实现自动化GUI交互。arXiv预印本 arXiv:2501.12326, 2025b.

P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, 和 A. T. Nguyen. 视觉语言模型是盲目的：未能将详细的视觉特征转化为文字，2025。网址 https://arxiv.org/abs/2407.06581。D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, 和 S. R. Bowman. Gpqa：一个研究生级别的谷歌级别的问答基准测试。发表于 2024 年第一届语言建模会议。

H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid 和 S. Savarese. 广义交并比：一种用于边界框回归的度量和损失。在 IEEE/CVF 计算机视觉与模式识别会议论文集，页码 658–666，2019。

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu 等. Deepseek- 数学：推动开放语言模型中数学推理的极限。arXiv 预印本 arXiv:2402.03300, 2024。G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin 和 C. Wu。Hybridflow：一个灵活高效的 RLHF 框架。ArXiv 预印本，abs/2409.19256, 2024。网址 https://arxiv.org/abs/2409.19256。

C. 团队。变色龙：混合模态早期融合基础模型。arXiv预印本 arXiv:2405.09818, 2024。

P. Tong, E. Brown, P. Wu, S. Woo, A. J. V. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, 等. Cambrian-1：一种完全开放、以视觉为中心的多模态大模型探索。神经信息处理系统进展，37：87310–87356，2024。

K. Wang, J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, 和 H. Li. 使用 math-vision 数据集衡量多模态数学推理。神经信息处理系统进展，37:95095–95169, 2024a.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

Y. Wang, B. Xu, Z. Yue, Z. Xiao, Z. Wang, L. Zhang, D. Yang, W. Wang, and Q. Jin. Timezero: Temporal video grounding with reasoning-guided lvlm. arXiv preprint arXiv:2503.13377, 2025.

Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems, 37:113569–113697, 2024c.

P. Wu and S. Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Z. Wu, Z. Wu, F. Xu, Y. Wang, Q. Sun, C. Jia, K. Cheng, Z. Ding, L. Chen, P. P. Liang, et al. Os-atlas: A foundation action model for generalist gui agents. arXiv preprint arXiv:2410.23218, 2024.

Y. Xiao, E. Sun, T. Liu, and W. Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. arXiv preprint arXiv:2407.04973, 2024.

L.-C.-T. Xiaomi. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. arXiv preprint arXiv:2505.07608, 2025.

T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems, 37:52040–52094, 2024.

T. Xie, J. Deng, X. Li, J. Yang, H. Wu, J. Chen, W. Hu, X. Wang, Y. Xu, Z. Wang, Y. Xu, J. Wang, D. Sahoo, T. Yu, and C. Xiong. Scaling computer-use grounding via user interface decomposition and synthesis, 2025. URL https://arxiv.org/abs/2505.13227.

H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying clip data. arXiv preprint arXiv:2309.16671, 2023.

Y. Xu, Z. Wang, J. Wang, D. Lu, T. Xie, A. Saha, D. Sahoo, T. Yu, and C. Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. arXiv preprint arXiv:2412.04454, 2024.

J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, and L. Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream.

L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 69–85. Springer, 2016.

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567, 2024a.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, 和 W. Chen. Mmlu-pro：一个更稳健、更具挑战性的多任务语言理解基准。在 A. Globersons、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. M. Tomczak 和 C. Zhang 编辑的《神经信息处理系统进展 38：2024 年神经信息处理系统年度会议》，NeurIPS 2024，加拿大不列颠哥伦比亚省温哥华，2024 年 12 月 10-15 日，2024b。网址 http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html。Y. Wang, B. Xu, Z. Yue, Z. Xiao, Z. Wang, L. Zhang, D. Yang, W. Wang 和 Q. Jin。Timezero：利用推理引导的 lvlm 进行时序视频定位。arXiv 预印本 arXiv:2503.13377，2025 年。

Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, 等. Charxiv: 在多模态大模型中绘制现实图表理解的差距。神经信息处理系统进展, 37:113569–113697, 2024c.

P. Wu 和 S. Xie. V*：作为多模态大模型核心机制的引导视觉搜索，2023。网址 https://arxiv.org/abs/2312.14135。

Z. Wu, Z. Wu, F. Xu, Y. Wang, Q. Sun, C. Jia, K. Cheng, Z. Ding, L. Chen, P. P. Liang, 等. Os-atlas：一种面向通用GUI代理的基础动作模型. arXiv预印本 arXiv:2410.23218, 2024.

Y. Xiao, E. Sun, T. Liu, 和 W. Wang. Logicvista：多模态大模型在视觉环境中的逻辑推理基准。arXiv 预印本 arXiv:2407.04973, 2024。

L.-C.-T. 小米。Mimo：释放语言模型的推理潜力——从预训练到后训练。arXiv预印本 arXiv:2505.07608，2025。

T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei 等. Osworld：在真实计算机环境中对多模态智能体进行开放式任务基准测试。神经信息处理系统进展，37：52040–52094，2024。

T. Xie, J. Deng, X. Li, J. Yang, H. Wu, J. Chen, W. Hu, X. Wang, Y. Xu, Z. Wang, Y. Xu, J. Wang, D. Sahoo, T. Yu 和 C. Xiong. 通过用户界面分解与合成实现计算机使用基础的扩展，2025。网址 https://arxiv.org/abs/2505.13227。

H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, 和 C. Feichtenhofer. 解密 clip 数据。arXiv 预印本 arXiv:2309.16671, 2023。

Y. Xu, Z. Wang, J. Wang, D. Lu, T. Xie, A. Saha, D. Sahoo, T. Yu, 和 C. Xiong. Aguvis：用于自主GUI交互的统一纯视觉代理。arXiv预印本 arXiv:2412.04454, 2024。

J. Ye, Z. Xie, L. Zheng, J. Gao, Z. Wu, X. Jiang, Z. Li, 和 L. Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream.

L. Yu, P. Poirson, S. Yang, A. C. Berg, 和 T. L. Berg. 在指称表达中建模上下文。收录于《计算机视觉–ECCV 2016：第十四届欧洲会议》，荷兰阿姆斯特丹，2016年10月11-14日，第二部分论文集，第14卷，第69–85页。Springer，2016。

X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, 等. Mmmu：一个面向专家级AGI的大规模多学科多模态理解与推理基准。在《IEEE/CVF计算机视觉与模式识别会议论文集》，第9556–9567页，2024a。

X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In CVPR, pages 9556–9567, 2024b.

X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2409.02813, 2024c.

K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li, and Z. Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772.

R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? ArXiv preprint, abs/2403.14624, 2024b.

Y.-F. Zhang, H. Zhang, H. Tian, C. Fu, S. Zhang, J. Wu, F. Li, K. Wang, Q. Wen, Z. Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257, 2024c.

L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In NeurIPS, 2023.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Conference on Robot Learning, pages 2165–2183. PMLR, 2023.

C. Zou, X. Guo, R. Yang, J. Zhang, B. Hu, and H. Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. arXiv preprint arXiv:2411.00836, 2024.

X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, 和 W. Chen. MMMU：一个面向专家AGI的大规模多学科多模态理解与推理基准。在CVPR会议上，第9556–9567页，2024b。

X. Yue, T. Zheng, Y. Ni, Y. Wang, K. Zhang, S. Tong, Y. Sun, B. Yu, G. Zhang, H. Sun 等. Mmmu-pro：一个更稳健的多学科多模态理解基准。arXiv 预印本 arXiv:2409.02813, 2024c。K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li 和 Z. Liu。Lmms-eval：对大型多模态模型评估的现实检验，2024a。网址 https://arxiv.org/abs/2407.12772。

R. Zhang, D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, P. Gao 等人。Mathverse：你的多模态大模型真的能理解视觉数学题中的图示吗？ArXiv预印本，abs/2403.14624, 2024b。

Y.-F. Zhang, H. Zhang, H. Tian, C. Fu, S. Zhang, J. Wu, F. Li, K. Wang, Q. Wen, Z. Zhang 等. Mme-realworld：你的多模态大模型能否挑战人类难以应对的高分辨率真实场景？arXiv 预印本 arXiv:2408.13257, 2024c。

L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, 和 I. Stoica. 使用 mt-bench 和 chatbot arena 评判 llm 作为裁判。发表于 NeurIPS，2023 年。J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, 和 L. Hou. 大型语言模型的指令遵循评估，2023 年。网址 https://arxiv.org/abs/2311.07911。B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid 等。Rt-2：视觉-语言-行动模型将网页知识转移到机器人控制中。发表于机器人学习会议，页码 2165–2183。PMLR，2023 年。C. Zou, X. Guo, R. Yang, J. Zhang, B. Hu 和 H. Zhang。Dynamath：一个用于评估视觉-语言模型数学推理鲁棒性的动态视觉基准。arXiv preprint arXiv:2411.00836，2024 年。

A Contributions and Acknowledgments

We would like to express our sincere gratitude to all contributors for their invaluable support and efforts, including the Xiaomi LLM-Plus, Mify, MiChat and CloudML teams, as well as those not explicitly listed in this paper. Authors within each role are listed in ***reverse order*** by their first names.

Core Contributors
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

Contributors
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
Bingquan Xia

A 贡献与致谢

我们衷心感谢所有贡献者的宝贵支持和努力，包括小米LLM-Plus、Mify、MiChat和CloudML团队，以及本文未明确列出的其他人员。每个角色中的作者按名字在***reverse order***中列出。

核心贡献者：岳子豪、林振儒、宋一凡、王伟坤、任书怀、谷书豪、李世成、李培点、赵亮、李雷、包开南、田浩、张刚、王大伟、朱慈慈、何宏红、叶博文、沈博文

贡献者
张子涵 江子涵
郑志贤 宋志超
罗振博 于悦 于东 王宇 袁媛
田宇 屠奕涵 严毅 黄旭 王新哲
许兴辰 宋星 张星 永
新 张 祥伟 邓
文宇 杨 文汉 马
伟伟 吕 伟基 庄
伟 刘 思锐 邓
烁 刘 世茂 陈
世华 于 少辉 刘
山德 王 瑞 马
千通 王 鹏 王
诺 陈 梦航 朱
康阳 周 康 周
凯 方 军 石 金
浩 董 杰宝 肖
嘉明 徐 华秋 刘
宏申 徐 恒 屈
豪辰 赵 航龙 吕
国安 王 多 张
东 张 迪 张 崇
马 昌 刘 灿 蔡
兵泉 夏

B Model Configuration of MiMo-VL-7B

In Table 4, the architecture and configuration of MiMo-VL-7B are detailed.

	Vision Encoder	Language Model
# Layers	32	36
# Heads	16	32
Hidden Size	1280	4096
Intermediate Size	3456	11008
Position Embedding	2D RoPE	MRoPE (Bai et al., 2025b)
Patch Size	14	-

Table 4 Configuration of MiMo-VL-7B.

C Evaluation Benchmarks

We evaluate our models across 50 diverse tasks, including:

General Visual Understanding AI2D (Kembhavi et al., 2016), BLINK (Fu et al., 2024b). CV-Bench (Tong et al., 2024), MMMU (Yue et al., 2024a), MMMU-Pro (Standard and Vision) (Yue et al., 2024c), Mantis (Jiang et al., 2024), MME-RealWorld (English and Chinese) (Zhang et al., 2024c), MMBench (English and Chinese) (Liu et al., 2024b), VibeEval (Padlewski et al., 2024), VL-RewardBench (Li et al., 2024b), V^* (Wu and Xie, 2023), and VLMs are Blind (Rahmanzadehgervi et al., 2025).

General Grounding and Counting RefCOCO (Yu et al., 2016; Nagaraja et al., 2016; Kazemzadeh et al., 2014), CountBench (Paiss et al., 2023), and PixmoCount (Deitke et al., 2024)

Document and Chart Understanding ChartQA (Masry et al., 2022), InfographicVQA (Mathew et al., 2022), DocVQA (Mathew et al., 2021), OCRBench (Liu et al., 2024c), SEED-Bench-2-Plus (Li et al., 2024a), and CharXiv (RQ/DQ) (Wang et al., 2024c).

Video Understanding and Localization Video-MME (Fu et al., 2024a), Video-MMMU (Hu et al., 2025), EgoSchema (Mangalam et al., 2023), and Charades-STA Gao et al. (2017).

GUI Understanding and Grounding WebSrc (Chen et al., 2021), VisualWebBench (Liu et al., 2024a), ScreenSpot (Cheng et al., 2024), ScreenSpot-V2 (Wu et al., 2024), ScreenSpot-Pro (Li et al., 2025), and OSWorld-G (Xie et al., 2025).

Text-only Benchmarks GPQA (Rein et al., 2024), SuperGPQA (Du et al., 2025), DROP (Dua et al., 2019), MMLU-Pro (Wang et al., 2024b), and IFEval (Zhou et al., 2023).

Multimodal Reasoning OlympiadBench (He et al., 2024), MathVision (Wang et al., 2024a), MathVerse (Vision Only) (Zhang et al., 2024b), DynaMath (Zou et al., 2024), WeMath (Qiao et al., 2024), LogicVista (Xiao et al., 2024), and MathVista (Lu et al., 2023).

MiMo-VL-7B的B模型配置

在表4中，详细介绍了MiMo-VL-7B的架构和配置。

	Vision Encoder	Language Model
# Layers	32	36
# Heads	16	32
Hidden Size	1280	4096
Intermediate Size	3456	11008
Position Embedding	2D RoPE	MRoPE (Bai et al., 2025b)
Patch Size	14	-

表4 MiMo-VL-7B的配置。

C 评估基准

我们在包括以下50个不同任务在内的多样任务中评估我们的模型:

通用视觉理解 AI2D (Kembhavi 等人，2016)，BLINK (Fu 等人，2024b)。CV- Bench (Tong 等人，2024)，MMMU (Yue 等人，2024a)，MMMU-Pro（标准版和视觉版） (Yue 等人，2024c)，Mantis (Jiang 等人，2024)，MME-RealWorld（英文和中文） (Zhang 等人，2024c)，MMBench（英文和中文） (Liu 等人，2024b)，VibeEval (Padlewski 等人，2024)，VL- RewardBench (Li 等人，2024b)，V^* (Wu 和 Xie，2023)，以及 VLMs 是盲的 (Rahmanzadehgervi 等人，2025)。

一般的基础和计数任务包括RefCOCO（Yu 等人，2016；Nagaraja 等人，2016；Kazemzadeh 等人，2014）、CountBench（Paiss 等人，2023）和PixmoCount（Deitke 等人，2024）

文档与图表理解：ChartQA（Masry 等，2022）、InfographicVQA（Mathew 等，2022）、DocVQA（Mathew 等，2021）、OCRBench（Liu 等，2024c）、SEED-Bench-2-Plus（Li 等，2024a）以及 CharXiv（RQ/DQ）（Wang 等，2024c）。

视频理解与本地化 Video-MME（Fu 等，2024a）、Video-MMMU（Hu 等，2025）、EgoSchema（Mangalam 等，2023）以及 Charades-STA Gao 等（2017）。

GUI 理解与定位 WebSrc (Chen et al., 2021)、VisualWebBench (Liu et al., 2024a)、ScreenSpot (Cheng et al., 2024)、ScreenSpot-V2 (Wu et al., 2024)、ScreenSpot-Pro (Li et al., 2025) 以及 OSWorld-G (Xie et al., 2025)。

仅文本基准测试 GPQA（Rein 等，2024）、SuperGPQA（Du 等，2025）、DROP（Dua 等，2019）、MMLU-Pro（Wang 等，2024b）和 IFEval（Zhou 等，2023）。

多模态推理奥林匹克平台（He 等，2024）、MathVision（Wang 等，2024a）、MathVerse（仅限视觉）（Zhang 等，2024b）、DynaMath（Zou 等，2024）、WeMath（Qiao 等，2024）、Logic Vista（Xiao 等，2024）以及MathVista（Lu 等，2023）。

Text Reasoning MATH500 (Hendrycks et al., 2021), AIME 2024 (MAA, 2024), and AIME 2025 (MAA, 2025).

D GUI Action Space

The syntax and definition for each action in the GUI action space are summarized in Table 5.

click	*Syntax:* `click{start_point:[x,y], text(option):text}`
	Definition: Click at (x, y) on an element with the given text.
scroll	*Syntax:* `scroll{direction:dir, scroll_distance(option):dist}`
	Definition: Scroll in the given direction by a specified distance.
finished	*Syntax:* `finished{status:status}`
	Definition: Mark the task as complete with a given `status`.
input	*Syntax:* `input{text:text,start_point(option):[x,y]}`
	Definition: Type the specified text at coordinates (x, y).
longpress	*Syntax:* `longpress{start_point:[x,y]}`
	Definition: Long press at coordinates (x,y).
open	*Syntax:* `open{app:app_name}`
	Definition: Open the specified application `app_name`.
press	*Syntax:* `press{keys:[key1, key2, ...]}`
	Definition: Press the specified hotkeys (`[key1, key2, ...]`).
drag	*Syntax:* `drag{start_point:[x,y], end_point:[x,y]}`
	Definition: Drag from the `start_point` to the `end_point`.
hover	*Syntax:* `hover{}`
	Definition: Hover the mouse over a location.
select	*Syntax:* `select{text:text}`
	Definition: Select the specified `text`.
wait	*Syntax:* `wait{}`
	Definition: Pause for a brief moment.
appswitch	*Syntax:* `appswitch{app:app_name}`
	Definition: Switch to the specified application `app_name`.

Table 5 Action Space Details.

E More Qualitative Examples

文本推理 MATH500（Hendrycks 等，2021）、AIME 2024（MAA，2024）和 AIME 2025（MAA，2025）。

D GUI 操作空间

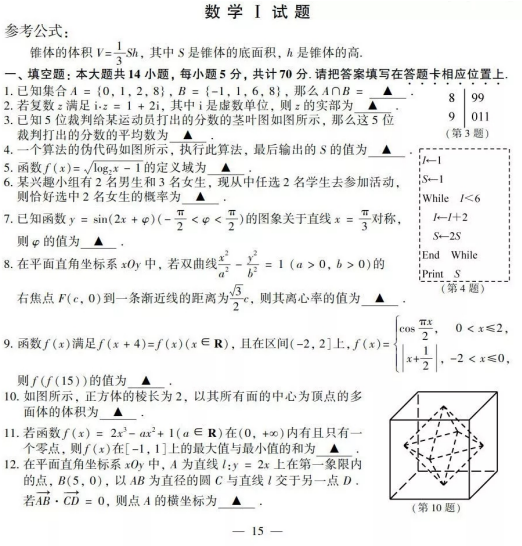
GUI动作空间中每个动作的语法和定义总结在表5中。

click	*Syntax:* `click{start_point:[x,y], text(option):text}` *Definition:* Click at (x, y) on an element with the given text.
scroll	*Syntax:* `scroll{direction:dir, scroll_distance(option):dist}` *Definition:* Scroll in the given direction by a specified distance.
finished	*Syntax:* `finished{status:status}` *Definition:* Mark the task as complete with a given `status`.
input	*Syntax:* `input{text:text,start_point(option):[x,y]}` *Definition:* Type the specified text at coordinates (x, y).
longpress	*Syntax:* `longpress{start_point:[x,y]}` *Definition:* Long press at coordinates (x,y).
open	*Syntax:* `open{app:app_name}` *Definition:* Open the specified application `app_name`.
press	*Syntax:* `press{keys:[key1, key2, ...]}` *Definition:* Press the specified hotkeys (`[key1, key2, ...]`).
drag	*Syntax:* `drag{start_point:[x,y], end_point:[x,y]}` *Definition:* Drag from the `start_point` to the `end_point`.
hover	*Syntax:* `hover{}` *Definition:* Hover the mouse over a location.
select	*Syntax:* `select{text:text}` *Definition:* Select the specified `text`.
wait	*Syntax:* `wait{}` *Definition:* Pause for a brief moment.
appswitch	*Syntax:* `appswitch{app:app_name}` *Definition:* Switch to the specified application `app_name`.

表5 行动空间详情。

E 更多定性示例

Input image:



数 学 Ⅰ 试 题

参考公式：

锥体的体积 $V=\frac{1}{3}Sh$，其中 S 是锥体的底面积，h 是锥体的高.

一、填空题：本大题共 14 小题，每小题 5 分，共计 70 分. 请把答案填写在答题卡相应位置上.

1. 已知集合 $A=\{0,1,2,8\}$，$B=\{-1,1,6,8\}$，那么 $A\cap B=$ ▲ .
2. 若复数 z 满足 $i\cdot z=1+2i$，其中 i 是虚数单位，则 z 的实部为 ▲ .
3. 已知 5 位裁判给某运动员打出的分数的茎叶图如图所示，那么这 5 位裁判打出的分数的平均数为 ▲ .

8	99
9	011

（第 3 题）

4. 一个算法的伪代码如图所示，执行此算法，最后输出的 S 的值为 ▲ .

```
I←1
S←1
While I<6
  I←I+2
  S←2S
End While
Print S
```

（第 4 题）

5. 函数 $f(x)=\sqrt{\log_2 x-1}$ 的定义域为 ▲ .
6. 某兴趣小组有 2 名男生和 3 名女生，现从中任选 2 名学生去参加活动，则恰好选中 2 名女生的概率为 ▲ .
7. 已知函数 $y=\sin(2x+\varphi)(-\frac{\pi}{2}<\varphi<\frac{\pi}{2})$ 的图象关于直线 $x=\frac{\pi}{3}$ 对称，则 φ 的值为 ▲ .
8. 在平面直角坐标系 xOy 中，若双曲线 $\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$（$a>0$，$b>0$）的右焦点 $F(c,0)$ 到一条渐近线的距离为 $\frac{\sqrt{3}}{2}c$，则其离心率的值为 ▲ .
9. 函数 $f(x)$ 满足 $f(x+4)=f(x)(x\in\mathbf{R})$，且在区间 $(-2,2]$ 上，$f(x)=\begin{cases}\cos\frac{\pi x}{2}, & 0<x\leqslant 2,\\ \left|x+\frac{1}{2}\right|, & -2<x\leqslant 0,\end{cases}$ 则 $f(f(15))$ 的值为 ▲ .
10. 如图所示，正方体的棱长为 2，以其所有面的中心为顶点的多面体的体积为 ▲ .

（第 10 题）

11. 若函数 $f(x)=2x^3-ax^2+1(a\in\mathbf{R})$ 在 $(0,+\infty)$ 内有且只有一个零点，则 $f(x)$ 在 $[-1,1]$ 上的最大值与最小值的和为 ▲ .
12. 在平面直角坐标系 xOy 中，A 为直线 $l:y=2x$ 上在第一象限内的点，$B(5,0)$，以 AB 为直径的圆 C 与直线 l 交于另一点 D. 若 $\overrightarrow{AB}\cdot\overrightarrow{CD}=0$，则点 A 的横坐标为 ▲ .

— 15 —

Instruction: 完成该试卷

MiMo-VL-7B: <think>...</think>

第**1**题：集合的交集 集合$A=\{0,1,2,8\}$，$B=\{-1,1,6,8\}$，交集为同时属于两个集合的元素，即$A\cap B=\{1,8\}$。

第**2**题：复数的实部 设$z=x+yi$（x,y 为实数），由$i\cdot z=1+2i$，即$i(x+yi)=-y+xi=1+2i$。分离实部和虚部：$-y=1$（实部），$x=2$（虚部）。故z 的实部为**2**。

第**3**题：茎叶图的平均数 茎叶图中，茎为十位，叶为个位。分数为$89,89,90,91,91$（共5个）。平均数为$\frac{89+89+90+91+91}{5}=\frac{450}{5}=90$。

第**4**题：伪代码的循环结果 初始化$I=1$，$S=1$。循环条件$I<6$：- 第1次循环：$I=1+2=3$，$S=2\times 1=2$；- 第2次循环：$I=3+2=5$，$S=2\times 2=4$；- 第3次循环：$I=5+2=7$（不满足$I<6$，退出循环）。最终$S=8$。

第**5**题：函数的定义域 函数$f(x)=\sqrt{\log_2 x-1}$ 需满足：1. 二次根式内非负：$\log_2 x-1\geq 0$；2. 对数有定义：$x>0$。由$\log_2 x\geq 1$ 得$x\geq 2^1=2$，故定义域为$[2,+\infty)$。

第**6**题：组合的概率 2男3女中选2人，总选法$C_5^2=10$；选中2女生的选法$C_3^2=3$。概率为$\frac{3}{10}$。

第**7**题：正弦函数的对称轴 函数$y=\sin(2x+\varphi)$ 关于$x=\frac{\pi}{3}$ 对称，即$2\cdot\frac{\pi}{3}+\varphi=k\pi+\frac{\pi}{2}$（$k\in\mathbb{Z}$）。取$k=0$，得$\varphi=-\frac{\pi}{6}$（满足$-\frac{\pi}{2}<\varphi<\frac{\pi}{2}$）。

第**8**题：双曲线的离心率 双曲线$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$，右焦点$F(c,0)$（$c=\sqrt{a^2+b^2}$），渐近线$bx-ay=0$。焦点到渐近线的距离为$\frac{|bc|}{\sqrt{a^2+b^2}}=b$，由题设$b=\frac{\sqrt{3}}{2}c$。结合$c^2=a^2+b^2$，得$a=\frac{c}{2}$，离心率$e=\frac{c}{a}=2$。

第**9**题：周期函数与复合函数 $f(x+4)=f(x)$ 说明周期为4。- $f(15)=f(15-4\times 3)=f(3)=f(-1)$（$3=-1+4$）；- $x\in(-2,0]$ 时，$f(-1)=|-1+\frac{1}{2}|=\frac{1}{2}$；- $f(\frac{1}{2})=\cos\left(\frac{\pi}{2}\cdot\frac{1}{2}\right)=\cos\frac{\pi}{4}=\frac{\sqrt{2}}{2}$（$\frac{1}{2}\in(0,2]$）。

第**10**题：多面体的体积 正方体棱长2，面中心坐标为$(\pm1,0,0),(0,\pm1,0),(0,0,\pm1)$，构成正八面体。正八面体体积为两个四棱锥之和，底面正方形边长$\sqrt{2}$，高1，体积为$2\times\frac{1}{3}\times 2\times 1=\frac{4}{3}$。

第**11**题：函数的极值与零点 函数$f(x)=2x^3-ax^2+1$ 在$(0,+\infty)$ 内有且只有一个零点，导数$f'(x)=6x^2-2ax$，临界点$x=0$ 和$x=\frac{a}{3}$。当$f(\frac{a}{3})=0$ 时（极小值为0，仅1个零点），解得$a=3$。$f(x)$ 在$[-1,1]$ 上，$f(0)=1$（最大值），$f(-1)=-4$（最小值），和为$1+(-4)=-3$。

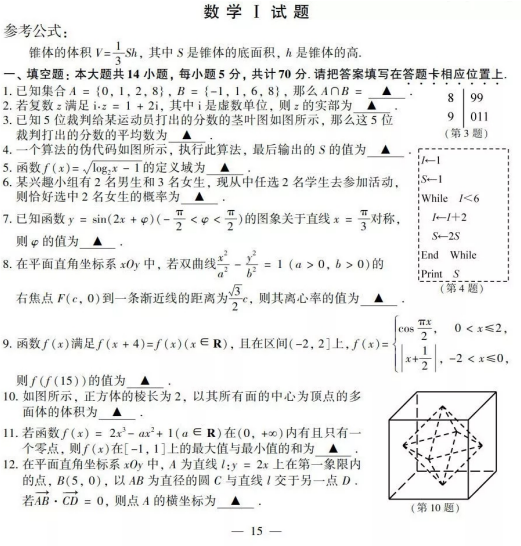
第**12**题：向量垂直的条件 设$A(x,2x)$，$AB=(5-x,-2x)$，圆心$C\left(\frac{x+5}{2},x\right)$，另一交点$D(1,2)$（由圆与直线交点唯一性得）。向量$\overrightarrow{CD}=\left(\frac{-x-3}{2},2-x\right)$，由$\overrightarrow{AB}\cdot\overrightarrow{CD}=0$，解得$x=3$。

答案

1. $\{1,8\}$ 2. **2** 3. **90** 4. **8** 5. $[2,+\infty)$ 6. $\frac{3}{10}$ 7. $-\frac{\pi}{6}$ 8. **2** 9. $\frac{\sqrt{2}}{2}$ 10. $\frac{4}{3}$ 11. **−3** 12. **3**

Figure 10 Example of MiMo-VL-7B-RL answering all questions within a whole test paper.

Input image:输入图像：



数 学 Ⅰ 试 题

参考公式：

锥体的体积 $V=\frac{1}{3}Sh$，其中 S 是锥体的底面积，h 是锥体的高.

一、填空题：本大题共 14 小题，每小题 5 分，共计 70 分. 请把答案填写在答题卡相应位置上.

1. 已知集合 $A=\{0,1,2,8\}$，$B=\{-1,1,6,8\}$，那么 $A\cap B=$ ▲ .
2. 若复数 z 满足 $i\cdot z=1+2i$，其中 i 是虚数单位，则 z 的实部为 ▲ .
3. 已知 5 位裁判给某运动员打出的分数的茎叶图如图所示，那么这 5 位裁判打出的分数的平均数为 ▲ .

8	99
9	011

（第 3 题）

4. 一个算法的伪代码如图所示，执行此算法，最后输出的 S 的值为 ▲ .

```
I←1
S←1
While I<6
  I←I+2
  S←2S
End While
Print S
```

（第 4 题）

5. 函数 $f(x)=\sqrt{\log_2 x-1}$ 的定义域为 ▲ .
6. 某兴趣小组有 2 名男生和 3 名女生，现从中任选 2 名学生去参加活动，则恰好选中 2 名女生的概率为 ▲ .
7. 已知函数 $y=\sin(2x+\varphi)\left(-\frac{\pi}{2}<\varphi<\frac{\pi}{2}\right)$ 的图象关于直线 $x=\frac{\pi}{3}$ 对称，则 φ 的值为 ▲ .
8. 在平面直角坐标系 xOy 中，若双曲线 $\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$（$a>0$，$b>0$）的右焦点 $F(c,0)$ 到一条渐近线的距离为 $\frac{\sqrt{3}}{2}c$，则其离心率的值为 ▲ .
9. 函数 $f(x)$ 满足 $f(x+4)=f(x)(x\in\mathbf{R})$，且在区间 $(-2,2]$ 上，$f(x)=\begin{cases}\cos\frac{\pi x}{2}, & 0<x\leqslant 2,\\ \left|x+\frac{1}{2}\right|, & -2<x\leqslant 0,\end{cases}$ 则 $f(f(15))$ 的值为 ▲ .
10. 如图所示，正方体的棱长为 2，以其所有面的中心为顶点的多面体的体积为 ▲ .

（第 10 题）

11. 若函数 $f(x)=2x^3-ax^2+1(a\in\mathbf{R})$ 在 $(0,+\infty)$ 内有且只有一个零点，则 $f(x)$ 在 $[-1,1]$ 上的最大值与最小值的和为 ▲ .
12. 在平面直角坐标系 xOy 中，A 为直线 $l:y=2x$ 上在第一象限内的点，$B(5,0)$，以 AB 为直径的圆 C 与直线 l 交于另一点 D. 若 $\overrightarrow{AB}\cdot\overrightarrow{CD}=0$，则点 A 的横坐标为 ▲ .

— 15 —

指令：完成该试卷

MiMo-VL-7B: <think>...</think>
第1题：集合的交集 集合$A=\{0,1,2,8\}$，$B=\{-1,1,6,8\}$，交集为同时属于两个集合的元素，即$A\cap B=\{1,8\}$。
第2题：复数的实部 设$z=x+yi$（x,y 为实数），由$i\cdot z=1+2i$，即$i(x+yi)=-y+xi=1+2i$。分离实部和虚部：$-y=1$（实部），$x=2$（虚部）。故z 的实部为2。
第3题：茎叶图的平均数 在茎叶图中，茎为十位，叶为个位。分数为89, 89, 90, 91, 91（共5个）。平均数 为 8 9+89+90+91+91 5 = 4505 = 90。
第4题：伪代码的循环结果 初始化$I=1$，$S=1$。循环条件$I<6$：- 第1次循环：$I=1+2=3$，$S=2\times1=2$；- 第2次循环：$I=3+2=5$，$S=2\times2=4$；- 第3次循环：$I=5+2=7$（不满足$I<6$，退出循环）。最终$S=8$。
第5题：函数的定义域 函数$f(x)=\sqrt{\log_2 x-1}$ 需满足：1. 二次根式内非负：$\log_2 x-1\geq0$；2. 对数有定义：$x>0$。由log 2翻译文本: 得$x\geq2^1=2$，故定义域为$[2,+\infty)$。
第6题：组合的概率 2男3女中选2人，总选法$C_5^2=10$；选中2女生的选法$C_3^2=3$。概率为$\frac{3}{10}$。
第7题：正弦函数的对称轴 函数$y=\sin(2x+\varphi)$ 关于$x=\frac{\pi}{3}$ 对称，即$2\cdot\frac{\pi}{3}+\varphi=k\pi+\frac{\pi}{2}$（$k\in\mathbb{Z}$）。取$k=0$，得$\varphi=-\frac{\pi}{6}$（满足$-\frac{\pi}{2}<\varphi<\frac{\pi}{2}$）。
第8题：双曲线的离心率 双曲线$\frac{x^2}{a^2}-\frac{y^2}{b^2}=1$，右焦点$F(c,0)$（$c=\sqrt{a^2+b^2}$），渐近线$bx-ay=0$。焦点到渐近线的距离为$\frac{|bc|}{\sqrt{a^2+b^2}}=b$，由题设$b=\frac{\sqrt{3}}{2}c$。结合$a^2+b^2=c^2$，得$a=\frac{c}{2}$，离心率$e=\frac{c}{a}=2$。
第9题：周期函数与复合函数 $f(x+4)=f(x)$ 说明周期为4。- $f(15)=f(15-4\times3)=f(3)=f(-1)$（$3=-1+4$）；- $x\in(-2,0]$ 时，$f(-1)=|-1+12|=\frac{1}{2}$；- $f(\frac{1}{2})=\cos\left(\frac{\pi}{2}\cdot\frac{1}{2}\right)=\cos\frac{\pi}{4}=\frac{\sqrt{2}}{2}$（$\frac{1}{2}\in(0,2]$）。
第10题：多面体的体积 正方体边长为2，面中心坐标为$(\pm1,0,0),(0,\pm1,0),(0,0,\pm1)$，构成正八面体。正八面体的体积为两个四棱锥之和，底面正方形边长$\sqrt{2}$，高为1，体积为$2\times\frac{1}{3}\times2\times1=\frac{4}{3}$。
第11题：函数的极值与零点 函数$f(x)=2x^3-ax^2+1$ 在$(0,+\infty)$ 内有且只有一个零点，导数$f'(x)=6x^2-2ax$，临界点$x=0$ 和$x=\frac{a}{3}$。当$f(\frac{a}{3})=0$ 时（极小值为0，仅1个零点），解得$a=3$。$f(x)$ 在$[-1,1]$ 上，$f(0)=1$（最大值），$f(-1)=-4$（最小值），和为$1+(-4)=-3$。
第12题：向量垂直的条件 设$A(x,2x)$，$AB=(5-x,-2x)$，圆心$C\left(\frac{x+5}{2},x\right)$，另一交点$D(1,2)$（由圆与直线交点唯一性得）。向量$\overrightarrow{CD}=\left(\frac{-x-3}{2},2-x\right)$，由$\overrightarrow{AB}\cdot\overrightarrow{CD}=0$，解得$x=3$。

答案

1. $\{1,8\}$ 2. 2 3. 90 4. 8 5. $[2,+\infty)$ 6. $\frac{3}{10}$ 7. $-\frac{\pi}{6}$ 8. 2 9. $\frac{\sqrt{2}}{2}$ 10. $\frac{4}{3}$ 11. -3 12. 3

图10 例子：MiMo-VL-7B-RL 在整个试卷中回答所有问题的示例。

Instruction: 这瓶可乐喝10罐会摄入多少克钠?
MiMo-VL-7B-RL: `<think>...</think>`要计算喝10罐可口可乐零度时摄入的钠量，需按以下步骤分析：

1. 步骤**1**：明确营养成分表信息 从图片中“营养成分表”可知：每**100**毫升该饮料中，钠的含量为**18**毫克。

2. 步骤**2**：确定单罐饮料容量（行业常见规格） 可口可乐零度的罐装规格通常为**330**毫升/罐（这是饮料行业的常见罐装容量，若实际容量不同需以产品标注为准，此处默认330ml）。

3. 步骤**3**：计算单罐钠含量 单罐（330ml）中钠的含量= 每100ml钠含量× 单罐容量÷ 100 即：18 毫克/100ml × 330 ml ÷ 100 = 59.4 毫克

4. 步骤**4**：计算**10**罐钠总量 10罐钠总量= 单罐钠含量× 10 即：59.4 毫克 × 10 = 594 毫克

因此，喝10罐该款可口可乐零度，约摄入**594**毫克钠。

Instruction: 该发票缺失的税额应该是多少?
MiMo-VL-7B-RL: `<think>...</think>`要计算该发票的税额，需结合税率和价税合计金额来推导：

1. 步骤**1**：明确关键信息 - 发票税率：13%（发票中“税率”栏显示）- 价税合计金额（大写）：壹拾壹万贰仟伍佰捌拾玖圆伍角贰分，对应小写为**112589.52**元

2. 步骤**2**：税额的计算逻辑 发票税额= 价税合计金额× （1 - 税率）× 税率（即：总价价税合计= 不含税金额+ 税额→ 税额= 价税合计× 税率÷ (1 + 税率)）

设不含税金额为 X，则：$X + X \times 13\% = 112589.52$ $X \times 1.13 = 112589.52$ $X = 112589.52 \div 1.13 = 99636.74$ 元【不含税金额】
税额= 不含税金额× 税率= $99636.74 \times 13\% =$ 12952.78 元

Figure 11 Example of MiMo-VL-7B-RL solving complex OCR problems with reasoning.

指令：这瓶可乐喝10罐会摄入多少克钠？ MiMo-VL-7B-RL: <think>...</think>要 计 算喝10罐可口可乐零度时摄入的钠量，需按以 下步骤分析：

1. 步骤1：明确营养成分表信息 从图片中“营养成分表”可知：每100毫升该饮料中，钠的含量为18毫克。

2. 步骤2：确定单罐饮料容量（行业常见规格） 可口可乐零度的罐装规格通常为330毫升/罐（这是饮料行业的常见罐装容量，若实际容量不同需以产品标注为准，此处默认330ml）。 3. 步骤3：计算单罐钠含量 单罐（330ml）中钠的含量= 每100ml钠含量× 单罐容量÷ 100 即：18毫克/100ml × 330ml ÷ 100 = 59.4毫克 4. 步骤4：计算10罐钠总量 10罐钠总量= 单罐钠含量× 10即：59.4毫克 × 10 = 594毫克 因此，喝10罐该款可口可乐零度，约摄入594毫克钠。

指令：该发票缺失的税额应该是多少？ MiMo-VL-7B-RL: <think>...</think>要 计 算该发票的税额，需结合税率和价税合计金额 来推导：

1. 步 骤1： 明 确 关 键 信 息 - 发 票 税 率：13%（发票中“税率”栏显示）- 价税合 计金额（大写）： 壹拾壹万贰仟伍佰捌拾 玖圆伍角贰分，对应小写为{v*}元

2. 步骤2：税额的计算逻辑 发票税额= 价税合计金额× （1 - 税率）× 税率（即：总价 价税合计= 不含税金额+ 税额→ 税额= 价 税合计× 税率÷ (1 + 税率) ）

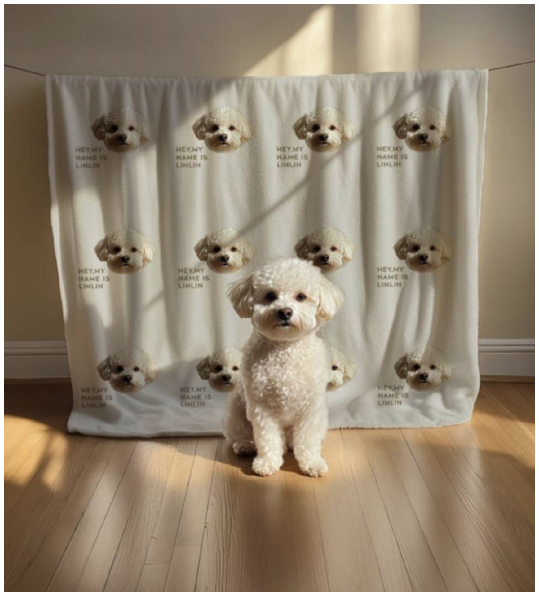
设 不 含 税 金 额 为 X， 则 ： $X + X \times 13\% = 112589.52$ $X \times 1.13 = 112589.52$ $X = 112589.52 \div 1.13 = 99636.74$ 元【不含税金 额】

税额= 不含税金额× 税率= $99636.74 \times 13\% = 12952.78$ 元

图11 采用MiMo-VL-7B-RL解决复杂OCR问题的推理示例。

Instruction: How many muffin in the image?
MiMo-VL-7B-RL: `<think>...</think>` To determine the number of muffins in the image, we analyze the 3×4 grid (12 cells total) and count muffins row by row:
- **Top row (row 1)**: Columns 1 and 3 are muffins → 2 muffins. - **Middle row (row 2)**: Columns 2 and 4 are muffins → 2 muffins. - **Bottom row (row 3)**: Columns 1 and 3 are muffins → 2 muffins.
Adding these, 2 + 2 + 2 = 6.
There are 6 muffins in the image.

Instruction: How many cats in the image?
MiMo-VL-7B-RL: `<think>...</think>` To determine the number of cats in the image, we can observe and count: - There is a black cat lying on the floor. - There is a white and brown cat sitting behind the black cat.
Thus, the number of cats in the image is 2.



Instruction: How many dogs in the image?
MiMo-VL-7B-RL: `<think>...</think>` To determine the number of dogs in the image, we analyze both the printed dogs on the blanket and the real dog in the foreground:

1. **Printed dogs on the blanket**: The blanket features a grid of dog faces. Counting these, we see 3 rows with 4 dogs each, so $3 \times 4 = 12$ printed dogs.

2. **Real dog in the foreground**: There is 1 real dog standing in front of the blanket.

Adding these together, 12 + 1 = 13.
Thus, the total number of dogs in the image is 13.

Figure 12 Example of MiMo-VL-7B-RL solving detailed perception problem.

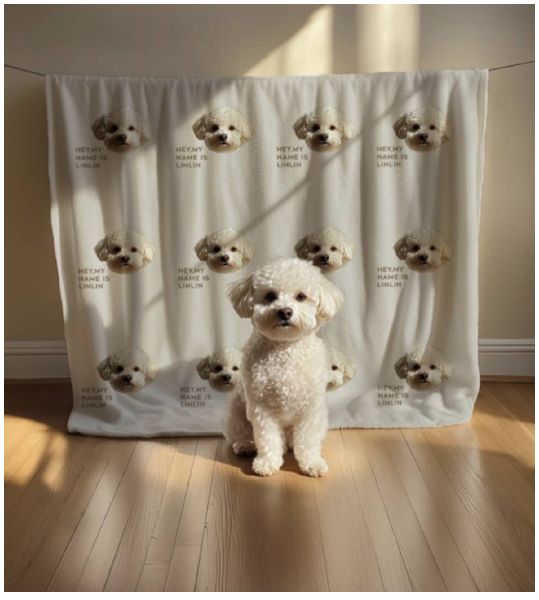
指令：图片中有多少松饼？MiMo-VL-7B-RL：<think>...</think> 为了确定图片中的松饼数量，我们分析3×4网格（共12个格子），并逐行数松饼：

- 顶行（第1行）：第1列和第3列是松饼 → 2个松饼。 - 中行（第2行）：第2列和第4列是松饼 → 2个松饼。 - 底行（第3行）：第1列和第3列是松饼 → 2个松饼。

加上这些，2 + 2 + 2 = 6。图片中有6个松饼。

指令：图片中有多少只猫？MiMo-VL-7B-RL：<think>...</think> 为了确定图片中的猫的数量，我们可以观察并数数：- 有一只黑色的猫躺在地板上。- 有一只白色和棕色的猫坐在黑色猫的后面。

因此，图像中的猫的数量是 {v*} 2。



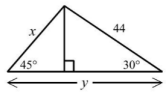
指令：图片中有多少只狗？MiMo-VL-7B-RL：<think>...</think> 为了确定图片中狗的数量，我们分析了毯子上的印刷狗和前景中的真实狗：

1. 印有狗狗的毯子：这条毯子上有一个狗脸的格子。数一数，我们看到有3行，每行4只狗，所以总共有3 × 4 = 12只印刷的狗。

2. 前景中的真实狗：有1只真实的狗站在毯子前面。将这些加在一起，12 + 1 = 13。因此，图片中狗的总数是13只。

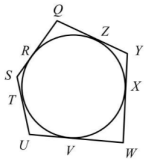
图12 MiMo-VL-7B-RL 解决详细感知问题的示例。

7. Find the value of x and y.

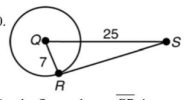


8. The length of the diagonal of a square is 22. What is the length of each side?

9. The circle is circumscribed by the pentagon as shown (not drawn to scale). If $QZ = 10$, $YX = 9$, $XW = 13$, $UW = 16$, and $SU = 17$, find the perimeter of the pentagon.

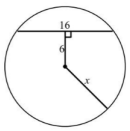


10.



Use the figure above. $\overline{SR}$ is tangent to $\odot Q$ at R. $SR =$ ____.

11. In the figure below, find the value of radius x.



12. R, S, and T are collinear. S is between R and T. $RS = 2w + 1$, $ST = w - 1$, and $RT = 18$. Solve for w. Then determine the length of $\overline{RS}$.

Instruction: answer all the questions.
MiMo-VL-7B-RL: `<think>...</think>`

1. Problem 7 To find x and y, we analyze the two right triangles formed by the height:

   - **Right triangle (30° angle)**: Hypotenuse $= 44$. The height h (opposite 30°) is $44 \cdot \sin(30°) = 44 \cdot 0.5 = 22$. The base segment y (adjacent to 30°) is $44 \cdot \cos(30°) = 44 \cdot \frac{\sqrt{3}}{2} = 22\sqrt{3}$.

   - **Right triangle (45° angle)**: This is a 45-45-90 triangle, so legs are equal. The height $h = 22$, so the other leg (left base segment) is also 22. Hypotenuse x (45-45-90 ratio: leg $\cdot\sqrt{2}$) is $22\sqrt{2}$.

2. Problem 8 For a square with diagonal $d = 22$, the side length s relates to the diagonal by $d = s\sqrt{2}$. Solving for s: $s = \frac{d}{\sqrt{2}} = \frac{22}{\sqrt{2}} = 11\sqrt{2}$ (rationalizing the denominator).

3. Problem 9 A pentagon circumscribed about a circle (tangential pentagon) has the property that the sum of the lengths of all sides equals 2× the sum of the tangent segments from each vertex. However, since the problem provides the side lengths $QZ = 10$, $YX = 9$, $XW = 13$, $UW = 16$, $SU = 17$, and the pentagon is tangential, the perimeter is the sum of these sides (as they represent the sides of the pentagon).

   Perimeter $= 10 + 9 + 13 + 16 + 17 = 65$.

4. Problem 10 $\overline{SR}$ is tangent to $\odot Q$ at R, so $QR \perp SR$ (tangent is perpendicular to radius). Thus, $\triangle QRS$ is a right triangle with $QR = 7$ (radius) and $QS = 25$ (hypotenuse). By the Pythagorean theorem: $SR = \sqrt{QS^2 - QR^2} = \sqrt{25^2 - 7^2} = \sqrt{625 - 49} = \sqrt{576} = 24$

5. Problem 11 A perpendicular from the circle's center to a chord bisects the chord. The chord length is 16, so half the chord is $\frac{16}{2} = 8$. The distance from the center to the chord is 6, forming a right triangle with legs 6 and 8, and hypotenuse x (radius). By the Pythagorean theorem: $x = \sqrt{6^2 + 8^2} = \sqrt{36 + 64} = \sqrt{100} = 10$

6. Problem 12 Since R, S, T are collinear and S is between R and T, $RS + ST = RT$. Substitute $RS = 2w + 1$, $ST = w - 1$, and $RT = 18$:

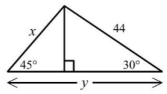
   $$(2w + 1) + (w - 1) = 18$$

   Simplify: $3w = 18 \implies w = 6$. Substitute $w = 6$ into RS: $RS = 2(6) + 1 = 13$

Final Answers 7. $x = 22\sqrt{2}$, $y = 22\sqrt{3}$ 8. $11\sqrt{2}$ 9. 65 10. 24 11. 10 12. $w = 6$, $\overline{RS} = 13$

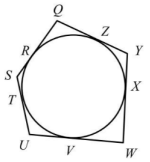
Figure 13 Example of MiMo-VL-7B-RL solving multiple geometry problems.

7. Find the value of x and y.

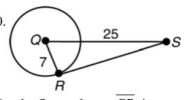


8. The length of the diagonal of a square is 22. What is the length of each side?

9. The circle is circumscribed by the pentagon as shown (not drawn to scale). If $QZ = 10$, $YX = 9$, $XW = 13$, $UW = 16$, and $SU = 17$, find the perimeter of the pentagon.

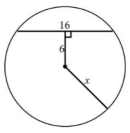


10.



Use the figure above. $\overline{SR}$ is tangent to $\odot Q$ at R. $SR =$ _____.

11. In the figure below, find the value of radius x.



12. R, S, and T are collinear. S is between R and T. $RS = 2w + 1$, $ST = w - 1$, and $RT = 18$. Solve for w. Then determine the length of $\overline{RS}$.

指令：回答所有问题。MiMo-VL-7B-RL：<think>...</think> 1. 问题7 为了找到 x 和 y，我们分析由高度形成的两个直角三角形：- **直角三角形（30°角）**：斜边 = 44。高度 h（对边 30°）为 $44 \cdot \sin(30^\circ) = 44 \cdot 0.5 = 22$。底边段 y（邻边 30°）为 $44 \cdot \cos(30^\circ) = 44 \cdot \frac{\sqrt{3}}{2} = 22\sqrt{3}$。- **直角三角形（45°角）**：这是一个 45-45-90 三角形，因此两条直角边相等。高度 $h = 22$，因此另一条直角边（左底边段）也为 22。斜边 x（45-45-90 比例：直角边 $\cdot\sqrt{2}$）为 $22\sqrt{2}$。 2. 问题8 对角线长度为 $d = 22$ 的正方形，其边长 s 与对角线的关系为 $d = s\sqrt{2}$。解出 s：$s = \frac{d}{\sqrt{2}} = \frac{22}{\sqrt{2}} = 11\sqrt{2}$（有理化分母）。 3. 问题9 一个内切圆的五边形（切圆五边形）具有这样一个性质：所有边长的总和等于 2 × 每个顶点的切线段总和。然而，由于题目提供的边长为 $QZ = 10$，$YX = 9$，$XW = 13$，$UW = 16$，$SU = 17$，并且五边形是切圆的，周长等于这些边的总和（因为它们代表五边形的边）。周长 = 10 + 9 + 13 + 16 + 17 = 65。 4. 问题10 SR 与 $\odot Q$ 在 R 相切，因此 $QR \perp SR$（切线与半径）垂直。因此，$\triangle QRS$ 是一个直角三角形，$QR = 7$（半径）和 $QS = 25$（斜边）。根据毕达哥拉斯定理：$SR = \sqrt{QS^2 - QR^2} = \sqrt{25^2 - 7^2} = \sqrt{625 - 49} = \sqrt{576} = 24$。 5. 问题11 从圆心到弦的垂线平分弦。弦长为 16，因此弦的一半为 $\frac{16}{2} = 8$。从圆心到弦的距离为 6，形成一个直角三角形，直角边为 6 和 8，斜边为 x（半径）。根据毕达哥拉斯定理：$x = \sqrt{6^2 + 8^2} = \sqrt{36 + 64} = \sqrt{100} = 10$。 6. 问题12 由于 R, S, T 共线且 S 位于 R 和 T 之间，$RS + ST = RT$。代入 $RS = 2w + 1$，$ST = w - 1$，$RT = 18$：

—

$$(2w + 1) + (w - 1) = 18$$

简化：$3w = 18 \implies w = 6$。将 $w = 6$ 代入 RS：$RS = 2(6) + 1 = 13$ 最终答案 7. $x = 22\sqrt{2}$，$y = 22\sqrt{3}$ 8. $11\sqrt{2}$ 9. 65 10. 24 11. 10 12. $w = 6$，$RS = 13$__

图13 MiMo-VL-7B-RL 解决多个几何问题示例。