KIMI-VL TECHNICAL REPORT

Kimi Team

ABSTRACT

We present **Kimi-VL**, an efficient open-source Mixture-of-Experts (MoE) vision-language model (VLM) that offers **advanced multimodal reasoning, long-context understanding, and strong agent capabilities**—all while activating only **2.8B** parameters in its language decoder (Kimi-VL-A3B).

Kimi-VL demonstrates strong performance across challenging domains: as a general-purpose VLM, Kimi-VL excels in multi-turn agent interaction tasks (*e.g.,* OSWorld), achieving state-of-the-art results comparable to flagship models. Furthermore, it exhibits remarkable capabilities across diverse challenging vision language tasks, including college-level image and video comprehension, optical character recognition (OCR), mathematical reasoning, multi-image understanding. In comparative evaluations, it effectively competes with cutting-edge efficient VLMs such as GPT-4o-mini, Qwen2.5-VL-7B, and Gemma-3-12B-IT, while surpassing GPT-4o in several specialized domains.

Kimi-VL also advances the pareto frontiers of multimodal models in processing long contexts and perceiving clearly: Equipped with a 128K extended context window, Kimi-VL can process long and diverse inputs, achieving impressive scores of 64.5 on LongVideoBench and 35.1 on MMLongBench-Doc. Its native-resolution vision encoder, MoonViT, further allows it to see and understand ultra-high-resolution visual inputs, achieving 83.2 on InfoVQA and 34.5 on ScreenSpot-Pro, while maintaining lower computational cost with common visual inputs and general tasks.

Building on this foundation, we introduce an advanced long-thinking variant: **Kimi-VL-Thinking**. Developed through long chain-of-thought (CoT) supervised fine-tuning (SFT) and reinforcement learning (RL), this model exhibits strong long-horizon reasoning capabilities. It achieves scores of 61.7 on MMMU, 36.8 on MathVision, and 71.3 on MathVista while maintaining the compact 2.8B activated LLM parameters, setting a new standard for efficient yet capable multimodal *thinking* models. Code and models are publicly accessible at https://github.com/MoonshotAI/Kimi-VL.

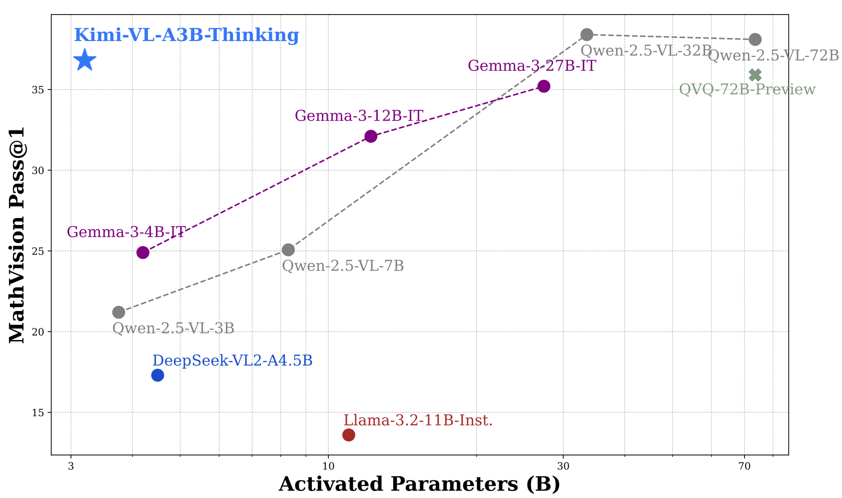


Figure 1: Comparison between **Kimi-VL-Thinking** and frontier open-source VLMs, including short-thinking VLMs (*e.g.* Gemma-3 series, Qwen2.5-VL series) and long-thinking VLMs (QVQ-72B-Preview), on MathVision benchmark. Our model achieves strong multimodal reasoning with just 2.8B LLM activated parameters.

KIMI-VL 技术报告

Kimi团队

摘要

我们介绍Kimi-VL，这是一种高效的开源专家混合（MoE）视觉-语言模型（VLM），提供先进的多模态推理、长上下文理解和强大的智能体能力——同时仅在其语言解码器中激活2.8B参数（Kimi-VL-A3B）。

Kimi-VL 在具有挑战性的领域中表现出色：作为一款通用的 VLM，Kimi-VL 在多轮代理交互任务中表现优异（*e.g.*, OSWorld），取得了与旗舰模型相媲美的最先进结果。此外，它在各种具有挑战性的视觉语言任务中展现了卓越的能力，包括大学级别的图像和视频理解、光学字符识别（OCR）、数学推理和多图像理解。在比较评估中，它有效地与前沿高效 VLM 竞争，如 GPT-4o-mini、Qwen2.5-VL-7B 和 Gemma-3-12B-IT，同时在多个专业领域超越了 GPT-4o。

Kimi-VL 还推动了多模态模型在处理长上下文和清晰感知方面的帕累托前沿：配备 128K 扩展上下文窗口，Kimi-VL 能够处理长且多样的输入，在 LongVideoBench 上取得了 64.5 的令人印象深刻的分数，在 MMLongBench-Doc 上取得了 35.1。其原生分辨率视觉编码器 MoonViT 进一步使其能够看到和理解超高分辨率的视觉输入，在 InfoVQA 上取得了 83.2，在 ScreenSpot-Pro 上取得了 34.5，同时在处理常见视觉输入和一般任务时保持较低的计算成本。

在此基础上，我们引入了一种先进的长思维变体：Kimi-VL-Thinking。通过长链思维（CoT）监督微调（SFT）和强化学习（RL）开发，该模型展现出强大的长远推理能力。它在 MMMU 上得分 61.7，在 MathVision 上得分 36.8，在 MathVista 上得分 71.3，同时保持紧凑的 2.8B 激活 LLM 参数，为高效而强大的多模态 *thinking* 模型设定了新标准。代码和模型可在 https://github.com/MoonshotAI/Kimi-VL 上公开获取。

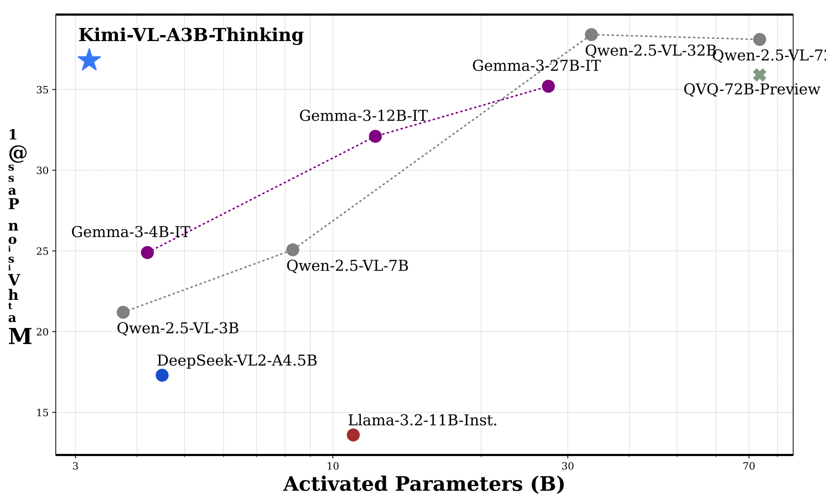


图1：Kimi-VL-Thinking与前沿开源VLM的比较，包括短思维VLM（*e.g.* Gemma-3系列，Qwen2.5-VL系列）和长思维VLM（QVQ-72B-Preview），在MathVision基准测试上的表现。我们的模型在仅激活2.8B LLM参数的情况下，实现了强大的多模态推理。

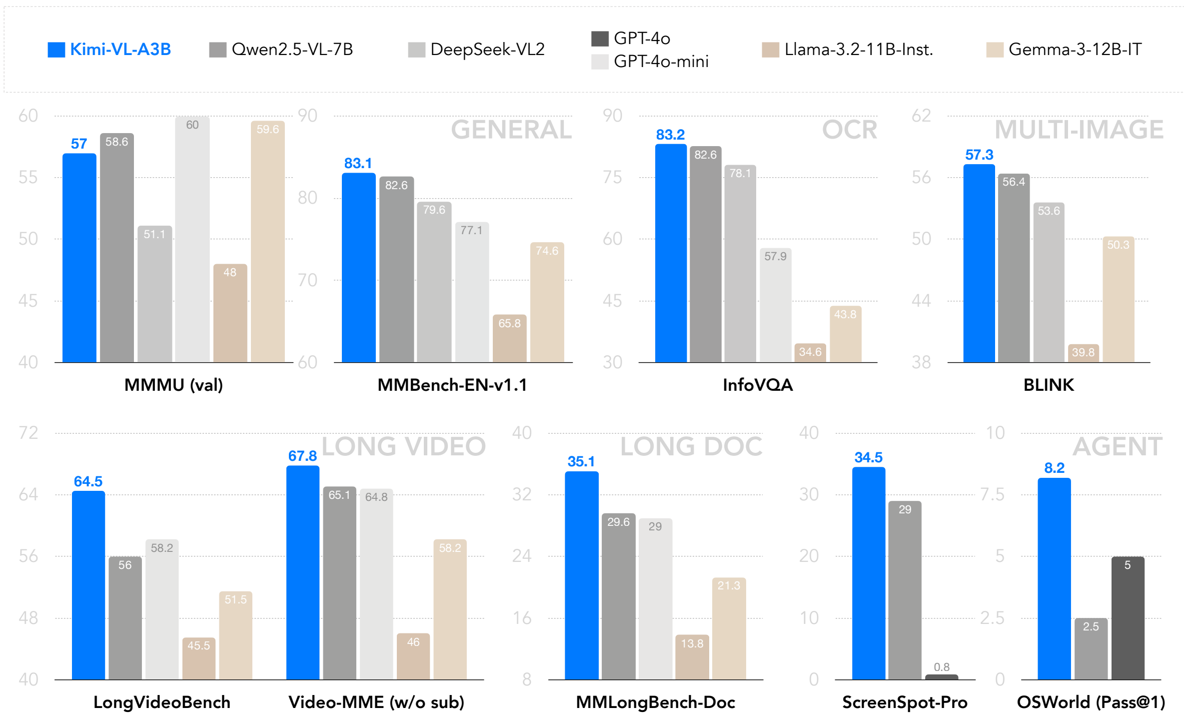


Figure 2: Highlights of **Kimi-VL** performance for a wide range of benchmarks like, general benchmarks (MMMU, MMBench), OCR (InfoVQA), multi-image (BLINK), long video (LongVideoBench, Video-MME), long document (MMLongBench-Doc), and agent (ScreenSpot-Pro and OSWorld). Detailed results are presented in Table 3.

1 Introduction

With the rapid advancement of artificial intelligence, human expectations for AI assistants have transcended traditional language-only interactions, increasingly aligning with the inherently multimodal nature of our world. To better understand and interact with these expectations, new generations of natively multimodal models, such as GPT-4o (OpenAI et al. 2024) and Google Gemini (Gemini Team et al. 2024), have emerged with the capability to seamlessly perceive and interpret visual inputs alongside language processing. Most recently, advanced multimodal models, pioneered by OpenAI o1 series (OpenAI 2024) and Kimi k1.5 (K. Team et al. 2025), have further pushed these boundaries by incorporating deeper and longer reasoning on multimodal inputs, thereby tackling more complex problems in the multimodal domain.

Nevertheless, development in large VLMs in the open-source community has significantly lagged behind their language-only counterparts, particularly in aspects of scalability, computational efficiency, and advanced reasoning capabilities. While language-only model DeepSeek R1 (DeepSeek-AI, D. Guo, et al. 2025) has already leveraged the efficient and more scalable mixture-of-experts (MoE) architecture and facilitated sophisticated long chain-of-thought (CoT) reasoning, most recent open-source VLMs, *e.g.* Qwen2.5-VL (Bai et al. 2025) and Gemma-3 (Gemma Team et al. 2025), continue to rely on dense architectures and do not support long-CoT reasoning. Early explorations into MoE-based vision-language models, such as DeepSeek-VL2 (Zhiyu Wu et al. 2024) and Aria (D. Li et al. 2024), exhibit limitations in other crucial dimensions. Architecturally, both models still adopt relatively traditional fixed-size vision encoders, hindering their adaptability to diverse visual inputs. From a capability perspective, DeepSeek-VL2 supports only a limited context length (4K), while Aria falls short in fine-grained visual tasks. Additionally, neither of them supports long-thinking abilities. Consequently, there remains a pressing need for an open-source VLM that effectively integrates structural innovation, stable capabilities, and enhanced reasoning through long-thinking.

In light of this, we present **Kimi-VL**, a vision-language model for the open-source community. Structurally, Kimi-VL consists of our Moonlight (J. Liu et al. 2025a) MoE language model with only **2.8B** activated (16B total) parameters, paired with a 400M native-resolution MoonViT vision encoder. In terms of capability, as illustrated in Figure 2, Kimi-VL can robustly handle diverse tasks (fine-grained perception, math, college-level problems, OCR, agent, *etc.*) across a broad spectrum of input forms (single-image, multi-image, video, long-document, *etc.*). Specifically, it features the following exciting abilities:

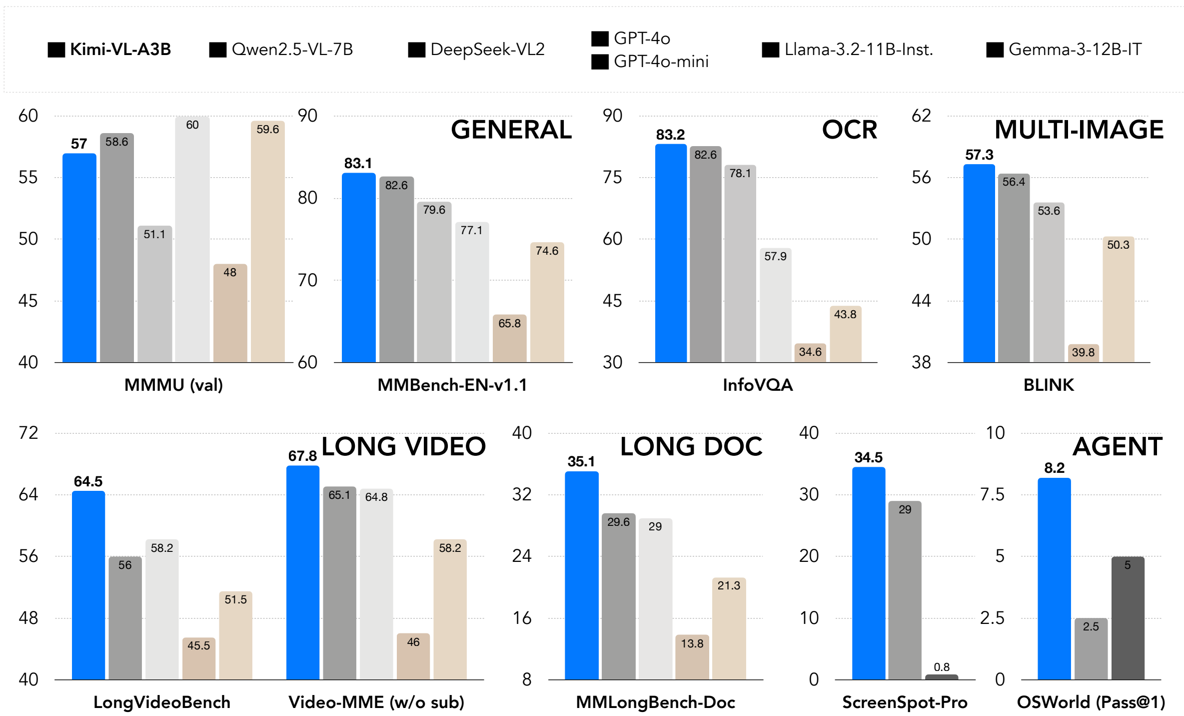


图2：Kimi-VL在广泛基准测试中的表现亮点，如一般基准（MMMU，MMBench）、OCR（InfoVQA）、多图像（BLINK）、长视频（LongVideoBench，Video-MME）、长文档（MMLongBench-Doc）和代理（ScreenSpot-Pro和OSWorld）。详细结果见表3。

1 引言

随着人工智能的快速发展，人们对AI助手的期望已经超越了传统的仅限语言的互动，越来越与我们世界固有的多模态特性相一致。为了更好地理解和互动这些期望，新一代原生多模态模型，如GPT-4o（OpenAI等，2024）和Google Gemini（Gemini团队等，2024），已经出现，具备无缝感知和解释视觉输入与语言处理的能力。最近，由OpenAI o1系列（OpenAI 2024）和Kimi k1.5（K.团队等，2025）开创的先进多模态模型，进一步推动了这些边界，通过对多模态输入进行更深层次和更长时间的推理，从而解决多模态领域中更复杂的问题。

尽管如此，开源社区中大型VLM的发展显著滞后于仅语言模型，特别是在可扩展性、计算效率和高级推理能力等方面。虽然仅语言模型DeepSeek R1（DeepSeek-AI，D. Guo等，2025）已经利用了高效且更具可扩展性的专家混合（MoE）架构，并促进了复杂的长链思维（CoT）推理，但最近的开源VLM，如*e.g.* Qwen2.5-VL（Bai等，2025）和Gemma-3（Gemma Team等，2025），仍然依赖于密集架构，并不支持长CoT推理。对基于MoE的视觉-语言模型的早期探索，如DeepSeek-VL2（Zhiyu Wu等，2024）和Aria（D. Li等，2024），在其他关键维度上也表现出局限性。在架构上，这两种模型仍然采用相对传统的固定大小视觉编码器，限制了它们对多样化视觉输入的适应性。从能力的角度来看，DeepSeek-VL2仅支持有限的上下文长度（4K），而Aria在细粒度视觉任务上表现不足。此外，它们都不支持长时间思考能力。因此，迫切需要一个有效整合结构创新、稳定能力和通过长时间思考增强推理的开源VLM。

鉴于此，我们推出了Kimi-VL，一个面向开源社区的视觉语言模型。从结构上看，Kimi-VL由我们的Moonlight (J. Liu et al. 2025a) MoE语言模型组成，该模型仅激活了2.8B参数（总计16B），并配备了一个400M原生分辨率的MoonViT视觉编码器。在能力方面，如图2所示，Kimi-VL能够稳健地处理多种任务（细粒度感知、数学、大学水平问题、OCR、代理、*etc.*），涵盖广泛的输入形式（单图像、多图像、视频、长文档、*etc.*）。具体而言，它具有以下令人兴奋的能力：

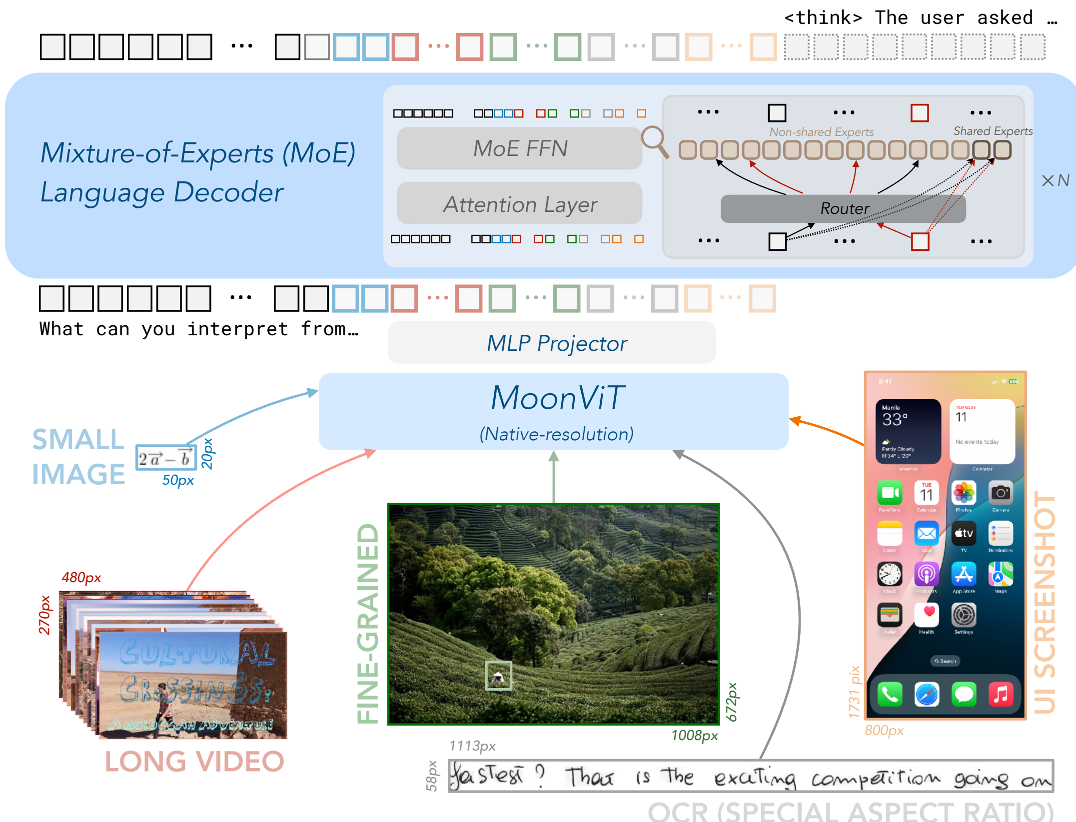


Figure 3: The model architecture of Kimi-VL and Kimi-VL-Thinking, consisting of a MoonViT that allows native-resolution images, an MLP projector, and a Mixture-of-Experts (MoE) language decoder.

1) Kimi-VL is smart: it has comparable text ability against efficient pure-text LLMs; without long thinking, Kimi-VL is already competitive in multimodal reasoning and multi-turn agent benchmarks, *e.g.,* MMMU, MathVista, OSWorld.

2) Kimi-VL processes long: it effectively tackles long-context understanding on various multimodal inputs within its 128K context window, far ahead of similar-scale competitors on long video benchmarks and MMLongBench-Doc.

3) Kimi-VL perceives clear: it shows all-round competitive ability over existing efficient dense and MoE VLMs in various vision-language scenarios: visual perception, visual world knowledge, OCR, high-resolution OS screenshot, *etc.*

Furthermore, with long-CoT activation and reinforcement learning (RL), we introduce the long-thinking version of Kimi-VL, **Kimi-VL-Thinking**, which further substantially improves performance on more complex multimodal reasoning scenarios. Despite its small scale, Kimi-VL-Thinking offers compelling performance on hard reasoning benchmarks (*e.g.,* MMMU, MathVision, MathVista), outperforming many state-of-the-art VLMs with even larger sizes.

2 Approach

2.1 Model Architecture

The architecture of Kimi-VL consists of three parts: a native-resolution vision encoder (MoonViT), an MLP projector, and an MoE language model, as depicted in Figure 3. We introduce each part in this section.

MoonViT: A Native-resolution Vision Encoder We design MoonViT, the vision encoder of Kimi-VL, to natively process images at their varying resolutions, eliminating the need for complex sub-image splitting and splicing operations, as employed in LLaVA-OneVision (B. Li et al. 2024). We incorporate the packing method from NaViT (Dehghani et al. 2023), where images are divided into patches, flattened, and sequentially concatenated into 1D sequences. These preprocessing operations enable MoonViT to share the same core computation operators and optimization as a language model, such as the variable-length sequence attention mechanism supported by FlashAttention (Dao et al. 2022), ensuring non-compromised training throughput for images of varying resolutions.

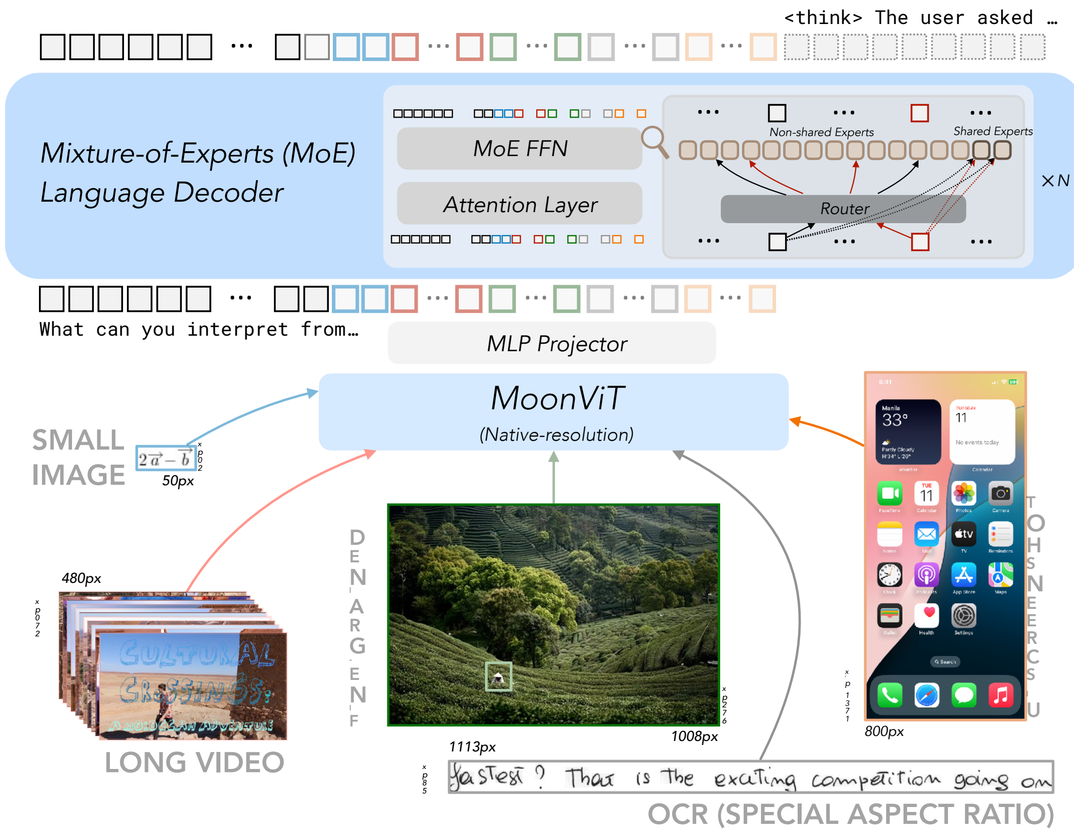


图3：Kimi-VL和Kimi-VL-Thinking的模型架构，包括一个允许原生分辨率图像的MoonViT，一个MLP投影器，以及一个混合专家（MoE）语言解码器。

1) Kimi-VL 很聪明：它在文本能力上与高效的纯文本 LLMs 相当；在没有长时间思考的情况下，Kimi-VL 在多模态推理和多轮代理基准测试中已经具有竞争力，*e.g.*, MMMU, MathVista, OSWorld。

2) Kimi-VL 处理长文本：它在其 128K 上下文窗口内有效地解决了对各种多模态输入的长上下文理解，远远领先于在长视频基准和 MMLongBench-Doc 上的同规模竞争对手。

3) Kimi-VL 清晰地表现出：在各种视觉-语言场景中，它展现了相较于现有高效稠密和 MoE VLM 的全面竞争能力：视觉感知、视觉世界知识、光学字符识别、高分辨率操作系统截图，*etc.*

此外，通过长链思维激活和强化学习（RL），我们引入了Kimi-VL的长思考版本Kimi-VL-Thinking，它在更复杂的多模态推理场景中进一步显著提高了性能。尽管规模较小，Kimi-VL-Thinking在困难的推理基准测试（*e.g.*, MMMU、MathVision、MathVista）上表现出色，超越了许多规模更大的最先进的VLM。

2 方法

2.1 模型架构

Kimi-VL的架构由三个部分组成：一个原生分辨率的视觉编码器（MoonViT）、一个MLP投影器和一个MoE语言模型，如图3所示。我们将在本节中介绍每个部分。

MoonViT：一种原生分辨率视觉编码器 我们设计了MoonViT，这是Kimi-VL的视觉编码器，能够原生处理不同分辨率的图像，消除了像LLaVA-OneVision（B. Li等，2024）中使用的复杂子图像拆分和拼接操作的需要。我们采用了NaViT（Dehghani等，2023）中的打包方法，将图像划分为补丁，展平，并顺序连接成1D序列。这些预处理操作使MoonViT能够与语言模型共享相同的核心计算操作和优化，例如由FlashAttention（Dao等，2022）支持的可变长度序列注意机制，确保对不同分辨率图像的训练吞吐量不受影响。

Figure 4: The pre-training stages of Kimi-VL consume a total of 4.4T tokens after text-only pre-training of its language model. To preserve text abilities, all stages that update the language model are joint training stages.

MoonViT is initialized from and continually pre-trained on SigLIP-SO-400M (Zhai et al. 2023), which originally employs learnable fixed-size absolute positional embeddings to encode spatial information. While we interpolate these original position embeddings to better preserve SigLIP's capabilities, these interpolated embeddings become increasingly inadequate as image resolution increases. To address this limitation, we incorporate 2D rotary positional embedding (RoPE) (J. Su et al. 2023) across the height and width dimensions, which improves the representation of fine-grained positional information, especially in high-resolution images. These two positional embedding approaches work together to encode spatial information for our model and seamlessly integrate with the flattening and packing procedures. This integration enables MoonViT to efficiently process images of varying resolutions within the same batch. The resulting continuous image features are then forwarded to the MLP projector and, ultimately, to the MoE language model for subsequent training stages.

MLP Projector We employ a two-layer MLP to bridge the vision encoder (MoonViT) and the LLM. Specifically, we first use a pixel shuffle operation to compress the spatial dimension of the image features extracted by MoonViT, performing 2×2 downsampling in the spatial domain and correspondingly expanding the channel dimension. We then feed the pixel-shuffled features into a two-layer MLP to project them into the dimension of LLM embeddings.

Mixture-of-Experts (MoE) Language Model The language model of Kimi-VL utilizes our Moonlight model (J. Liu et al. 2025a), an MoE language model with 2.8B activated parameters, 16B total parameters, and an architecture similar to DeepSeek-V3 (DeepSeek-AI, A. Liu, et al. 2025). For our implementation, we initialize from an intermediate checkpoint in Moonlight's pre-training stage—one that has processed 5.2T tokens of pure text data and activated an 8192-token (8K) context length. We then continue pre-training it using a joint recipe of multimodal and text-only data totaling 2.3T tokens, as detailed in Sec. 2.3.

2.2 Muon Optimizer

We use an enhanced Muon optimizer (J. Liu et al. 2025b) for model optimization. Compared to the original Muon optimizer (Jordan et al. 2024), we add weight decay and carefully adjust the per-parameter update scale. Additionally, we develop a distributed implementation of Muon following the ZeRO-1 (Rajbhandari et al. 2020) optimization strategy, which achieves optimal memory efficiency and reduced communication overhead while preserving the algorithm's mathematical properties. This enhanced Muon optimizer is used throughout the entire training process to optimize all model parameters, including the vision encoder, the projector, and the language model.

2.3 Pre-Training Stages

As illustrated in Figure 4 and Table 1, after loading the intermediate language model discussed above, Kimi-VL's pre-training comprises a total of 4 stages consuming 4.4T tokens overall: first, standalone ViT training to establish a robust native-resolution visual encoder, followed by three joint training stages (pre-training, cooldown, and long-context activation) that simultaneously enhance the model's language and multimodal capabilities. The details are as follows.

ViT Training Stages The MoonViT is trained on image-text pairs, where the text components consist of a variety of targets: image alt texts, synthetic captions, grounding bboxes, and OCR texts. The training incorporates two objectives: a SigLIP (Zhai et al. 2023) loss $\mathcal{L}_{siglip}$ (a variant of contrastive loss) and a cross-entropy loss $\mathcal{L}_{caption}$ for caption generation conditioned on input images. Following CoCa's approach (J. Yu et al. 2022), the final loss function is formulated as $\mathcal{L} = \mathcal{L}_{siglip} + \lambda\mathcal{L}_{caption}$, where $\lambda = 2$. Specifically, the image and text encoders compute the contrastive loss, while the text decoder performs next-token prediction (NTP) conditioned on features from the image encoder. To accelerate training, we initialized both encoders with SigLIP SO-400M (Zhai et al. 2023) weights and implemented a progressive resolution sampling strategy to gradually allow larger size; the text decoder is initialized from a tiny decoder-only language model. During training, we observed an emergence in the caption loss while scaling up OCR

图4：Kimi-VL的预训练阶段在仅进行文本预训练后消耗了总共4.4T的标记。为了保持文本能力，所有更新语言模型的阶段都是联合训练阶段。

MoonViT 是从 SigLIP-SO-400M（Zhai 等，2023）初始化并持续进行预训练的，该模型最初采用可学习的固定大小绝对位置嵌入来编码空间信息。虽然我们对这些原始位置嵌入进行插值以更好地保留 SigLIP 的能力，但随着图像分辨率的增加，这些插值嵌入变得越来越不足。为了解决这一限制，我们在高度和宽度维度上引入了 2D 旋转位置嵌入（RoPE）（J. Su 等，2023），这改善了细粒度位置信息的表示，特别是在高分辨率图像中。这两种位置嵌入方法共同作用于我们的模型，以编码空间信息，并与展平和打包过程无缝集成。这种集成使 MoonViT 能够高效处理同一批次中不同分辨率的图像。生成的连续图像特征随后被转发到 MLP 投影器，最终传递给 MoE 语言模型以进行后续训练阶段。

MLP 投影器 我们采用一个两层的 MLP 来连接视觉编码器 (MoonViT) 和 LLM。具体来说，我们首先使用像素重排操作来压缩 MoonViT 提取的图像特征的空间维度，在空间域中执行 2x2 下采样，并相应地扩展通道维度。然后，我们将像素重排后的特征输入到一个两层的 MLP 中，以将其投影到 LLM 嵌入的维度。

混合专家（MoE）语言模型 Kimi-VL 的语言模型利用了我们的 Moonlight 模型（J. Liu 等，2025a），这是一个具有 28 亿激活参数、160 亿总参数的 MoE 语言模型，其架构类似于 DeepSeek-V3（DeepSeek-AI，A. Liu 等，2025）。在我们的实现中，我们从 Moonlight 预训练阶段的一个中间检查点初始化——该检查点处理了 5.2T 纯文本数据，并激活了 8192 令牌（8K）的上下文长度。然后，我们继续使用多模态和仅文本数据的联合配方进行预训练，总计 2.3T 令牌，详细信息见第 2.3 节。

2.2 介子优化器

我们使用增强版的Muon优化器（J. Liu等，2025b）进行模型优化。与原始的Muon优化器（Jordan等，2024）相比，我们增加了权重衰减，并仔细调整每个参数的更新规模。此外，我们开发了一个遵循ZeRO-1（Rajbhandari等，2020）优化策略的Muon分布式实现，达到了最佳的内存效率和减少的通信开销，同时保留了算法的数学特性。这个增强版的Muon优化器在整个训练过程中用于优化所有模型参数，包括视觉编码器、投影器和语言模型。

2.3 预训练阶段

如图4和表1所示，在加载上述中间语言模型后，Kimi-VL的预训练总共包括4个阶段，整体消耗4.4T的tokens：首先是独立的ViT训练，以建立一个强大的原生分辨率视觉编码器，随后是三个联合训练阶段（预训练、冷却和长上下文激活），同时增强模型的语言和多模态能力。具体细节如下。

ViT 训练阶段 MoonViT 在图像-文本对上进行训练，其中文本组件由多种目标组成：图像替代文本、合成标题、定位框和 OCR 文本。训练包含两个目标：SigLIP (Zhai et al. 2023) 损失 $\mathcal{L}_{siglip}$（一种对比损失的变体）和用于基于输入图像生成标题的交叉熵损失 $\mathcal{L}_{caption}$。遵循 CoCa 的方法 (J. Yu et al. 2022)，最终损失函数被表述为 $\mathcal{L} = \mathcal{L}_{siglip} + \lambda\mathcal{L}_{caption}$，其中 $\lambda = 2$。具体而言，图像和文本编码器计算对比损失，而文本解码器在图像编码器的特征基础上执行下一个标记预测 (NTP)。为了加速训练，我们用 SigLIP SO-400M (Zhai et al. 2023) 权重初始化了两个编码器，并实施了渐进分辨率采样策略以逐步允许更大的尺寸；文本解码器则从一个小型解码器语言模型初始化。在训练过程中，我们观察到在扩大 OCR 时标题损失的出现。

Table 1: Overview of training stages: data composition, token volumes, sequence lengths, and trainable components.

Stages	**ViT Training**	**Joint Pre-training**	**Joint Cooldown**	**Joint Long-context**
Data	Alt text Synthesis Caption Grounding OCR	+ Text, Knowledge Interleaving Video, Agent	+ High-quality Text High-quality Multimodal Academic Sources	+ Long Text Long Video Long Document
Tokens	2T + 0.1T	1.4T	0.6T	0.3T
Sequence length	8192	8192	8192	32768->131072
Training	ViT	ViT & LLM	ViT & LLM	ViT & LLM

Table 2: **Needle-in-a-Haystack (NIAH)** test on text/video haystacks, where needles are uniformly distributed at various positions within the haystack. We report recall accuracy across different haystack lengths up to 131,072 tokens (128K).

Haystack Length	$(0, 2048]$	$(2048, 4096]$	$(4096, 8192]$	$(8192, 16384]$	$(16384, 32768]$	$(32768, 65536]$	$(65536, 131072]$
- *text haystack*	100.0	100.0	100.0	100.0	100.0	100.0	87.0
- *video haystack*	100.0	100.0	100.0	100.0	100.0	100.0	91.7

data, indicating that the text decoder had developed some OCR capabilities. After training the ViT in the CoCa-alike stage with 2T tokens, we align the MoonViT to the MoE language model using another 0.1T tokens, where only MoonViT and MLP projector are updated. This alignment stage significantly reduces the initial perplexity of MoonViT embeddings in the language model, allowing a smoother joint pre-training stage as follows.

Joint Pre-training Stage In the joint pre-training stage, we train the model with a combination of pure text data (sampled from the same distribution as the initial language model) and a variety of multimodal data (as discussed in Sec. 3.1). We continue training from the loaded LLM checkpoint using the same learning rate scheduler, consuming an additional 1.4T tokens. The initial steps utilize solely language data, after which the proportion of multimodal data gradually increases. Through this progressive approach and the previous alignment stage, we observe that joint pre-training preserves the model's language capabilities while successfully integrating visual comprehension abilities.

Joint Cooldown Stage The stage following the pre-training stage is a multimodal cooldown phase, where the model is continue trained with high-quality language and multimodal datasets to ensure superior performance. For the language part, through empirical investigation, we observe that the incorporation of synthetic data during the cooling phase yields significant performance improvements, particularly in mathematical reasoning, knowledge-based tasks, and code generation. The general text components of the cooldown dataset are curated from high-fidelity subsets of the pre-training corpus. For math, knowledge, and code domains, we employ a hybrid approach: utilizing selected pre-training subsets while augmenting them with synthetically generated content. Specifically, we leverage existing mathematical knowledge and code corpora as source material to generate question-answer (QA) pairs through a proprietary language model, implementing rejection sampling techniques to maintain quality standards (Yue, Qu, et al. 2023; D. Su et al. 2024). These synthesized QA pairs undergo comprehensive validation before being integrated into the cooldown dataset. For the multimodal part, in addition to the two strategies as employed in text cooldown data preparation, *i.e.* question-answer synthesis and high-quality subset replay, to allow more comprehensive visual-centric perception and understanding (B. Li et al. 2024; Tong et al. 2024; J. Guo et al. 2024), we filter and rewrite a variety of academic visual or vision-language data sources to QA pairs. Unlike post-training stages, these language and multimodal QA pairs in the cooldown stage are only included for activating specific abilities and henceforth facilitating learning high-quality data, thus, we keep their ratio at a low portion to avoid overfitting these QA patterns. The joint cooldown stage significantly improves both language and multimodal abilities of the model.

Joint Long-context Activation Stage In the final pre-training stage, we extend the context length of the model from 8192 (8K) to 131072 (128K), with the inverse frequency of its RoPE (J. Su et al. 2023) embeddings reset from 50,000 to 800,000. The joint long-context stage is conducted in two sub-stages, where each one extends the model's context length by four times. For data composition, we filter and upsample the ratio of long data to 25% in each sub-stage, while using the remaining 75% tokens to replay shorter data in its previous stage; our exploration confirms that this composition allows the model to effectively learn long-context understanding while maintaining short-context ability.

表1：训练阶段概述：数据组成、标记量、序列长度和可训练组件。

Stages	ViT Training	Joint Pre-training	Joint Cooldown	Joint Long-context
Data	Alt text Synthesis Caption Grounding OCR	+ Text, Knowledge Interleaving Video, Agent	+ High-quality Text High-quality Multimodal Academic Sources	+ Long Text Long Video Long Document
Tokens	2T + 0.1T	1.4T	0.6T	0.3T
Sequence length	8192	8192	8192	32768->131072
Training	ViT	ViT & LLM	ViT & LLM	ViT & LLM

表2：针在干草堆中（NIAH）测试文本/视频干草堆，其中针均匀分布在干草堆的不同位置。我们报告了在不同干草堆长度（最多131,072个标记（128K））下的召回准确率。

Haystack Length	(0, 2048]	(2048, 4096]	(4096, 8192]	(8192, 16384]	(16384, 32768]	(32768, 65536]	(65536, 131072]
- *text haystack*	100.0	100.0	100.0	100.0	100.0	100.0	87.0
- *video haystack*	100.0	100.0	100.0	100.0	100.0	100.0	91.7

数据表明，文本解码器已经发展出了一些OCR能力。在使用2T标记训练ViT的CoCa类似阶段后，我们使用另外0.1T标记将MoonViT与MoE语言模型对齐，此时仅更新MoonViT和MLP投影器。这个对齐阶段显著降低了MoonViT嵌入在语言模型中的初始困惑度，从而使得后续的联合预训练阶段更加顺利。

联合预训练阶段 在联合预训练阶段，我们使用纯文本数据（从与初始语言模型相同的分布中抽样）和各种多模态数据（如第3.1节所讨论）结合来训练模型。我们从加载的LLM检查点继续训练，使用相同的学习率调度器，消耗额外的1.4T标记。最初的步骤仅使用语言数据，之后多模态数据的比例逐渐增加。通过这种渐进的方法和之前的对齐阶段，我们观察到联合预训练保留了模型的语言能力，同时成功整合了视觉理解能力。

联合冷却阶段 在预训练阶段之后的阶段是一个多模态冷却阶段，在此阶段，模型继续使用高质量的语言和多模态数据集进行训练，以确保卓越的性能。通过实证研究，我们观察到在冷却阶段引入合成数据显著提高了性能，特别是在数学推理、基于知识的任务和代码生成方面。冷却数据集的一般文本组件来自于预训练语料库的高保真子集。对于数学、知识和代码领域，我们采用混合方法：利用选定的预训练子集，同时用合成生成的内容进行增强。具体而言，我们利用现有的数学知识和代码语料作为源材料，通过专有语言模型生成问答（QA）对，实施拒绝采样技术以保持质量标准（Yue, Qu等，2023；D. Su等，2024）。这些合成的QA对在整合到冷却数据集之前经过全面验证。对于多模态部分，除了在文本冷却数据准备中采用的两种策略，即*i.e.*问答合成和高质量子集重放，以允许更全面的视觉中心感知和理解（B. Li等，2024；Tong等，2024；J. Guo等，2024），我们过滤并重写各种学术视觉或视觉语言数据源为QA对。与后训练阶段不同，这些冷却阶段的语言和多模态QA对仅用于激活特定能力，从而促进学习高质量数据，因此，我们将其比例保持在较低的部分，以避免过拟合这些QA模式。联合冷却阶段显著提高了模型的语言和多模态能力。

联合长上下文激活阶段 在最终的预训练阶段，我们将模型的上下文长度从8192（8K）扩展到131072（128K），其RoPE（J. Su et al. 2023）嵌入的逆频率从50000重置为800000。联合长上下文阶段分为两个子阶段，每个子阶段将模型的上下文长度扩展四倍。对于数据组成，我们在每个子阶段中过滤并上采样长数据的比例至25%，同时使用剩余的75%标记在其前一阶段中重放短数据；我们的探索确认这种组成方式使模型能够有效学习长上下文理解，同时保持短上下文能力。

Figure 5: The post-training stages of Kimi-VL and Kimi-VL-Thinking, including two stages of joint SFT in 32K and 128K context, and further long-CoT SFT and RL stages to activate and enhance long thinking abilities.

To allow the model to activate long-context abilities on both pure-text and multimodal inputs, the long data used in Kimi-VL's long-context activation consists of not only long text, but also long multimodal data, including long interleaved data, long videos, and long documents. Similar as cooldown data, we also synthesize a small portion of QA pairs to augment the learning efficiency of long-context activation. After the long-context activations, the model can pass needle-in-a-haystack (NIAH) evaluations with either long pure-text or long video haystack, proving its versatile long-context ability. We provide the NIAH recall accuracy on various range of context length up to 128K in Table 2.

2.4 Post-Training Stages

Joint Supervised Fine-tuning (SFT) In this phase, we fine-tune the base model of Kimi-VL with instruction-based fine-tuning to enhance its ability to follow instructions and engage in dialogue, culminating in the creation of the interactive Kimi-VL model. This is achieved by employing the ChatML format (Openai, 2024), which allows for a targeted instruction optimization while maintaining architectural consistency with Kimi-VL. We optimize the language model, MLP projector, and vision encoder using a mixture of pure-text and vision-language SFT data, which will be described in Sec 3.2. Supervision is applied only to answers and special tokens, with system and user prompts being masked. The model is exposed to a curated set of multimodal instruction-response pairs, where explicit dialogue role tagging, structured injection of visual embeddings, and preservation of cross-modal positional relationships are ensured through the format-aware packing. Additionally, to guarantee the model's comprehensive proficiency in dialogue, we incorporate a mix of multimodal data and pure text dialogue data used in Moonlight, ensuring its versatility across various dialogue scenarios.

We first train the model at the sequence length of 32k tokens for 1 epoch, followed by another epoch at the sequence length of 128k tokens. In the first stage (32K), the learning rate decays from 2×10^{-5} to 2×10^{-6}, before it re-warmups to 1×10^{-5} in the second stage (128K) and finally decays to 1×10^{-6}. To improve training efficiency, we pack multiple training examples into each single training sequence.

Long-CoT Supervised Fine-Tuning With the refined RL prompt set, we employ prompt engineering to construct a small yet high-quality long-CoT warmup dataset, containing accurately verified reasoning paths for both text and image inputs. This approach resembles rejection sampling (RS) but focuses on generating long-CoT reasoning paths through prompt engineering. The resulting warmup dataset is designed to encapsulate key cognitive processes that are fundamental to human-like reasoning, such as **planning**, where the model systematically outlines steps before execution; **evaluation**, involving critical assessment of intermediate steps; **reflection**, enabling the model to reconsider and refine its approach; and **exploration**, encouraging consideration of alternative solutions. By performing a lightweight SFT on this warm-up dataset, we effectively prime the model to internalize these multimodal reasoning strategies. As a result, the fine-tuned long-CoT model demonstrates improved capability in generating more detailed and logically coherent responses, which enhances its performance across diverse reasoning tasks.

Reinforcement Learning To further advance the model's reasoning abilities, we then train the model with reinforcement learning (RL), enabling the model to autonomously generate structured CoT rationales. Specifically, similar as Kimi k1.5 (K. Team et al. 2025), we adopt a variant of online policy mirror descent as our RL algorithm, which iteratively refines the policy model π_θ to improve its problem-solving accuracy. During the i-th training iteration, we treat the current model as a reference policy model and optimize the following objective, regularized by relative entropy to stabilize policy updates:

$$\max_\theta \mathbb{E}_{(x,y^*)\sim\mathcal{D}} \left[\mathbb{E}_{(y,z)\sim\pi_\theta}\left[r(x,y,y^*)\right] - \tau \mathrm{KL}(\pi_\theta(x)||\pi_{\theta_i}(x))\right] , \tag{1}$$

where r is a reward model that justifies the correctness of the proposed answer y for the given problem x, by assigning a value $r(x,y,y^*) \in \{0,1\}$ based on the ground truth y^*, and $\tau > 0$ is a parameter controlling the degree of regularization.

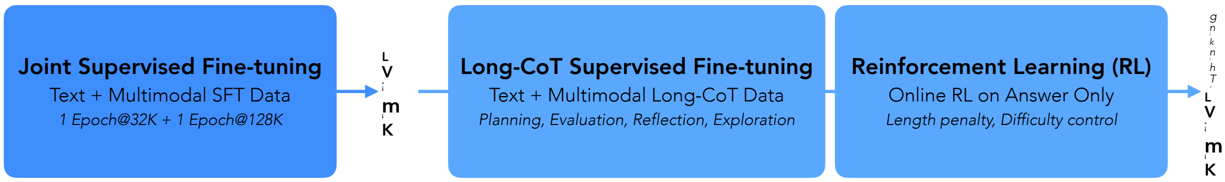


图5：Kimi-VL 和 Kimi-VL-Thinking 的后训练阶段，包括 32K 和 128K 上下文的两个联合 SFT 阶段，以及进一步的长 CoT SFT 和 RL 阶段，以激活和增强长时间思考能力。

为了使模型能够在纯文本和多模态输入上激活长上下文能力，Kimi-VL的长上下文激活所使用的长数据不仅包括长文本，还包括长多模态数据，包括长交错数据、长视频和长文档。与冷却数据类似，我们还合成了一小部分问答对，以增强长上下文激活的学习效率。在长上下文激活之后，模型可以通过长纯文本或长视频的针在干草堆（NIAH）评估，证明其多功能的长上下文能力。我们在表2中提供了各种上下文长度（最高可达128K）的NIAH召回准确率。

2.4 训练后阶段

联合监督微调 (SFT) 在这个阶段，我们通过基于指令的微调来优化 Kimi-VL 的基础模型，以增强其遵循指令和参与对话的能力，最终创建出互动式 Kimi-VL 模型。这是通过采用 ChatML 格式 (Openai, 2024) 实现的，该格式允许在保持 Kimi-VL 架构一致性的同时进行针对性的指令优化。我们使用纯文本和视觉语言 SFT 数据的混合来优化语言模型、MLP 投影器和视觉编码器，这将在第 3.2 节中描述。监督仅应用于答案和特殊标记，系统和用户提示被屏蔽。模型接触到一组精心策划的多模态指令-响应对，其中通过格式感知打包确保了明确的对话角色标记、视觉嵌入的结构化注入以及跨模态位置关系的保留。此外，为了确保模型在对话中的全面能力，我们结合了 Moonlight 中使用的多模态数据和纯文本对话数据，确保其在各种对话场景中的多样性。

我们首先在32k个标记的序列长度上训练模型1个周期，然后在128k个标记的序列长度上再训练1个周期。在第一阶段（32K），学习率从2×10^{-5} 下降到2×10^{-6}，然后在第二阶段（128K）重新升温到1×10^{-5}，最后下降到1×10^{-6}。为了提高训练效率，我们将多个训练示例打包到每个单一的训练序列中。

长-CoT 监督微调 通过精炼的 RL 提示集，我们采用提示工程构建一个小而高质量的长-CoT 热身数据集，包含针对文本和图像输入的准确验证推理路径。这种方法类似于拒绝采样（RS），但专注于通过提示工程生成长-CoT 推理路径。生成的热身数据集旨在概括对人类推理至关重要的关键认知过程，例如规划，其中模型在执行之前系统地列出步骤；评估，涉及对中间步骤的批判性评估；反思，使模型能够重新考虑和完善其方法；以及探索，鼓励考虑替代解决方案。通过对这个热身数据集进行轻量级的 SFT，我们有效地使模型内化这些多模态推理策略。因此，微调后的长-CoT 模型在生成更详细和逻辑连贯的响应方面表现出更强的能力，从而提升了其在各种推理任务中的表现。

强化学习 为了进一步提升模型的推理能力，我们接着使用强化学习（RL）训练模型，使其能够自主生成结构化的CoT推理。具体而言，类似于Kimi k1.5（K. Team等，2025），我们采用在线策略镜像下降的变体作为我们的RL算法，迭代地优化策略模型 π_θ 以提高其问题解决的准确性。在第 i 次训练迭代中，我们将当前模型视为参考策略模型，并优化以下目标，通过相对熵进行正则化以稳定策略更新：

$$\max_\theta \mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\mathbb{E}_{(y,z)\sim\pi_\theta}\left[r(x,y,y^*)\right] - \tau \mathrm{KL}(\pi_\theta(x)||\pi_{\theta_i}(x))\right] , \tag{1}$$

其中 r 是一个奖励模型，它通过根据真实情况 y^* 为给定问题 x 提出的答案 y 的正确性分配一个值 $r(x, y, y^*) \in \{0, 1\}$ 来证明其正确性，而 $\tau > 0$ 是一个控制正则化程度的参数。

Each training iteration begins by sampling a problem batch from the dataset $\mathcal{D}$, and the model parameters are updated to θ_{i+1} using the policy gradient derived from (1), with the optimized policy model subsequently assuming the role of reference policy for the subsequent iteration. To enhance RL training efficiency, we implement a length-based reward to penalize excessively long responses, mitigating the overthinking problem where the model generates redundant reasoning chains. Besides, we employ two sampling strategies including curriculum sampling and prioritized sampling, which leverage difficulty labels and per-instance success rates to focus training effort on the most pedagogically valuable examples, thereby optimizing the learning trajectory and improving training efficiency.

Through large-scale reinforcement learning training, we can derive a model that harnesses the strengths of both basic prompt-based CoT reasoning and sophisticated planning-enhanced CoT approaches. During inference, the model maintains standard autoregressive sequence generation, eliminating the deployment complexities associated with specialized planning algorithms that require parallel computation. Simultaneously, the model develops essential meta-reasoning abilities including error detection, backtracking, and iterative solution refinement by effectively utilizing the complete history of explored reasoning paths as contextual information. With endogenous learning from its complete reasoning trace history, the model can effectively encode planned search procedures into its parametric knowledge.

2.5 Infrastructure

Storage We utilize S3 (Amazon Web Services 2023) compatible object storage from cloud service vendors to store our visual-text data. To minimize the time between data preparation and model training, we store visual data in its original format and have developed an efficient and flexible data loading system. This system provides several key benefits:

- Supports on-the-fly data shuffling, mixing, tokenization, loss masking and packing during training, allowing us to adjust data proportions as needed;
- Enables random augmentation of both visual and text data, while preserving the correctness of 2D coordinate and orientation information during transformations;
- Ensures reproducibility by strictly controlling random states and other states across different data loader workers, guaranteeing that any interrupted training can be resumed seamlessly—the data sequence after resumption remains identical to an uninterrupted run;
- Delivers high-performance data loading: through multiple caching strategies, our system reliably supports training on large scale clusters while maintaining controlled request rates and throughput to the object storage.

Additionally, to ensure consistent dataset quality control, we developed a centralized platform for data registration, visualization, compiling statistics, synchronizing data across cloud storage systems, and managing dataset lifecycles.

Parallelism We adopt a 4D parallelism strategy—Data Parallelism (S. Li et al. 2020), Expert Parallelism (Fedus et al. 2022), Pipeline Parallelism (Y. Huang et al. 2019; Narayanan et al. 2021), and Context Parallelism (Jacobs et al. 2023; H. Liu et al. 2023)—to accelerate the speed of Kimi-VL . After optimizing parallel strategies, the resulting training throughput of our model is around 60% higher than a 7B dense VLM (*e.g.* VLMs based on Qwen2.5-7B).

- **Data Parallelism (DP).** DP replicates the model across multiple devices, each processing different micro-batches. This setup allows larger effective batch sizes by simply increasing the number of devices.
- **Expert Parallelism (EP).** EP distributes expert modules in the MoE layer across multiple devices. When combined with DP, experts on a given device can handle tokens from different DP groups, enhancing computational efficiency.
- **Pipeline Parallelism (PP).** PP splits the model into multiple layer-based stages. To minimize pipeline bubbles, we allocate the Vision Tower (VT) and several decoder layers to the first stage, place the output layer and additional decoder layers in the last stage, and distribute the remaining decoder layers evenly across intermediate stages based on their time overhead.
- **Context Parallelism (CP).** CP addresses long-sequence training by splitting sequences across different CP ranks in conjunction with flash attention (Dao et al. 2022). This substantially reduces peak memory usage and relieves the memory pressure from attention computations.

Beyond these four parallel strategies, we incorporate ZeRO1 (Rajbhandari et al. 2020) and Selective Checkpointing Activation (T. Chen et al. 2016; Korthikanti et al. 2022) to further optimize memory usage. ZeRO1 reduces optimizer state overhead by using a distributed optimizer while avoiding extra communication costs. Selective Checkpointing Activation trades time for space by recomputing only those layers that have low time overhead but high memory consumption, striking a balance between computation efficiency and memory demands. For extremely long sequences, we expand recomputation to a broader set of layers to prevent out-of-memory errors.

每次训练迭代开始时，从数据集 $\mathcal{D}$ 中抽取一个问题批次，并使用从 (1) 中推导出的策略梯度将模型参数更新为 θ_{i+1}，随后优化后的策略模型作为后续迭代的参考策略。为了提高强化学习训练的效率，我们实施了一种基于长度的奖励，以惩罚过长的响应，从而减轻模型生成冗余推理链的过度思考问题。此外，我们采用了两种采样策略，包括课程采样和优先采样，这两种策略利用难度标签和每个实例的成功率，将训练重点放在最具教学价值的示例上，从而优化学习轨迹并提高训练效率。

通过大规模强化学习训练，我们可以推导出一个模型，该模型结合了基本基于提示的链式推理和复杂的规划增强链式推理方法的优点。在推理过程中，该模型保持标准的自回归序列生成，消除了与需要并行计算的专门规划算法相关的部署复杂性。同时，该模型通过有效利用探索过的推理路径的完整历史作为上下文信息，发展了基本的元推理能力，包括错误检测、回溯和迭代解决方案的优化。通过对其完整推理轨迹历史的内生学习，该模型能够有效地将规划搜索过程编码到其参数知识中。

2.5 基础设施

存储 我们利用与 S3（亚马逊网络服务 2023）兼容的对象存储来自云服务供应商来存储我们的视觉-文本数据。为了最小化数据准备与模型训练之间的时间，我们以原始格式存储视觉数据，并开发了一个高效且灵活的数据加载系统。该系统提供了几个关键好处：

- 支持在训练过程中动态数据洗牌、混合、标记化、损失掩蔽和打包，使我们能够根据需要调整数据比例；
- 启用视觉和文本数据的随机增强，同时在变换过程中保持2D坐标和方向信息的正确性；
- 通过严格控制不同数据加载器工作者的随机状态和其他状态，确保可重复性，保证任何中断的训练可以无缝恢复——恢复后的数据序列与未中断的运行保持一致；
- 提供高性能数据加载：通过多种缓存策略，我们的系统可靠地支持在大规模集群上进行训练，同时保持对对象存储的请求速率和吞吐量的控制。

此外，为了确保数据集质量控制的一致性，我们开发了一个集中平台，用于数据注册、可视化、统计汇编、在云存储系统之间同步数据以及管理数据集生命周期。

并行性 我们采用了一种4D并行策略——数据并行（S. Li等，2020），专家并行（Fedus等，2022），流水线并行（Y. Huang等，2019；Narayanan等，2021），以及上下文并行（Jacobs等，2023；H. Liu等，2023）——以加速Kimi-VL的速度。在优化并行策略后，我们模型的训练吞吐量比7B稠密VLM（基于Qwen2.5-7B的 *e.g.* VLM）高出约60%。

- 数据并行性（DP）。DP在多个设备上复制模型，每个设备处理不同的微批次。此设置通过简单地增加设备数量来允许更大的有效批量大小。
- 专家并行性（EP）。EP在MoE层中将专家模块分布到多个设备上。当与DP结合时，给定设备上的专家可以处理来自不同DP组的令牌，从而提高计算效率。
- 管道并行性（PP）。PP将模型分成多个基于层的阶段。为了最小化管道气泡，我们将视觉塔（VT）和几个解码器层分配给第一阶段，将输出层和额外的解码器层放置在最后阶段，并根据它们的时间开销将剩余的解码器层均匀分布在中间阶段。
- 上下文并行性（CP）。CP通过在不同的CP等级之间拆分序列，并结合闪存注意力（Dao等，2022），来解决长序列训练的问题。这大大减少了峰值内存使用，并减轻了注意力计算的内存压力。

除了这四种并行策略，我们还结合了ZeRO1（Rajbhandari等，2020）和选择性检查点激活（T. Chen等，2016；Korthikanti等，2022）来进一步优化内存使用。ZeRO1通过使用分布式优化器来减少优化器状态开销，同时避免额外的通信成本。选择性检查点激活通过仅重新计算那些时间开销低但内存消耗高的层来以时间换取空间，在计算效率和内存需求之间取得平衡。对于极长的序列，我们将重新计算扩展到更广泛的层集，以防止内存溢出错误。

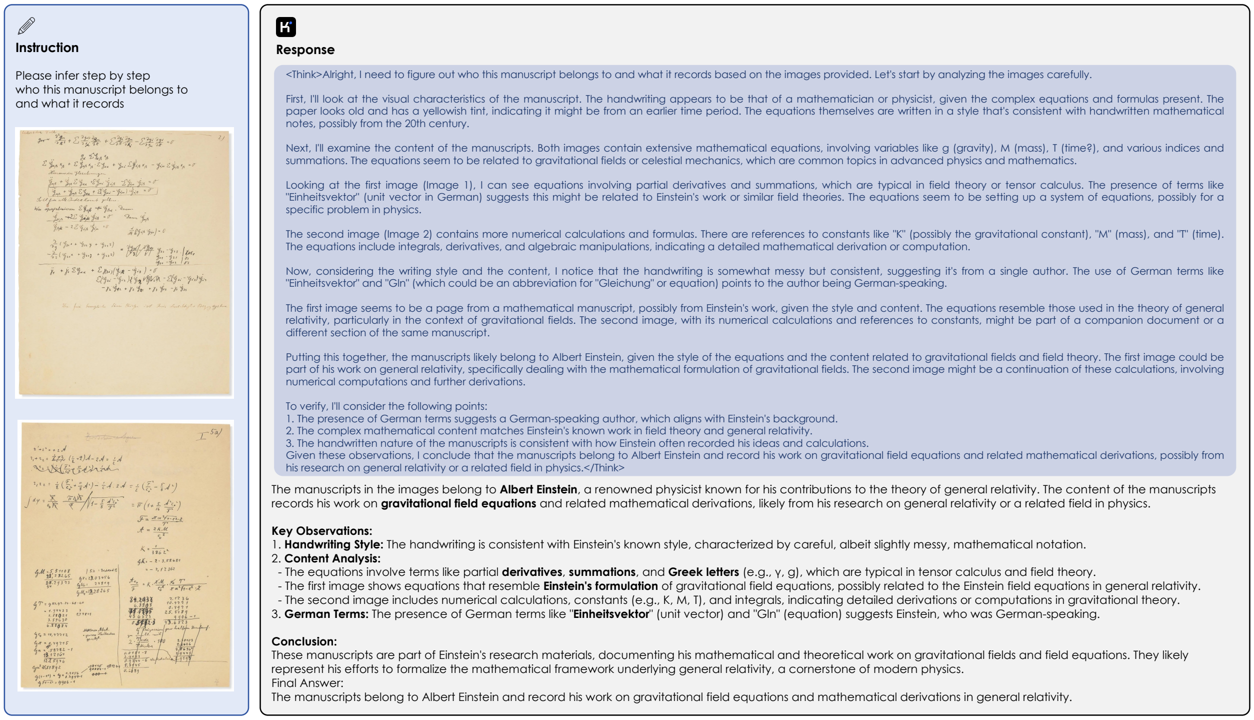


Figure 6: Manuscript reasoning visualization. Kimi-VL-Thinking demonstrates the ability to perform historical and scientific inference by analyzing handwritten manuscripts step by step. In this example, our model identifies the author as Albert Einstein based on handwriting style, content analysis, and language cues. It reasons that the manuscripts relate to gravitational field equations, consistent with Einstein's contributions to general relativity.

3 Data Construction

3.1 Pre-Training Data

Our multimodal pre-training corpus is designed to provide high-quality data that enables models to process and understand information from multiple modalities, including text, images, and videos. To this end, we have also curated high-quality data from six categories – caption, interleaving, OCR, knowledge, video, and agent – to form the corpus.

When constructing our training corpus, we developed several multimodal data processing pipelines to ensure data quality, encompassing filtering, synthesis, and deduplication. Establishing an effective multimodal data strategy is crucial during the joint training of vision and language, as it both preserves the capabilities of the language model and facilitates alignment of knowledge across diverse modalities.

We provide a detailed description of these sources in this section, which is organized into the following categories:

Caption Data Our caption data provides the model with fundamental modality alignment and a broad range of world knowledge. By incorporating caption data, the multimodal LLM gains wider world knowledge with high learning efficiency. We have integrated various open-source Chinese and English caption datasets like (Schuhmann et al. 2022; Gadre et al. 2024) and also collected substantial in-house caption data from multiple sources. However, throughout the training process, we strictly limit the proportion of synthetic caption data to mitigate the risk of hallucination stemming from insufficient real-world knowledge.

For general caption data, we follow a rigorous quality control pipeline that avoids duplication and maintain high image-text correlation. We also vary image resolution during pre-training to ensure that the vision tower remains effective when processing images of both high- and low-resolution.

Image-text Interleaving Data During the pre-training phase, the model benefits from interleaving data for many aspects. For example, multi-image comprehension ability can be boosted by interleaving data; interleaving data always provides detailed knowledge for the given image; a longer multimodal context learning ability can also be gained by interleaving data. What's more, we also find that interleaving data can contribute positively to maintaining the model's

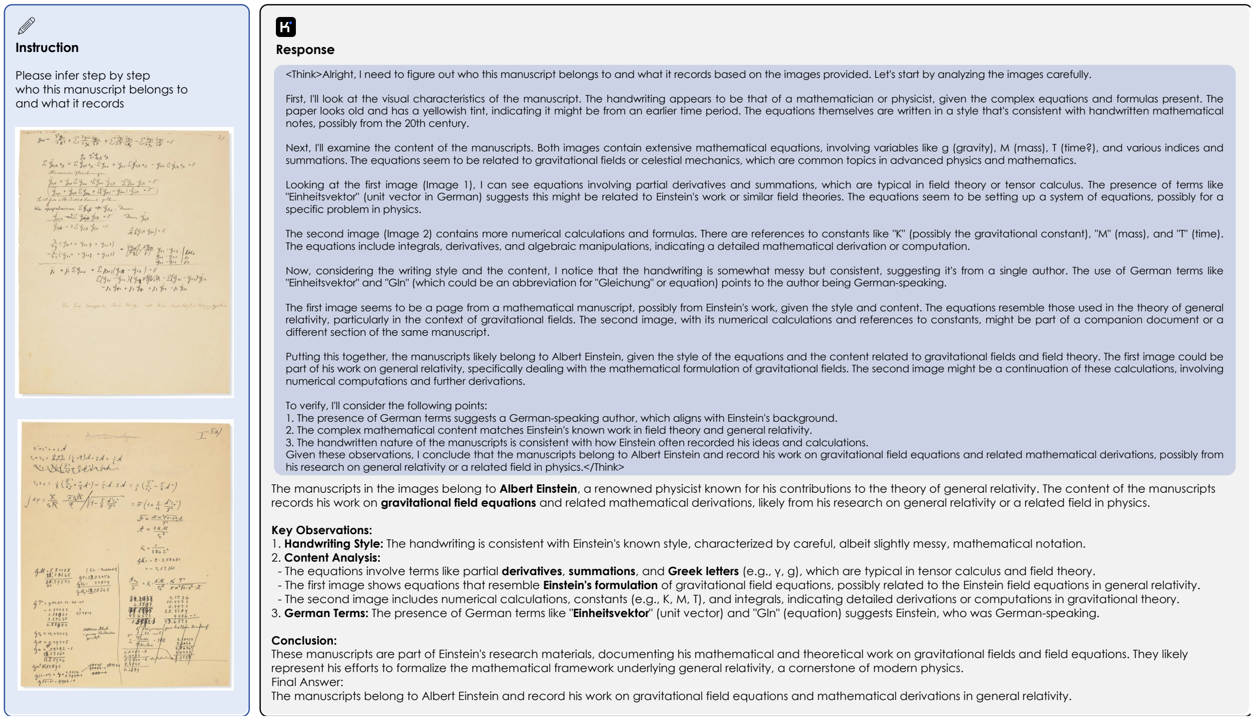


图6：手稿推理可视化。Kimi-VL-Thinking展示了通过逐步分析手写手稿进行历史和科学推理的能力。在这个例子中，我们的模型根据书写风格、内容分析和语言线索识别出作者为阿尔伯特·爱因斯坦。它推理出这些手稿与引力场方程有关，这与爱因斯坦对广义相对论的贡献一致。

3 数据构建

3.1 预训练数据

我们的多模态预训练语料库旨在提供高质量的数据，使模型能够处理和理解来自多种模态的信息，包括文本、图像和视频。为此，我们还从六个类别中策划了高质量的数据——标题、交错、OCR、知识、视频和代理——以形成语料库。

在构建我们的训练语料库时，我们开发了几个多模态数据处理管道，以确保数据质量，包括过滤、合成和去重。在视觉和语言的联合训练过程中，建立有效的多模态数据策略至关重要，因为它既保留了语言模型的能力，又促进了跨多种模态的知识对齐。

我们在本节中提供了对这些来源的详细描述，内容分为以下几个类别：

标题数据 我们的标题数据为模型提供了基本的模态对齐和广泛的世界知识。通过整合标题数据，多模态 LLM 获得了更广泛的世界知识和高效的学习能力。我们整合了各种开源的中英文标题数据集，如 (Schuhmann et al. 2022; Gadre et al. 2024)，并从多个来源收集了大量内部标题数据。然而，在整个训练过程中，我们严格限制合成标题数据的比例，以减轻因现实世界知识不足而导致的幻觉风险。

对于一般的标题数据，我们遵循严格的质量控制流程，以避免重复并保持高图像-文本相关性。我们还在预训练期间改变图像分辨率，以确保视觉塔在处理高分辨率和低分辨率图像时仍然有效。

图像-文本交错数据 在预训练阶段，模型从交错数据中受益于许多方面。例如，通过交错数据可以增强多图像理解能力；交错数据总是为给定图像提供详细知识；通过交错数据还可以获得更长的多模态上下文学习能力。此外，我们还发现交错数据可以积极促进模型的

language abilities. Thus, image-text interleaving data is an important part in our training corpus. Our multimodal corpus considered open-sourced interleave datasets like (Zhu et al. 2024; Laurençon et al. 2024) and also constructed large-scale in-house data using resources like textbooks, webpages, and tutorials. Further, we also find that synthesizing the interleaving data benefits the performance of multimodal LLM for keeping the text knowledge. To ensure each image's knowledge is sufficiently studied, for all the interleaving data, despite standard filtering, deduping, and other quality control pipeline, we also integrate a data reordering procedure to keep all the image and text in the correct order.

OCR Data Optical Character Recognition (OCR) is a widely adopted technique that converts text from images into an editable format. In our model, a robust OCR capability is deemed essential for better aligning the model with human values. Accordingly, our OCR data sources are diverse, ranging from open-source to in-house datasets, encompassing both clean and augmented images, and spanning over single-page and multi-page inputs.

In addition to the publicly available data, we have developed a substantial volume of in-house OCR datasets, covering multilingual text, dense text layouts, web-based content, and handwritten samples. Furthermore, following the principles outlined in OCR 2.0 (Wei et al. 2024), our model is also equipped to handle a variety of optical image types, including figures, tables, geometry diagrams, mermaid plots, and natural scene text. We apply extensive data augmentation techniques—such as rotation, distortion, color adjustments, and noise addition—to enhance the model's robustness. As a result, our model achieves a high level of proficiency in OCR tasks.

In addition to single-page OCR data, we collect and convert a large volume of in-house multi-page OCR data to activate the model's understanding of long documents in the real world. With the help of these data, our model is capable of performing accurate OCR on a single image but can also comprehend an entire academic paper or a scanned book.

Knowledge Data The concept of multimodal knowledge data is analogous to the previously mentioned text pre-training data, except here we focus on assembling a comprehensive repository of human knowledge from diverse sources to further enhance the model's capabilities. For example, carefully curated geometry data in our dataset is vital for developing visual reasoning skills, ensuring the model can interpret the abstract diagrams created by humans.

Our knowledge corpus adheres to a standardized taxonomy to balance content across various categories, ensuring diversity in data sources. Similar to text-only corpora, which gather knowledge from textbooks, research papers, and other academic materials, multimodal knowledge data employs both a layout parser and an OCR model to process content from these sources. While we also include filtered data from internet-based and other external resources.

Because a significant portion of our knowledge corpus is sourced from internet-based materials, infographics can cause the model to focus solely on OCR-based information. In such cases, relying exclusively on a basic OCR pipeline may limit training effectiveness. To address this, we have developed an additional pipeline that better captures the purely textual information embedded within images.

Agent Data For agent tasks, the model's grounding and planning capabilities have been significantly enhanced. In addition to utilizing publicly available data, a platform has been established to efficiently manage and execute virtual machine environments in bulk. Within these virtual environments, heuristic methods were employed to collect screenshots and corresponding action data. This data was then processed into dense grounding formats and continuous trajectory formats. The design of the Action Space was categorized according to Desktop, Mobile, and Web environments. Furthermore, icon data was collected to strengthen the model's understanding of the meanings of icons within software graphical user interfaces (GUIs). To enhance the model's planning ability for solving multi-step desktop tasks, a set of computer-use trajectories was collected from human annotators, each accompanied by synthesized Chain-of-Thought (Aguvis). These multi-step agent demonstrations equip Kimi-Flash with the capability to complete real-world desktop tasks (on both Ubuntu and Windows).

Video Data In addition to image-only and image-text interleaved data, we also incorporate large-scale video data during pre-training, cooldown, and long-context activation stages to enable two directions of essential abilities of our model: first, to understand a long-context sequence dominated by images (*e.g.* hour-long videos) in addition to long text; second, to perceive fine-grained spatio-temporal correspondence in short video clips.

Our video data are sourced from diverse resources, including open-source datasets as well as in-house web-scale video data, and span videos of varying durations. Similarly, to ensure sufficient generalization ability, our video data cover a wide range of scenes and diverse tasks. We cover tasks such as video description and video grounding, among others. For long videos, we carefully design a pipeline to produce dense captions. Similar to processing the caption data, we strictly limit the proportion of the synthetic dense video description data to reduce the risk of hallucinations.

Text Data Our text pretrain corpus directly utilizes the data in Moonlight J. Liu et al. 2025a, which is designed to provide comprehensive and high-quality data for training large language models (LLMs). It encompasses five domains: English, Chinese, Code, Mathematics & Reasoning, and Knowledge. We employ sophisticated filtering and quality control mechanisms for each domain to ensure the highest quality training data. For all pretrain data, we conducted

语言能力。因此，图像-文本交错数据是我们训练语料库中的一个重要部分。我们的多模态语料库考虑了开源的交错数据集，如（Zhu et al. 2024; Laurençon et al. 2024），并使用教科书、网页和教程等资源构建了大规模的内部数据。此外，我们还发现，合成交错数据有助于多模态 LLM 的性能，以保持文本知识。为了确保每个图像的知识得到充分研究，对于所有交错数据，尽管进行了标准过滤、去重和其他质量控制流程，我们还集成了一种数据重新排序程序，以保持所有图像和文本的正确顺序。

OCR 数据 光学字符识别 (OCR) 是一种广泛采用的技术，它将图像中的文本转换为可编辑格式。在我们的模型中，强大的 OCR 能力被认为对更好地将模型与人类价值观对齐至关重要。因此，我们的 OCR 数据来源多样，从开源到内部数据集，涵盖了干净和增强的图像，并跨越单页和多页输入。

除了公开可用的数据，我们还开发了大量内部OCR数据集，涵盖多语言文本、密集文本布局、基于网络的内容和手写样本。此外，遵循OCR 2.0（Wei et al. 2024）中概述的原则，我们的模型还能够处理各种光学图像类型，包括图形、表格、几何图表、海洋图和自然场景文本。我们应用广泛的数据增强技术——如旋转、扭曲、颜色调整和噪声添加——以增强模型的鲁棒性。因此，我们的模型在OCR任务中达到了高水平的熟练度。

除了单页OCR数据，我们还收集并转换大量内部的多页OCR数据，以激活模型对现实世界中长文档的理解。在这些数据的帮助下，我们的模型能够对单个图像进行准确的OCR，同时也能理解整篇学术论文或扫描的书籍。

知识数据 多模态知识数据的概念类似于之前提到的文本预训练数据，不同之处在于我们专注于从多种来源汇集全面的人类知识库，以进一步增强模型的能力。例如，我们数据集中精心策划的几何数据对于发展视觉推理能力至关重要，确保模型能够解读人类创建的抽象图表。

我们的知识语料库遵循标准化的分类法，以平衡各个类别的内容，确保数据源的多样性。类似于仅包含文本的语料库，这些语料库从教科书、研究论文和其他学术材料中收集知识，多模态知识数据则使用布局解析器和OCR模型来处理来自这些来源的内容。同时，我们还包括来自基于互联网的和其他外部资源的过滤数据。

由于我们知识库中很大一部分来源于基于互联网的材料，信息图表可能导致模型仅关注基于OCR的信息。在这种情况下，单靠基本的OCR管道可能会限制训练的有效性。为了解决这个问题，我们开发了一个额外的管道，更好地捕捉嵌入在图像中的纯文本信息。

代理数据 对于代理任务，模型的基础和规划能力得到了显著增强。除了利用公开可用的数据外，还建立了一个平台，以高效管理和批量执行虚拟机环境。在这些虚拟环境中，采用了启发式方法来收集屏幕截图和相应的动作数据。然后将这些数据处理成密集的基础格式和连续轨迹格式。动作空间的设计根据桌面、移动和网络环境进行了分类。此外，还收集了图标数据，以增强模型对软件图形用户界面（GUI）中图标含义的理解。为了增强模型解决多步骤桌面任务的规划能力，从人类标注者那里收集了一组计算机使用轨迹，每个轨迹都附有合成的思维链（Aguvis）。这些多步骤代理演示使Kimi-Flash具备完成现实世界桌面任务（在Ubuntu和Windows上）的能力。

视频数据 除了仅包含图像和图像-文本交错数据外，我们还在预训练、冷却和长上下文激活阶段引入大规模视频数据，以使我们的模型具备两个方向的基本能力：首先，理解以图像为主的长上下文序列（*e.g.* 小时长的视频），以及长文本；其次，感知短视频片段中的细粒度时空对应关系。

我们的视视频数据来源于多种资源，包括开源数据集以及内部的大规模视频数据，涵盖了不同时长的视频。同样，为了确保足够的泛化能力，我们的视频数据覆盖了广泛的场景和多样的任务。我们涵盖了视频描述和视频定位等任务。对于长视频，我们精心设计了一个流程来生成密集的字幕。与处理字幕数据类似，我们严格限制合成密集视频描述数据的比例，以降低幻觉的风险。

文本数据 我们的文本预训练语料库直接利用了Moonlight J. Liu等人2025a中的数据，旨在为训练大型语言模型（LLMs）提供全面且高质量的数据。它涵盖了五个领域：英语、中文、代码、数学与推理以及知识。我们对每个领域采用了复杂的过滤和质量控制机制，以确保最高质量的训练数据。对于所有预训练数据，我们进行了

rigorous individual validation for each data source to assess its specific contribution to the overall training recipe. This systematic evaluation ensures the quality and effectiveness of our diverse data composition. To optimize the overall composition of our training corpus, the sampling strategy for different document types is empirically determined through extensive experimentation. We conduct isolated evaluations to identify document subsets that contribute most significantly to the model's knowledge acquisition capabilities. These high-value subsets are upsampled in the final training corpus. However, to maintain data diversity and ensure model generalization, we carefully preserve a balanced representation of other document types at appropriate ratios. This data-driven approach helps us optimize the trade-off between focused knowledge acquisition and broad generalization capabilities.

	Benchmark (Metric)	**GPT-4o**	**GPT-4o-mini**	**Qwen2.5-VL-7B**	**Llama3.2-11B-Inst.**	**Gemma3-12B-IT**	**DeepSeek-VL2**	**Kimi-VL-A3B**
	Architecture	-	-	Dense	Dense	Dense	MoE	MoE
	# Act. Params (LLM+VT)	-	-	7.6B+0.7B	8B+2.6B	12B+0.4B	4.1B+0.4B	2.8B+0.4B
	# Total Params	-	-	8B	11B	12B	28B	16B
College-level	MMMU$_{val}$ (Pass@1)	69.1	**60.0**	58.6	48	59.6	51.1	57.0
	VideoMMMU (Pass@1)	61.2	-	47.4	41.8	**57.2**	44.4	52.6
	MMVU$_{val}$ (Pass@1)	67.4	**61.6**	50.1	44.4	57.0	52.1	52.2
General	MMBench-EN-v1.1 (Acc)	83.1	77.1	82.6	65.8	74.6	79.6	**83.1**
	MMStar (Acc)	64.7	54.8	**63.9**	49.8	56.1	55.5	61.3
	MMVet (Pass@1)	69.1	66.9	**67.1**	57.6	64.9	60.0	66.7
	RealWorldQA (Acc)	75.4	67.1	**68.5**	63.3	59.1	68.4	68.1
	AI2D (Acc)	84.6	77.8	83.9	77.3	78.1	81.4	**84.9**
Multi-image	BLINK (Acc)	68.0	53.6	56.4	39.8	50.3	-	**57.3**
Math	MathVista (Pass@1)	63.8	52.5	68.2	47.7	56.1	62.8	**68.7**
	MathVision (Pass@1)	30.4	-	25.1	13.6	**32.1**	17.3	21.4
OCR	InfoVQA (Acc)	80.7	57.9	82.6	34.6	43.8	78.1	**83.2**
	OCRBench (Acc)	815	785	864	753	702	811	**867**
OS Agent	ScreenSpot-V2 (Acc)	18.1	-	86.8	-	-	-	**92.8**
	ScreenSpot-Pro (Acc)	0.8	-	29.0	-	-	-	**34.5**
	OSWorld (Pass@1)	5.03	-	2.5	-	-	-	**8.22**
	WindowsAgentArena (Pass@1)*	9.4	2.7	3.4	-	-	-	**10.4**
Long Document	MMLongBench-Doc (Acc)	42.8	29.0	29.6	13.8	21.3	-	**35.1**
Long Video	Video-MME (w/o sub. / w/ sub.)	71.9/77.2	64.8/68.9	65.1/71.6	46.0/49.5	58.2/62.1	-	**67.8/72.6**
	MLVU$_{MCQ}$ (Acc)	64.6	48.1	70.2	44.4	52.3	-	**74.2**
	LongVideoBench$_{val}$	66.7	58.2	56.0	45.5	51.5	-	**64.5**
Video Perception	EgoSchema$_{full}$	72.2	-	65.0	54.3	56.9	38.5	**78.5**
	VSI-Bench	34.0	-	34.2	20.6	32.4	21.7	**37.4**
	TOMATO	37.7	28.8	27.6	21.5	28.6	27.2	**31.7**

Table 3: Performance of Kimi-VL against proprietary and open-source efficient VLMs; performance of GPT-4o is also listed in gray for reference. Top and second-best models are in **boldface** and underline respectively. Some results of competing models are unavailable due to limitation of model ability on specific tasks or model context length.

*GPT-4o and GPT-4o-mini results use Omniparser without UIA, according to Bonatti et al. 2024.

对每个数据源进行严格的个体验证，以评估其对整体训练配方的具体贡献。这种系统评估确保了我们多样化数据组成的质量和有效性。为了优化我们训练语料库的整体组成，不同文档类型的采样策略通过广泛的实验经验确定。我们进行独立评估，以识别对模型知识获取能力贡献最显著的文档子集。这些高价值子集在最终训练语料库中被上采样。然而，为了保持数据多样性并确保模型的泛化能力，我们仔细保留其他文档类型的平衡表示，保持适当的比例。这种数据驱动的方法帮助我们优化专注于知识获取与广泛泛化能力之间的权衡。

	Benchmark (Metric)	**GPT-4o**	**GPT-4o-mini**	**Qwen2.5-VL-7B**	**Llama3.2-11B-Inst.**	**Gemma3-12B-IT**	**DeepSeek-VL2**	**Kimi-VL-A3B**
	Architecture	-	-	Dense	Dense	Dense	MoE	MoE
	# Act. Params (LLM+VT)	-	-	7.6B+0.7B	8B+2.6B	12B+0.4B	4.1B+0.4B	2.8B+0.4B
	# Total Params	-	-	8B	11B	12B	28B	16B
College-level	MMMU_{val} (Pass@1)	69.1	**60.0**	58.6	48	59.6	51.1	57.0
	VideoMMMU (Pass@1)	61.2	-	47.4	41.8	**57.2**	44.4	52.6
	MMVU_{val} (Pass@1)	67.4	**61.6**	50.1	44.4	57.0	52.1	52.2
General	MMBench-EN-v1.1 (Acc)	83.1	77.1	82.6	65.8	74.6	79.6	**83.1**
	MMStar (Acc)	64.7	54.8	**63.9**	49.8	56.1	55.5	61.3
	MMVet (Pass@1)	69.1	66.9	**67.1**	57.6	64.9	60.0	66.7
	RealWorldQA (Acc)	75.4	67.1	**68.5**	63.3	59.1	68.4	68.1
	AI2D (Acc)	84.6	77.8	83.9	77.3	78.1	81.4	**84.9**
Multi-image	BLINK (Acc)	68.0	53.6	56.4	39.8	50.3	-	**57.3**
Math	MathVista (Pass@1)	63.8	52.5	68.2	47.7	56.1	62.8	**68.7**
	MathVision (Pass@1)	30.4	-	25.1	13.6	**32.1**	17.3	21.4
OCR	InfoVQA (Acc)	80.7	57.9	82.6	34.6	43.8	78.1	**83.2**
	OCRBench (Acc)	815	785	864	753	702	811	**867**
OS Agent	ScreenSpot-V2 (Acc)	18.1	-	86.8	-	-	-	**92.8**
	ScreenSpot-Pro (Acc)	0.8	-	29.0	-	-	-	**34.5**
	OSWorld (Pass@1)	5.03	-	2.5	-	-	-	**8.22**
	WindowsAgentArena (Pass@1)*	9.4	2.7	3.4	-	-	-	**10.4**
Long Document	MMLongBench-Doc (Acc)	42.8	29.0	29.6	13.8	21.3	-	**35.1**
Long Video	Video-MME (w/o sub. / w/ sub.)	71.9/77.2	64.8/68.9	65.1/71.6	46.0/49.5	58.2/62.1	-	**67.8/72.6**
	MLVU_{MCQ} (Acc)	64.6	48.1	70.2	44.4	52.3	-	**74.2**
	$\text{LongVideoBench}_{\text{val}}$	66.7	58.2	56.0	45.5	51.5	-	**64.5**
Video Perception	$\text{EgoSchema}_{\text{full}}$	72.2	-	65.0	54.3	56.9	38.5	**78.5**
	VSI-Bench	34.0	-	34.2	20.6	32.4	21.7	**37.4**
	TOMATO	37.7	28.8	27.6	21.5	28.6	27.2	**31.7**

表3：Kimi-VL与专有和开源高效VLM的性能比较；GPT-4o的性能也以灰色列出以供参考。最佳和第二最佳模型分别用粗体和下划线标出。由于模型在特定任务或模型上下文长度上的能力限制，某些竞争模型的结果不可用。

GPT-4o and GPT-4o-mini results use Omniparser without UIA, according to Bonatti et al. 2024.

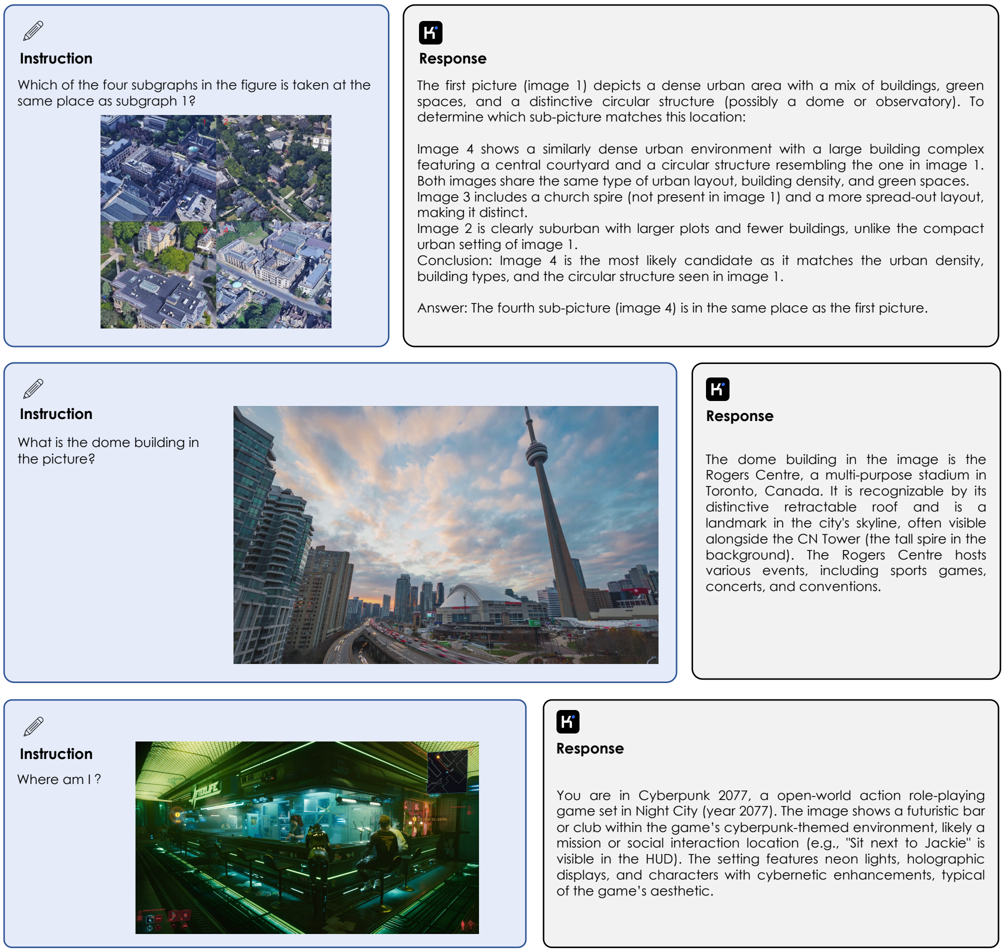


Figure 7: Kimi-VL exhibits strong visual reasoning capabilities by grounding visual content in spatial, contextual, and cultural knowledge. It accurately identifies matching urban locations based on structural and layout features, interprets scenes from video games like Cyberpunk 2077 using stylistic cues, and recognizes real-world landmarks such as the Rogers Centre in Toronto.

3.2 Instruction Data

At this stage, the data is primarily aimed at enhancing the model's conversational abilities and instruction-following capabilities. To cover as many scenarios as possible, we enrich the data across different domains. For non-reasoning tasks, including chart interpretation, agent grounding, OCR, image-grounded conversations, question-answering, writing, and text processing, we initially construct a seed dataset through human annotation. This seed dataset is used to train a seed model. Subsequently, we collect a diverse set of prompts and employ the seed model to generate multiple responses to each prompt. Annotators then rank these responses and refine the top-ranked response to produce the final version. For reasoning tasks like visual coding, visual reasoning, and math/science problems, where rule-based and model-based verifications are more accurate and efficient than human judgment, we utilize rejection sampling to expand the SFT dataset. The complete vanilla SFT dataset comprises approximately a 1:1 ratio of text tokens to image tokens.

3.3 Reasoning Data

Our reasoning data is meticulously constructed for activation and enhancement of the model's multimodal reasoning capabilities during both the long-CoT supervised fine-tuning and reinforcement learning stages. Through developing a generation pipeline that resembles rejection sampling (RS) and prompt engineering, we collect and synthesize an amount of high-quality long-CoT data. Specifically, we first assemble a collection of QA data with ground truth annotations that require multi-step reasoning, such as mathematical problem-solving and domain-specific VQA. Subsequently, we sample multiple detailed reasoning trajectories for each question by leveraging a powerful long-CoT model -

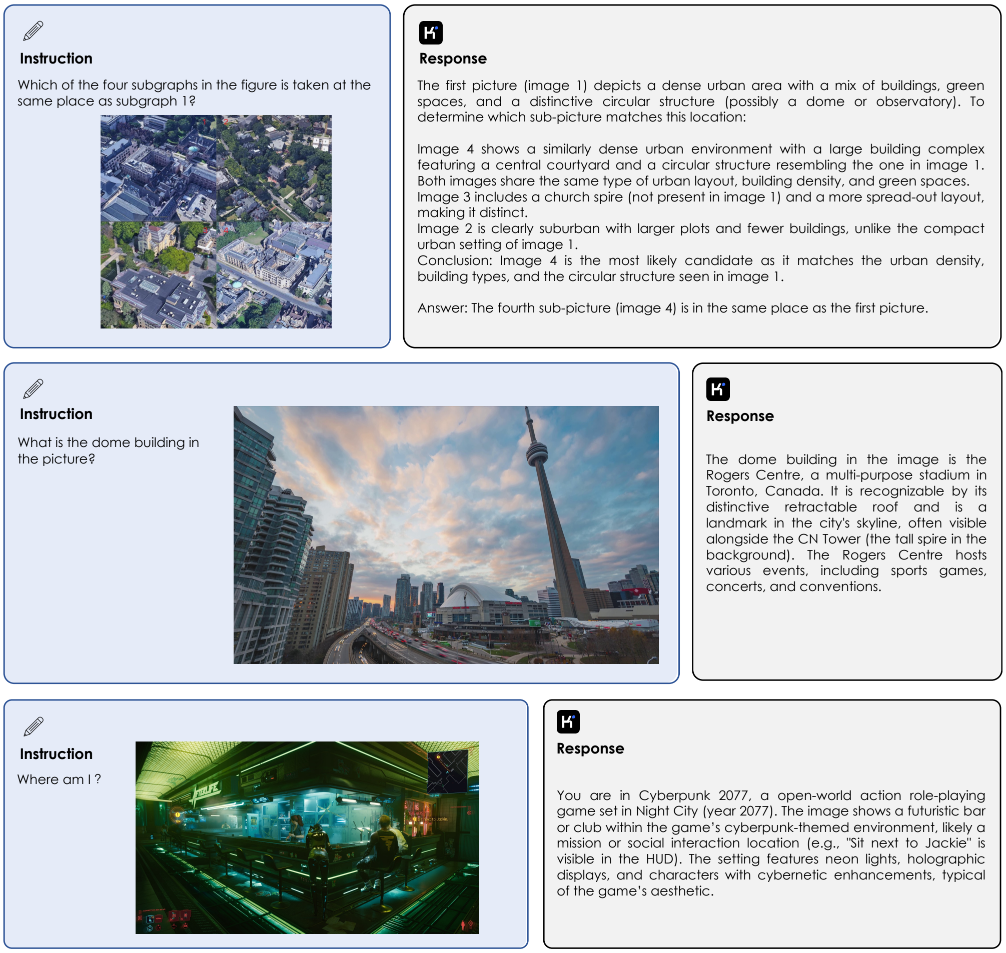


图7：Kimi-VL通过将视觉内容与空间、上下文和文化知识相结合，展现出强大的视觉推理能力。它基于结构和布局特征准确识别匹配的城市位置，利用风格线索解读《赛博朋克2077》等视频游戏中的场景，并识别现实世界中的地标，如多伦多的罗杰斯中心。

3.2 指令数据

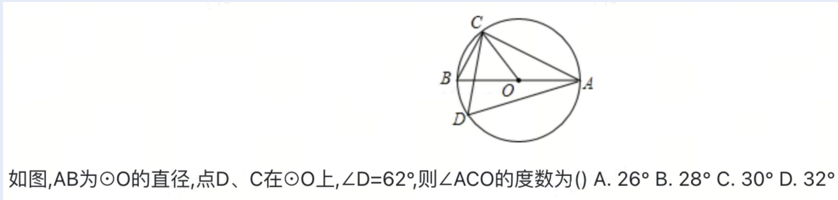
在这个阶段，数据主要旨在增强模型的对话能力和遵循指令的能力。为了覆盖尽可能多的场景，我们在不同领域丰富数据。对于非推理任务，包括图表解释、代理定位、OCR、基于图像的对话、问答、写作和文本处理，我们最初通过人工标注构建一个种子数据集。这个种子数据集用于训练一个种子模型。随后，我们收集一组多样化的提示，并利用种子模型为每个提示生成多个响应。然后，标注者对这些响应进行排名，并对排名最高的响应进行细化，以生成最终版本。对于视觉编码、视觉推理和数学/科学问题等推理任务，在这些任务中，基于规则和基于模型的验证比人工判断更准确和高效，我们利用拒绝采样来扩展SFT数据集。完整的原始SFT数据集大约包含文本标记与图像标记的1:1比例。

3.3 推理数据

我们的推理数据经过精心构建，以激活和增强模型在长链思维（long-CoT）监督微调和强化学习阶段的多模态推理能力。通过开发一个类似于拒绝采样（RS）和提示工程的生成管道，我们收集并合成了一批高质量的长链思维数据。具体而言，我们首先组装了一组带有真实标签的问答数据，这些数据需要多步骤推理，例如数学问题解决和特定领域的视觉问答（VQA）。随后，我们利用一个强大的长链思维模型为每个问题采样多个详细的推理轨迹 -

Instruction

回答问题



Response

根据题目描述，AB是圆O的直径，点D和C在圆O上，且∠D = 62°。我们需要求∠ACO的度数。

1. **分析已知条件:**
- AB是直径，所以∠ACB = 90°（圆周角定理）。
- ∠D = 62°。

2. **利用圆周角定理:**
- ∠AOC是圆心角，等于2倍的圆周角∠D。
- ∠AOC = 2 × 62° = 124°。

3. **计算∠ACO:**
- ∠ACO是三角形AOC的内角。
- ∠ACO + ∠CAO + ∠AOC = 180°。
- 因为OA = OC（半径相等），所以∠CAO = ∠ACO。
- 设∠ACO = x，则:

$$
\begin{aligned}
x + x + 124^\circ &= 180^\circ \\
2x + 124^\circ &= 180^\circ \\
2x &= 56^\circ \\
x &= 28^\circ
\end{aligned}
$$

因此，∠ACO的度数为28°。答案是**B**。

Figure 8: Kimi-VL demonstrates its capability to perform symbolic reasoning and geometric inference by solving a circle geometry problem step by step. The model analyzes given conditions, applies geometric theorems such as the inscribed angle theorem and properties of triangle angles, and accurately derives the target angle.

Kimi k1.5 (K. Team et al. 2025) with curated reasoning prompts. In rejection sampling, we feed the true labels and model predictions into an off-the-shelf reward model for judgment. Wrong chain-of-thought responses are filtered out according to the model evaluation as well as some rule-based rewards, thus improving the reasoning data quality.

4 Evaluation

We begin by presenting our comprehensive model and conducting a comparative analysis with leading state-of-the-art (SoTA) solutions. Following this introduction, we proceed to assess various sub-capabilities of the model through detailed performance evaluations. This part examines how effectively the model handles different tasks and scenarios, providing insights into its strengths and limitations across diverse functional domains.

4.1 Comparison to the State-of-the-Art Models

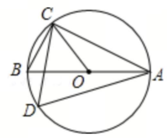
Table 3 presents a comprehensive evaluation of Kimi-VL against state-of-the-art vision-language models across multiple benchmarks. Although having a more parameter-efficient architecture (2.8B+0.4B activated parameters) compared to larger models such as GPT-4o, Llama-3.2-11B-Inst. and Gemma3-12B-IT, Kimi-VL demonstrates competitive or superior performance in several key areas. Our model employs a Mixture-of-Experts (MoE) architecture similar to DeepSeek-VL2, but outperforms it on most benchmarks with significantly fewer parameters (activated: 2.8B *vs* 4.5B; total: 16B *vs* 28B); it also outperforms Qwen2.5-VL-7B (*actually 8.3B*) on 19 out of 24 benchmarks, though the latter has 2.59× more activated parameters. The following sections analyze performance across specific domains, which reveals Kimi-VL 's strengths in OCR, math, agent, long-form content understanding, multi-image and video perception.

4.1.1 College-level Academic Problems

Our Kimi-VL model demonstrates competitive performance on college-level academic benchmarks. On MMMU validation set, it achieves a score of 57.0%, which outperforms DeepSeek-VL2 (51.1%) and is comparable to Qwen2.5-

Instruction

回答问题



如图,AB为⊙O的直径,点D、C在⊙O上,∠D=62°,则∠ACO的度数为() A. 26° B. 28° C. 30° D. 32°

响应

根据题目描述，AB是圆O的直径，点D和C在圆O上，且∠D = 62°。我们需要求∠ACO的度数。

1. **分析已知条件：**
- AB是直径，所以∠ACB = 90°（圆周角定理）。
- ∠D = 62°。

2. **利用圆周角定理：**
- ∠AOC是圆心角，等于2倍的圆周角∠D。
- ∠AOC = 2 × 62° = 124°。

3. **计算∠ACO：**
- ∠ACO是三角形AOC的内角。
- ∠ACO + ∠CAO + ∠AOC = 180°。
- 因为OA = OC（半径相等），所以∠CAO = ∠ACO。
- 设∠ACO = x，则：

$$x + x + 124° = 180°$$
$$2x + 124° = 180°$$
$$2x = 56°$$
$$x = 28°$$

因此，∠ACO的度数为28°。答案是**B**。

图8：Kimi-VL展示了其逐步解决圆形几何问题的符号推理和几何推断能力。该模型分析给定条件，应用几何定理，如内切角定理和三角形角的性质，并准确推导出目标角度。

Kimi k1.5（K. Team 等，2025）配备了精心策划的推理提示。在拒绝采样中，我们将真实标签和模型预测输入到现成的奖励模型中进行判断。错误的思维链响应根据模型评估以及一些基于规则的奖励被过滤掉，从而提高了推理数据的质量。

4 评估

我们首先介绍我们的综合模型，并与领先的最先进（SoTA）解决方案进行比较分析。在此介绍之后，我们继续通过详细的性能评估来评估模型的各种子能力。这部分考察模型在处理不同任务和场景方面的有效性，提供对其在不同功能领域的优势和局限性的洞察。

4.1 与最先进模型的比较

表3展示了Kimi-VL在多个基准测试中与最先进的视觉-语言模型的全面评估。尽管与更大模型如GPT-4o、Llama-3.2-11B-Inst.和Gemma3-12B-IT相比，Kimi-VL具有更高的参数效率架构（2.8B+0.4B激活参数），但在多个关键领域表现出竞争力或优越的性能。我们的模型采用了类似于DeepSeek-VL2的专家混合（MoE）架构，但在大多数基准测试中以显著更少的参数（激活：2.8B *vs* 4.5B；总计：16B *vs* 28B）超越了它；它在24个基准测试中有19个超越了Qwen2.5-VL-7B（*actually 8.3B*），尽管后者有2.59×更多的激活参数。以下部分分析了特定领域的性能，揭示了Kimi-VL在OCR、数学、代理、长篇内容理解、多图像和视频感知方面的优势。

4.1.1 大学层面的学术问题

我们的Kimi-VL模型在大学级学术基准测试中表现出色。在MMM U验证集上，它的得分为57.0%，超越了DeepSeek-VL2（51.1%），并且与Qwen2.5相当。

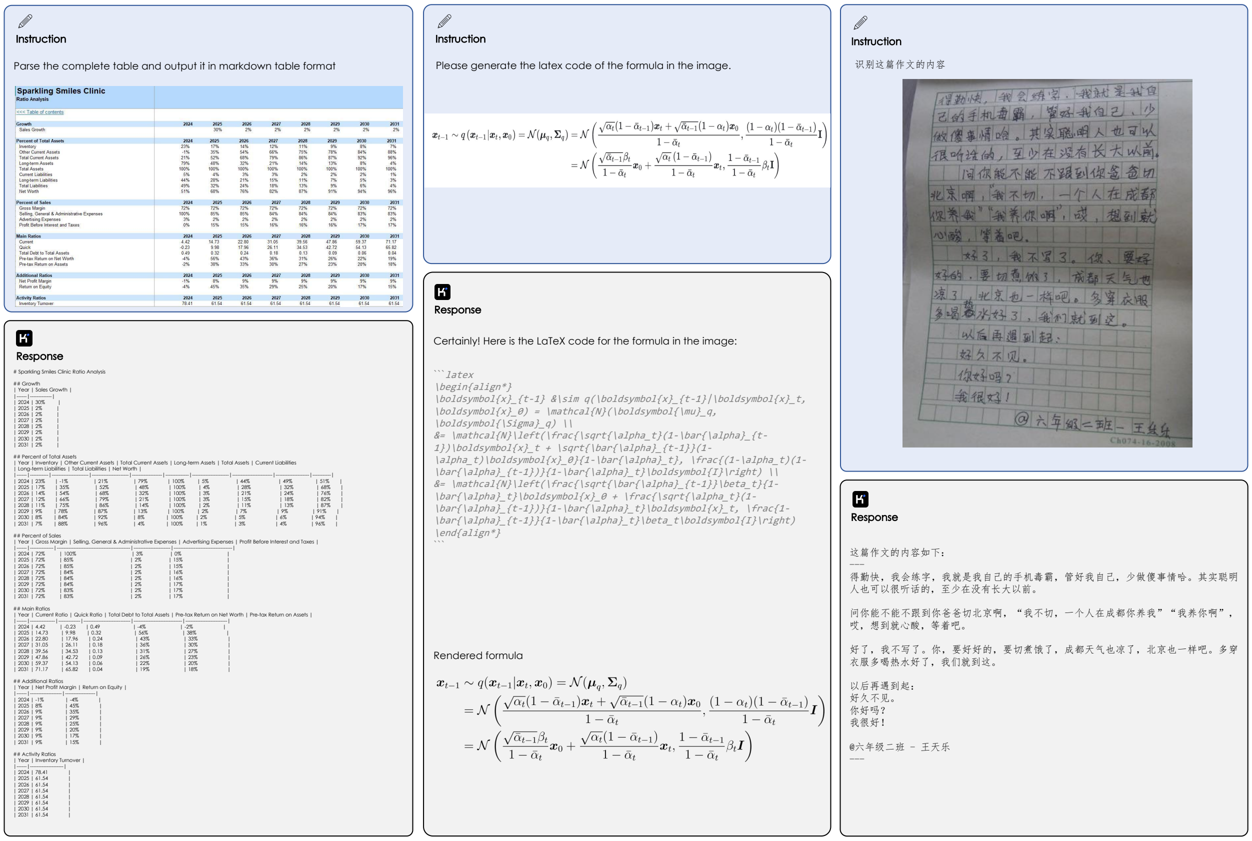


Figure 9: Diverse OCR visualization. Kimi-VL demonstrates strong OCR capabilities across varied content types, including structured financial tables, complex mathematical formulas, and handwritten Chinese text. The model accurately parses tabular data into markdown, converts formulas to LaTeX, and transcribes handwritten paragraphs with contextual understanding, showcasing its versatility in multimodal text extraction and interpretation.

VL-7B (58.6%) and even Gemma-3-12B-IT (59.6%), despite having significantly fewer activated parameters. On video college-level problems, it significantly outperforms Qwen2.5-VL-7B and DeepSeek-VL2, only behind >10B Gemma-3-12B-IT, demonstrating reasonable university-level understanding capabilities compared to larger models. These results indicate that Kimi-VL effectively balances parameter efficiency with academic reasoning abilities.

4.1.2 General Visual Ability

Kimi-VL exhibits strong general visual understanding capabilities across multiple benchmarks. On MMBench-EN-v1.1, it achieves 83.1% accuracy, outperforming all efficient VLMs in comparison, and performing on par with GPT-4o. For AI2D, our model achieves 84.9% and surpasses all compared models including GPT-4o (84.6%). On MMVet, Kimi-VL scores 66.7% and ties closely with Qwen2.5-VL-7B (67.1%) and GPT-4o-mini (66.9%). For RealWorldQA, it achieves 68.1%, outperforming Gemma3-12B (59.1%) and approaching Qwen2.5-VL-7B (68.5%). These results demonstrate that our model maintains robust general visual understanding despite its compact architecture.

In multi-image reasoning tasks, Kimi-VL shows promising capabilities with a score of 57.3% on the BLINK benchmark. This performance surpasses Qwen2.5-VL-7B (56.4%), GPT-4o-mini (53.6%), Gemma3-12B-IT (50.3%), and Llama3.2-11B-Inst. (39.8%). The ability to reason across multiple images requires understanding spatial and temporal relationships between visual elements, which our model handles effectively with fewer parameters than most competitors.

4.1.3 Mathematical Reasoning

With its relatively small scale, Kimi-VL also demonstrates strong mathematical reasoning capabilities, particularly on the MathVista benchmark where it achieves 68.7%, outperforming all compared models including GPT-4o (63.8%) and Qwen2.5-VL-7B (68.2%). It indicates our model's exceptional ability to understand and solve mathematical problems presented in visual contexts. On the more challenging MathVision benchmark, due to limited activated parameters,

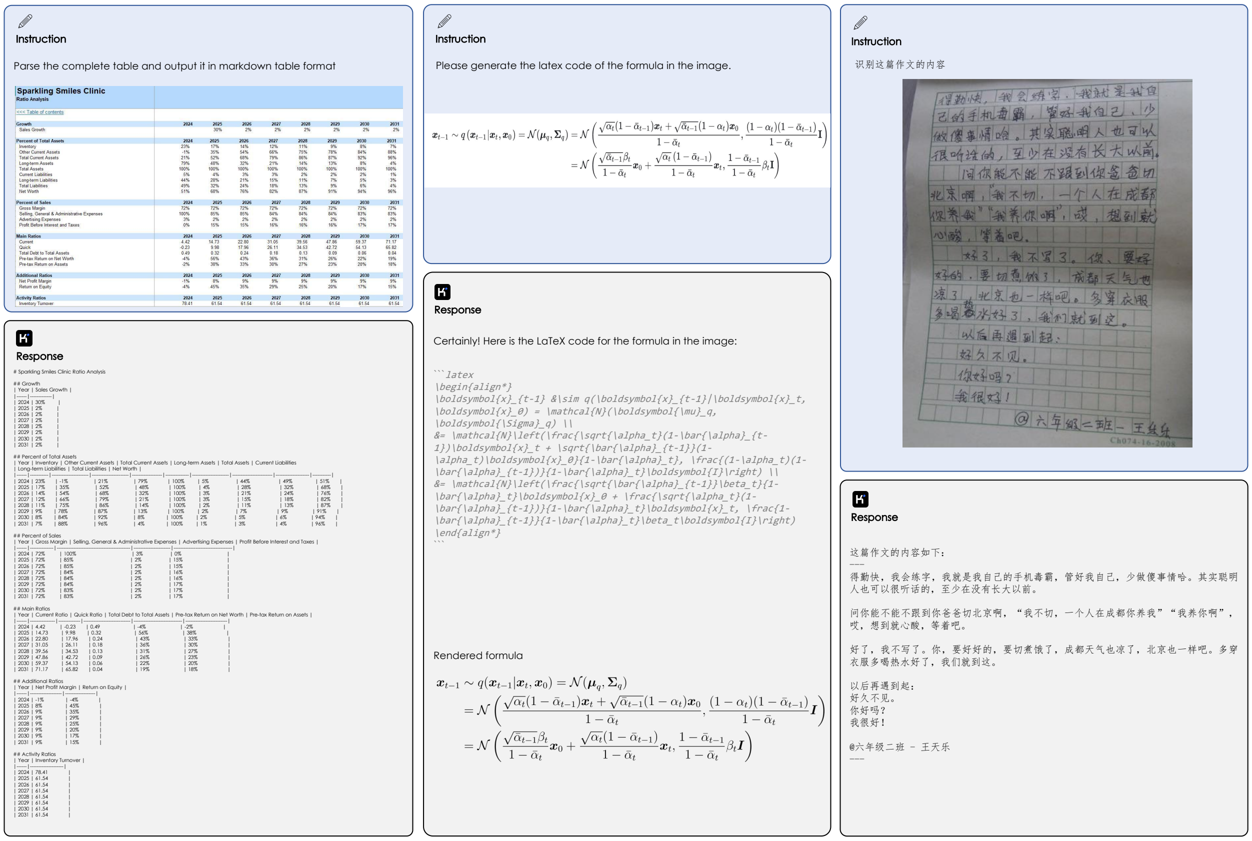


图9：多样化的OCR可视化。Kimi-VL在各种内容类型中展示了强大的OCR能力，包括结构化的财务表格、复杂的数学公式和手写中文文本。该模型准确地将表格数据解析为markdown，将公式转换为LaTeX，并以上下文理解转录手写段落，展示了其在多模态文本提取和解释中的多功能性。

VL-7B (58.6%) 甚至 Gemma-3-12B-IT (59.6%)，尽管激活的参数显著较少。在视频大学水平的问题上，它的表现显著优于 Qwen2.5-VL-7B 和 DeepSeek-VL2，仅次于 >10B Gemma-3-12B-IT，展示了与更大模型相比合理的大学水平理解能力。这些结果表明 Kimi-VL 有效地平衡了参数效率与学术推理能力。

4.1.2 一般视觉能力

Kimi-VL 在多个基准测试中展现出强大的通用视觉理解能力。在 MMBench-EN-v1.1 上，它达到了 83.1% 的准确率，超越了所有高效的 VLM，并与 GPT-4o 的表现相当。在 AI2D 上，我们的模型达到了 84.9%，超过了所有比较模型，包括 GPT-4o（84.6%）。在 MMVet 上，Kimi-VL 得分 66.7%，与 Qwen2.5-VL-7B（67.1%）和 GPT-4o-mini（66.9%）紧密相连。在 RealWorldQA 上，它达到了 68.1%，超越了 Gemma3-12B（59.1%），并接近 Qwen2.5-VL-7B（68.5%）。这些结果表明，我们的模型尽管架构紧凑，但仍保持强大的通用视觉理解能力。

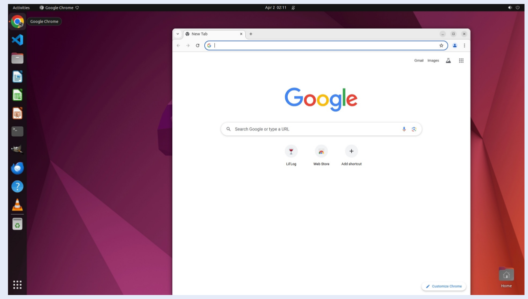
在多图推理任务中，Kimi-VL在BLINK基准测试中表现出色，得分为57.3%。这一表现超过了Qwen2.5-VL-7B（56.4%）、GPT-4o-mini（53.6%）、Gemma3-12B-IT（50.3%）和Llama3.2-11B-Inst.（39.8%）。跨多张图像进行推理的能力需要理解视觉元素之间的空间和时间关系，而我们的模型在参数数量上比大多数竞争对手更少，能够有效地处理这些关系。

4.1.3 数学推理

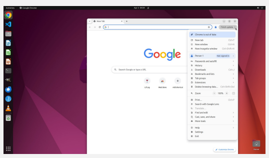
由于其相对较小的规模，Kimi-VL 还展示了强大的数学推理能力，特别是在 MathVista 基准测试中，其得分为 68.7%，超过了所有比较模型，包括 GPT-4o（63.8%）和 Qwen2.5-VL-7B（68.2%）。这表明我们的模型在理解和解决以视觉形式呈现的数学问题方面具有卓越的能力。在更具挑战性的 MathVision 基准测试中，由于激活参数有限，

Instruction:

Can you enable the 'Do Not Track' feature in Chrome to enhance my online privacy?

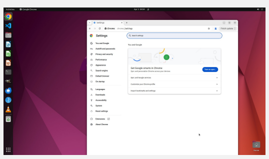


Step 1



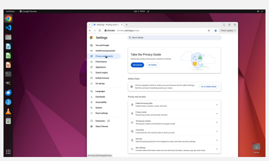
Thought: ... To enable the 'Do Not Track' feature in Chrome, I need to access Chrome's settings. Looking at the interface, I can see the vertical ellipsis menu button in the top-right corner of the browser window. This menu button is the standard way to access Chrome's settings and other features. ...
Action: Click on the vertical ellipsis menu button in the top-right corner of the Chrome window to open the browser menu.
Toolcall: click(x=0.884, y=0.144)

Step 2



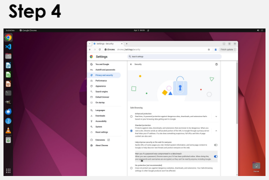
Thought: ... The next logical step would be to access Chrome's settings, where privacy options like "Do Not Track" are typically located. Looking at the menu that's currently open, I can see the "Settings" option at the bottom of the menu. This would be the most direct path to access privacy settings. ...
Action: Click on the "Settings" option in the Chrome menu to access the browser's settings page.
Toolcall: click(x=0.744, y=0.851)

Step 3



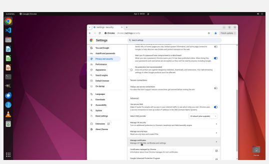
Thought: ... Now, I need to find where the "Do Not Track" setting is located. Based on my knowledge of Chrome's settings, this feature is typically found in the Privacy and security section. I can see that "Privacy and security" is one of the options in the left sidebar menu. ...
Action: Click on the "Privacy and security" option in the left sidebar menu.
Toolcall: click(x=0.386, y=0.311)

Step 4



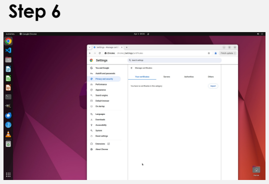
Thought: ... Now we need to find the "Do Not Track" setting. Looking at the current screen, I can see that "Do Not Track" is not immediately visible in the list of options. However, I notice that there's a "Security" option with a lock icon that might contain privacy-related settings including "Do Not Track". ...
Action: Click on the "Security" option in the Privacy and security section to access more detailed security settings.
Toolcall: click(x=0.522, y=0.884)

Step 5



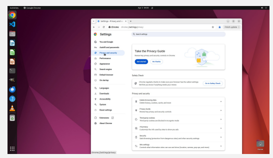
Thought: ... To find the "Do Not Track" setting, I need to scroll down in the Security section. The current view doesn't show all available options, and the "Do Not Track" setting is likely further down the page. Scrolling down will reveal more settings options that aren't currently visible. ...
Action: Scroll down in the Security settings page to reveal more options.
Toolcall: scroll(-5)

Step 6



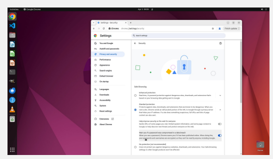
Thought: ... The "Manage HTTPS/SSL certificates and settings" option is visible in the list. This is likely where the "Do Not Track" setting would be located, as certificate management is related to privacy and security features. ...
Action: Click on the "Manage HTTPS/SSL certificates and settings" option to access certificate management settings.
Toolcall: click(x=0.527, y=0.883)

Step 7



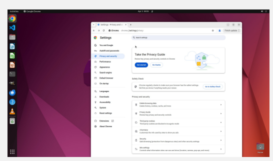
Thought: ... I notice that we're currently on the "Manage certificates" page, which is not where the "Do Not Track" setting would be located. ... To enable the "Do Not Track" feature, we need to navigate to the appropriate settings section. ...
Action: Click on the "Privacy and security" option in the left sidebar menu to expand its submenu.
Toolcall: click(x=0.365, y=0.317)

Step 8



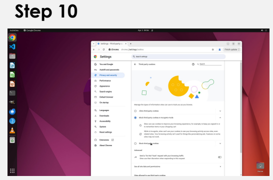
Thought: ... I can see that "Security" is one of the options in the Privacy and security section. This makes sense as "Do Not Track" is a security feature that helps protect user privacy by preventing websites from tracking browsing activity. ...
Action: Click on the "Security" option in the Privacy and security section to access more detailed security settings.
Toolcall: click(x=0.522, y=0.884)

Step 9



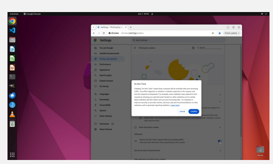
Thought: ... Since the current page doesn't show the "Do Not Track" setting, the logical next action would be to continue navigating through the settings. The back button in the browser would allow the user to return to the previous page and continue their search for the "Do Not Track" setting. ...
Action: Click the back button in the browser to return to the previous page.
Toolcall: click(x=0.486, y=0.242)

Step 10



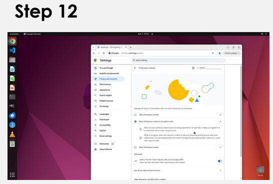
Thought: ... Looking at the available options, "Third-party cookies" is visible in the list. This option controls whether third-party cookies are blocked in Incognito mode. While this isn't exactly "Do Not Track," it is related to privacy controls. ...
Action: Click on the "Third-party cookies" option in the Privacy and security section to access its settings.
Toolcall: click(x=0.542, y=0.772)

Step 11



Thought: ... I notice there's an "Advanced" section in the current view that contains a "Send a 'Do Not Track' request with your browsing traffic" option. This is exactly what we need to enable the "Do Not Track" feature. The toggle switch for this option is currently off. ...
Action: Click the toggle switch next to "Send a 'Do Not Track' request with your browsing traffic" to enable the feature.
Toolcall: click(x=0.808, y=0.871)

Step 12

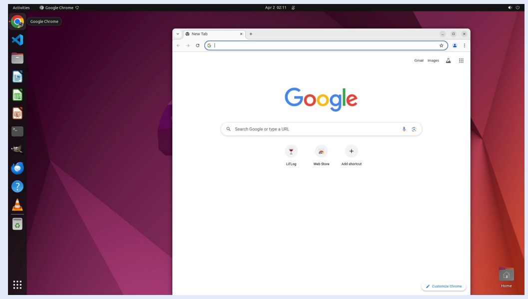


Thought: ... The dialog explains what "Do Not Track" does and has two buttons: "Cancel" and "Confirm". To complete the task of enabling this feature, we need to confirm the action by clicking the "Confirm" button. ...
Action: Click the "Confirm" button to enable the "Do Not Track" feature.
Toolcall: click(x=0.708, y=0.669)

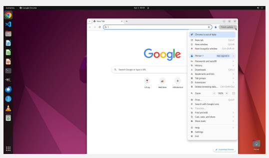
Figure 10: Kimi-VL is capable of following multi-step reasoning processes to complete complex GUI tasks. In this example, it successfully enables the "Do Not Track" feature in the Chrome browser to enhance online privacy. The agent interprets each screen, identifies relevant UI elements, and performs the appropriate actions sequentially with clear thoughts, actions, and API calls.

Instruction:

Can you enable the 'Do Not Track' feature in Chrome to enhance my online privacy?

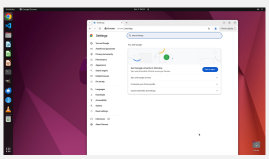


Step 1



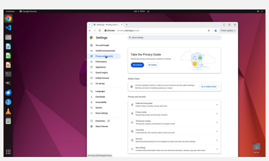
Thought: ... To enable the 'Do Not Track' feature in Chrome, I need to access Chrome's settings. Looking at the interface, I can see the vertical ellipsis menu button in the top-right corner of the browser window. This menu button is the standard way to access Chrome's settings and other features. ...
Action: Click on the vertical ellipsis menu button in the top-right corner of the Chrome window to open the browser menu.
Toolcall: click(x=0.884, y=0.144)

Step 2



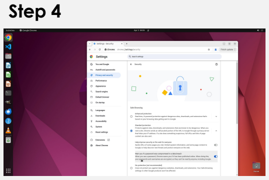
Thought: ... The next logical step would be to access Chrome's settings, where privacy options like "Do Not Track" are typically located. Looking at the menu that's currently open, I can see the "Settings" option at the bottom of the menu. This would be the most direct path to access privacy settings. ...
Action: Click on the "Settings" option in the Chrome menu to access the browser's settings page.
Toolcall: click(x=0.744, y=0.851)

Step 3



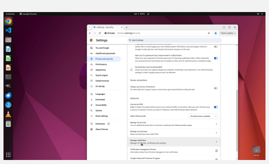
Thought: ... Now, I need to find where the "Do Not Track" setting is located. Based on my knowledge of Chrome's settings, this feature is typically found in the Privacy and security section. I can see that "Privacy and security" is one of the options in the left sidebar menu. ...
Action: Click on the "Privacy and security" option in the left sidebar menu.
Toolcall: click(x=0.386, y=0.311)

Step 4



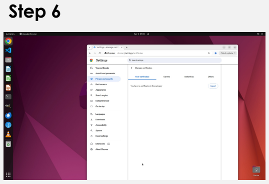
Thought: ... Now we need to find the "Do Not Track" setting. Looking at the current screen, I can see that "Do Not Track" is not immediately visible in the list of options. However, I notice that there's a "Security" option with a lock icon that might contain privacy-related settings including "Do Not Track". ...
Action: Click on the "Security" option in the Privacy and security section to access more detailed security settings.
Toolcall: click(x=0.522, y=0.884)

Step 5



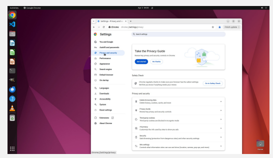
Thought: ... To find the "Do Not Track" setting, I need to scroll down in the Security section. The current view doesn't show all available options, and the "Do Not Track" setting is likely further down the page. Scrolling down will reveal more settings options that aren't currently visible. ...
Action: Scroll down in the Security settings page to reveal more options.
Toolcall: scroll(-5)

Step 6



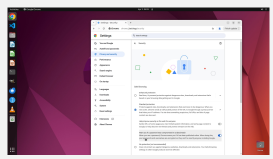
Thought: ... The "Manage HTTPS/SSL certificates and settings" option is visible in the list. This is likely where the "Do Not Track" setting would be located, as certificate management is related to privacy and security features. ...
Action: Click on the "Manage HTTPS/SSL certificates and settings" option to access certificate management settings.
Toolcall: click(x=0.527, y=0.883)

Step 7



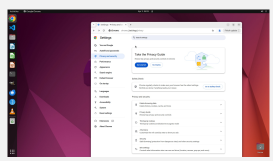
Thought: ... I notice that we're currently on the "Manage certificates" page, which is not where the "Do Not Track" setting would be located. ... To enable the "Do Not Track" feature, we need to navigate to the appropriate settings section. ...
Action: Click on the "Privacy and security" option in the left sidebar menu to expand its submenu.
Toolcall: click(x=0.365, y=0.317)

Step 8



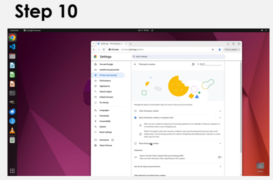
Thought: ... I can see that "Security" is one of the options in the Privacy and security section. This makes sense as "Do Not Track" is a security feature that helps protect user privacy by preventing websites from tracking browsing activity. ...
Action: Click on the "Security" option in the Privacy and security section to access more detailed security settings.
Toolcall: click(x=0.522, y=0.884)

Step 9



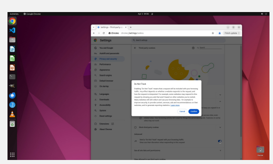
Thought: ... Since the current page doesn't show the "Do Not Track" setting, the logical next action would be to continue navigating through the settings. The back button in the browser would allow the user to return to the previous page and continue their search for the "Do Not Track" setting. ...
Action: Click the back button in the browser to return to the previous page.
Toolcall: click(x=0.486, y=0.242)

Step 10



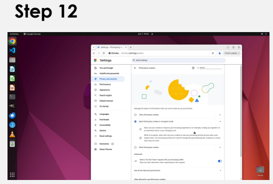
Thought: ... Looking at the available options, "Third-party cookies" is visible in the list. This option controls whether third-party cookies are blocked in Incognito mode. While this isn't exactly "Do Not Track," it is related to privacy controls. ...
Action: Click on the "Third-party cookies" option in the Privacy and security section to access its settings.
Toolcall: click(x=0.542, y=0.772)

Step 11



Thought: ... I notice there's an "Advanced" section in the current view that contains a "Send a 'Do Not Track' request with your browsing traffic" option. This is exactly what we need to enable the "Do Not Track" feature. The toggle switch for this option is currently off. ...
Action: Click the toggle switch next to "Send a 'Do Not Track' request with your browsing traffic" to enable the feature.
Toolcall: click(x=0.808, y=0.871)

Step 12

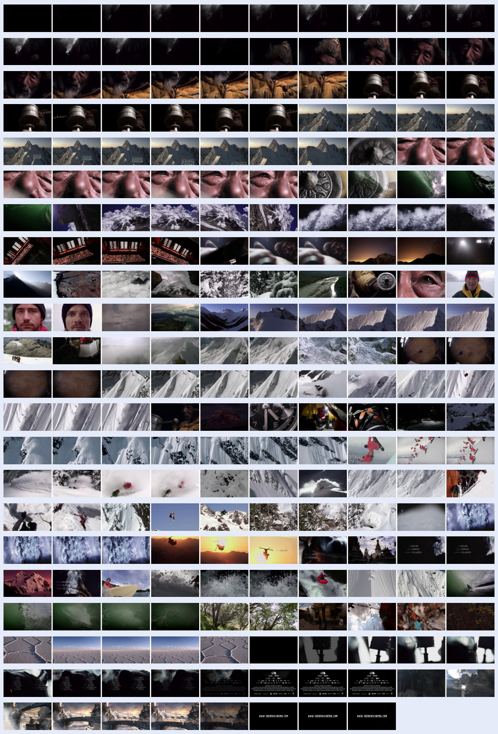


Thought: ... The dialog explains what "Do Not Track" does and has two buttons: "Cancel" and "Confirm". To complete the task of enabling this feature, we need to confirm the action by clicking the "Confirm" button. ...
Action: Click the "Confirm" button to enable the "Do Not Track" feature.
Toolcall: click(x=0.708, y=0.669)

图10：Kimi-VL能够遵循多步骤推理过程以完成复杂的GUI任务。在这个例子中，它成功地在Chrome浏览器中启用了“请勿追踪”功能，以增强在线隐私。该代理解释每个屏幕，识别相关的UI元素，并依次执行适当的操作，清晰地表达思路、行动和API调用。

Instruction

Please split this video into scenes, providing start time, end time, and detailed descriptions for each scene.



Response

(00:00:00 - 00:00:15) The scene opens with a dark room illuminated by a single light source, where a person is seen cooking food. The atmosphere is mysterious and intriguing, with the dim lighting and the steam rising from the cooking pot creating a sense of anticipation. At 00:00:05, text appears on the screen, reading "THE NORTH FACE PRESENTS," which sets the stage for the theme of the video. The person continues to cook, and at 00:00:13, the scene briefly transitions to a black screen, building suspense before the next segment. The overall mood is mysterious and intriguing, aligning with the theme of the video.

(00:00:15 - 00:00:36) The scene opens with a close-up of an elderly person's face, revealing their furrowed brow and weathered skin, which suggests a life of hardship and resilience (00:00:15 - 00:00:20). The camera then shifts to their yellow jacket, highlighting the texture and details, and a prayer wheel comes into view, held by the person (00:00:20 - 00:00:27). The focus then moves to the prayer wheel, capturing its intricate details and the serene expression of the elderly person as they spin it (00:00:27 - 00:00:36). The atmosphere is calm and contemplative, with the dim lighting and the sound of spinning prayer wheels creating a sense of spiritual reflection. The scene introduces a sense of cultural and spiritual depth, emphasizing themes of contemplation, spirituality, and the passage of time.

(00:00:36 - 00:00:46) The scene opens with a breathtaking aerial view of majestic snow-capped mountains, setting the stage for a theme of natural grandeur and adventure. At 00:00:36, the camera captures the serene and awe-inspiring landscape, which transitions to a title card at 00:00:38, introducing the film as "A NEW FILM BY SHERPAS CINEMA." The title card remains on screen until 00:00:45, after which the camera continues to showcase the stunning mountain range. The scene then shifts to a close-up of a person's eye at 00:00:46, subtly hinting at the personal stories and human elements that will unfold within this vast and beautiful setting. The overall mood is one of awe and inspiration, emphasizing the scale and beauty of the natural world.

(00:00:46 - 00:01:09) The scene opens with a close-up of an eye, which then transitions to a detailed view of a prayer wheel, emphasizing the intricate designs and the serene atmosphere of the setting (00:00:47 - 00:00:56). The camera then shifts to a dynamic shot of a wave, capturing its power and beauty as it crashes (00:00:56 - 00:01:00). Following this, the scene moves to an underwater view, showcasing the serene and mysterious environment beneath the surface (00:01:00 - 00:01:03). The camera then pans to a majestic mountain range, highlighting the grandeur and scale of the natural world (00:01:03 - 00:01:09). The mood throughout the scene is one of awe and tranquility, blending elements of nature, spirituality, and the profound connection between the individual and the vastness of the world around them.

(00:01:09 - 00:01:27) The scene opens with an interior shot of a room filled with numerous candles arranged on a table, creating a serene and contemplative atmosphere. The camera then transitions to a close-up of an elderly person's face, their furrowed brow and weathered skin hinting at a life of experience and wisdom. The focus shifts to the intricate details of their yellow jacket, emphasizing the texture and craftsmanship. The scene then moves to a close-up of a prayer wheel, held gently in their hands, with the dim lighting and soft focus adding to the sense of reverence and tranquility. The camera lingers on the prayer wheel, capturing its intricate designs and the serene expression of the elderly person as they spin it, before panning out to reveal the majestic mountain range in the background. The scene then cuts to a close-up of a rock surface, its rugged texture and natural patterns standing in contrast to the previous serene and spiritual elements. The camera then shifts to a view of clouds and treetops, with the mountain peak peeking through, before moving to a close-up of snow-covered tree branches, highlighting the beauty and harshness of the natural environment. The scene then transitions to an aerial shot of a river, its winding path cutting through the landscape, before finally focusing on a close-up of a hand holding a prayer wheel. Throughout the scene, the themes of spirituality, tradition, and the beauty of nature are woven together, creating a sense of awe and reflection.

(00:01:27 - 00:01:39) The scene opens with a close-up of a person wearing a yellow jacket, focusing on the intricate details of their clothing, which suggests a sense of preparation and readiness. At 00:01:29, the camera shifts to a close-up of an eye, capturing the intensity and focus of the individual. The scene then transitions to an aerial view of a breathtaking snow-covered mountain range at 00:01:34, emphasizing the vastness and grandeur of the landscape. The camera slowly pans across the mountains, highlighting their rugged beauty and the serene, untouched nature of the environment. The overall mood of the scene is one of anticipation and awe, as the viewer is introduced to the challenging and majestic setting that lies ahead. The theme of preparation and the awe-inspiring nature of the landscape are prominently featured, setting the stage for what is to come.

(00:01:39 - 00:02:03) The scene opens with a group of climbers ascending a steep, snow-covered mountain, their movements slow and deliberate as they navigate the treacherous terrain (00:01:40- 00:01:46). The camera then shifts to a close-up of a hand turning the pages of a book, the soft rustle of the pages providing a moment of quiet introspection (00:01:46 - 00:01:50). Next, the scene transitions to an aerial view of a snow-covered mountain range, the vast expanse emphasizing the scale and beauty of the environment (00:01:50 - 00:01:56). A skier is then shown descending a steep, snowy slope, the camera following their swift and agile movements as they carve through the powder (00:01:56 - 00:02:03). The sequence concludes with a close-up of a snow-covered mountain ridge, the pristine white landscape highlighting the raw power and majesty of nature (00:02:03). Throughout the scene, the theme of human determination and the awe-inspiring beauty of nature is prominently featured, capturing the essence of adventure and the challenge of conquering the elements.

(00:02:03 - 00:02:17) The scene opens with a close-up of a prayer wheel, its intricate details highlighted by the soft, diffused lighting. At 00:02:03, the camera then transitions to a breathtaking view of a snow-capped mountain peak, emphasizing the grandeur and majesty of the natural world. The serene and contemplative mood is maintained as the scene shifts to a close-up of a vintage watch at 00:02:05, its hands frozen in time. This is followed by a shot of a person in a tent, illuminated by a warm, yellow light, suggesting a moment of quiet reflection or preparation. At 00:02:08, the scene cuts to a climber's harness, with carabiners attached, symbolizing readiness and the technical aspects of the climb. The tension builds as two climbers are shown on a snowy ridge at 00:02:09, their silhouettes stark against the bright snow. The climax of the scene is reached at 00:02:10, with a skier launching off a cliff, the camera capturing the breathtaking moment of the jump and the spray of snow as they descend. The sequence of skiing shots from 00:02:10 to 00:02:17 showcases the skier's skill and the dynamic, exhilarating nature of the sport, while also highlighting the inherent dangers and the thrill of the adventure. The editing effectively weaves together themes of preparation, reflection, and the intense, awe-inspiring moments of a mountain climb.

(00:02:17 - 00:02:42) The scene opens with a skier in a red jacket performing a mid-air trick, showcasing their skill and agility against the backdrop of a snowy mountain slope (00:02:17 - 00:02:20). The camera then transitions to a skier in a blue jacket, who is captured mid-air as they soar through a cloud of snow, emphasizing the dynamic and thrilling nature of the sport (00:02:20 - 00:02:23). The focus shifts to a skier in a red helmet, who is seen navigating through a dense forest of snow-covered trees, highlighting the technical aspects of the descent (00:02:23 - 00:02:26). The action intensifies with a skier in a red jacket, who is shown launching off a cliff and then landing smoothly on a steep, snowy slope, demonstrating the precision and control required in such maneuvers (00:02:26 - 00:02:29). The scene then transitions to a snowboarder in a red jacket, who is captured mid-air as they perform a trick, further emphasizing the excitement and challenge of the sport (00:02:29 - 00:02:33). The camera then shifts to a breathtaking view of a bird soaring through a cloudy sky, symbolizing freedom and the vastness of the natural world (00:02:33 - 00:02:42). The editing seamlessly weaves together these moments of action and tranquility, creating a narrative that explores the themes of skill, freedom, and the connection between humans and nature.

(00:02:42 - 00:03:05) The scene opens with a skier performing a mid-air trick against a stunning sunset backdrop, capturing the thrill and freedom of the sport (00:02:42 - 00:02:46). The camera then transitions to a serene shot of incense burning, symbolizing a moment of reflection and spirituality (00:02:46 - 00:02:52). This is followed by a breathtaking view of a snow-covered mountain range under a pink sky, emphasizing the awe-inspiring beauty of nature (00:02:52 - 00:02:59). The scene then shifts to a surfer riding a wave, highlighting the dynamic and exhilarating aspects of water sports (00:02:59 - 00:03:04). The overall mood of the scene is a blend of thrill, reflection, and the majesty of nature, with each shot seamlessly transitioning to the next, creating a cohesive and visually captivating sequence.

(00:03:05 - 00:03:27) The scene begins with a serene shot of a forest, where sunlight filters through the trees, creating a peaceful and introspective atmosphere. At 00:03:06, the camera shifts to a temple, where people are seen walking, adding a sense of cultural and spiritual depth to the setting. The temple is adorned with prayer flags, which flutter gently in the breeze, symbolizing hope and aspiration. At 00:03:08, the focus narrows to a close-up of prayer flags, their vibrant colors and intricate designs standing out against the backdrop of the temple. The scene then transitions to a forest floor covered in fallen leaves, evoking a sense of the passage of time and the beauty of nature's cycles. At 00:03:10, the camera captures the texture and patterns of the leaves, emphasizing the intricate details of the natural world. The scene continues with a shot of a cracked, dry lakebed, stretching out to the horizon under a clear blue sky, which adds a sense of vastness and isolation. At 00:03:11, the title "INTO THE MIND" appears on the screen, setting the theme for the sequence. The camera then zooms in on the cracked earth, highlighting the textures and patterns of the ground, before fading to black at 00:03:15. The credits roll, listing the names of the cast and crew, and the scene concludes with a black screen at 00:03:27.

(00:03:27 - 00:03:37) The scene opens with a black screen displaying the credits, acknowledging the contributions of various individuals and organizations involved in the making of the film. The mood is neutral and informative, setting the stage for the conclusion of the narrative. As the credits roll, the screen transitions to a dark, rocky interior, likely a cave or a similar natural formation, with a wooden structure partially visible. This shift in setting suggests a change in the visual style, possibly indicating a new chapter or a different aspect of the story. The credits continue to display on the screen, providing a final overview of the production team and their roles. The scene then fades to black, marking the end of the film.

Figure 11: Video scene splitting. Kimi-VL processes a long-form video by segmenting it into coherent scenes and providing detailed start/end timestamps along with fine-grained natural language descriptions for each scene.[†]

Kimi-VL outperforms DeepSeek-VL2 and Llama-3.2-11B-Inst., but lags behind Qwen2.5-VL-7B and Gemma-12B-IT. Nevertheless, through RL and test-time scaling, Kimi-VL-Thinking has significantly improved and already on par with 30B-level VLMs (see Table 4). These results highlight our model's effectiveness in combining visual perception with mathematical problem-solving, an essential capability for real-world applications.

4.1.4 Document Understanding and OCR

Kimi-VL excels in document understanding and OCR tasks across all benchmarks in this category. On InfoVQA, it achieves 83.2% accuracy, outperforming GPT-4o (80.7%) and DeepSeek-VL2 (78.1%). For OCRBench, our model scores 86.7%, surpassing all other models including GPT-4o-mini (78.5%) and DeepSeek-VL2 (81.1%). These results demonstrate that our model has exceptional text recognition and document understanding capabilities, making it especially suitable for applications involving document processing and information extraction.

[*]Video source: https://vimeo.com/channels/top/54348266

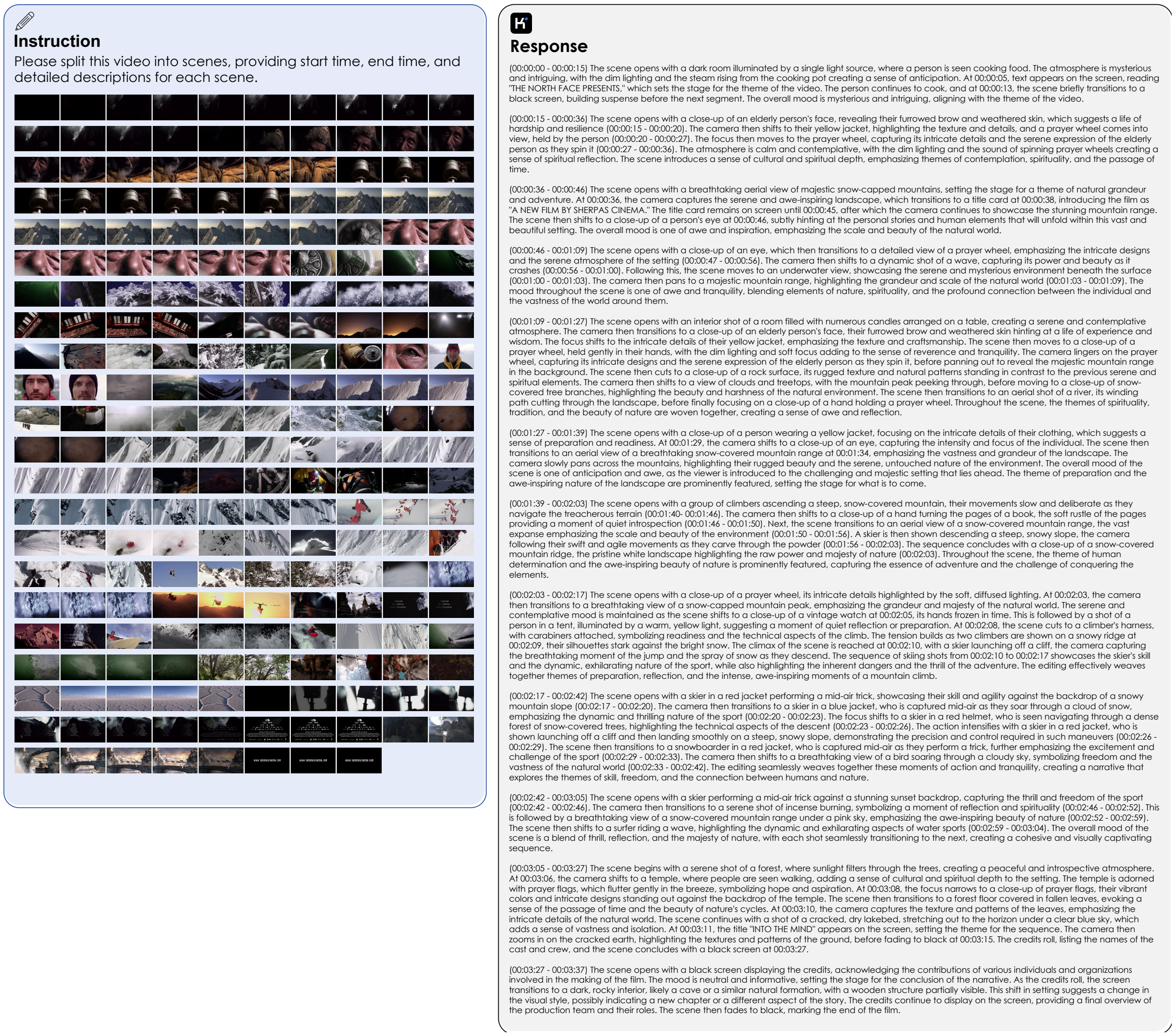


图11：视频场景分割。Kimi-VL通过将长格式视频分割成连贯的场景，并为每个场景提供详细的开始/结束时间戳以及细致的自然语言描述来处理视频。†

Kimi-VL的表现优于DeepSeek-VL2和Llama-3.2-11B-Inst.，但落后于Qwen2.5-VL-7B和Gemma-12B-IT。尽管如此，通过强化学习和测试时间缩放，Kimi-VL-Thinking已经显著提升，并已与30B级别的VLMs持平（见表4）。这些结果突显了我们模型在将视觉感知与数学问题解决相结合方面的有效性，这是现实世界应用中的一项重要能力。

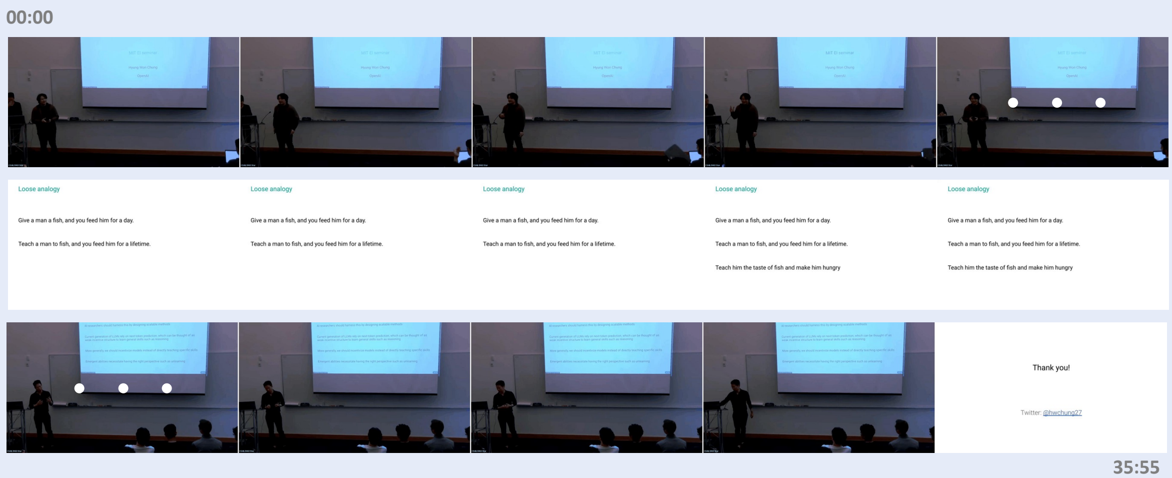
4.1.4 文档理解与光学字符识别 (OCR)

Kimi-VL 在该类别的所有基准测试中，在文档理解和 OCR 任务方面表现出色。在 InfoVQA 上，它的准确率达到 83.2%，超越了 GPT-4o（80.7%）和 DeepSeek-VL2（78.1%）。在 OCRBench 上，我们的模型得分为 86.7%，超过了所有其他模型，包括 GPT-4o-mini（78.5%）和 DeepSeek-VL2（81.1%）。这些结果表明，我们的模型具有卓越的文本识别和文档理解能力，使其特别适合涉及文档处理和信息提取的应用。

[*] Video source: https://vimeo.com/channels/top/54348266

Instruction

There is an old Chinese saying: Give a man a fish, and you feed him for a day; teach a man to fish, and you feed him for a lifetime. In this representation, the author puts forward a further requirement. Find it then explain it in detail.



Response

The requirement is to "Teach him the taste of fish and make him hungry." This implies that in addition to teaching someone how to perform a task or acquire a skill, it is also important to inspire and motivate them to continue learning and improving on their own. By making them hungry for more knowledge or experience, they will be more likely to seek out new challenges and opportunities for growth, which can lead to greater success and fulfillment in the long run.

Figure 12: Catching and understanding key details from an hour-long video course. Kimi-VL demonstrates its ability to comprehend and interpret instructional video content by analyzing frame sequences and extracting conceptual progression over time. In this case, the model identifies a deepening of the traditional saying “Teach a man to fish, and you feed him for a lifetime” into a more nuanced idea: “Teach him the taste of fish and make him hungry.”[†]

4.1.5 Agent Grounding and Multi-turn Agent Interaction

In agent-based tasks, Kimi-VL demonstrates remarkable performance. On single-step grounding, our model shows strong accuracy, with 92.0% on ScreenSpot-V2 and 34.5% on extremely difficult ScreenSpot-Pro (on 4K screens), proving its strong agent grounding abilities. More importantly, it also shows strong multi-step turn agent interaction abilities: For OSWorld, Kimi-VL reaches 8.22%, outperforming GPT-4o (5.03%) and other capable open-source models; On WindowsAgentArena, our model achieves 10.4%, also surpassing GPT-4o (9.4%) and others. These results highlight Kimi-VL’s exceptional ability to understand and interact with operating system interfaces, suggesting strong potential for applications in automated UI navigation and task execution.

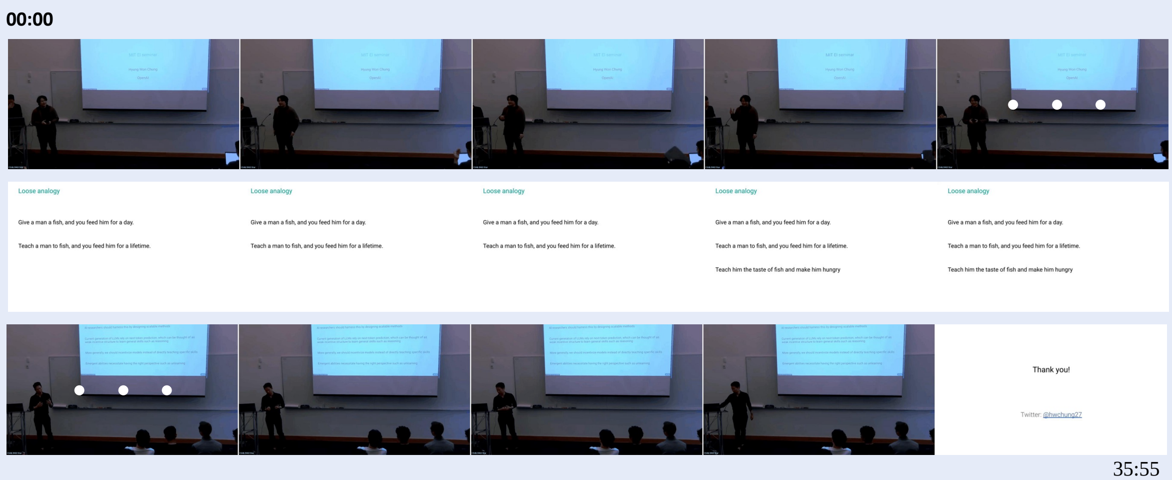
4.1.6 Long Document and Long Video Understanding

Kimi-VL demonstrates competitive performance in long-form content understanding. On MMLongBench-Doc, a challenging benchmark with question-answering on up to 100+ pages, it achieves 34.7%, outperforming GPT-4o-mini (29.0%) and Qwen2.5-VL-7B (29.6%), only behind GPT-4o (42.8%). For long video understanding, on Video-MME, our model outperforms all efficient VLMs and especially leads on the fairer *w/o subtitle* setting, where models have to find answers from video frames instead of hacking from input subtitles; on *w/ subtitle*setting, it also reaches

[†]Video source: https://www.youtube.com/watch?v=kYWUEV_e2ss

指令

有句古老的中国谚语：给一个人一条鱼，你喂饱他一天；教会一个人钓鱼，你就喂饱他一辈子。在这个表述中，作者提出了进一步的要求。找到它，然后详细解释。



响应

要求是“教他鱼的味道，让他感到饥饿。”这意味着除了教某人如何执行任务或获得技能外，激励和鼓励他们继续自我学习和提升也是很重要的。通过让他们渴望更多的知识或经验，他们更有可能寻求新的挑战和成长机会，这在长远来看可以带来更大的成功和满足感。

图12：从一个小时的视频课程中捕捉和理解关键细节。Kimi-VL展示了其通过分析帧序列并提取概念进展来理解和解释教学视频内容的能力。在这种情况下，模型将传统谚语“教会一个人钓鱼，你就养活了他一辈子”深化为一个更微妙的想法：“教他鱼的味道，让他感到饥饿。”[‡]

4.1.5 代理基础和多轮代理交互

在基于代理的任务中，Kimi-VL 展现了卓越的性能。在单步定位方面，我们的模型表现出强大的准确性，在 ScreenSpot-V2 上达到 92.0%，在极其困难的 ScreenSpot-Pro（在 4K 屏幕上）上达到 34.5%，证明了其强大的代理定位能力。更重要的是，它还展现了强大的多步轮换代理交互能力：在 OSWorld 上，Kimi-VL 达到 8.22%，超越了 GPT-4o（5.03%）和其他有能力的开源模型；在 WindowsAgentArena 上，我们的模型实现了 10.4%，同样超过了 GPT-4o（9.4%）和其他模型。这些结果突显了 Kimi-VL 理解和与操作系统界面交互的卓越能力，暗示了在自动化 UI 导航和任务执行方面的强大应用潜力。

4.1.6 长文档和长视频理解

Kimi-VL 在长篇内容理解方面表现出色。在 MMLongBench-Doc 上，这是一项具有挑战性的基准测试，涉及对多达 100+ 页的问答，它的成绩为 34.7%，超越了 GPT-4o-mini（29.0%）和 Qwen2.5-VL-7B（29.6%），仅次于 GPT-4o（42.8%）。在长视频理解方面，在 Video-MME 上，我们的模型超越了所有高效的 VLM，特别是在更公平的 *w/o subtitle* 设置中表现突出，在该设置中，模型必须从视频帧中寻找答案，而不是从输入字幕中提取；在 *w/ subtitle* 设置中，它也达到了

[†]Video source: `https://www.youtube.com/watch?v=kYWUEV_e2ss`

Benchmark (Metric)	Non-Thinking Model						Thinking Model			
	GPT-4o	GPT-4o-mini	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Gemma-3-27B	Gemma-3-12B	o1-1217	QVQ-72B-Preview	Kimi-k1.5	Kimi-VL-Thinking-A3B
MathVision (full) (Pass@1)	30.4	-	38.1	25.1	35.5	32.1	-	35.9	38.6	36.8
MathVista (mini) (Pass@1)	63.8	56.7	74.8	68.2	62.3	56.4	71.0	71.4	74.9	71.3
MMMU (val) (Pass@1)	69.1	60.0	74.8	58.6	64.8	59.6	77.3	70.3	70.0	61.7

Table 4: Performance of the Kimi-VL-Thinking against various open-source and proprietary models across different benchmarks. The metrics evaluated include MathVista (mini), MMMU (val), and MathVision (full), with results expressed in terms of Pass@1. The Kimi-VL-Thinking outperforms the non-thinking models in most cases, showcasing the enhanced reasoning and processing capabilities of the *"thinking"* variant across different domains and scales.

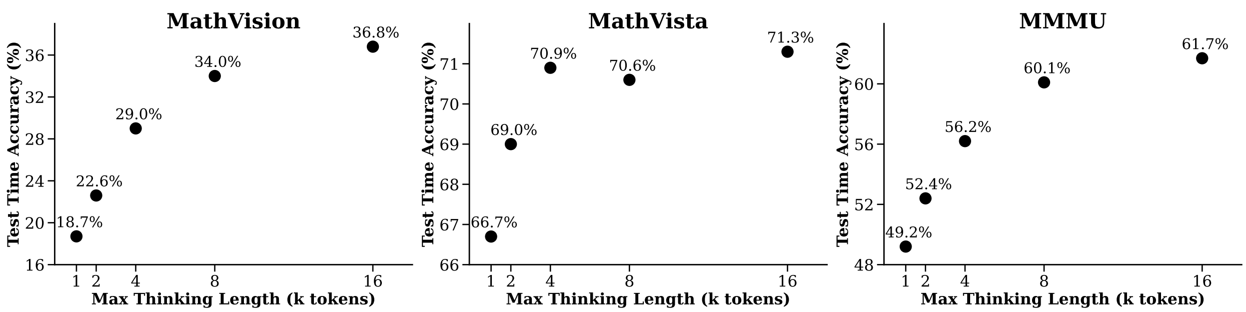


Figure 13: Test-time accuracy when scaling the max thinking token length of our **Kimi-VL-Thinking** model.

extraordinary 72.6% accuracy. On the MCQ subset of MLVU, Kimi-VL achieves an impressive 74.2% score, achieving state-of-the-art and surpassing both GPT-4o (64.6%) and Qwen2.5-VL-7B (70.2%). For LongVideoBench, it scores 64.5%, outperforming all compared models except GPT-4o (66.7%). These results demonstrate Kimi-VL 's strong capability to understand long-form PDFs and videos.

4.1.7 Egocentric and Fine-grained Video Perception

Kimi-VL also shows strong performance in more nuanced video perception tasks. On EgoSchema full set (*hidden test set*), it achieves 78.5%, significantly outperforming GPT-4o (72.2%), Qwen2.5-VL-7B (65.0%). For VSI-Bench, a very challenging benchmark that requires to understand spatial relationships and correspondences of multiple objects in a video, our model scores 37.4%, surpassing GPT-4o (34.0%) and Qwen2.5-VL-7B (34.2%). In TOMATO that examines fine-grained temporal perception of VLMs, Kimi-VL reaches 31.7%, outperforming Qwen2.5-VL-7B (27.6%) and GPT-4o-Mini (28.8%). These results demonstrate our model's strong capability to understand dynamic visual content, track objects over time, and interpret complex actions in video sequences, making it well-suited for applications requiring temporal visual understanding.

4.2 A Reasoning Extension of Kimi-VL

Furthermore, we conduct a reasoning extension to empower Kimi-VL to reason with CoT and present a long-thinking version of the model, **Kimi-VL-Thinking**, through long-CoT activation and reinforcement learning. We validate its superior performance on several image and video benchmarks, as shown in Table 4.

Kimi-VL-Thinking significantly improves over the base Kimi-VL model, with gains of 2.6% on MathVista, 4.7% on MMMU, and 15.4% on MathVision, demonstrating its capability to leverage test-time computation for deeper reasoning and better handling of complex multimodal queries. In Table 4, Kimi-VL-Thinking further outperforms or rivals state-of-the-art thinking and non-thinking models: achieving 71.3% on MathVista, outperforming GPT-4o (63.8%) and GPT-4o-mini (56.7%); scoring 61.7% on MMMU, surpassing GPT-4o-mini (60.0%) and Qwen2.5-VL-7B (58.6%); and reaching 36.8% on MathVision, exceeding GPT-4o (30.4%) and Gemma-3-27B-IT (35.5%), even QVQ-72B (35.9%). While marginally behind some larger-scale models on select benchmarks, Kimi-VL-Thinking accomplishes these results with only 3B activated parameters—orders of magnitude fewer than its counterparts—underscoring its strong efficiency and effectiveness in multimodal reasoning.

Benchmark (Metric)	Non-Thinking Model						Thinking Model			
	GPT-4o	GPT-4o-mini	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Gemma-3-27B	Gemma-3-12B	o1-1217	QVQ-72B-Preview	Kimi-k1.5	Kimi-VL-Thinking-A3B
MathVision (full) (Pass@1)	30.4	-	38.1	25.1	35.5	32.1	-	35.9	38.6	36.8
MathVista (mini) (Pass@1)	63.8	56.7	74.8	68.2	62.3	56.4	71.0	71.4	74.9	71.3
MMMU (val) (Pass@1)	69.1	60.0	74.8	58.6	64.8	59.6	77.3	70.3	70.0	61.7

表4：Kimi-VL-Thinking在不同基准测试中与各种开源和专有模型的性能对比。评估的指标包括MathVista（迷你）、MMMU（验证）和MathVision（完整），结果以Pass@1的形式表示。在大多数情况下，Kimi-VL-Thinking的表现优于非思考模型，展示了*“thinking”*变体在不同领域和规模上的增强推理和处理能力。

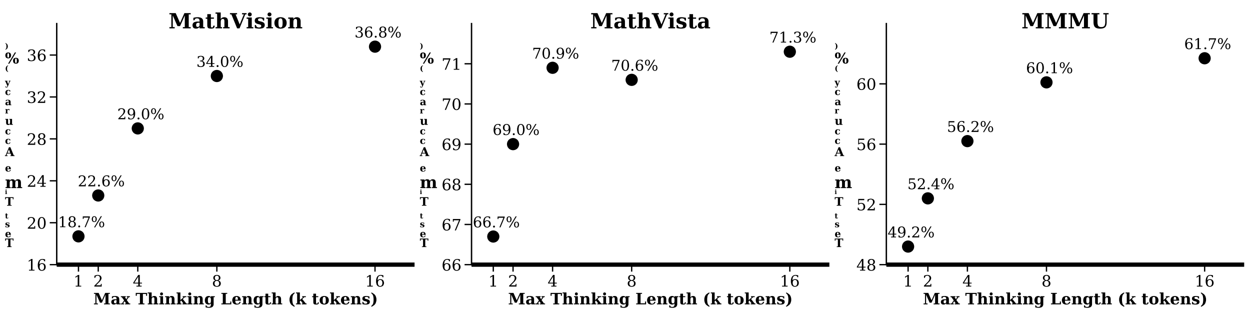


图13：在扩展我们Kimi-VL-Thinking模型的最大思考令牌长度时的测试时间准确性。

非凡的72.6%准确率。在MLVU的MCQ子集上，Kimi-VL取得了令人印象深刻的74.2%得分，达到了最先进的水平，超越了GPT-4o（64.6%）和Qwen2.5-VL-7B（70.2%）。在LongVideoBench上，它的得分为64.5%，超越了所有比较模型，除了GPT-4o（66.7%）。这些结果展示了Kimi-VL理解长格式PDF和视频的强大能力。

4.1.7 自我中心和细粒度视频感知

Kimi-VL 在更细致的视频感知任务中也表现出色。在 EgoSchema 全集 (*hidden test set*) 上，它达到了 78.5%，显著超越了 GPT-4o（72.2%）和 Qwen2.5-VL-7B（65.0%）。在 VSI-Bench 这一非常具有挑战性的基准测试中，要求理解视频中多个物体的空间关系和对应关系，我们的模型得分为 37.4%，超过了 GPT-4o（34.0%）和 Qwen2.5-VL-7B（34.2%）。在 TOMATO 中，考察 VLM 的细粒度时间感知，Kimi-VL 达到了 31.7%，超越了 Qwen2.5-VL-7B（27.6%）和 GPT-4o-Mini（28.8%）。这些结果展示了我们模型理解动态视觉内容、随时间跟踪物体以及解释视频序列中复杂动作的强大能力，使其非常适合需要时间视觉理解的应用。

4.2 Kimi-VL的推理扩展

此外，我们进行推理扩展，以使Kimi-VL能够通过长链推理（CoT）进行推理，并通过长链推理激活和强化学习呈现模型的长思考版本Kimi-VL-Thinking。我们在多个图像和视频基准测试中验证了其卓越的性能，如表4所示。

Kimi-VL-Thinking 在基础 Kimi-VL 模型上显著提升，MathVista 上提高了 2.6%，MMMU 上提高了 4.7%，MathVision 上提高了 15.4%，展示了其利用测试时计算进行更深层次推理和更好处理复杂多模态查询的能力。在表 4 中，Kimi-VL-Thinking 进一步超越或与最先进的思维和非思维模型相媲美：在 MathVista 上达到 71.3%，超越 GPT-4o（63.8%）和 GPT-4o-mini（56.7%）；在 MMMU 上得分 61.7%，超过 GPT-4o-mini（60.0%）和 Qwen2.5-VL-7B（58.6%）；在 MathVision 上达到 36.8%，超过 GPT-4o（30.4%）和 Gemma-3-27B-IT（35.5%），甚至 QVQ-72B（35.9%）。尽管在某些基准测试上略微落后于一些大规模模型，Kimi-VL-Thinking 仅用 3B 激活参数就取得了这些结果——比其对应模型少几个数量级——突显了其在多模态推理中的强大效率和有效性。

Our Kimi-VL-Thinking model also exhibits strong test-time scaling properties, as shown in Figure 13. Specifically, increasing the max thinking token length at inference time consistently improves test-time accuracy across all three benchmarks. For example, on **MathVision**, the accuracy rises steadily from 18.7% at 1k tokens to 36.8% at 16k tokens, and similar upward trend is also observed on **MMMU**, indicating that the model is able to utilize longer reasoning chains for better performance. However, not all benchmarks benefit equally from longer thinking lengths. On **MathVista**, performance saturates early, with accuracy reaching 70.9% at 4k tokens and no further significant gains observed as the token length increases to 16k. It suggests that for this task, the necessary reasoning depth is already captured within a relatively short context, and additional computation does not yield further improvements.

5 Conclusion, Limitation, and Future Work

We introduce Kimi-VL, a VLM designed with a balanced approach to cover both multimodal and text-only pre-training/post-training, underpinned by an MoE-based architecture for scalable efficiency. Its 128K extended context window enables precise retrieval in lengthy texts and videos, while the native-resolution encoder MoonViT helps maintain high accuracy with low computational overhead in ultra-high-resolution visual tasks. Additionally, Kimi-VL-Thinking facilitates effective long-chain reasoning in complex image and video inference. Overall, Kimi-VL demonstrates robust adaptability and efficiency across multimodal, long-context, and high-resolution tasks, indicating substantial potential for future research and industrial applications.

However, Kimi-VL still faces several challenges:

1. Although the current model size performs effectively for many standard tasks, it remains too limited to address highly specialized or domain-specific problems, or problems that are strongly dependent on language abilities, restricting Kimi-VL's ability to handle extremely complex scenarios.
2. While the reasoning capability is already strong for typical use cases, it has yet to reach its theoretical upper bound, particularly for intricate tasks requiring multi-step inference or deeper contextual understanding.
3. Despite providing a 128K extended context window, due to limited parameters in its attention layers (which is only comparable to a 3B model), its long-context abilities is still insufficient for certain advanced applications that involve extremely long sequences or high-volume contextual information.

In the future, we will tackle these challenges by scaling up the model size, expanding pre-training data, and enhancing post-training algorithms. Our next steps include optimizing Kimi-VL and releasing larger versions, as well as refining post-training and test-time scaling mechanisms for a better thinking model. These efforts will pave the way for more advanced applications in both research and industry.

References

Amazon Web Services. *Amazon Simple Storage Service (Amazon S3)*. Web. Available at: https://aws.amazon.com/s3/. 2023. URL: https://aws.amazon.com/s3/ (visited on 12/15/2023).

Bai, Shuai et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: https://arxiv.org/abs/2502.13923.

Bonatti, Rogerio et al. *Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale*. 2024. arXiv: 2409.08264 [cs.AI]. URL: https://arxiv.org/abs/2409.08264.

Chen, Lin et al. "Are We on the Right Way for Evaluating Large Vision-Language Models?" In: *arXiv preprint arXiv:2403.20330* (2024).

Chen, Tianqi et al. *Training Deep Nets with Sublinear Memory Cost*. 2016. arXiv: 1604.06174 [cs.LG]. URL: https://arxiv.org/abs/1604.06174.

Cheng, Kanzhi et al. "Seeclick: Harnessing gui grounding for advanced visual gui agents". In: *arXiv preprint arXiv:2401.10935* (2024).

Dao, Tri et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. arXiv: 2205.14135 [cs.LG]. URL: https://arxiv.org/abs/2205.14135.

DeepSeek-AI, Daya Guo, et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

DeepSeek-AI, Aixin Liu, et al. *DeepSeek-V3 Technical Report*. 2025. arXiv: 2412.19437 [cs.CL]. URL: https://arxiv.org/abs/2412.19437.

Dehghani, Mostafa et al. *Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution*. 2023. arXiv: 2307.06304 [cs.CV]. URL: https://arxiv.org/abs/2307.06304.

我们的 Kimi-VL-Thinking 模型在测试时也表现出强大的扩展性，如图 13 所示。具体而言，在推理时增加最大思考令牌长度始终能提高所有三个基准的测试准确性。例如，在 MathVision 上，准确率从 1k 令牌的 18.7 % 稳步上升至 16k 令牌的 36.8%，在 MMMU 上也观察到了类似的上升趋势，这表明模型能够利用更长的推理链以获得更好的性能。然而，并非所有基准都能从更长的思考长度中获得同等的好处。在 MathVista 上，性能很早就达到饱和，准确率在 4k 令牌时达到 70.9%，随着令牌长度增加到 16k，未观察到进一步的显著提升。这表明对于这个任务，所需的推理深度已经在相对较短的上下文中捕获，额外的计算并未带来进一步的改善。

5 结论、局限性与未来工作

我们介绍Kimi-VL，这是一种VLM，采用平衡的方法来覆盖多模态和仅文本的预训练/后训练，基于MoE架构以实现可扩展的效率。其128K扩展上下文窗口能够在冗长的文本和视频中实现精确检索，而原生分辨率编码器MoonViT则有助于在超高分辨率视觉任务中保持高准确性和低计算开销。此外，Kimi-VL-Thinking促进了复杂图像和视频推理中的有效长链推理。总体而言，Kimi-VL在多模态、长上下文和高分辨率任务中表现出强大的适应性和效率，显示出未来研究和工业应用的巨大潜力。

然而，Kimi-VL仍然面临几个挑战：

1. 尽管当前模型的规模在许多标准任务中表现有效，但它仍然过于有限，无法解决高度专业化或特定领域的问题，或那些强烈依赖语言能力的问题，这限制了Kimi-VL处理极其复杂场景的能力。 2. 尽管推理能力对于典型用例已经很强，但它尚未达到其理论上限，特别是对于需要多步骤推理或更深层次上下文理解的复杂任务。 3. 尽管提供了128K的扩展上下文窗口，但由于其注意力层中的参数有限（仅可与3B模型相媲美），其长上下文能力仍不足以满足某些涉及极长序列或高容量上下文信息的高级应用。

在未来，我们将通过扩大模型规模、扩展预训练数据和增强后训练算法来应对这些挑战。我们的下一步包括优化 Kimi-VL 并发布更大版本，以及改进后训练和测试时间缩放机制，以实现更好的思维模型。这些努力将为研究和工业中的更高级应用铺平道路。

参考文献

亚马逊网络服务。*Amazon Simple Storage Service (Amazon S3)*。网络。可在以下网址获取：https://aws.amazon.com/s3/。2023年。网址：https://aws.amazon.com/s3/（访问于2023年12月15日）。白帅等。 *Qwen2.5-VL Technical Report*。2025年。arXiv: 2502.13923 [cs.CV]。网址：https://arxiv.org/abs/2502.13923。博纳提，罗杰里奥等。*Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale*。2024年。arXiv: 2409.08264 [cs.AI]。网址：https://arxiv.org/abs/2409.08264。陈林等。“我们在评估大型视觉-语言模型的正确道路上吗？”在：*arXiv preprint arXiv:2403.20330* (2024)。陈天奇等。 *Training Deep Nets with Sublinear Memory Cost*。2016年。arXiv: 1604.06174 [cs.LG]。网址：https://arxiv.org/abs/1604.06174。程侃之等。“Seeclick：利用GUI基础为高级视觉GUI代理提供支持”。在：*arXiv preprint arXiv:2401.10935* (2024)。道，Tri等。*FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*。2022年。arXiv: 2205.14135 [cs.LG]。网址：https://arxiv.org/abs/2205.14135。DeepSeek-AI，郭大亚等。 *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*。2025年。arXiv: 2501.12948 [cs.CL]。网址：https://arxiv.org/abs/2501.12948。DeepSeek-AI，刘爱新等。*DeepSeek-V3 Technical Report*。2025年。arXiv: 2412.19437 [cs.CL]。网址：https://arxiv.org/abs/2412.19437。德赫加尼，穆斯塔法等。 *Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution*。2023年。arXiv: 2307.06304 [cs.CV]。网址：https://arxiv.org/abs/2307.06304。

Fedus, William, Barret Zoph, and Noam Shazeer. *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. 2022. arXiv: 2101.03961 [cs.LG]. URL: https://arxiv.org/abs/2101.03961.

Fu, Chaoyou et al. "Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis". In: *arXiv:2405.21075* (2024).

Fu, Xingyu et al. "Blink: Multimodal large language models can see but not perceive". In: *European Conference on Computer Vision*. Springer. 2024, pp. 148–166.

Gadre, Samir Yitzhak et al. "Datacomp: In search of the next generation of multimodal datasets". In: *Advances in Neural Information Processing Systems* 36 (2024).

Grauman, Kristen et al. "Ego4d: Around the world in 3,000 hours of egocentric video". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 18995–19012.

Guo, Jarvis et al. *MAmmoTH-VL: Eliciting Multimodal Reasoning with Instruction Tuning at Scale*. 2024. arXiv: 2412.05237 [cs.CL]. URL: https://arxiv.org/abs/2412.05237.

Hu, Kairui et al. "Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos". In: *arXiv preprint arXiv:2501.13826* (2025).

Huang, Yanping et al. *GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism*. 2019. arXiv: 1811.06965 [cs.CV]. URL: https://arxiv.org/abs/1811.06965.

Jacobs, Sam Ade et al. *DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models*. 2023. arXiv: 2309.14509 [cs.LG]. URL: https://arxiv.org/abs/2309.14509.

Jordan, Keller et al. *Muon: An optimizer for hidden layers in neural networks*. 2024. URL: https://kellerjordan.github.io/posts/muon/.

Kembhavi, Aniruddha et al. "A diagram is worth a dozen images". In: *European conference on computer vision*. Springer. 2016, pp. 235–251.

Korthikanti, Vijay et al. *Reducing Activation Recomputation in Large Transformer Models*. 2022. arXiv: 2205.05198 [cs.LG]. URL: https://arxiv.org/abs/2205.05198.

Laurençon, Hugo et al. "Obelics: An open web-scale filtered dataset of interleaved image-text documents". In: *Advances in Neural Information Processing Systems* 36 (2024).

Li, Bo et al. *LLaVA-OneVision: Easy Visual Task Transfer*. 2024. arXiv: 2408.03326 [cs.CV]. URL: https://arxiv.org/abs/2408.03326.

Li, Dongxu et al. *Aria: An Open Multimodal Native Mixture-of-Experts Model*. 2024. arXiv: 2410.05993 [cs.CV]. URL: https://arxiv.org/abs/2410.05993.

Li, Kaixin et al. "ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use". In: *Workshop on Reasoning and Planning for Large Language Models*. 2025.

Li, Shen et al. *PyTorch Distributed: Experiences on Accelerating Data Parallel Training*. 2020. arXiv: 2006.15704 [cs.DC]. URL: https://arxiv.org/abs/2006.15704.

Liu, Hao, Matei Zaharia, and Pieter Abbeel. *Ring Attention with Blockwise Transformers for Near-Infinite Context*. 2023. arXiv: 2310.01889 [cs.CL]. URL: https://arxiv.org/abs/2310.01889.

Liu, Jingyuan et al. "Muon is Scalable for LLM Training". In: *arXiv preprint arXiv:2502.16982* (2025).

– "Muon is Scalable for LLM Training". In: *arXiv preprint arXiv:2502.16982* (2025).

Liu, Yuan et al. "MMBench: Is Your Multi-modal Model an All-around Player?" In: *arXiv:2307.06281* (2023).

Liu, Yuliang et al. "On the hidden mystery of ocr in large multimodal models". In: *arXiv e-prints* (2023), arXiv–2305.

Lu, Pan et al. "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts". In: *arXiv preprint arXiv:2310.02255* (2023).

Mangalam, Karttikeya, Raiymbek Akshulakov, and Jitendra Malik. "Egoschema: A diagnostic benchmark for very long-form video language understanding". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 46212–46244.

Mathew, Minesh et al. "Infographicvqa". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1697–1706.

Narayanan, Deepak et al. *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM*. 2021. arXiv: 2104.04473 [cs.CL]. URL: https://arxiv.org/abs/2104.04473.

OpenAI. "Learning to reason with LLMs". In: (2024). URL: https://openai.com/index/learning-to-reason-with-llms/.

OpenAI et al. *GPT-4o System Card*. 2024. arXiv: 2410.21276 [cs.CL]. URL: https://arxiv.org/abs/2410.21276.

Rajbhandari, Samyam et al. "Zero: Memory optimizations toward training trillion parameter models". In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2020, pp. 1–16.

Schuhmann, Christoph et al. "Laion-5b: An open large-scale dataset for training next generation image-text models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25278–25294.

Fedus, William, Barret Zoph, 和 Noam Shazeer. *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. 2022. arXiv: 2101.03961 [cs.LG]. URL: https://arxiv.org/abs/2101.03961. Fu, Chaoyou 等. “视频-MME：多模态 LLM 在视频分析中的首次全面评估基准”。在：*arXiv:2405.21075* (2024)。Fu, Xingyu 等. “Blink：多模态大型语言模型可以看见但无法感知”。在：*European Conference on Computer Vision*。Springer. 2024, 第 148–166 页。Gadre, Samir Yitzhak 等. “Datacomp：寻找下一代多模态数据集”。在：*Advances in Neural Information Processing Systems* 36 (2024)。Grauman, Kristen 等. “Ego4d：在 3000 小时的自我中心视频中环游世界”。在：*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*。2022, 第 18995–19012 页。Guo, Jarvis 等. *MAmmoTH-VL: Eliciting Multimodal Reasoning with Instruction Tuning at Scale*. 2024. arXiv: 2412.05237 [cs.CL]. URL: https://arxiv.org/abs/2412.05237. Hu, Kairui 等. “视频-MMMU：评估来自多学科专业视频的知识获取”。在：*arXiv preprint arXiv:2501.13826* (2025)。Huang, Yanping 等. *GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism*. 2019. arXiv: 1811.06965 [cs.CV]. URL: https://arxiv.org/abs/1811.06965. Jacobs, Sam Ade 等. *DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models*. 2023. arXiv: 2309.14509 [cs.LG]. URL: https://arxiv.org/abs/2309.14509. Jordan, Keller 等. *Muon: An optimizer for hidden layers in neural networks*. 2024. URL: https://kellerjordan.github.io/posts/muon/. Kembhavi, Aniruddha 等. “一张图胜过十张图像”。在：*European conference on computer vision*。Springer. 2016, 第 235–251 页。Korthikanti, Vijay 等. *Reducing Activation Recomputation in Large Transformer Models*. 2022. arXiv: 2205.05198 [cs.LG]. URL: https://arxiv.org/abs/2205.05198. Laurençon, Hugo 等. “Obelics：一个开放的网络规模过滤数据集，包含交错的图像-文本文档”。在：*Advances in Neural Information Processing Systems* 36 (2024)。Li, Bo 等. *LLaVA-OneVision: Easy Visual Task Transfer*. 2024. arXiv: 2408.03326 [cs.CV]. URL: https://arxiv.org/abs/2408.03326. Li, Dongxu 等. *Aria: An Open Multimodal Native Mixture-of-Experts Model*. 2024. arXiv: 2410.05993 [cs.CV]. URL: https://arxiv.org/abs/2410.05993. Li, Kaixin 等. “ScreenSpot-Pro：专业高分辨率计算机使用的 GUI 定位”。在：*Workshop on Reasoning and Planning for Large Language Models*。2025。Li, Shen 等. *PyTorch Distributed: Experiences on Accelerating Data Parallel Training*. 2020. arXiv: 2006.15704 [cs.DC]. URL: https://arxiv.org/abs/2006.15704. Liu, Hao, Matei Zaharia, 和 Pieter Abbeel. *Ring Attention with Blockwise Transformers for Near-Infinite Context*. 2023. arXiv: 2310.01889 [cs.CL]. URL: https://arxiv.org/abs/2310.01889. Liu, Jingyuan 等. “Muon 可扩展用于 LLM 训练”。在：*arXiv preprint arXiv:2502.16982* (2025)。– “Muon 可扩展用于 LLM 训练”。在：*arXiv preprint arXiv:2502.16982* (2025)。Liu, Yuan 等. “MMBench：你的多模态模型是全能选手吗？”在：*arXiv:2307.06281* (2023)。Liu, Yuliang 等. “关于大型多模态模型中 OCR 的隐藏奥秘”。在：*arXiv e-prints* (2023)，arXiv–2305。Lu, Pan 等. “Mathvista：评估基础模型在视觉上下文中的数学推理”。在：*arXiv preprint arXiv:2310.02255* (2023)。Mangalam, Karttikeya, Raiymbek Akshulakov, 和 Jitendra Malik. “Egoschema：一个用于非常长视频语言理解的诊断基准”。在：*Advances in Neural Information Processing Systems* 36 (2023), 第 46212–46244 页。Mathew, Minesh 等. “Infographicvqa”。在：*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*。2022, 第 1697–1706 页。Narayanan, Deepak 等. *Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM*. 2021. arXiv: 2104.04473 [cs.CL]. URL: https://arxiv.org/abs/2104.04473. OpenAI. “学习用 LLM 推理”。在：（2024）。URL: https://openai.com/index/learning-to-reason-with-llms/. OpenAI 等. *GPT-4o System Card*. 2024. arXiv: 2410.21276 [cs.CL]. URL: https://arxiv.org/abs/2410.21276。Rajbhandari, Samyam 等. “Zero：训练万亿参数模型的内存优化”。在：*SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*。IEEE. 2020, 第 1–16 页。Schuhmann, Christoph 等. “Laion-5b：一个开放的大规模数据集，用于训练下一代图像-文本模型”。在：*Advances in Neural Information Processing Systems* 35 (2022), 第 25278–25294 页。

Shangguan, Ziyao et al. "TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models". In: *International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=fCi4o83Mfs.
Su, Dan et al. "Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset". In: *arXiv preprint arXiv:2412.02595* (2024).
Su, Jianlin et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. 2023. arXiv: 2104.09864 [cs.CL]. URL: https://arxiv.org/abs/2104.09864.
Team, Gemini et al. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. 2024. arXiv: 2403.05530 [cs.CL]. URL: https://arxiv.org/abs/2403.05530.
Team, Gemma et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.
Team, Kimi et al. "Kimi k1. 5: Scaling reinforcement learning with llms". In: *arXiv preprint arXiv:2501.12599* (2025).
Tong, Shengbang et al. *Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs*. 2024. arXiv: 2406.16860 [cs.CV]. URL: https://arxiv.org/abs/2406.16860.
Wang, Ke et al. "Measuring multimodal mathematical reasoning with math-vision dataset". In: *arXiv preprint arXiv:2402.14804* (2024).
Wei, Haoran et al. "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model". In: *arXiv preprint arXiv:2409.01704* (2024).
Wu, Haoning et al. "Longvideobench: A benchmark for long-context interleaved video-language understanding". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 28828–28857.
Wu, Zhiyong et al. "Os-atlas: A foundation action model for generalist gui agents". In: *arXiv preprint arXiv:2410.23218* (2024).
Wu, Zhiyu et al. *DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding*. 2024. arXiv: 2412.10302 [cs.CV]. URL: https://arxiv.org/abs/2412.10302.
x.ai. "Grok-1.5 Vision Preview". In: (2024). URL: https://x.ai/news/grok-1.5v.
Xie, Tianbao et al. "Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 52040–52094.
Yang, Jihan et al. "Thinking in space: How multimodal large language models see, remember, and recall spaces". In: *arXiv preprint arXiv:2412.14171* (2024).
Yu, Jiahui et al. *CoCa: Contrastive Captioners are Image-Text Foundation Models*. 2022. arXiv: 2205.01917 [cs.CV]. URL: https://arxiv.org/abs/2205.01917.
Yu, Weihao et al. "Mm-vet: Evaluating large multimodal models for integrated capabilities". In: *International conference on machine learning*. PMLR. 2024.
Yue, Xiang, Yuansheng Ni, 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*. 2024, pp. 9556–9567.
Yue, Xiang, Xingwei Qu, et al. "Mammoth: Building math generalist models through hybrid instruction tuning". In: *arXiv preprint arXiv:2309.05653* (2023).
Zhai, Xiaohua et al. *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303.15343 [cs.CV]. URL: https://arxiv.org/abs/2303.15343.
Zhao, Yilun et al. "MMVU: Measuring Expert-Level Multi-Discipline Video Understanding". In: *arXiv preprint arXiv:2501.12380* (2025).
Zhou, Junjie et al. "Mlvu: A comprehensive benchmark for multi-task long video understanding". In: *arXiv preprint arXiv:2406.04264* (2024).
Zhu, Wanrong et al. "Multimodal c4: An open, billion-scale corpus of images interleaved with text". In: *Advances in Neural Information Processing Systems* 36 (2024).

商冠, 子耀 等. “TOMATO: 评估多模态基础模型的视觉时间推理能力”. 在:
International Conference on Learning Representations. 2025. URL: https://openreview.net/forum?id=fCi4o83Mfs.
苏, 丹 等. “Nemotron-CC: 将公共爬虫转化为精炼的长时间预训练数据集”. 在:
arXiv preprint arXiv:2412.02595 (2024). 苏, 建林 等.
RoFormer: Enhanced Transformer with Rotary Position Embedding. 2023. arXiv: 2104 . 09864 [cs.CL]. URL: https://arxiv.org/abs/2104.09864. 团队, Gemini 等.
Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 2024. arXiv: 2403.05530 [cs.CL]. URL: https://arxiv.org/abs/2403.05530. 团队, Gemma 等. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786. 团队, Kimi 等. “Kimi k1.5: 用 llms 扩展强化学习”. 在:
arXiv preprint arXiv:2501.12599 (2025). 唐, 盛邦 等.
Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. 2024. arXiv: 2406.16860 [cs.CV]. URL: https://arxiv.org/abs/2406.16860. 王, 可 等. “使用数学视觉数据集测量多模态数学推理”. 在:
arXiv preprint arXiv:2402.14804 (2024). 魏, 浩然 等. “通用 OCR 理论: 通过统一的端到端模型迈向 OCR-2.0”. 在: *arXiv preprint arXiv:2409.01704* (2024). 吴, 昊宁 等. “Longvideobench: 一个用于长上下文交错视频语言理解的基准”. 在: *Advances in Neural Information Processing Systems* 37 (2024), 第 28828–28857 页. 吴, 志勇 等. “Os-atlas: 一种通用 GUI 代理的基础动作模型”. 在: *arXiv preprint arXiv:2410.23218* (2024). 吴, 志宇 等.
DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. 2024. arXiv: 2412.10302 [cs.CV]. URL: https://arxiv.org/abs/2412.10302. x.ai. “Grok-1.5 视觉预览”. 在: (2024). URL: https://x.ai/news/grok-1.5v. 谢, 天宝 等. “Osworld: 在真实计算环境中对开放式任务的多模态代理进行基准测试”. 在: *Advances in Neural Information Processing Systems* 37 (2024), 第 52040–52094 页. 杨, 继涵 等. “在空间中思考: 多模态大型语言模型如何看待、记忆和回忆空间”. 在: *arXiv preprint arXiv:2412.14171* (2024). 余, 嘉辉 等.
CoCa: Contrastive Captioners are Image-Text Foundation Models. 2022. arXiv: 2205.01917 [cs.CV]. URL: https://arxiv.org/abs/2205.01917. 余, 伟豪 等. “Mm-vet: 评估大型多模态模型的综合能力”. 在: *International conference on machine learning*. PMLR. 2024. 岳, 祥, 袁生 尼 等. “Mmmu: 一个针对专家 AGI 的大规模多学科多模态理解和推理基准”. 在: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, 第 9556–9567 页. 岳, 祥, 兴伟 曲 等. “Mammoth: 通过混合指令调优构建数学通用模型”. 在:
arXiv preprint arXiv:2309.05653 (2023). 翟, 小华 等. *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303 . 15343 [cs.CV]. URL: https://arxiv.org/abs/2303.15343. 赵, 逸伦 等. “MMVU: 测量专家级多学科视频理解”. 在: *arXiv preprint arXiv:2501.12380* (2025). 周, 军杰 等. “Mlvu: 一个综合性的多任务长视频理解基准”. 在: *arXiv preprint arXiv:2406.04264* (2024). 朱, 万荣 等. “多模态 C4: 一个开放的、十亿规模的图像与文本交错的语料库”. 在: *Advances in Neural Information Processing Systems* 36 (2024).

Appendix

A Contributions

Core Contributors

Bohong Yin
Bowei Xing
Cheng Chen
Chu Wei
Dehao Zhang
Dongliang Wang
Haoning Wu*
Haotian Yao
Haoyu Lu*
Hao Yang
Lin Sui
Xinyuan Wang#
Xinyu Zhou
Yang Li
Y. Charles*
Yiping Bao
Yimin Chen
Yuxin Wu
Zaida Zhou
Zhaowei Li
Zhiqi Huang
Zhilin Yang
Ziwei Chen

Contributors

Angang Du
Bowen Qu
Bowen Wang#
Chenlin Zhang
Chenzhuang Du
Congcong Wang
Dikang Du
Enming Yuan
Enzhe Lu
Fang Li
Flood Sung
Guangda Wei
Guokun Lai
Han Zhu
Hao Ding
Hao Zhang
Heng Wang
Hongcheng Gao
Huabin Zheng
Jiaming Li
Jianlin Su
Jianzhou Wang
Jiaqi Deng#
Jiezhong Qiu
Jin Xie
Jinhong Wang
Jingyuan Liu
Junjie Yan
Liang Chen
Longhui Yu
Mengfan Dong
Mengnan Dong
Nuo Xu
Runjie Zhou
Shaowei Liu
Sihan Cao
Tao Yu#
Tianhui Song
Tongtong Bai
Weiran He
Weixiao Huang
Weixin Xu
Xiaokun Yuan
Xingzhe Wu
Xingcheng Yao
Xinxing Zu
Yangyang Hu
Yan Zhong
Yanru Chen
Yibo Miao
Yejie Wang
Yibo Liu
Yidao Qin
Yiqin Wang
Yongsheng Kang
Yulun Du
Yuzhi Wang
Yuzi Yan
Zhejun Jiang
Zheng Zhang
Zihao Huang
Zijia Zhao

* Project lead(s).
# The University of Hong Kong, Moonshot.ai
The listing of authors is in alphabetical order based on their first names.

附录

A 贡献

核心贡献者

博鸿 尹博伟 邢
承辰 处伟 德浩
张东亮 王浩宁
吴* 浩天 姚浩宇
陆* 昊阳 林隋
新源 王# 新宇
周扬 李 Y. 查尔
斯* 逸平 包怡敏
陈宇欣 吴在达
周兆伟 李志奇
黄志林 杨子维
陈

贡献者

安刚 杜 博文 曲
博文 王# 陈林
张 辰庄 杜 从聪
王 迪康 杜 恩铭
袁 恩哲 陆 方
李 洪 浪 宋 光
达 魏 国坤 赖
汉 朱 浩 丁 浩
张 恒 王
洪城 高 华彬 郑
佳铭 李 建林 苏
建洲 王 佳琪 邓
杰中 邱 锦 谢
金红 王 靖远 刘
军杰 闫 梁 陈
龙辉 余 梦凡 董
梦南 董 诺 许
润杰 周 少伟 刘
思涵 曹 涛 余#
天辉 宋 通通 白
伟然 何 伟晓 黄
伟欣 许 晓坤 袁
星哲 吴 兴成 姚
新星 祖 杨阳 胡
燕 钟 闫如 陈
逸博 苗 叶杰 王
逸博 刘 逸道 秦
依琴 王 永生 康
玉伦 杜 余志 王
余子 闫 哲俊 江
正 张 子豪 黄
子佳 赵

* Project lead(s).
# The University of Hong Kong, Moonshot.ai
The listing of authors is in alphabetical order based on their first names.

B Evaluation Details

B.1 Image Benchmark

MMMU (Yue, Ni, et al. 2024) encompasses a carefully curated collection of 11.5K multimodal questions sourced from college exams, quizzes, and textbooks. These questions span six major academic fields: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.

MMBench-EN-v1.1 (Yuan Liu et al. 2023) is a fine-grained benchmark that contains 2974 multiple-choice questions, covering 20 ability dimensions. It incorporate perception and reasoning as the top-level ability dimensions in its ability taxonomy, leading to different levels of evaluation in various ability dimensions.

MMStar (Lin Chen et al. 2024) is an elite vision-indispensable multimodal benchmark comprising 1,500 challenge samples meticulously selected by humans. It is designed to benchmark 6 core capabilities and 18 detailed axes, aiming to evaluate the multimodal capacities of LVLMs with a carefully balanced and purified selection of samples.

MMVet (W. Yu et al. 2024) is designed based on the insight that the intriguing ability to solve complicated tasks is often achieved by a generalist model being able to integrate different core vision-language capabilities. It defines 6 core VL capabilities and examines the 16 integrations of interest derived from the capability combination.

RealWorldQA (x.ai 2024) is a benchmark designed to evaluate the real-world spatial understanding capabilities of multimodal models. It assesses how well the models comprehend physical environments. The benchmark consists of over 700 images, each accompanied by a question and a verifiable answer, and these images are drawn from various real-world scenarios.

AI2D (Kembhavi et al. 2016) is a dataset of over 5000 grade school science diagrams with over 150000 rich annotations, their ground truth syntactic parses, and more than 15000 corresponding multiple choice questions.

MathVision (K. Wang et al. 2024) is a carefully curated collection of 3,040 high-quality mathematical problems with visual contexts that are sourced from real math competitions. It covers 16 distinct mathematical disciplines and is graded across 5 levels of difficulty. This dataset offers a comprehensive and diverse set of challenges, making it ideal for evaluating the mathematical reasoning abilities of LMMs.

MathVista (P. Lu et al. 2023) is a benchmark that integrates challenges from a variety of mathematical and visual tasks, demanding participants to exhibit fine-grained, deep visual understanding along with compositional reasoning to successfully complete the tasks.

BLINK (X. Fu et al. 2024) is a benchmark designed to evaluate multi-image visual cognition, encompassing tasks related to depth relationships, feature matching, digital forensics, and spatiotemporal reasoning. It features a diverse set of multi-image perceptual similarity tasks, validated through standardized protocols.

InfoVQA (Mathew et al. 2022) is a dataset specifically designed to assess models' capabilities in interpreting and reasoning with complex infographics that integrate text, graphics, and visual elements. Model performance on this dataset is evaluated using the ANLS metric on the test set.

OCRBench (Yuliang Liu et al. 2023) evaluates the OCR capabilities of MLLMs across five tasks: text recognition, scene text VQA, document VQA, key information extraction, and handwritten math expression recognition. The benchmark is scored out of a maximum of 1000 points.

B.2 Video and Long Document Benchmark

VideoMMMU (K. Hu et al. 2025) is a video benchmark designed to evaluate the college-level knowledge acquisition capabilities of large multimodal models. It consists of 300 expert-level videos and 900 human-annotated questions. The videos span six diverse academic disciplines: Art, Humanities, Medicine, Business, Science, and Engineering. The questions are structured to align with three cognitive stages: Perception, Comprehension, and Adaptation.

MMVU (Y. Zhao et al. 2025) is a video benchmark designed to evaluate the expert-level video understanding ability. The benchmark contains 3,000 expert-annotated questions over 1,529 videos, which span 27 subjects from four core disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering.

Video-MME (C. Fu et al. 2024) is a video benchmark that consists of 900 manually selected videos (totaling 254 hours length), and 2,700 QA pairs. The videos, varying in duration, are categorized into 30 fine-grained classes across six diverse domains: Knowledge, Film & Television, Sports Competition, Artistic Performance, Life Record, and Multilingual content. Evaluations are conducted under two different settings: with and without subtitles.

B 评估细节

B.1 图像基准测试

MMMU（Yue, Ni 等，2024）包含了一组精心策划的11.5K多模态问题，这些问题来源于大学考试、测验和教科书。这些问题涵盖六个主要学术领域：艺术与设计、商业、科学、健康与医学、人文学科与社会科学，以及技术与工程。

MMBench-EN-v1.1（Yuan Liu等，2023）是一个细粒度基准，包含2974个多项选择题，涵盖20个能力维度。它将感知和推理作为其能力分类中的顶级能力维度，从而在不同的能力维度中导致不同级别的评估。

MMStar（林晨等，2024）是一个精英视觉不可或缺的多模态基准，包含1,500个由人类精心挑选的挑战样本。它旨在基准测试6个核心能力和18个详细轴线，旨在通过精心平衡和纯化的样本选择来评估LVLM的多模态能力。

MMVet（W. Yu 等，2024）是基于这样一个见解而设计的：解决复杂任务的有趣能力通常是通过一个能够整合不同核心视觉-语言能力的通用模型来实现的。它定义了6个核心视觉-语言能力，并考察了从能力组合中衍生出的16个感兴趣的整合。

RealWorldQA (x.ai 2024) 是一个旨在评估多模态模型在现实世界空间理解能力的基准。它评估模型对物理环境的理解程度。该基准包含超过 700 张图像，每张图像都附有一个问题和一个可验证的答案，这些图像来自各种现实世界场景。

AI2D（Kembhavi 等，2016）是一个包含超过 5000 个小学科学图表的数据集，拥有超过 150000 个丰富的注释、它们的真实语法解析，以及超过 15000 个相应的多项选择题。

MathVision (K. Wang et al. 2024) 是一个精心策划的集合，包含 3,040 个高质量的数学问题，这些问题具有视觉背景，来源于真实的数学竞赛。它涵盖了 16 个不同的数学学科，并分为 5 个难度等级。该数据集提供了一套全面而多样的挑战，非常适合评估 LMM 的数学推理能力。

MathVista（P. Lu 等，2023）是一个基准，整合了来自各种数学和视觉任务的挑战，要求参与者展示细致的深度视觉理解以及组合推理，以成功完成任务。

BLINK（X. Fu 等，2024）是一个旨在评估多图像视觉认知的基准，涵盖与深度关系、特征匹配、数字取证和时空推理相关的任务。它具有一套多样化的多图像感知相似性任务，通过标准化协议进行了验证。

InfoVQA（Mathew 等，2022）是一个专门设计的数据库，用于评估模型在解释和推理复杂信息图方面的能力，这些信息图结合了文本、图形和视觉元素。模型在该数据集上的表现通过在测试集上使用 ANLS 指标进行评估。

OCRBench（刘宇亮等，2023）评估了多语言大模型（MLLMs）在五个任务上的OCR能力：文本识别、场景文本视觉问答（VQA）、文档视觉问答、关键信息提取和手写数学表达式识别。该基准的满分为1000分。

B.2 视频和长文档基准测试

VideoMMMU（K. Hu 等，2025）是一个视频基准，旨在评估大型多模态模型的大学水平知识获取能力。它由300个专家级视频和900个人工标注的问题组成。这些视频涵盖六个不同的学科：艺术、人文学科、医学、商业、科学和工程。问题的结构与三个认知阶段相对应：感知、理解和适应。

MMVU（Y. Zhao 等，2025）是一个视频基准，旨在评估专家级视频理解能力。该基准包含3,000个专家标注的问题，涵盖1,529个视频，涉及四个核心学科的27个主题：科学、医疗保健、人文学科与社会科学以及工程。

Video-MME（C. Fu 等，2024）是一个视频基准，包含900个手动选择的视频（总时长254小时）和2,700个问答对。这些视频时长各异，分为六个不同领域中的30个细分类别：知识、影视、体育竞赛、艺术表演、生活记录和多语言内容。评估在两种不同的设置下进行：有字幕和无字幕。

MLVU (J. Zhou et al. 2024) is designed to evaluate the model performance in comprehending long videos from multiple aspects. It consists of 1,730 videos along with 3,102 corresponding question-answer pairs (2,593 in dev set and 509 in test set). Videos of this benchmark are collected from multiple scenarios, including Sport, Ego-centric, Life Record, Tutorial, etc. The close-ended task set of MLVU comprises 7 different tasks: Action Order, Action Count, Topic Reasoning, Anomaly Recognition, Plot QA, Ego Reasoning, and Needle QA.

LongVideoBench (H. Wu et al. 2024) is a video question-answering benchmark designed to evaluate the long-form multimodal perception and relation capability of large multimodal models. The benchmark includes 3,763 web-collected videos spanning various lengths and themes, along with their corresponding subtitles. It includes 6,678 human-annotated multiple-choice questions, distributed across 17 fine-grained categories, which accesses different aspects of video-language understanding.

EgoSchema (Mangalam et al. 2023) is a video benchmark designed to evaluate the long-form video understanding capabilities within the ego-centric scenario. Derived from Ego4D (Grauman et al. 2022), the benchmark comprises over 5,031 multiple choice question-answer pairs spanning more than 250 hours real-world videos with a semi-automatic data pipeline.

VSI-Bench (Yang et al. 2024) is designed to evaluate the visual-spatial comprehensive capabilities of large multimodal models. It consists of over 5,000 question-answer pairs across around 290 real indoor-scene videos.

TOMATO (Shangguan et al. 2025) is a video benchmark comprises 1,484 human-annotated question-answer pairs and 1,417 videos. TOMATO focuses on evaluating the temporal reasoning capabilities of large multimodal models, including action counting, direction prediction, rotation analysis, shape & trend detection, velocity & frequency estimation, and visual cue interpretation.

B.3 Agent Benchmark

ScreenSpot V2 (Zhiyong Wu et al. 2024) is an enhanced version of the ScreenSpot (K. Cheng et al. 2024) benchmark, which focuses on evaluating the performance of GUI grounding models across multiple platforms, including web, desktop, and mobile interfaces. This updated version addresses several issues identified in the original ScreenSpot dataset, such as incorrect or ambiguous annotations, spelling mistakes, and mislabeled bounding boxes.

ScreenSpot Pro (K. Li et al. 2025) is a benchmark for evaluating GUI grounding in high-resolution, complex UI environments. It contains 1,581 real-world, high-resolution images and expert-annotated tasks from diverse professional domains. Including domain-specific interface conventions that challenge models to understand professional-grade interfaces beyond consumer applications.

OSWorld (T. Xie et al. 2024) is a pioneering scalable, real computer environment designed for multimodal agents, facilitating task setup, execution-based evaluation, and interactive learning across multiple operating systems, including Ubuntu, Windows, and macOS. It serves as a unified platform for evaluating open-ended computer tasks that involve arbitrary applications, addressing the limitations of existing benchmarks that often lack interactive environments or are confined to specific applications or domains.

WindowsAgentArena (Bonatti et al. 2024) is a benchmark designed to evaluate multimodal agents in realistic Windows environments. Built on the OSWorld framework, it allows agents to interact with a full range of applications and web tools. The benchmark is scalable and can complete evaluations in under 20 minutes on Azure. It offers insights into agent performance, highlighting the potential for future research in agent development and task automation.

MLVU（J. Zhou 等，2024）旨在从多个方面评估模型在理解长视频方面的表现。它由1,730个视频和3,102个相应的问题-答案对（开发集2,593个，测试集509个）组成。该基准的视频来自多个场景，包括体育、以自我为中心、生活记录、教程等。MLVU的封闭式任务集包括7个不同的任务：动作顺序、动作计数、主题推理、异常识别、情节问答、自我推理和针问答。

LongVideoBench（H. Wu 等，2024）是一个视频问答基准，旨在评估大型多模态模型的长篇多模态感知和关系能力。该基准包括 3,763 个从网络收集的视频，涵盖各种长度和主题，以及相应的字幕。它包含 6,678 个人工标注的多项选择题，分布在 17 个细分类别中，涉及视频语言理解的不同方面。

EgoSchema（Mangalam 等，2023）是一个视频基准，旨在评估自我中心场景中的长视频理解能力。该基准源自 Ego4D（Grauman 等，2022），包含超过 5,031 对多项选择题问答，涵盖超过 250 小时的真实世界视频，并采用半自动数据处理流程。

VSI-Bench（Yang et al. 2024）旨在评估大型多模态模型的视觉-空间综合能力。它由超过5000对问答对组成，涵盖约290个真实室内场景视频。

TOMATO（Shangguan 等，2025）是一个视频基准，包含 1,484 对人类标注的问题-答案对和 1,417 个视频。TOMATO 侧重于评估大型多模态模型的时间推理能力，包括动作计数、方向预测、旋转分析、形状与趋势检测、速度与频率估计以及视觉线索解释。

B.3 代理基准

ScreenSpot V2（吴志勇等，2024）是ScreenSpot（K. Cheng等，2024）基准的增强版本，专注于评估GUI定位模型在多个平台上的性能，包括网页、桌面和移动界面。此更新版本解决了原始ScreenSpot数据集中发现的几个问题，例如不正确或模糊的注释、拼写错误和标记错误的边界框。

ScreenSpot Pro (K. Li et al. 2025) 是一个用于评估高分辨率复杂用户界面环境中 GUI 定位的基准。它包含 1,581 张真实世界的高分辨率图像和来自不同专业领域的专家注释任务。包括特定领域的界面规范，挑战模型理解超越消费应用的专业级界面。

OSWorld (T. Xie et al. 2024) 是一个开创性的可扩展真实计算机环境，旨在为多模态智能体提供支持，促进任务设置、基于执行的评估和跨多个操作系统（包括 Ubuntu、Windows 和 macOS）的互动学习。它作为一个统一的平台，用于评估涉及任意应用程序的开放式计算任务，解决了现有基准测试的局限性，这些基准测试通常缺乏互动环境或仅限于特定应用程序或领域。

WindowsAgentArena（Bonatti 等，2024）是一个旨在评估多模态代理在真实 Windows 环境中的基准测试。它基于 OSWorld 框架构建，允许代理与全范围的应用程序和网络工具进行交互。该基准测试具有可扩展性，并且可以在 Azure 上在 20 分钟内完成评估。它提供了对代理性能的洞察，突显了未来在代理开发和任务自动化方面研究的潜力。