GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning

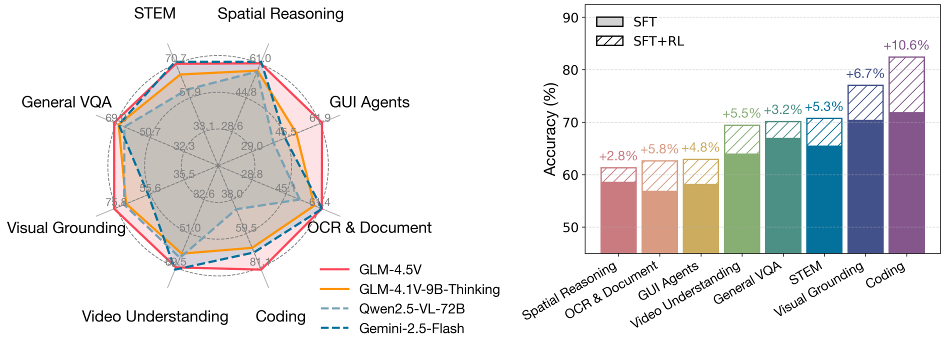
GLM-V Team

Zhipu AI & Tsinghua University

(For the complete list of authors, please refer to the Contribution section)

Abstract

We present GLM-4.1V-Thinking and GLM-4.5V, a family of vision-language models (VLMs) designed to advance general-purpose multimodal understanding and reasoning. In this report, we share our key findings in the development of the reasoning-centric training framework. We first develop a capable vision foundation model with significant potential through large-scale pre-training, which arguably sets the upper bound for the final performance. We then propose **R**einforcement **L**earning with **C**urriculum **S**ampling (**RLCS**) to unlock the full potential of the model, leading to comprehensive capability enhancement across a diverse range of tasks, including STEM problem solving, video understanding, content recognition, coding, grounding, GUI-based agents, and long document interpretation. In a comprehensive evaluation across 42 public benchmarks, GLM-4.5V achieves state-of-the-art performance on nearly all tasks among open-source models of similar size, and demonstrates competitive or even superior results compared to closed-source models such as Gemini-2.5-Flash on challenging tasks including Coding and GUI Agents. Meanwhile, the smaller GLM-4.1V-9B-Thinking remains highly competitive—achieving superior results to the much larger Qwen2.5-VL-72B on 29 benchmarks. We open-source both GLM-4.1V-9B-Thinking and GLM-4.5V. Code, models and more information are released at `https://github.com/zai-org/GLM-V`.



(A) Comparison with baselines.

(B) Reinforcement learning gains.

Figure 1: (A) GLM-4.5V achieves efficient scaling based on its compact predecessor, GLM-4.1V-9B-Thinking, and compares favorably with Gemini-2.5-Flash, according to benchmark assessments. Table 2 presents full performance comparison. (B) Reinforcement learning substantially boosts the model's performance, with gains of up to +10.6% when experimented with GLM-4.5V.

1 Introduction

Vision-language models (VLMs) have become a crucial cornerstone of modern intelligent systems, enabling the perception and understanding of visual information beyond text. Over the past decade, as the intelligence level of models has advanced dramatically [36; 47; 20; 4], the complexity of corresponding multimodal intelligence tasks has increased accordingly. From solving scientific problems [67; 32; 34] to developing autonomous agents [21; 62; 43], the demands on VLMs have far surpassed simple visual content perception [31], with an increasing emphasis on advanced reasoning abilities. Recently, numerous studies have shown that long-form reasoning [60] and scalable reinforcement learning [42] can significantly enhance the ability of large language models (LLMs) to solve complex problems [23; 17]. Several previous works have attempted to enhance the reasoning capabilities of VLMs using similar paradigms [46; 33], but they mainly focus on specific domains. The open-source community currently also lacks a multimodal reasoning model that consistently outperforms traditional non-thinking models of comparable parameter scale across a broad range of scenarios and tasks.

In this report, we share our key findings in the development of GLM-4.1V-Thinking and GLM-4.5V, a family of VLMs designed to advance general-purpose multimodal reasoning. *Our training framework is structured around a unified objective: to comprehensively enhance the model's reasoning capabilities through scalable reinforcement learning.* For pre-training, we curate a broad and diverse corpus of knowledge-intensive multimodal data to equip the model with strong foundational capabilities, including (a) massive image-text pairs with accurate factual knowledge; (b) a self-curated academic corpus with interleaved image and text; (c) annotated documents and diagrams, instructional videos, and grounding data spanning both natural and synthetic images. This foundation model serves as a high-potential multimodal reasoning base for subsequent reinforcement learning. In the supervised fine-tuning phase, we construct carefully designed, domain-specific datasets that teach the model to perform effective reasoning with a standardized format across a wide range of tasks. Finally, we introduce **R**einforcement **L**earning with **C**urriculum **S**ampling (**RLCS**) to drive large-scale, cross-domain reasoning capabilities. RLCS is a multi-domain reinforcement learning framework that combines curriculum learning with difficulty-aware sampling to improve training efficiency by selecting tasks and samples suited to the model's current competence. Our reinforcement learning process enhances training effectiveness and stability, and systematically improves the model's reasoning abilities through interaction and feedback across diverse domains.

To advance research in this field, we open-source GLM-4.1V-9B-Thinking (9 billion parameters) and GLM-4.5V (106B-A12B: 106 billion total parameters, 12 billion activated parameters), both of which achieve state-of-the-art performance among models of comparable size. In a comprehensive evaluation across 42 public benchmarks, GLM-4.5V achieves state-of-the-art performance on nearly all tasks, consistently outperforming strong open-source models such as Step-3 (321B-A38B) and Qwen-2.5-VL-72B, and achieves comparable or even superior performance on 22 benchmarks relative to the closed-source Gemini-2.5-Flash. Notably, GLM-4.5V advances the state-of-the-art for open-source VLMs of comparable size by roughly 10% or more across a wide range of tasks, including general VQA (MMStar, GeoBench), STEM (MMMU Pro, MathVerse, WeMath), chart understanding (ChartQAPro, ChartMuseum), long document understanding (MMLongBench-Doc), visual grounding (TreeBench, Ref-L4-test), spatial reasoning (ERQA), GUI agents (OSWorld, AndroidWorld, WebVoyagerSom, WebQuest), VLM coding (Design2Code, Flame-React-Eval), and video understanding (VideoMMMU, LVBench, MotionBench). GLM-4.1V-9B-Thinking also demonstrates competitive or superior performance compared to much larger models such as Qwen2.5-VL-72B on 29 benchmarks. We further open-source the pre-trained base model, GLM-4.1V-9B-Base, to provide a strong foundation for all researchers to develop and extend their own models.

We summarize our key findings from the development process below and provide more detailed explanations in the following sections.

- **Multi-domain reinforcement learning demonstrates robust cross-domain generalization and mutual facilitation.** Training on one domain boosts performance in others, and joint training across domains yields even greater improvements in each. (See Section 6.3)

- **Dynamically selecting the most informative rollout problems is essential for both efficiency and performance.** Therefore, we propose strategies including Reinforcement Learning with Curriculum

Sampling (RLCS) and dynamic sampling expansion via ratio-based Exponential Moving Average (EMA). (See Section 5.3)

- **A robust and precise reward system is critical for multi-domain RL.** When training a unified VLM across diverse skills, even a slight weakness in the reward signal for one capability can collapse the entire process. (See Section 5.2)

In summary, our contributions are as follows:

- We present GLM-4.1V-Thinking and GLM-4.5V, a family of VLMs developed to advance general-purpose multimodal reasoning. Notably, GLM-4.5V natively supports both "thinking" and "non-thinking" modes, enabling flexible trade-offs between performance and efficiency. We introduce the model design and the reasoning-centric training framework, along with key insights and challenges encountered during the development process.
- We open-source GLM-4.1V-9B-Thinking, GLM-4.1V-9B-Base, GLM-4.5V, and other useful components such as domain-specific reward systems, to facilitate further research in this area. Code, models and more information are released at `https://github.com/zai-org/GLM-V`.
- Comprehensive experiments demonstrate the superiority of the proposed models: GLM-4.5V and GLM-4.1V-9B-Thinking achieve state-of-the-art performance among models of comparable size, with GLM-4.1V-9B-Thinking even surpassing much larger models on several benchmarks. Furthermore, GLM-4.5v matches or outperforms Gemini-2.5-Flash across multiple tasks.

2 Overview and Architecture

Figure 2 shows the shared architecture of GLM-4.1V-Thinking and GLM-4.5V, composed of three core components: a vision encoder, an MLP adapter, and a large language model (LLM) as the decoder. We employ AIMv2-Huge [9] as the initialization of the vision encoder. For the LLM component, we use GLM-4-9B-0414 [13] for the GLM-4.1V-Thinking model, and GLM-4.5-Air [13] for the GLM-4.5V model. Within the vision encoder, we adopt a strategy similar to Qwen2-VL [57], replacing the original 2D convolutions with 3D convolutions. This enables temporal downsampling by a factor of two for video inputs, thereby improving model efficiency. For single-image inputs, the image is duplicated to maintain consistency.

To enable our underlying Vision Transformer (ViT) to support arbitrary image resolutions and aspect ratios, we introduce two adaptations. First, we integrate 2D-RoPE [44] into the ViT's self-attention layers, enabling the model to effectively process images with extreme aspect ratios (over 200:1) or high resolutions (beyond 4K). Second, to preserve the foundational capabilities of the pre-trained ViT, we retain its original learnable absolute position embeddings. During training, these embeddings are dynamically adapted to variable-resolution inputs via bicubic interpolation. Specifically, for an input image divided into a grid of $H_p \times W_p$ patches, the integer coordinates $\mathbf{g} = (w, h)$ of each patch are first normalized to a continuous grid $\mathbf{g}_{\text{norm}}$ spanning $[-1, 1]$:

$$\mathbf{g}_{\text{norm}} = (w_{\text{norm}}, h_{\text{norm}}) = 2 \cdot \left(\frac{w + 0.5}{W_p}, \frac{h + 0.5}{H_p} \right) - 1 \tag{1}$$

These normalized coordinates are then used to sample from the original position embedding table P_{orig} using a bicubic interpolation function $\mathcal{I}_{\text{bicubic}}$ to generate the final adapted embedding P_{adapted} for that patch:

$$P_{\text{adapted}}(\mathbf{g}) = \mathcal{I}_{\text{bicubic}}(P_{\text{orig}}, \mathbf{g}_{\text{norm}}) \tag{2}$$

To further enhance spatial awareness on the language side, we extend RoPE to 3D-RoPE in the LLM. This extension provides superior spatial understanding for multimodal contexts, while preserving the original model's text-related capabilities.

After addressing spatial adaptation, we turn to temporal modeling in video inputs. For videos, we insert a time index token after each frame token, where the time index is implemented by encoding each frame's timestamp as a string. Unlike multi-image inputs, video frames form a temporally coherent sequence. This design explicitly informs the model of the real-world timestamps and temporal distances between frames, thereby boosting its temporal understanding and grounding capabilities.