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

GLM-V Team
Zhipu AI & Tsinghua University

Paper Review

2025. 8. 22. Fri.

중앙대학교 첨단영상대학원 메타버스융합학과 FoV LAB

Hongseok Cho

>> Contents

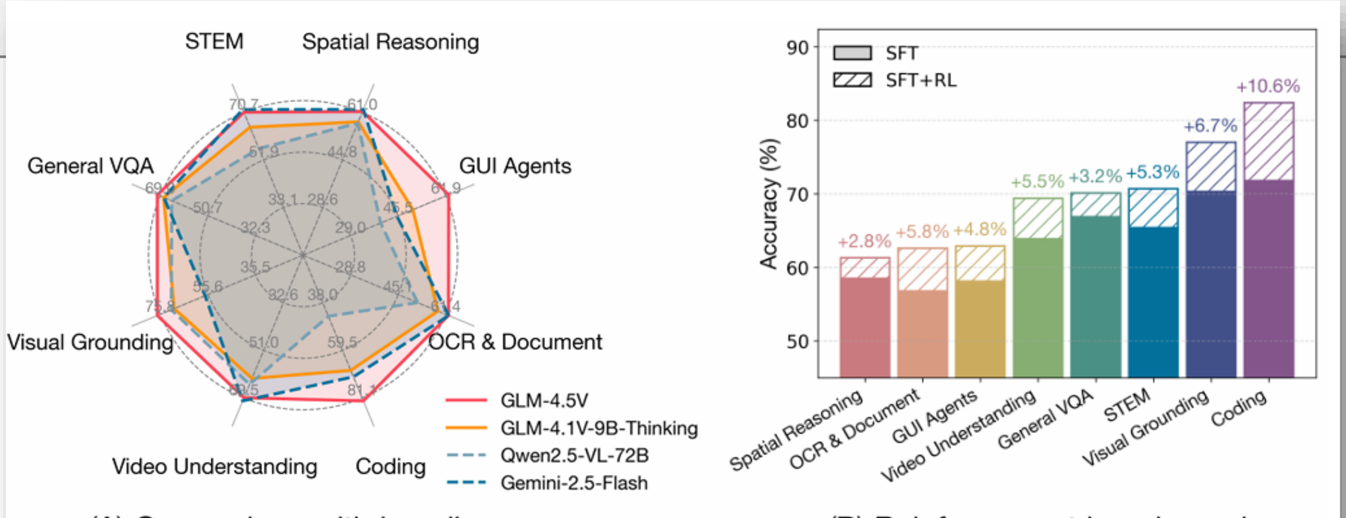
>> Overview

✓ GLM-4.5V&GLM-4.1V-Thinking

- **This paper proposes GLM-4.5V, a multimodal model capable of complex reasoning, along with a smaller variant, GLM-4.1V-Thinking**

- **Using a reinforcement learning technique called RLCS, the models are trained efficiently by selecting data matched to their current capabilities**

- **Despite its smaller size, GLM-4.1V-Thinking outperforms existing SOTA models across multiple benchmarks**

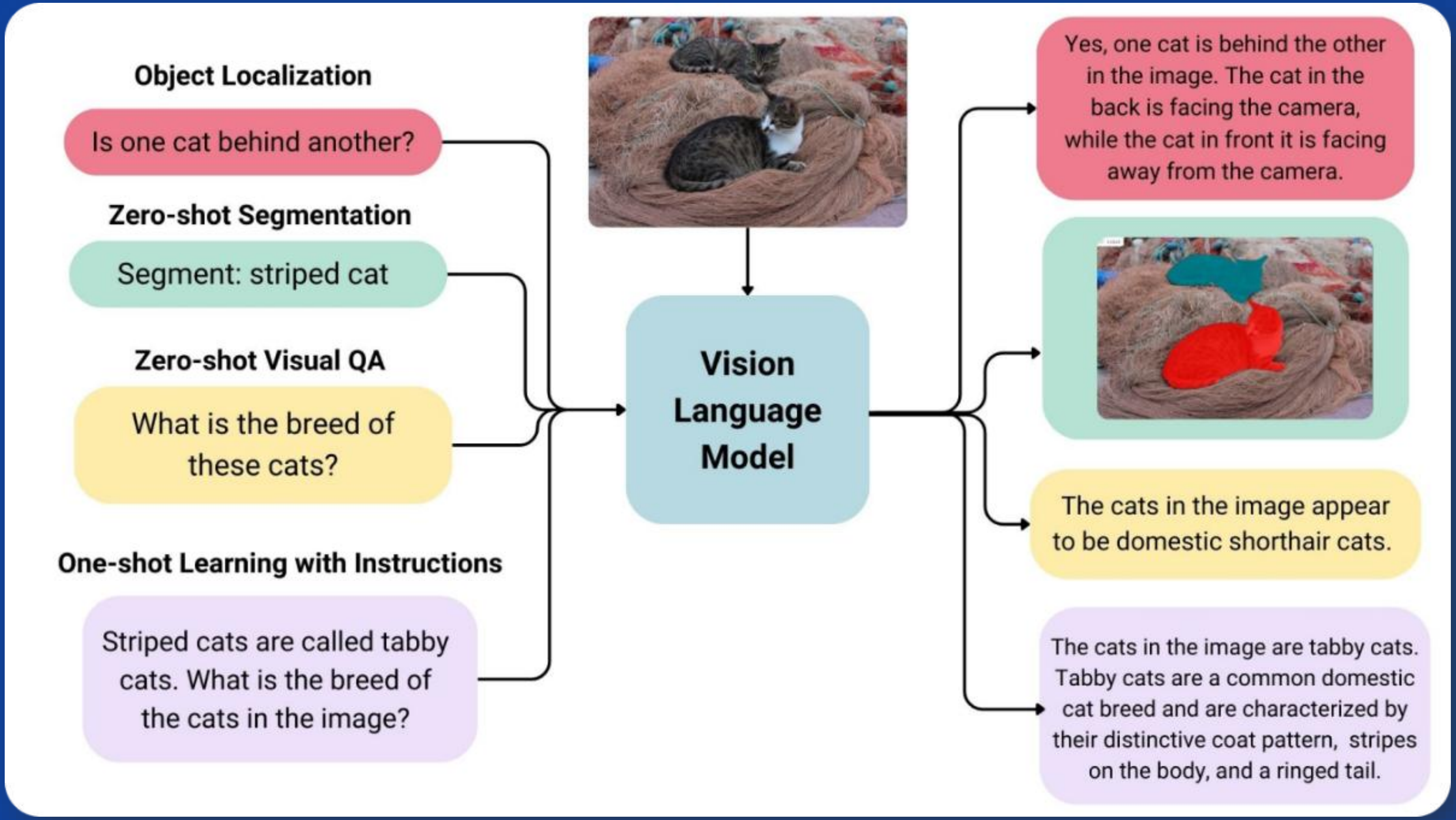


(A) Comparison with baselines. (B) Reinforcement learning gains.

>> Introduction

❑ Background

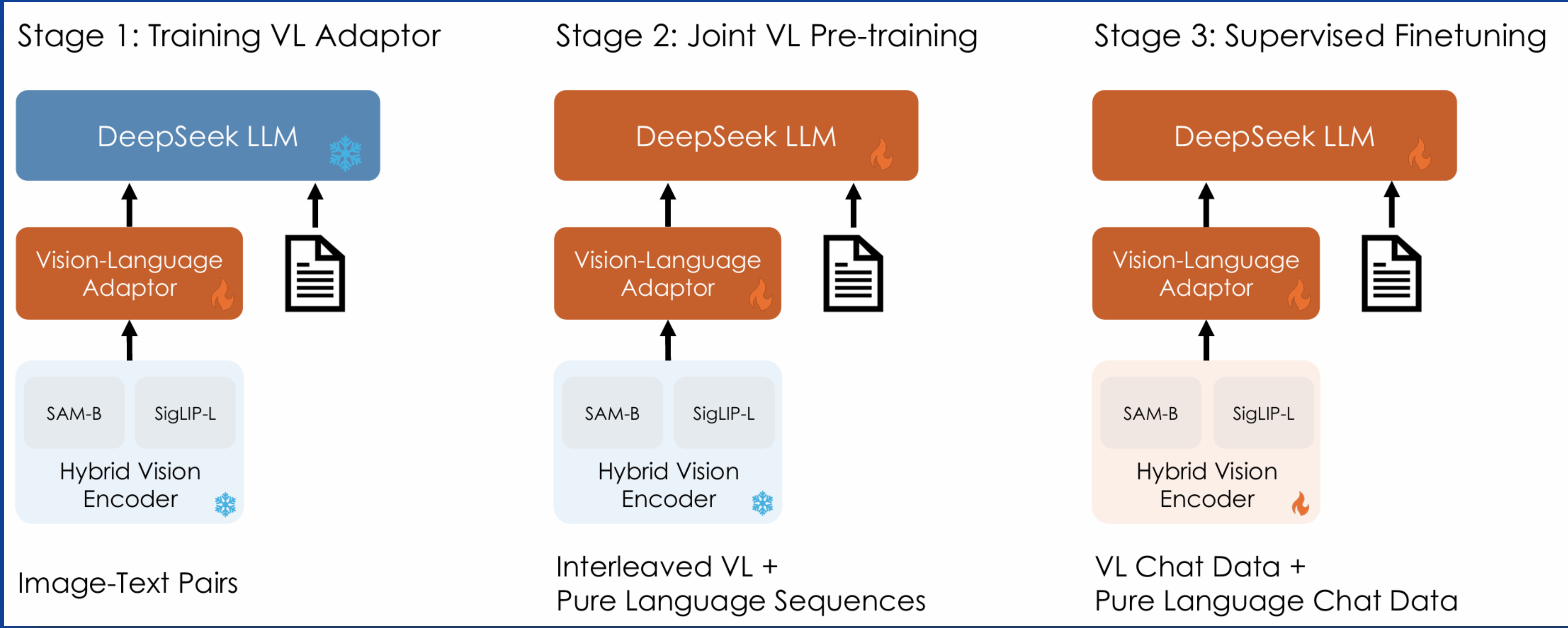
✓ VLM



>> Introduction

❑ Background

✓ VLM training pipelines

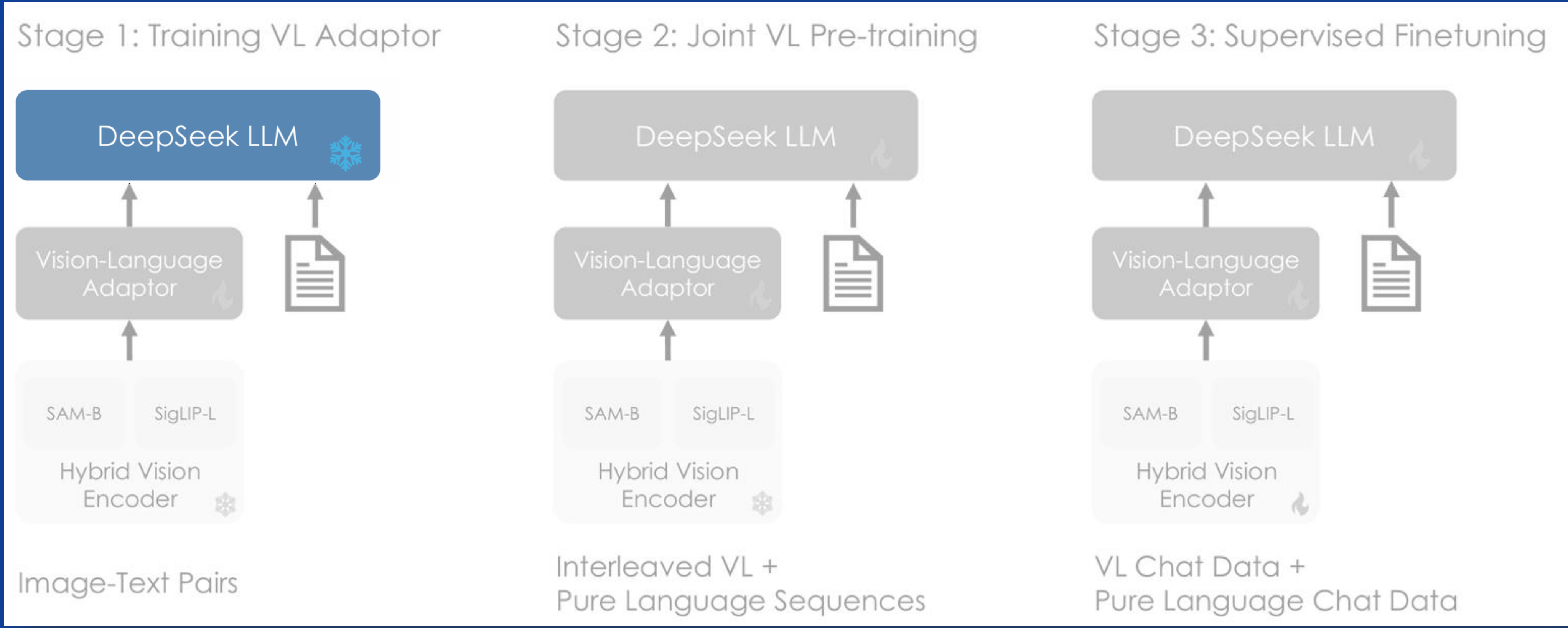


[Example of Deepseek-VL]

>> Introduction

❑ Background

✓ VLM training pipelines

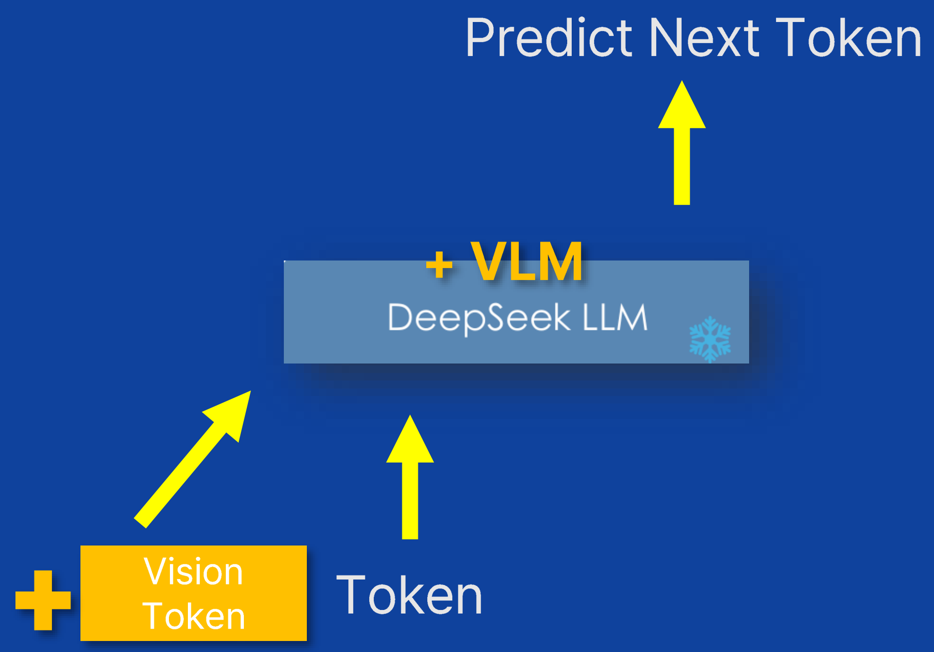


[Example of Deepseek-VL]

>> Introduction

❑ Background

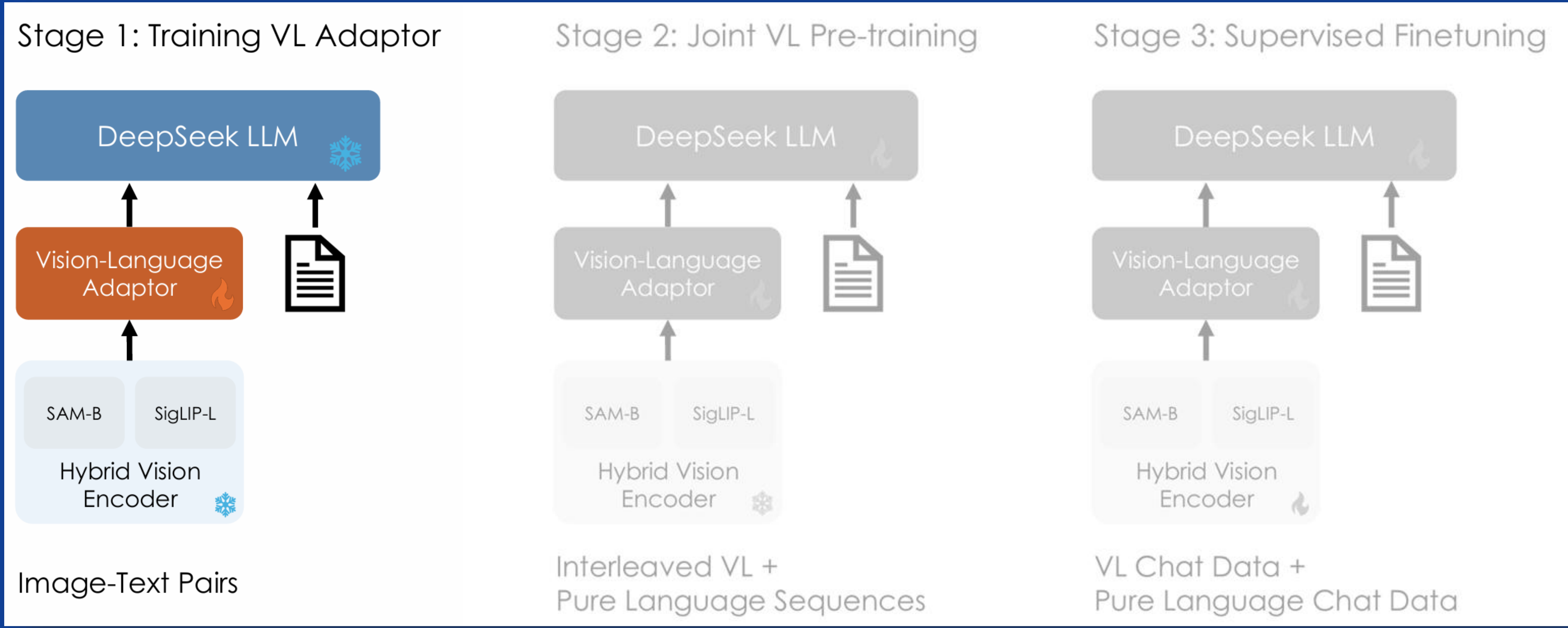
✓ **VLM training pipelines**



>> Introduction

❑ Background

✓ VLM training pipelines

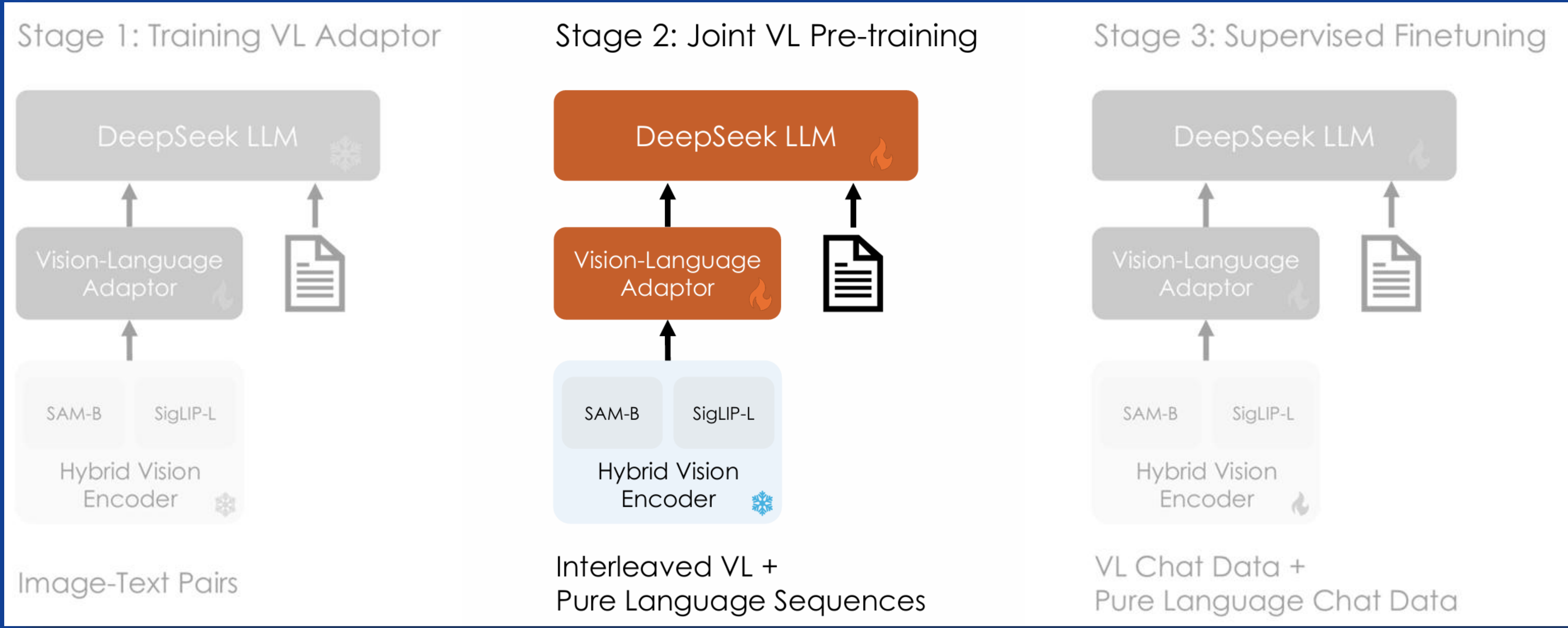


[Example of Deepseek-VL]

>> Introduction

❑ Background

✓ VLM - Architecture

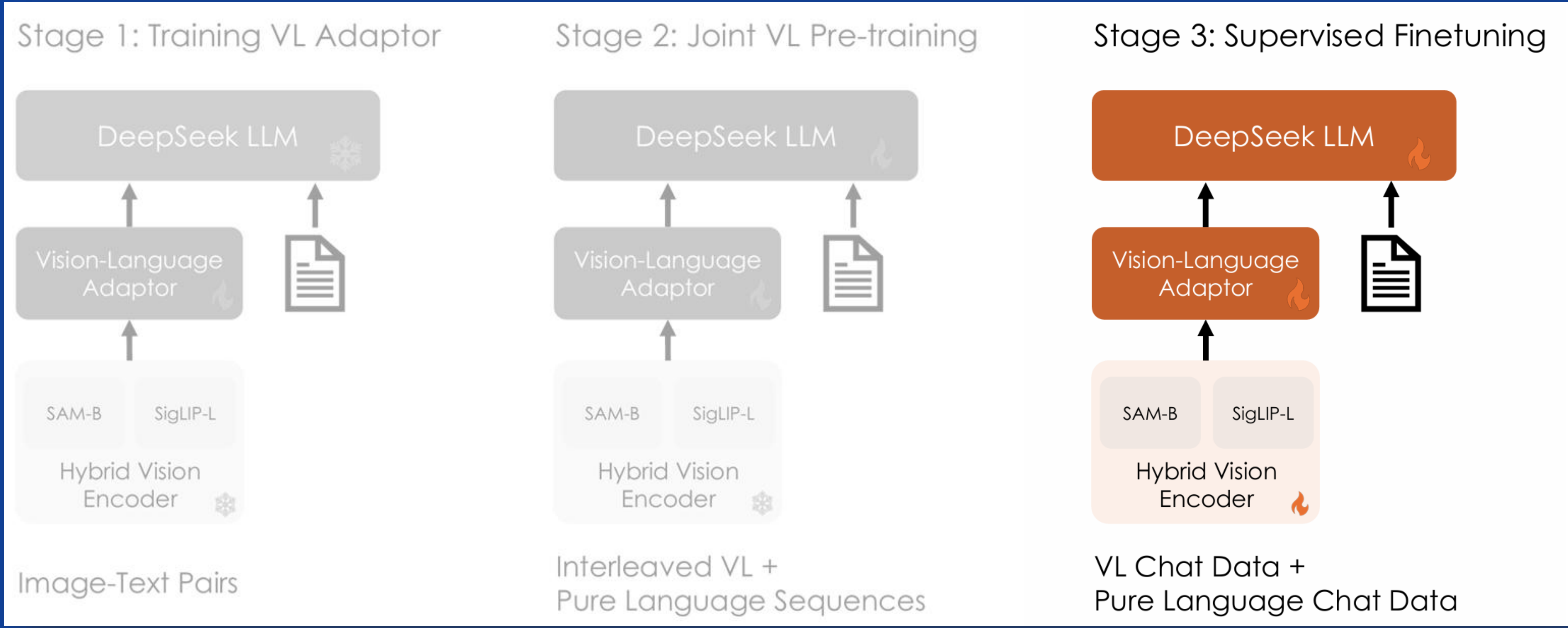


[Example of Deepseek-VL]

>> Introduction

❑ Background

✓ VLM - Architecture

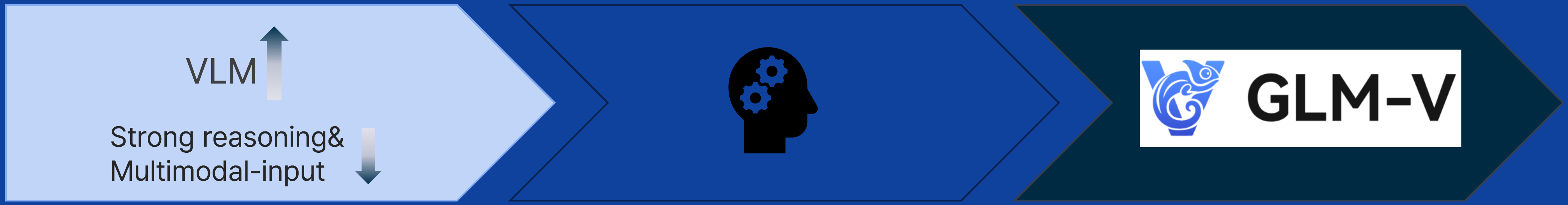


[Example of Deepseek-VL]

>> Introduction

❑ Motivation&Research goals

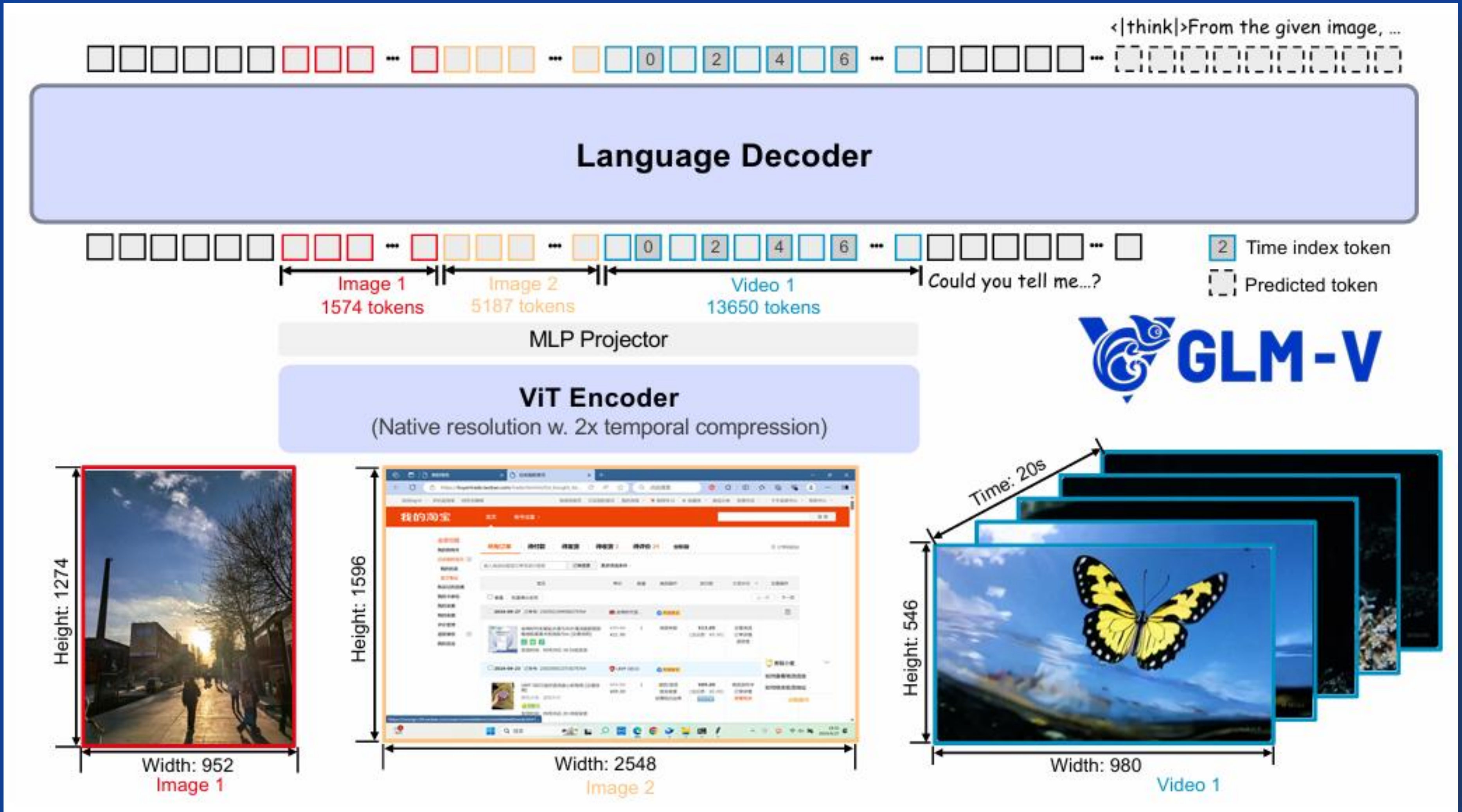
- Focus on efficient training and strong reasoning
- Support complex, real-world multimodal tasks



reason over complex, diverse inputs

>> Overview and Architecture

✓ Model Overview



>> Method

❑ Training Pipeline Summary

✓ **3-stage pipeline :**

1. Pre-training

2. Supervised Fine-tuning(SFT)

3. Reinforcement Learning with Curriculum Sampling(RLCS)

>> Method

❑ Pre-training

✓ Data Sources

Data Type	Description
📷 Image-Caption	10B+ pairs, CLIP filter, concept balancing, recaptioning
📚 Interleaved(text+image fusion)	Web (MINT/MMC4), **100M+ books (STEM)**, noise filtered
📰 OCR	220M doc images, natural scene text, arXiv parsing
🧭 Grounding	LAION, GLIPv2, GUI screenshots → 140M QA pairs
🎥 Video	Academic + web + proprietary video, annotated
💬 Instruction	Task-diverse tuning set, contamination filtering

>> Method

❑ Training Recipe

- **Two sequential stages:**

✓ **Multimodal Pre-training Configuration**

Model	Parallelism	Seq. Len	Batch	Steps
GLM-4.1V-Thinking	Tensor = 2	8,192	1,536	120,000
GLM-4.5V	Expert = 8, Pipeline = 4	8,192	1,536	120,000

- Lossless routing & scalable setup
- General-purpose multimodal capability building

✓ **Long-Context Extension Phase**

- After base training, expand to longer inputs & higher complexity
- Settings:
  - Seq Length → 32,768, Context Parallel = 4
  - +10,000 steps on high-res images, videos, long texts

>> Method

❑ Supervised Fine-tuning(SFT)

✓ Data Curation Strategy

- High-quality, long CoT examples focused on **verifiable tasks**
- **Standardized output formatting** required
- Iterative improvement: **RL-sampled** examples added to initial dataset to improve quality and difficulty

✓ Training Configuration

Setting	Value
Full fine-tuning	All parameters
Sequence length	32,768 tokens
Global batch size	32
Data types	Multimodal + Long-form Text

- GLM-4.5V supports both “thinking” and “non-thinking” modes
- Language understanding is retained through long-form text exposure

>> Method

❑ Reinforcement Learning : What Is Challenging and What Works

Combined Approaches	Reward System Design	Training Enhancements via RLCS	Infrastructure Optimization
• RLHF(Human Feedback) + RLVR (Verifiable Rewards) • Applied across diverse multimodal tasks	Domain-specific verifiers for robust reward computation (STEM, Chart QA, OCR, Grounding, GUI agents, Video QA)	• Curriculum-based dynamic Sampling (ratio EMA) • Improves stability and sample efficiency	Developing high-performance, stable RL infrastructure for large-scale RL training

>> Experimental Results

❑ Comprehensive Evaluation

- **GLM-4.5V**
  - Outperforms most open-source models of similar scale
  - Competitive with closed-source **Gemini-2.5-Flash** on several tasks
- **GLM-4.1V-Thinking (9B)**
  - Outperforms **Qwen2.5-VL-72B** on 29 benchmarks
  - Achieves SOTA on 23/28 benchmarks among models ≤10B

👀 Cross-Domain Effects

- RL in one domain → improves performance in others
- **Mix-all RL** → boosts performance across multiple tasks

>> Experimental Results

- GLM-4.5V was directly compared with competing models across various multimodal tasks, including VQA, STEM, OCR, Visual Grounding, GUI Agents, and Video QA
- The thinking mode of GLM-4.5V demonstrates superior performance on nearly all benchmarks, with a clear advantage observed in OCR, STEM, WebQA, and Codin

Task	Benchmark	GLM-4.1V	GLM-4.5V	GLM-4.5V	Step-3	Qwen2.5-VL	Kimi-VL-2506	Gemma-3
Size		9B	106B (A12B)	106B (A12B)	321B (A38B)	72B	16B (A3B)	27B
Mode		thinking	non-thinking	thinking	thinking	non-thinking	thinking	non-thinking
General VQA	MMBench V1.1	85.8	86.7	**88.2**	81.1*	88.0	84.4	80.1*
	MMBench V1.1 (CN)	84.7	86.5	**88.3**	81.5*	86.7*	80.7*	84.8*
	MMStar	72.9	73.4	**75.3**	69.0*	70.8	70.4	60.0*
	BLINK (Val)	65.1	63.7	**65.3**	62.7*	58.0*	53.5*	52.9*
	MUIRBENCH	74.7	71.1	**75.3**	75.0*	62.9*	63.8*	50.3*
	HallusionBench	63.2	59.1	**65.4**	64.2	56.8*	59.8*	45.8*
	ZeroBench (sub)	19.2	21.9	**23.4**	23.0	19.5*	16.2*	17.7*
	GeoBench[1]	76.0	78.4	**79.7**	72.9*	74.3*	48.0*	57.5*
STEM	MMMU (Val)	68.0	68.4	**75.4**	74.2	70.2	64.0	62.0*
	MMMU Pro	57.1	59.8	**65.2**	58.6	51.1	46.3	37.4*
	MathVista	80.7	78.2	**84.6**	79.2*	74.8	80.1	64.3*
	MathVision	54.4	52.5	**65.6**	64.8	38.1	54.4*	39.8*
	MathVerse	68.4	65.4	**72.1**	62.7*	47.8*	54.6*	34.0*
	DynaMath	42.5	44.1	**53.9**	50.1	36.1*	28.1*	28.5*
	LogicVista	60.4	54.8	**62.4**	60.2*	56.2*	51.4*	47.3*
	AI2D	87.9	86.6	**88.1**	83.7*	87.6*	81.9*	80.2*
	WeMath	63.8	58.9	**68.8**	59.8	46.0*	42.0*	37.9*
Long Document, OCR & Chart	MMLongBench-Doc	42.4	41.1	**44.7**	31.8*	35.2*	42.1	28.4*
	OCRBench	84.2	**87.2**	86.5	83.7*	85.1*	86.9	75.9*
	ChartQAPro	59.5	54.2	**64.0**	56.4*	46.7*	23.7*	37.6*
	ChartMuseum	48.8	47.1	**55.3**	40.0*	39.6*	33.6*	23.9*
Visual Grounding	RefCOCO-avg (val)	85.3	**91.5**	91.3	20.2*	90.3	33.6*	2.4*
	TreeBench	37.5	47.9	**50.1**	41.3*	42.3	41.5*	33.8*
	Ref-L4-test	86.8	**89.5**	**89.5**	12.2*	80.8*	51.3*	2.5*
Spatial Reco & Reasoning	OmniSpatial	47.7	49.6	**51.0**	47.0*	47.9	37.3*	40.8*
	CV-Bench	85.0	86.5	**87.3**	80.9*	82.0*	79.1*	74.6*
	ERQA	45.8	46.5	**50.0**	44.5*	44.8*	36.0*	37.5*
	All-Angles Bench	52.7	54.3	**56.9**	52.4*	54.4*	48.9*	48.2*
GUI Agents	OSWorld[2]	14.9	31.8	**35.8**	-	8.8	8.2	6.2*
	AndroidWorld	41.7	**57.0**	**57.0**	-	35.0	-	4.4*
	WebVoyager[2]	69.0	75.9	**84.4**	-	40.4*	-	34.8*
	Webquest-SingleQA	72.1	73.3	**76.9**	58.7*	60.5*	35.6*	31.2*
	Webquest-MultiQA	54.7	53.8	**60.6**	52.8*	52.1*	11.1*	36.5*
Coding	Design2Code	64.7	**84.5**	82.2	34.1*	41.9*	38.8*	16.1*
	Flame-React-Eval	72.5	78.8	**82.5**	63.8*	46.3*	36.3*	27.5*
Video Understanding	VideoMME (w/o sub)	68.2	74.3	**74.6**	-	73.3	67.8	58.9*
	VideoMME (w/sub)	73.6	80.0	**80.7**	-	79.1	71.9	68.4*
	MMVU	59.4	64.8	**68.7**	-	62.9	57.5	57.7*
	VideoMMMU	61.0	67.5	**72.4**	-	60.2	65.2	54.5*
	LVBench	44.0	**56.2**	53.8	-	47.3	47.6*	45.9*
	MotionBench	59.0	61.8	**62.4**	-	56.1*	54.3*	47.8*
	MVBench	68.4	**73.4**	73.0	-	70.4	59.7*	43.5*

>> Discussion

- **Strengths**
  - **Achieves SOTA-level performance even with small-scale models**
  - **Enhances general reasoning ability across diverse domains**
  - **Enables efficient and stable training with RLCS**

- **Limitations**
  - **Correct answers may still include errors in reasoning process**
  - **Training stability is sensitive in RL settings**
  - **Weakness in handling complex visual conditions (occlusion, ambiguity)**

>> Conclusion & Future Work

- Conclusion
  - GLM-4.1V-Thinking and GLM-4.5V → Successfully enhanced multimodal reasoning
  - Demonstrated the effectiveness of curriculum-based reinforcement learning (RLCS)

- Future Work
  - Develop evaluation metrics for intermediate reasoning processes
  - Improve training stability
  - Strengthen robustness under complex visual conditions

감사합니다