arXiv:2509.14033v2 [cs.CV] 18 Sep 2025

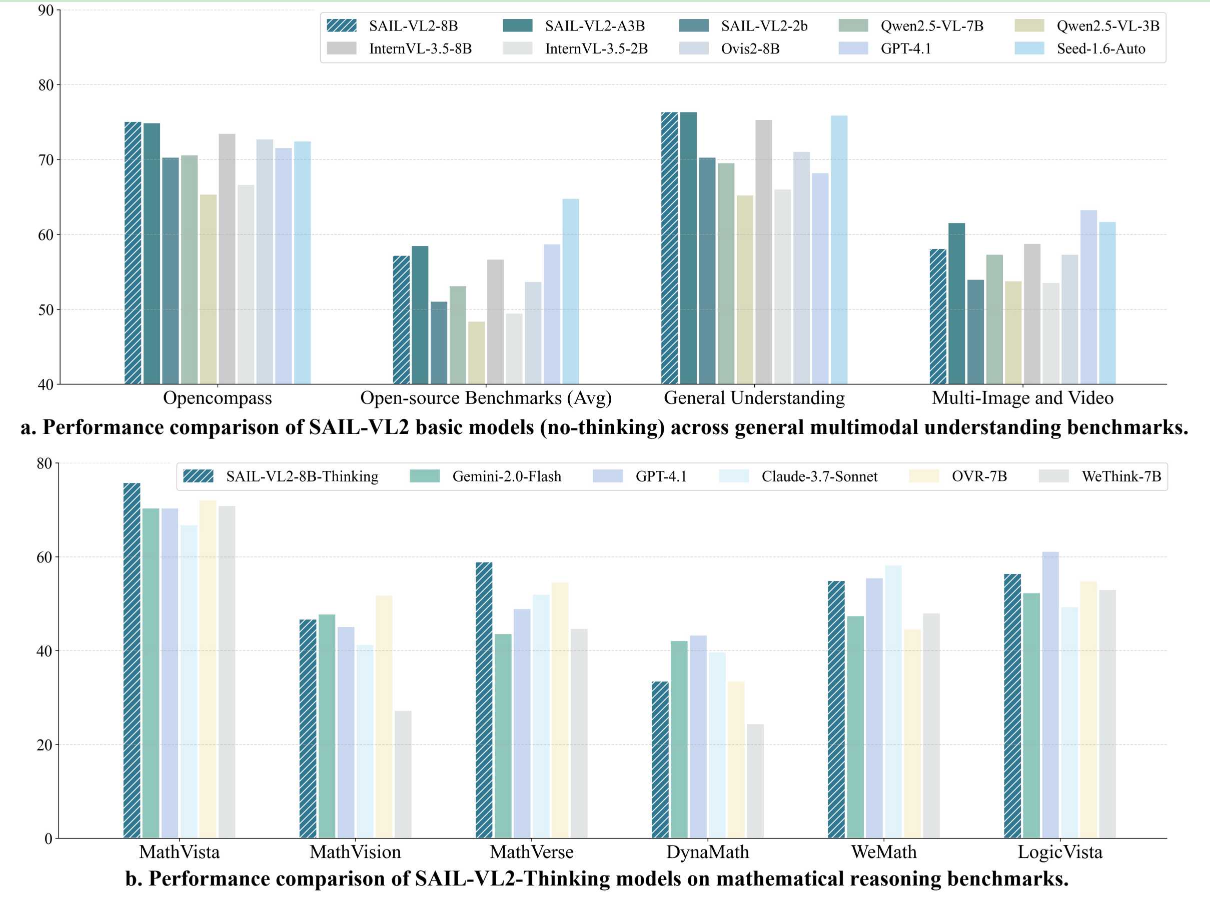
SAIL-VL2 Technical Report

Douyin SAIL Team, LV-NUS Lab

https://huggingface.co/BytedanceDouyinContent
https://github.com/BytedanceDouyinContent/SAIL-VL2

Abstract

We introduce SAIL-VL2, an open-suite vision-language foundation model (LVM) for comprehensive multimodal understanding and reasoning. As the successor to SAIL-VL, SAIL-VL2 achieves state-of-the-art performance at the 2B and 8B parameter scales across diverse image and video benchmarks, demonstrating strong capabilities from fine-grained perception to complex reasoning. Its effectiveness is driven by three core innovations. First, a large-scale data curation pipeline with scoring and filtering strategies enhances both quality and distribution across captioning, OCR, QA, and video data, improving training efficiency. Second, a progressive training framework begins with a powerful pre-trained vision encoder (SAIL-ViT), advances through multimodal pre-training, and culminates in a thinking-fusion SFT–RL hybrid paradigm that systematically strengthens model capabilities. Third, architectural advances extend beyond dense LLMs to efficient sparse Mixture-of-Experts (MoE) designs. With these contributions, SAIL-VL2 demonstrates competitive performance across 106 datasets and achieves state-of-the-art results on challenging reasoning benchmarks such as MMMU and MathVista. Furthermore, on the OpenCompass leaderboard, SAIL-VL2-2B ranks first among officially released open-source models under the 4B parameter scale, while serving as an efficient and extensible foundation for the open-source multimodal community.



a. Performance comparison of SAIL-VL2 basic models (no-thinking) across general multimodal understanding benchmarks.

b. Performance comparison of SAIL-VL2-Thinking models on mathematical reasoning benchmarks.

Figure 1: Performance comparison between **SAIL-VL2 (2B/8B/A3B)** and other LVMs (both open-source and large-scale closed-source). SAIL-VL2 demonstrates strong performance across multiple dimensions.

1 Introduction

Large-scale Vision-Language Models (LVMs) (Bai et al., 2025; Chen et al., 2024b; Zhu et al., 2025; Xiaomi, 2025; Team et al., 2025) bridge the gap between vision and language by integrating visual representations with linguistic descriptions in a shared semantic space. This shift from unimodal to cross-modal and multimodal understanding mirrors natural human interactions with the world. Driven by advances in large language models (LLMs) (Brown et al., 2020; Grattafiori et al., 2024; Yang et al., 2025) and visual representation techniques (Radford et al., 2021; Li et al., 2023), LVMs have evolved from coarse-grained visual understanding to fine-grained multimodal reasoning. Concurrently, training paradigms have progressed from teacher-forcing supervised learning to hybrid methods, integrating supervised fine-tuning (SFT) and reinforcement learning (RL) (Ouyang et al., 2022; Shao et al., 2024; Yu et al., 2025) for self-improvement. With continuous advancements in technology and scaling-up training data and model parameters, LVMs are steadily advancing toward Artificial General Intelligence (AGI).

Scaling up model parameters and training data to make LVMs 'larger' has emerged as a pivotal approach for pushing the performance boundaries of LVMs (Guo et al., 2025; Bai et al., 2025; Zhu et al., 2025). While this paradigm has yielded substantial performance gains, it also imposes considerable challenges in terms of computational demands and training, as well as deployment costs. In contrast, our SAIL-VL series focuses on developing efficient LVMs, aiming to explore *'how knowledge can be effectively injected through efficient architectures and training strategies,'* thereby establishing an open-source model family that embodies the principle of *'small model, strong performance'*.

To advance the development of more powerful yet efficient LVMs, we present **SAIL-VL2**, the latest iteration of our research, building upon our previous work, SAIL-VL (Dong et al., 2025a). SAIL-VL2 introduces substantial upgrades in architecture, training strategies, and data quality, achieving state-of-the-art performance across diverse benchmarks at a comparable parameter scale. These advancements endow the model with superior capabilities, ranging from multimodal understanding to complex reasoning.

From the data perspective, we design a comprehensive scoring and filtering pipeline that covers the full spectrum of multimodal inputs, ranging from captioning to QA and from images to videos. This pipeline systematically enhances both quality and distribution, thereby improving data efficiency across pre-training and post-training stages. In terms of training, we develop a progressive and efficient framework: beginning with a powerful pre-trained vision encoder (SAIL-ViT), advancing through basic multimodal pre-training, and culminating in a thinking-fusion SFT–RL hybrid paradigm, which enables systematic capability enhancement. Architecturally, SAIL-VL2 moves beyond conventional dense LLMs by adopting more efficient sparse Mixture-of-Experts (MoE) designs.

With the comprehensive upgrades and designs described above, the core capabilities and highlights of SAIL-VL2 can be summarized as follows (Figure 1):

- **SAIL-VL2 is powerful yet efficient:** With training on 776B tokens, SAIL-VL2 has verified its effectiveness across 106 datasets, achieving state-of-the-art results on a broad spectrum of influential benchmarks under the 2B-parameter scale. Remarkably, even without specialized prompting, the base SAIL-VL2 model delivers highly competitive performance on challenging reasoning benchmarks such as MMMU and MathVista, demonstrating strong out-of-the-box capabilities.
- **SAIL-VL2 as a deep thinker:** Many real-world tasks demand sophisticated reasoning and multi-step thought processes, which remain challenging for standard LVMs. To address this, we develop *SAIL-VL2-Thinking*, a specialized variant trained with advanced Chain-of-Thought (CoT) and reinforcement learning (RL) strategies. This design substantially improves performance on complex reasoning benchmarks, often matching or even surpassing models with far larger parameter scales, thereby setting a new standard for efficient architectures in high-level reasoning.
- **SAIL-VL2 perceives with clarity:** Fine-grained visual understanding is a critical challenge for multimodal models. SAIL-VL2 delivers high-fidelity perception in tasks such as OCR, high-resolution document layout analysis, and complex chart interpretation, achieving detailed visual grounding beyond models of similar scale.

In summary, SAIL-VL2 represents a comprehensive advancement in the design of efficient large vision-language models, integrating innovations in architecture, training strategies, and data curation. To foster openness and collaboration, we will release the full SAIL-VL2 model suite along with its inference code. We envision SAIL-VL2 as an efficient and extensible foundation that not only advances state-of-the-art performance but also empowers the broader open-source multimodal ecosystem.