Scalable Vision Language Model Training via High Quality Data Curation

Hongyuan Dong*, **Zijian Kang***, **Weijie Yin***, **Xiao Liang**, **Chao Feng**†, **Jiao Ran**
ByteDance Douyin Content Group
{donghongyuan.dousia, zijian.kang, yinweijie}
{liangxiao.ilx, chaofeng.zz, ranjiao}@bytedance.com

Abstract

In this paper, we introduce ***SAIL**-VL* (***S**c**A**lable Vision Language Model Tra**I**ning via High Qua**L**ity Data Curation*), an open-source vision language model (VLM) series achieving state-of-the-art (SOTA) performance in 2B and 8B parameters. The following three key improvements contribute to SAIL-VL's leading performance: (1) Scalable high-quality visual understanding data construction: We implement a data construction pipeline to enable hundred-million-scale high-quality recaption data annotation. The resulted dataset SAIL-Caption is validated to be of the highest data quality compared with opensource datasets. (2) Scalable Pretraining with High-Quality Visual Understanding Data: We scale SAIL-VL's pretraining budget up to 655B tokens and show that even a 2B VLM benefits from scaled up training data sizes, exhibiting logarithmic data size scaling laws in benchmark performance. (3) Scalable SFT via data quantity and complexity scaling: We curate a high-quality SFT dataset collection with leading data quantity scaling effectiveness and demonstrate that training with progressively higher-complexity data surpasses baseline one-stage training by a large margin.

SAIL-VL series models achieve the highest average score in 18 widely used VLM benchmarks in our evaluation, with the 2B model takes the top position over VLMs of comparable sizes on OpenCompass 2024 (https://rank.opencompass.org.cn/leaderboard-multimodal), demonstrating robust visual comprehension abilities. SAIL-VL series models are released at HuggingFace (https://huggingface.co/BytedanceDouyinContent).

1 Introduction

Researches in large vision language models (VLMs) (Liu et al., 2024a; Li et al., 2024a; Yao et al., 2024; Wang et al., 2024b; Gu et al., 2024; Chen et al., 2024d,c) have made significant progress in recent years, facilitating various vision tasks via language interactions. Due to the memory and computational constraints in model deployment, training compact VLMs with robust visual comprehension performance has become a popular research field recently (Marafioti et al., 2024; Chen et al., 2023b; Yao et al., 2024; Li et al., 2024c; Gao et al., 2024). However, how to make optimal use of publicly available resources to unlock the potential of compact VLMs remains an unanswered question. We attribute the suboptimal performance of recent lightweight vision language models to their limited fundamental visual understanding abilities and unsatisfactory instruction following performance.

The fundamental visual understanding abilities of VLMs are typically established via large-scale pretraining, which necessitates not only substantial training budgets, but also a sufficient amount of high-quality visual understanding data to take effect. Recently proposed VLMs, such as LLaVA series (Liu et al., 2024a; Li et al., 2024a; Chen et al., 2024a), conduct light-weight pretraining with a limited amount of low-quality caption data, and therefore suffer from suboptimal visual understanding abilities which hinder subsequent visual instruction tuning. MiniCPM-V-2.5 (Yao et al., 2024) and Qwen2-VL (Wang et al., 2024b) allocate hundreds of billions of tokens' computation budgets to the pretraining stage, but the limited visual understanding data quality undermines their visual understanding performance. More importantly, despite the large amount of resources consumed in pretraining, existing works do not provide reliable conclusions to understand how pretraining budgets and data quality influence VLM performance.

During the supervised fine-tuning (SFT) stage, VLM's visual understanding capabilities are generalized to instruction following tasks. However, how to make optimal use of high-quality visual

* Equal contribution.
†Email corresponding

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33272–33293

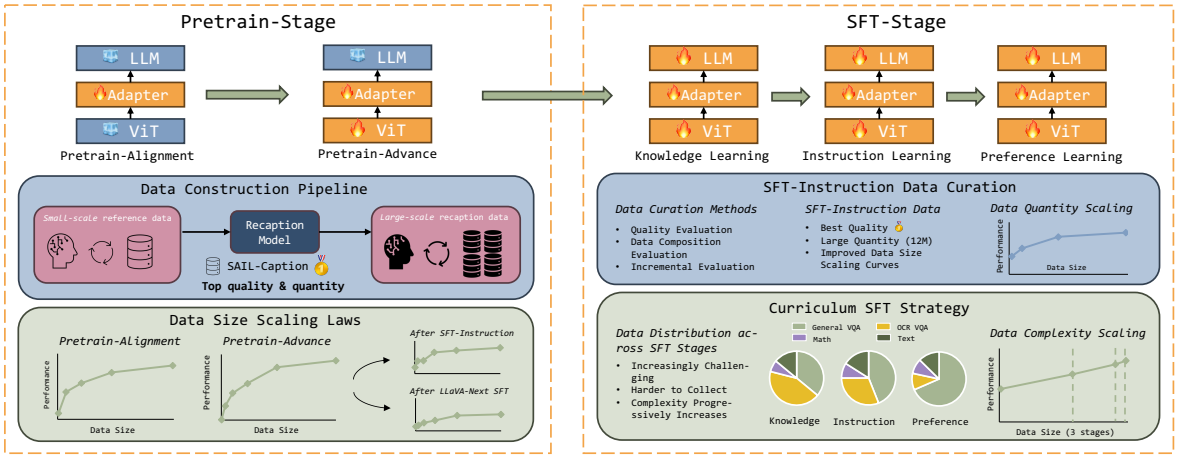

Figure 1: SAIL-VL's overall data construction and model training pipeline, as well as data size scaling laws observed in our large-scale VLM training experiments.

instruction tuning datasets remains unexplored. To obtain SFT data collections with higher quality, recent works focus on adjusting the data distribution across various domains and formats (Li et al., 2024a; Chen et al., 2024d,c; Yao et al., 2024). Infinity-MM (Gu et al., 2024) further explores enhancing the data efficiency of visual instruction tuning datasets with a multi-stage SFT strategy, obtaining promising performance scaling results. Despite the promising results of these works, there still lacks widely acknowledged methodologies to determine the distribution of SFT dataset collections or allocation of SFT stages.

To address the above issues, we propose SAIL-VL, an opensource vision language model series in 2B and 8B parameters with state-of-the-art (SOTA) performance. SAIL-VL is trained through several pretraining and SFT stages. We first establish SAIL-VL's basic visual understanding abilities via large-scale pretraining. To explore how pretraining computation budgets and data quality influence VLM performance, we scale up VLM pretraining to 655B tokens with *SAIL-Caption*, our synthesized large-scale detail caption dataset with top data quality compared to opensource alternatives. During the following SFT stages, we train SAIL-VL on our customized SFT data collection which outperforms opensource datasets markedly in data quality. SAIL-VL is trained in a curriculum learning paradigm of three stages, leading to improved data efficiency and model performance. The resulting SAIL-VL-2B and 8B models achieve new SOTA performance in 18 widely used VLM benchmarks.

We summarize the key contribution of this research as below:

(1) We implement a data construction pipeline for scalable high-quality visual understanding data construction, equipped with which we construct SAIL-Caption, which is of large quantity and the highest quality compared with opensource datasets.

(2) We scale up SAIL-VL's pretraining data size to 655B tokens, and report logarithmic model performance scaling laws w.r.t. training data sizes. To the best of our knowledge, this is the first time that data size scaling laws for VLM pretraining are proposed and discussed.

(3) We elaborate on the methodologies for high-quality SFT data curation, and demonstrate the effectiveness of the curriculum SFT strategy. Our SAIL-VL series models achieve top-ranked performance in our evaluation on 18 opensource VLM benchmarks.

2 Model Training Pipeline

In this section, we introduce SAIL-VL's training strategy as shown in Fig 1. Starting from InternViT (Chen et al., 2023c) and Qwen-2.5 (Team, 2024b) series models, SAIL-VL is pretrained for visual understanding and adapted to instruction following tasks in a total of five training stages.

2.1 Pretrain

During pretraining, we gradually open model parameters for larger-scale pretraining to develop SAIL-VL's visual understanding abilities. We start from a randomly initialized multi-layer perceptron (MLP) module as the vision-to-language projector, and train it with approximately 131B tokens of detail caption and OCR data in the *Pretrain-Alignment* stage. After warming up, we unlock

Dataset	Language	# Sample	Avg. Len.	Quality	Uni. 2-gram	Uni. 3-gram	Uni. Noun	Uni. Verb	Uni. Adj.
SAIL-Caption$_{EN}$	EN	225,000,000	87.86	-	14.34	33.40	0.6980	0.0722	0.4719
SAIL-Caption$_{CN}$	CN	75,000,000	156.95	-	11.71	32.10	1.032	0.3238	0.0625
DataComp-LLaVA-Caption (2024b)	EN	940,891,257	48.37	70.0	8.400	18.23	0.4728	0.0333	0.2801
SAIL-Caption-DataComp$_{Subset}$	EN	10,000	83.08	87.2	13.20	30.49	0.6354	0.0616	0.4627
SA1B-QwenVL-Caption (2024b)	CN	8,631,495	130.3	74.6	7.450	22.08	0.5797	0.1828	0.0378
SAIL-Caption-SA1B$_{Subset}$	CN	10,000	156.8	88.2	7.742	22.74	0.5872	0.1688	0.0359
BLIP3-KALE (2024)	CN	235,125,090	66.16	73.2	21.08	42.83	0.8686	0.0930	0.7012
SAIL-Caption-KALE$_{Subset}$	CN	10,000	63.53	80.6	16.79	32.08	0.9107	0.0634	0.5768

Table 1: Statistics of SAIL-Caption and other opensource datasets. "Quality" refers to quality scores evaluated by human annotators. We employ NLTK (Bird, 2006) and Jieba (Sun) to perform text segmentation and part-of-speech tagging for English and Chinese captions, respectively. "Avg. Len." stands for "average length" and "Uni." denotes "unique" items per sample. Statistics of SAIL-Caption subsets are marked with .

the visual encoder of SAIL-VL for larger model capacity during the following *Pretrain-Advance* stage, and train the model through approximately 524B tokens. Note that we do not use the entire SAIL-Caption dataset but a subset with an even distribution instead to ensure the diversity in data distribution. For OCR data, we use several high-quality OCR datasets repeatedly instead of incorporating diverse but relatively low-quality data. The advantage of using repeated-yet-high-quality data is shown in Section 5.1. For SAIL-VL-8B, we allocate 20B- and 32B-token training budgets in the two pretraining stages for efficiency.

2.2 SFT

We train all parameters of SAIL-VL in a curriculum learning fashion with progressively higher-complexity training data in SFT stages. In the first *SFT-Knowledge* stage, SAIL-VL learns basic instruction-following abilities and ingests world knowledge from Infinity-MM Stage2 (Gu et al., 2024) data. During the subsequent *SFT-Instruction* stage, we further optimize SAIL-VL towards enhanced visual instruction following capabilities with our customized 12M-sample high-quality visual instruction tuning dataset. For the final *SFT-Preference* stage, we train SAIL-VL on a small amount of complex visual instruction tuning data, including LLaVA (Li et al., 2024a) SFT, Molmo Caption (Deitke et al., 2024), and Infinity-MM Stage4 (Gu et al., 2024) data, enabling SAIL-VL to tackle a wider range of complex instruction following tasks. We refer to Section 5.2 for detailed data distribution of the three stages.

3 Towards Scalable VLM Training

In this section, we introduce our scalable high-quality data construction pipeline and elaborate on the model performance scaling laws observed in both pretraining and SFT stages.

3.1 Scalable High-Quality Visual Understanding Data Construction

Our scalable data construction pipeline is shown in Figure 1, consisting of the following four steps.

Data collection. We collect source data from a wide range of public image datasets to ensure data distribution diversity. Our source datasets include LAION-COCO (Schuhmann et al., 2022), TextCaps (Sidorov et al., 2020), SA1B (Kirillov et al., 2023), and several other large-scale datasets.

Reference data curation. We curate a small amount of reference data to train a compact VLM for efficient data annotation at scale. We first select a subset of source images with a balanced distribution, and then task GPT4-O-20240513 (OpenAI, 2024) deployed by Azure to annotate detail captions. Following previous works (Yu et al., 2024a; Hong et al., 2024), alt-texts are provided if available for supplementary world knowledge and enhanced reference data quality.

Captioner model training. Equipped with the high-quality reference data, we train an InternVL2-8B (Team, 2024a) model on the reference data to generate high-quality data at scale, which is called SAIL-Captioner. Similarly, alt-texts are optionally included in the caption generation prompt, enabling SAIL-Captioner to perform both captioning and recaptioning tasks.

Scalable high-quality data construction. In the final stage, we deploy SAIL-Captioner with LMDeploy (Contributors, 2023) for large-scale detail caption data construction. We implement a multi-task, multi-node, and multi-processing asynchronous annotation pipeline, enabling flexible computation resource allocation.

SAIL-Caption. Equipped with the aforementioned data construction pipeline, we construct

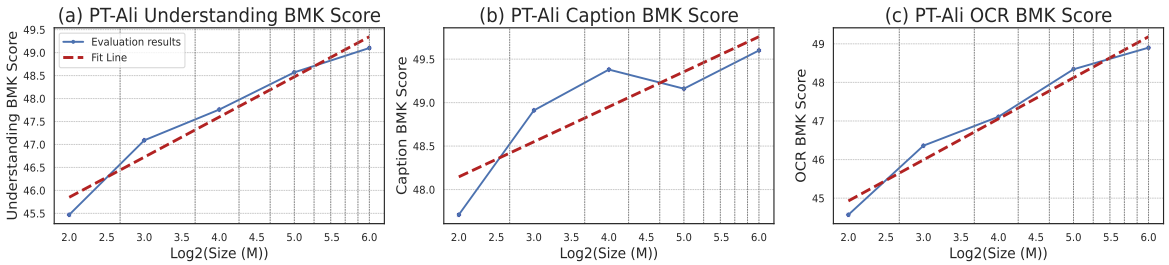

Figure 2: Scaling curves of SAIL-VL-2B's performance dynamics in the pretrain-alignment (PT-Ali) stage. We show model performance on all understanding benchmarks, caption tasks and OCR tasks, respectively. "BMK Score" stands for average benchmark scores.

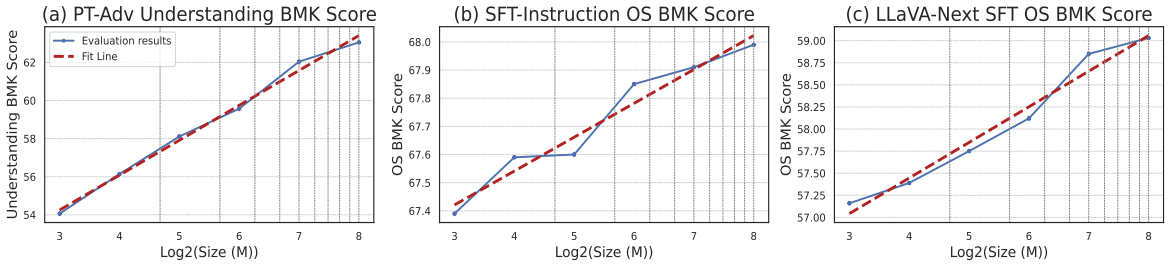

Figure 3: Scaling curves of SAIL-VL-2B's performance dynamics in the pretrain-advance (PT-Adv) stage. We show pretrained and SFT model performance on understanding benchmarks and OS (opensource) VLM benchmarks, respectively. "BMK Score" stands for average benchmark scores.

SAIL-Caption, a detail caption dataset with 300M image samples from various sources. To validate the data quality of SAIL-Caption, we randomly sample 10,000 cases from SAIL-Caption and other opensource caption datasets for comparison, and the statistics are shown in Table 1. Results show that SAIL-Caption is not only of large quantity, but also demonstrates leading richness of visual elements, for example, unique n-grams, nouns, verbs, and adjectives in caption texts. These statistics indicate that SAIL-Caption encompasses more visual elements and exhibits greater linguistic diversity in caption texts. Moreover, SAIL-Caption receives higher quality scores from human annotators, surpassing existing opensource datasets by a large margin. We refer to Appendix D for detailed caption quality evaluation procedure and SAIL-Caption showcases.

3.2 Scalable VLM Pretraining with High-Quality Visual Understanding Data

In this part, we introduce the data size scaling laws observed in SAIL-VL-2B large-scale pretraining. For model checkpoints obtained at different pretraining steps, we conduct lightweight annealing training with 2M identically distributed data for improved convergence and evaluation stability.

3.2.1 Improving VLM Visual Understanding Performance via Data Size Scaling

SAIL-VL-2B is trained through 131B and 524B tokens during the two pretraining stages, respectively, during which we investigate model performance dynamics. To evaluate the visual understanding performance of SAIL-VL, we establish an evaluation suite which covers fundamental visual understanding tasks such as detail caption generation (Dong et al., 2024) and OCR detection (Biten et al., 2022; Wang et al., 2020; Gupta et al., 2016; Kim et al., 2022). Details can be found in Appendix E.

As shown in Figure 2, SAIL-VL's visual understanding performance in each domain improves steadily in the pretrain-alignment stage. As the training data size scales up exponentially, the model performance exhibits a linear growth trend. We also show the understanding performance dynamics in the pretrain-advance stage. SAIL-VL's understanding benchmark scores improve markedly in this stage, which we attribute to the large capacity of the vision encoder optimized for visual understanding. In Figure 3 (a), a similar linear performance scaling curve is observed, unveiling a promising prospect to scale up VLM pretraining data sizes for improved model performance.

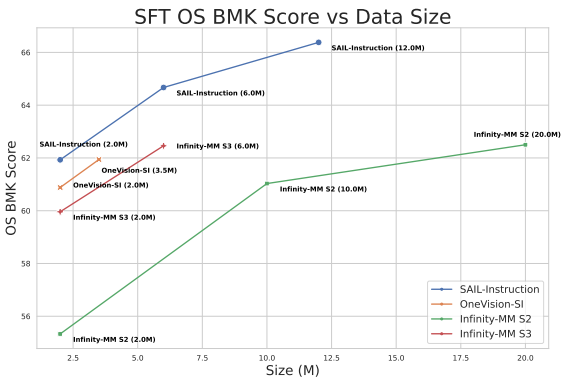

Figure 4: Scaling curves of model performance trained on our SAIL-Instruct dataset, LLaVA-OneVision (Li et al., 2024a) single image SFT data, and datasets from Infinity-MM (Gu et al., 2024). Model performance is shown as an average score across 18 benchmarks.

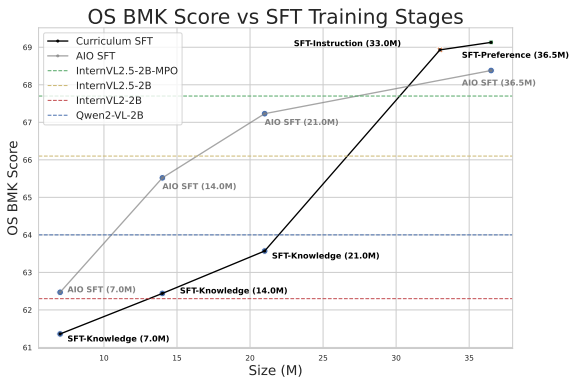

Figure 5: Model performance dynamics of the quality scaling and all-in-one (AIO) training strategy. "AIO learning" incorporates all three-stage SFT data into a single training loop. Model performance is shown as an average score across 18 benchmarks.

3.2.2 Generalizing Visual Understanding Abilities to Instruction Following Tasks

To further investigate the effectiveness of SAIL-VL's large-scale pretraining, we conduct SFT with different data collections for pretrain-advance model checkpoints trained with different data sizes.

As shown in Figure 3 (b)(c), the overall performance dynamics of SFT models can be plotted as a near-linear curve on an exponential horizontal axis, exhibiting smooth data size scaling laws on open-source VLM benchmarks. We conduct experiments with both our SFT-Instruction data and opensource LLaVA-Next SFT data. Despite the different data composition and final benchmark scores, similar scaling curves can be observed in both experiment sets. We further investigate pretrained and SFT model performance correlation in Appendix G.

3.3 Scaling up Visual Instruction Tuning

Despite the abundance of publicly available visual instruction tuning data, high-quality training data is still scarce. We first introduce guidelines for our high-quality SFT data curation in Section 3.3.1, and demonstrate model performance scaling laws of the curriculum SFT strategy in Section 3.3.2.

3.3.1 High-Quality SFT Data Curation for Data Quantity Scaling

In this part, we elaborate on the methodologies for visual instruction tuning data curation and demonstrate their effectiveness in SAIL-VL training.

High-quality visual instruction tuning dataset curation. To judge the quality of different SFT data collections efficiently, we start with the *Quick*

SFT Stage	Size ↓	Diff. ↑	Comp. ↑	Rel. ↑
SFT-Knowledge	21M	1.90	2.44	3.94
SFT-Instruction	12M	2.15	2.62	4.45
SFT-Preference	3.5M	2.20	2.74	4.55

Table 2: Sizes and quality evaluation results of the three-stage SFT data. "Diff.", "Comp.", and "Rel." stand for task difficulty, data complexity, and imaget-text relevance, respectively.

Quality Evaluation strategy. This strategy assesses the quality of a given SFT data collection by training with its 2M-sample subset. The resulting model performance reflects the training data quality, enabling efficient data quality evaluation and comparison. In this strategy, we assume that models trained on different datasets maintain a consistent performance ranking across varying training data sizes. This assumption is validated by experiment results shown in Figure 4.

We then propose the *Composition Evaluation* strategy to judge the quality of existing SFT data components. In composition evaluation, we start with existing SFT data collections, for example, LLaVA-OneVision (Li et al., 2024a), Cauldron (Laurençon et al., 2024). We then categorize the datasets based on their format and distribution, resulting in a series of data groups, including closed-form VQA, open-ended VQA, document VQA, math&reasoning, and pure text QA data [1]. To optimize the proportion of different data components, we halve each data group and judge the quality of the resulting data collection with our

[1]Closed-form and open-ended VQA data refer to natural image VQA data requiring specific choice answers and open-ended responses, respectively

		Pretrain-Alignment		Pretrain-Advance		SFT-Knowledge		SFT-Instruction		SFT-Preference	
		2B	8B	2B	8B	2B	8B	2B	8B	2B	8B
Vision	Resolution	$448 \times \{\{1 \times 1\}, ..., \{2 \times 4\}\}$		$448 \times \{\{1 \times 1\}, ..., \{2 \times 4\}\}$		$448 \times \{\{1 \times 1\}, ..., \{2 \times 5\}\}$		$448 \times \{\{1 \times 1\}, ..., \{2 \times 5\}\}$		$448 \times \{\{1 \times 1\}, ..., \{2 \times 5\}\}$	
	# Max Visual Token	2048		2048		2560		2560		2560	
Data	Data Composition	SAIL-Caption & OCR		SAIL-Caption & OCR		Curated VQA Data		Curated VQA Data		Curated VQA Data	
	Dataset Size	64M	10M	256M	16M	21M		12M		3.5M	
Training	Trainable Module	Projector		Vision & Projector		Full Model		Full Model		Full Model	
	Trainable Parameter	8.65M	27.52M	313M	332M	1.85B	7.95B	1.85B	7.95B	1.85B	7.95B
	Batch Size	1920	512	2048	2048	512		512		512	
	Learning Rate	1×10^{-4}	1×10^{-3}	4×10^{-5}	4×10^{-5}	2×10^{-5}	1×10^{-5}	2×10^{-5}	1×10^{-5}	2×10^{-5}	1×10^{-5}

Table 3: Details of the training pipeline of SAIL-VL-2B and SAIL-VL-8B.

quick quality evaluation method. Once the model performance improves, the downward adjustment of the data proportion is retained.

For incoming datasets to be incorporated into the SFT data, we conduct *Incremental Evaluation*. Each new dataset is included in the SFT data collection, with the resulting data quality evaluated via lightweight model training. Datasets improving the model performance are regarded as beneficial for the data quality, and are therefore incorporated into our data collection. We also incorporate datasets which maintain the model performance, as they help expand the data scale for improved results.

Data Quantity Scaling. We curate our SAIL-Instruction data collection (used in SFT-Instruction stage) with the methodologies described above. To validate its advantage in data quality, we train the SAIL-VL model with our SAIL-Instruction data and other opensource SFT data collections at varying data scales. As shown in Figure 4, the performance of SAIL-VL scales up stably as the model training proceeds, depicting a logarithmic performance scaling curve. Compared with other opensource SFT data collections, our SAIL-Instruction data achieves the highest model performance at every data point. It is also worth noticing that the performance ranking of models trained with different datasets remains consistent across the training process. This observation validates our quick quality evaluation method introduced above.

3.3.2 Multi-stage Instruction Tuning for Data Complexity Scaling

In this part, we introduce data complexity scaling, a curriculum learning strategy for VLM SFT for enhanced model performance.

Curriculum SFT with progressively improving data quality. As elaborated in Section 2.2, we train SAIL-VL through three SFT stages, and the data collections used in later stages differ from previous ones in the following aspects: (1) Datasets are harder to collect and therefore of smaller quantity. (2) Training tasks become increasingly challenging, and the questions in the training data are more difficult to answer. (3) Data complexity progressively increases, requiring more fine-grained understanding of the visual elements and in-depth reasoning. We validate our design by quantifying the data distribution variance across the three stages via human evaluation. As shown in Table 6, the task difficulty, data complexity, and image-text relevance increase monotonously across the three stages. SFT data in the later stages is of higher overall quality, but is also more challenging for the model to learn from, which coincides with our curriculum SFT design. We refer to Appendix F for the detailed definition of these data quality dimensions and the full instruction for human evaluation.

Data Complexity Scaling. To demonstrate the effectiveness of our curriculum SFT strategy with progressively higher-complexity data, we show model performance dynamics derived from the three SFT stages in comparison with an all-in-one (AIO) training strategy in Figure 5. The model trained with our curriculum SFT strategy exhibits a near-linear performance scaling curve across training stages, outperforming the logarithmic scaling curve of AIO training baseline. This result validates the marked effectiveness of the curriculum SFT strategy. Training with small and high-complexity SFT data in later stages yields more promising performance scaling curves.

4 Experiments

4.1 Experiment Setup

Model Training. We start from InternViT-300M (Chen et al., 2023c), Qwen2.5-2B and Qwen2.5-7B (Team, 2024b) for model training. Detailed model training recipes are elaborated in Table 3. As training progresses, the input image resolution gradually increases, with a 2×2 pixel

Benchmark	SAIL-VL	Qwen2-VL	InternVL2.5-MPO	DeepSeekVL-2	SAIL-VL	Qwen2-VL	InternVL2.5-MPO	DeepSeekVL-2
	2B Model				8B Model			
Overall Performance								
Opensource Average	**69.1**	64.4	67.7	67.0	**74.5**	73.0	74.3	72.7
General VQA	60.4	58.3	**63.1**	59.4	68.3	68.5	**71.2**	66.8
OCR VQA	**75.9**	72.5	71.1	74.4	**79.8**	79.6	76.3	79.0
Math&Knowledge	**79.0**	59.0	75.3	71.3	**83.3**	71.0	83.2	79.0
Hallucination	**66.2**	62.9	64.5	63.6	68.7	67.5	**69.7**	65.3
General VQA								
MMStar (2024b)	**55.1**	46.3	54.3	49.9	64.2	58.3	**65.3**	57.7
$\text{MMBench}_{\text{DEV}}$ (2024c)	72.4	68.8	**72.5**	68.3	79.5	79.5	**83.3**	78.1
MMMU_{VAL} (2024)	40.1	39.9	**41.2**	39.6	48.2	50.9	**52.8**	47.6
MME (2023)	1969	1923	**2123**	1910	2244	2321	2321	2149
$\text{SEEDBench}_{\text{IMG}}$ (2023a)	**74.7**	72.0	73.2	72.5	75.5	75.3	**76.9**	76.8
RealWorldQA (2024)	63.8	60.9	60.7	**64.8**	**71.9**	69.7	70.2	70.2
MMVet (2024b)	46.1	51.2	**64.0**	52.8	58.3	62.6	**66.8**	60.3
OCR VQA								
$\text{AI2D}_{\text{TEST}}$ (2016)	**79.0**	72.3	75.3	74.6	83.7	82.9	**84.1**	82.0
$\text{DocVQA}_{\text{VAL}}$ (2021)	**89.2**	88.7	87.8	88.6	92.2	**93.7**	92.1	92.3
$\text{InfoVQA}_{\text{VAL}}$ (2022)	**67.2**	63.4	61.6	63.8	75.2	75.9	**76.2**	72.5
$\text{ChartQA}_{\text{TEST}}$ (2022)	81.0	70.6	70.9	**81.2**	84.6	81.6	77.6	84.6
$\text{TextVQA}_{\text{VAL}}$ (2019)	75.7	78.8	77.2	**80.5**	77.7	**83.8**	79.2	83.3
$\text{OCRVQA}_{\text{TEST}}$ (2019)	**58.5**	54.3	40.0	51.4	**61.4**	56.2	36.7	54.5
OCRBench (2024d)	806	794	**846**	808	835	833	**880**	834
Math&Knowledge								
$\text{MathVista}_{\text{MINI}}$ (2023)	**62.8**	45.0	55.3	54.5	68.4	57.3	**68.5**	61.8
$\text{ScienceQA}_{\text{VAL}}$ (2022)	95.3	73.0	95.3	88.1	**98.2**	84.6	97.9	96.2
Hallucination								
HallusionBench (2024)	**45.7**	38.3	39.2	38.4	**52.2**	48.5	50.3	41.2
POPE (2023)	86.7	87.6	**89.8**	88.8	85.2	86.5	89.1	**89.4**

Table 4: Evaluation results of SAIL-VL and other opensource VLM with comparable sizes. "Opensource average" includes all opensource benchmarks listed in the table. Bold numbers indicate the best performance among models of comparable sizes, while underlined ones are those ranked as the second.

shuffle (Chen et al., 2024d) module employed in the projector, maintaining a balance between efficiency and performance. For SAIL-VL-8B, we use smaller batch sizes and larger learning rates in pretraining stages to improve training efficiency. During SFT stages, the 8B model is trained with a smaller learning rate, mitigating the instability in full model training with larger LLMs.

Baselines. We compare our SAIL-VL models with previous SOTA VLM baselines of comparable sizes, including Qwen2-VL (Wang et al., 2024b), InternVL2.5-MPO (Chen et al., 2024c), DeepSeekVL-2(Wu et al., 2024), etc. Evaluation results against more existing baseline models are shown in Appendix H.2.

Evaluation. We evaluate SAIL-VL and baseline VLMs on a series of widely used benchmarks, including General VQA, OCR VQA, Math&Knowledge, and Hallucination. These categories cover VQA tasks on natural images/videos, OCR-related documents, as well as those involving complicated reasoning abilities and world knowledge to tackle. We use a customized version of VLMEvalKit (Duan et al., 2024) for evaluations.

4.2 Benchmark Results

SAIL-VL-2B ourperforms previous SOTA VLMs with comparable sizes significantly. We list the performance of SAIL-VL along with other opensource VLMs in Table 4. As the results show, SAIL-VL-2B outperforms previous SOTA VLMs by a large margin, scoring 1.4 (2.06% ↑) higher average performance than InternVL2.5-MPO-2B. SAIL-VL-2B achieves new SOTA performance on 3 out of 4 subfields except for General VQA. We attribute it to the instability lying in benchmarks requiring long text generation, such as MMVet.

SAIl-VL-8B achieves leading performance over opensource baselines. As shown in Table 4, SAIL-VL-8B also achieves leading visual comprehension performance over Qwen2-VL, DeepSeekVL-2, and even InternVL2.5-MPO-8B, which requires an additional reinforcement learning stage in model training. We admit the shrunk performance advantage of SAIL-VL-8B

Caption Data	OCR Data	Overall	Caption	OCR
SAIL-Caption	HQ	54.36	51.80	55.38
SA1B-QwenVL-Caption	HQ	48.43	39.57	51.97
DataComp-LLaVA-Caption	HQ	49.08	42.70	51.63
BLIP3-KALE	HQ	53.06	46.00	55.89
SAIL-Caption	HQ+LQ	52.13	51.22	52.50
SAIL-Caption	HQ (RP)	54.05	52.63	54.62

Table 5: Visual understanding performance of model checkpoints pretrained with different data sources. We report models performance on our visual understanding benchmarks. "HQ", "LQ", and "RP" indicates high-quality, low-quality, and repeated data, respectively.

over SOTA baselines, which may be caused by the relatively small data sizes used for model training. We take these results as an early attempt for larger VLM training, and more competitive large VLMs will be released in our SAIL-VL series.

5 Analysis

5.1 Pretrain Data Quality Determines Pretrained Model Performance

We explore pretraining SAIL-VL-2B with varying data quality. Specifically, we conduct lightweight 16B-token training in the pretrain-advance stage, starting from the model checkpoint after the same alignment pretraining. We fix the data distribution across different data types, and modify data composition with varying-quality data.

As shown in Table 5, the model trained with SAIL-Caption achieves significantly higher performance than those trained on other opensource caption datasets, which is consistent with data quality evaluation results as shown in Appendix 8. It is also worth noticing that the model trained with repeated-yet-high-quality OCR data yields better results than incorporating diverse but relatively low-quality data for model training. We attribute this result to our frozen-LLM pretraining setting, which mitigates the potential overfitting problem lying in repeated training data.

5.2 SFT Data Quality Analysis

To further validate our data quality evaluation results shown in Table 1, we select 2M-sample subsets from each SFT stage to train the pretrained SAIL-VL-2B model. Performance evaluation results are shown in Table 6. A significant performance advantage is observed in the model trained with SFT-Instruction data collection, validating the effectiveness of the proposed data curation methods. This result coincides with the data quality evaluation results given in Table 1, where SFT-Instruction data collection exhibits advanced task difficulty, data complexity, and image-text relevance. It is also worth noticing that despite the improved data quality of the SFT-Preference data, it fails to further improve model performance in Table 6. We attribute it to its excessively high data complexity, which may hinder effective model learning. This observation further validates the proposed curriculum VLM SFT strategy as discussed in Section 3.3.2.

Training Data	Overall	General	OCR	Math.	Hall.
SFT-Knowledge	57.8	53.2	60.9	56.9	63.9
SFT-Instruction	61.9	55.8	67.1	65.4	61.7
SFT-Preference	61.3	57.1	65.8	59.5	61.3

Table 6: Performance evaluation results of models trained with SFT data from each stage. We denote "Math." as Math &Knowledge benchmarks in evaluation. "Hall." denotes Hallucination benchmarks as defined in Section 4.1.

6 Related Works

6.1 Visual Understanding Data

Visual understanding data consists of vision modality contents and corresponding language depictions, and is regarded as the keystone to various vision and language model applications. Whether it is representation learning models like CLIP and its derivatives (Radford et al., 2021; Jia et al., 2021; Shen et al., 2022; Cherti et al., 2023; Fang et al., 2023), generative models (Wang et al., 2022; Li et al., 2021; Bao et al., 2022; Yu et al., 2022; Li et al., 2023b), or recent vision language models (Li et al., 2023b; Liu et al., 2024b; Team, 2024a; Bai et al., 2023), all of these methods are built upon large scale high-quality visual understanding data. LAION (Schuhmann et al., 2021, 2022), TaiSu (Liu et al., 2022), Coyo (Byeon et al., 2022), DataComp (Gadre et al., 2024), and etc. provide relatively low-quality alt-texts paired with source images. Subsequent works such as ShareGPT4V (Chen et al., 2023a) and ALLaVA (Chen et al., 2024a) annotate small scale high-quality caption data with powerful VLM APIs. To produce high-quality detail caption data at scale, CapsFusion (Yu et al., 2024a), World2Seq (Wang et al., 2024a), CAPTURE (Dong et al., 2024), SA1B-Recaption (Data, 2024b), DataComp-Recaption (Li et al., 2024b), and BLIP3-KALE (Awadalla et al., 2024) employ

recaptioner models for efficient data annotation. The resulting datasets are widely used in recent VLM research.

6.2 Vision Language Model Pretrain

VLM pretraining benefits from higher-quality and larger-scale visual understanding data effectively. Previous works, such as BLIP2 (Li et al., 2023b) and LLaVA (Liu et al., 2024b), pretrain the model with relatively low-quality caption datasets (Li et al., 2023b). Subsequent works, such as MiniCPM-V (Yao et al., 2024), InternVL (Chen et al., 2024d; Team, 2024a; Chen et al., 2024c), and QwenVL (Bai et al., 2023; Wang et al., 2024b) series, explore expanding high-quality visual understanding data sizes to improve model performance. In this work, we further reveal model performance dynamics w.r.t. SFT data quality and size, which are largely unexplored in previous works.

6.3 Visual Instruction Tuning

LLaVA (Liu et al., 2024b) first defines visual instruction tuning and provides a baseline for VLM SFT data curation. Subsequent LLaVA series models (Liu et al., 2023, 2024a; Li et al., 2024a) refine the visual instruction tuning datasets and achieve significantly better model performance. BLIP3 (Xue et al., 2024) incorporates image-text interleaved data into visual instruction tuning, while CogVLM (Hong et al., 2024), InternVL (Chen et al., 2024d; Team, 2024a; Chen et al., 2024c) and QwenVL series (Wang et al., 2024b) models explore using video question answering data for VLM SFT. In this paper, we elaborate the guidelines for the design of visual instruction datasets, providing valuable references for VLM training.

7 Conclusions

In this work, we introduce SAIL-VL, an opensource vision language model series with SOTA performance. We propose a scalable caption data construction pipeline and curate SAIL-Caption, a large-scale caption dataset with the highest quality among opensource alternatives. Equipped with SAIL-Caption, we conduct large-scale pretraining with up to 655B tokens and demonstrate that even compact VLMs can benefit from scaled up training data size. We further present data size scaling laws that SAIL-VL's visual comprehension performance improves logarithmically as training data size increases. For visual instruction tuning stages, we elaborate on several key guidelines for high-quality SFT data curation, guided by which we curate our SFT-Instrcution dataset, a high-quality SFT data collection exhibiting improved model performance scaling curves than opensource alternatives during model training. The phased SFT strategy used in SAIL-VL SFT further improves the scaling curves from logarithmic to near-linear. We evaluate SAIL-VL on 18 opensource VLM benchmarks, and our model outperforms existing VLMs of comparable sizes consistently either in overall performance or domain-specific abilities, depicting promising prospects in real-world applications.

8 Limitations

Despite the leading performance of SAIL-VL among VLMs of comparable sizes, we acknowledge the potential insights that could be gained from experimenting with larger models. We intend to explore this avenue in future work to enhance the robustness of the presented data size scaling laws and other findings. Additionally, our exploration of data size scaling laws has been confined to a specific data magnitude. Although model performance is observed to be saturating at this data quantity, it remains uncertain whether there is room for further improvement under optimized training settings.

We also point out that although SAIL-VL's training process is designed carefully, models may generate hallucinated, biased, or harmful information under certain circumstances, which will be further discussed and mitigated in our future works.

References

Huawei ascend.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947.

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. 2024. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large

vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Ali Furkan Biten, Ruben Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. 2022. Ocr-idl: Ocr annotations for industry document library dataset. In *European Conference on Computer Vision*, pages 241–252. Springer.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023b. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

LMDeploy Contributors. 2023. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning (2023). *arXiv preprint arXiv:2307.08691*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Tongyi Data. 2024a. Qwen-vl-chat-finetuned-dense-captioner. https://modelscope.cn/models/Tongyi-DataEngine/Qwen-VL-Chat-Finetuned-Dense-Captioner.

Tongyi Data. 2024b. Sa1b-dense-caption dataset. https://www.modelscope.cn/datasets/Tongyi-DataEngine/SA1B-Dense-Caption.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. 2024. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Huawei. 2023. Ascend 910b npu. https://carrier.huawei.com/~/media/CNBG/Downloads/Product/Fixed%20Network/carrierip-router/ATN%20910B-%E4%B8%AD%E6%96%87%E7%89%88-%E9%AB%98%E7%B2%BE%E5%BA%A6%E5%8D%B0%E5%88%B7%E6%96%87%E4%BB%B6.pdf.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong

Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. 2024b. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024c. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Andres Marafioti, Merve Noyan, Miquel Farré, Elie Bakouch, and Pedro Cuenca. 2024. Smolvlm - small yet mighty vision language model. https://huggingface.co/blog/smolvlm.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

NVIDIA. 2024. Nvidia cuda toolkit. https://developer.nvidia.com/cuda-toolkit.

OpenAI. 2024. Gpt-4o(mini) system card.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush

Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. 2022. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

J Jieba Sun. Chinese text segmentation: Built to be the best python chinese word segmentation module. https://github.com/fxsjy/jieba.

OpenGVLab Team. 2024a. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Jiacong Wang, Bohong Wu, Haiyong Jiang, Zhou Xun, Xin Xiao, Haoyuan Guo, and Jun Xiao. 2024a. World to code: Multi-modal data generation via self-instructed compositional captioning and filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4608–4623.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. *arXiv preprint arXiv:2010.11685*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

xAI Team. 2024. Grok-1.5 vision preview.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Li Yifan, Du Yifan, Zhou Kun, Wang Jinpeng, Zhao Wayne Xin, and Wen Ji-Rong. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. 2024a. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024b. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

A Authorship and Credit Attribution

Data Construction *Hongyuan Dong*, *Weijie Yin*

Pretraining *Hongyuan Dong*, *Weijie Yin*

SFT *Zijiang Kang*, *Weijie Yin*

Evaluation *Weijie Yin*

Project Lead *Xiao Liang*, *Chao Feng*, and *Jiao Ran*

B SAIL-VL Model Card

We provide a simplified model card for the proposed SAIL-VL-2B and SAIL-VL-8B model.

	SAIL-VL-2B	SAIL-VL-8B
# Parameter	1.85B	7.95B
LLM	Qwen2.5-1.5B	Qwen2.5-7B
Vision Encoder	InternViT-300M	
Resolution	$448 \times \{\{1 \times 1\}, ..., \{2 \times 5\}\}$	
Adapter	2-layer MLP	
Token Merge	2×2	
# Max Visual Token	2560	

Table 7: SAIL-VL-2B and SAIL-VL-8B model card.

C SAIL-VL Showcases

We task SAIL-VL-8B to tackle vision-based language interactions in Figure 6. The images are selected from the internet, and input questions are set to cover common queries in real-world human interactions. SAIL-VL demonstrates marked capabilities in language interactions in both English and Chinese. It also exhibits marked visual comprehension abilities for various input visual element types. Our model recognizes famous landmarks, buildings, and artworks, demonstrating a vast reservoir of world knowledge. It is also worth noticing that SAIL-VL also performs well in meme understanding. It not only perceives the visual elements accurately, but also points out the contrast that makes the meme humorous, exhibiting powerful visual comprehension and language interaction abilities.

D SAIL-Caption

D.1 Caption Data Quality Assessment

Datasets. To evaluate the caption data quality of SAIL-Caption and opensource alternatives, we curate an evaluation subset for each recaption dataset as the test set. Specifically, we randomly select 500 samples from three recaption datasets listed in Table 8. SA1B-QwenVL-Caption employs a finetuned QwenVL (Data, 2024a) model to annotate Chinese dense captions on the SA1B dataset. DataComp-LLaVA-Caption, on the other hand, trains a customized version of LLaVA (Liu et al., 2023) model to perform annotation. BLIP3-KALE tasks CogVLM-18B (Wang et al., 2023) and Mistral-8B-Instruct (Jiang et al., 2024) to generate knowledge-augmented detail captions, and then distill this pipeline into a 2B VLM to annotate the DataComp-1B (Gadre et al., 2024) dataset.

Dataset	Language	Captioner	GPT Eval	Human Eval
DataComp-LLaVA-Caption	EN	LLaVA-Captioner	51.14	70.0
SAIL-Caption-DataComp	EN	SAIL-Captioner	**61.50**	**87.2**
SA1B-QwenVL-Caption	CN	QwenVL-Captioner	63.82	74.6
SAIL-Caption-SA1B	CN	SAIL-Captioner	**71.36**	**88.2**
BLIP3-KALE	EN	2B VLM	59.88	73.2
SAIL-Caption-KALE	EN	SAIL-Captioner	**61.50**	**80.6**

Table 8: Data quality evaluation results of SAIL-Caption and other opensource caption datasets. The evaluation results of our SAIL-Captioner are marked with . The quality scores from GPT and human evaluation are rescaled to [0, 100] for simplicity.

We instruct Azure GPT4O-20240513 (OpenAI, 2024) to generate ground truth captions for the curated evaluation sets. We also use our SAIL-Captioner to generate detail captions on the evaluation sets. Captions from the original recaption dataset and those generated by our SAIL-Captioner model are then compared with the ground truth ones for evaluation.

GPT Evaluation. We first conduct GPT evaluation for efficiency. We feed the candidate captions and ground truth ones to Azure GPT4O API, and ask the model to judge the candidate caption quality based on the precision and recall of visual elements. The detailed prompt used for GPT evaluation is shown in Figure 7.

Human Evaluation. To further validate the advantage of the data quality of our SAIL-Caption dataset, we also task human experts to evaluate the data quality. We randomly select a 100-sample subset from each dataset and instruct the annotators to judge candidate caption quality based on the original image. As shown in Figure 7, human annotators are given two candidate captions from both the baseline dataset and SAIL-Caption simultaneously. The experts are required to provide quality scores for the given captions, as well as a Good-Same-Bad (GSB) judgment reflecting more fine-grained data quality differences.

In the inspection of 10% annotated samples, we

Figure 6: SAIL-VL-8B showcases. We include both English and Chinese queries with various input images.

observe a 95%+ accuracy, verifying the reliability of our human evaluation results.

Evaluation Results. As shown in Table 8, SAIL-Caption-DataComp, SAIL-Caption-SA1B, and SAIL-Caption-KALE achieve significantly higher scores than previous baseline datasets in both GPT and human evaluation. These results demonstrate the leading performance of our SAIL-Captioner model and the advantage in SAIL-Caption's data quality.

Caption Evaluation Instructions

GPT Evaluation

Provide feedback for the detail image description generated by an AI assistant. Below, I will give you the description of the image provided by the AI assistant and the ground truth description of this image.

- AI assistant's description of the image: {}
- Ground truth description of the image: {}

Please rate the precision and completeness of the AI assistant's description based on the content of the image. A high-quality image description should cover all key visual elements of the image, such as objects, attributes and relations, and do not contain hallucinations or errors. The score range is from 1 to 10, with higher scores indicating higher description quality. Output a single line, containing only the score of the AI assistant's description. No additional explanations are required.

Human Evaluation

Each image is paired with two captions, you should rate them based on the following criteria:

- Hallucination: Is there any Entity/Attribute/Relationship described in the caption but NOT present in the image?
- Inaccuracy: Is there any Entity/Attribute/Relationship present in the image but incorrectly described?
- Omission: Is there any Entity/Attribute/Relationship missing in the caption but clearly visible in the image?
- Language Fluency: Does the caption violate standard syntax or use unnatural expressions?

First, compare the two captions using GSB criteria:

- Win (1 > 2): The former caption is significantly better than the latter one
- Lose (2 > 1): The latter caption is significantly better than the former one
- Tie (1 = 2): Both captions are comparable in quality

Subsequently, assign absolute scores (1-5) to each caption independently:

- 5: Caption fully captures all visual elements (entities/attributes/relationships) with zero errors or omissions. Language is fluent and natural.
- 4: Contains minor errors (≤2 inaccuracies) or omits 1-2 non-critical elements. Language is mostly fluent.
- 3: Has moderate issues: 3-4 errors/omissions or noticeable hallucinations. Language is readable but awkward.
- 2: Frequent errors (≥5) or omits key elements. Language is confusing but partially understandable.
- 1: Severely mismatched with the image or unreadable (e.g., grammatical chaos, major hallucinations).

Output Format:
1. GSB:
2. Absolute Score (1-5):

Figure 7: GPT and human evaluation prompts for SAIL-Caption and other opensource caption datasets.

We also show the GSB evaluation results in Figure 8. The GSB comparison reflects more fine-grained caption quality differences in candidate captions than a single rating. In the GSB evaluation, SAIL-Captioner achieves 87%, 91%, and 79% win+tie rates against SA1B-QwenVL-Caption, DataComp-LLaVA-Caption, and BLIP3-KALE, respectively, exhibiting marked quality advantages.

We attribute the leading performance of SAIL-Captioner to the simple-yet-effective data distillation pipeline. SAIL-Captioner develops visual understanding abilities effectively from reference data annotated by powerful VLM APIs, enabling large-scale high-quality data generation with limited resources.

D.2 SAIL-Caption Showcases

We curate several image samples from SA1B (Kirillov et al., 2023), DataComp (Gadre et al., 2024), and BLIP3-KALE (Awadalla et al., 2024) as demonstrations to compare the quality of SAIL-Caption with existing opensource caption datasets. We compare SAIL-Caption with SA1B-QwenVL-

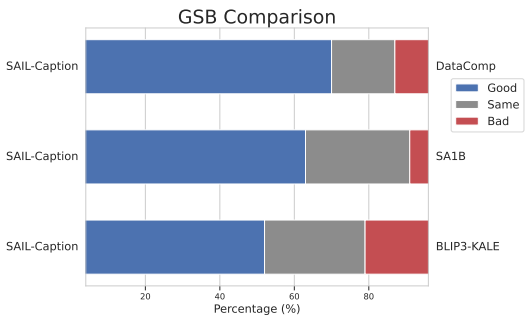

Figure 8: GPT and human evaluation prompts for SAIL-Caption and other opensource caption datasets.

Benchmark	Task type	Language	# Sample
Caption			
DetailCaps-4870-EN (Dong et al., 2024)	Caption	EN	4870
DetailCaps-4870-CN (Dong et al., 2024)	Caption	CN	4870
OCR			
IDL-WDS (Biten et al., 2022)	OCR	EN	1000
DocStruct (Wang et al., 2020)	OCR	EN	1000
SynthText (Gupta et al., 2016)	OCR	EN	1000
SynthDog-EN (Kim et al., 2022)	OCR	EN	1000
SynthDog-ZH (Kim et al., 2022)	OCR	CN	1000

Table 9: Basic information of the visual understanding benchmarks used in our experiments.

Caption, DataComp-LLaVA-Caption, and BLIP3-KALE. Showcases are shown in Figure 9. As the demonstrations show, SAIL-Caption encompasses more detailed visual elements than other alternative datasets in both English and Chinese. Observations drawn from these showcases coincide with quantified caption quality evaluation results shown in Section D.1, underscoring the leading data quality of SAIL-Caption.

A subset of the SAIL-Caption dataset with considerable data size will be released to promote opensource VLM research.

E Visual Understanding Benchmark

To inspect SAIL-VL's visual understanding performance during pretraining stages, we curate a series of visual understanding benchmarks for evaluation. To be specific, we focus on evaluating model performance in detailed captioning and OCR tasks in both English and Chinese, which are also the optimization objectives of SAIL-VL and opensource VLMs' pretraining stages. We list the basic information of the selected visual understanding benchmarks in Table 9.

Benchmarks. DetailCaps-4870 (Dong et al., 2024) encompasses images from a wide range of publicly available datasets, including COYO (Byeon et al., 2022), LAION (Schuhmann et al., 2021), CC (Changpinyo et al., 2021), Flickr (Young et al., 2014), SBU (Ordonez et al., 2011), and COCO (Chen et al., 2015), as well as ground truth detail captions generated by powerful VLM APIs. We use the human-refined version of DetailCaps-4870 for evaluation and adopt both the corrected Chinese captions and translated English captions for multilingual evaluation.

The remaining OCR benchmarks consist of images with a diverse distribution. IDL-WDS (Biten et al., 2022) consists of document pages with abundant text information; DocStruct (Wang et al., 2020) contains both document pages but also illustrative images rendered from tables and charts; SynthText (Gupta et al., 2016) is composed of images with a single word, but the fonts vary from one sample to another; SynthDog-EN and SynthDog-ZH (Kim et al., 2022) are compositional datasets comprised of natural image backgrounds and foreground word pieces.

Metrics. We evaluate SAIL-VL's caption performance on the DetailCaps-4870 benchmark with GPT evaluation. Provided with three ground truth captions and a candidate caption, GPT is tasked to score the candidate caption based on the precision and recall of the visual elements. For OCR tasks, we compute the ANLS score (Biten et al., 2019) between the predicted OCR contents and the ground truth ones, resulting in scores ranging from 0 to 1. The higher score indicates better prediction quality.

All benchmark data is curated from left-out subsets to avoid data leakage between model training and evaluation. We select a 500-case subset randomly from each benchmark to evaluate the pretrained model checkpoints for efficiency.

F SFT Data Quality Evaluation

In this section, we show the detailed instructions for SFT data quality evaluation. As shown in Figure 10, human experts to annotate the challenging, complexity, and relevance scores for our three-stage SFT data.

G Generalizing Visual Understanding Abilities to Instruction Following Tasks

In this part, we investigate the correlation between VLM's visual understanding and instruction following abilities. We train SAIL-VL pretrained checkpoints from the pretrain-advance stage with exponentially larger training data sizes, and train them

Figure 9: SAIL-Caption showcases versus SA1B-QwenVL-Caption, DataComp-LLaVA-Caption, and BLIP3-KALE. Images are curated from SA1B, DataComp and BLIP3-KALE.

SFT Data Evaluation Instructions

You are given a Q&A conversation and a corresponding image. Your goal is to:

1. *Evaluate Complexity Score*: Rate the combined information richness and complexity of the Q&A conversation and image on a scale from 1 to 5. Consider factors like the amount of detail, depth of content, and how well the conversation and image complement each other in conveying comprehensive information.
   - 1: Minimal detail, shallow content, limited interaction between conversation and image.
   - 2: Some detail, moderate depth, a basic connection between the conversation and image.
   - 3: Good amount of detail, some complexity, moderate complementarity between conversation and image.
   - 4: High level of detail, deep content, the conversation and image work well together to convey a thorough understanding.
   - 5: Very rich in detail, highly complex, the conversation and image are seamlessly integrated to provide a comprehensive, insightful picture.

2. *Evaluate Challenging Score*: Rate how difficult the question is to answer based on the image and conversation on a scale from 1 to 5. Consider factors like the need for nuanced visual analysis, contextual reasoning, ambiguity in the image, and the level of inference required.
   - 1: Very easy or unanswerable, requires only superficial observation or literal interpretation of the image. On the other hand, if there is no meaningful answer for the question, we also score the question with the lowest one, e.g., how many sands are in the beach.
   - 2: Easy, involves basic reasoning or simple inference with minimal ambiguity.
   - 3: Moderately challenging, requires combining multiple visual elements or contextual clues.
   - 4: Difficult, demands complex analysis, abstract reasoning, or resolving significant ambiguity.
   - 5: Extremely challenging, involves expert-level interpretation, synthesizing subtle details, or tackling high ambiguity/abstract concepts.

3. *Evaluate Relevance*: Rate the relevance of the conversation to the image on a scale from 1 to 5.
   - 1: Very low relevance, the conversation and image are almost unrelated.
   - 2: Low relevance, the conversation and image share some overlap, but one is mostly independent of the other.
   - 3: Moderate relevance, there is some connection, but the conversation could stand alone without the image or vice versa.
   - 4: High relevance, the conversation and image are closely tied, and both contribute significantly to each other's meaning.
   - 5: Very high relevance, the conversation and image are inseparable, and one cannot fully be understood without the other.

Output Format:
1. Complexity Score (1-5):
2. Challenging Score (1-5):
3. Relevance Score (1-5):

Figure 10: Detailed instructions for human experts to judge SFT data quality.

through either LLaVA-Next (Liu et al., 2024a) SFT data or our SFT-instruction data.

We plot the correlation of pretrained models' visual understanding performance and SFT models' opensource benchmark performance in Figure 11. A notable correlation is observed across different training strategies. As the VLM gains stronger visual understanding abilities during pretraining, its visual instruction following performance after SFT is improved accordingly, even if trained with different visual instruction tuning datasets. We quantify this correlation with Pearson correlation (ρ) and coefficient of determination (R^2). It turns out that SAIL-VL's pretrained visual understanding performance and SFT visual instruction following performance share a significant correlation. In experiments with SFT-Instruction data collection, pretrained model performance and SFT model scores share a Pearson correlation coefficient $\rho = 0.97$ and a coefficient of determination $R^2 = 0.94$. For LLaVA-Next SFT data experiments, we observe an even stronger correlation with $\rho = 0.99$ and $R^2 = 0.98$. These correlation results illustrate the generalization of model abilities across training stages and tasks, validating the necessity of pretraining VLMs for more robust visual understanding abilities.

H Experiment Details

H.1 Experiment settings

Storage. The training data of SAIL-VL's pretraining and SFT stages is stored on our Hadoop file system (HDFS) for persistent storage. Training data is fetched in a stream fashion during model training, making possible training with large scale distributed data storage.

Training framework. We use PyTorch (Paszke et al., 2019; Ansel et al., 2024) version 2.1.0 with CUDA (NVIDIA, 2024) 12.1 for model training. Deepspeed (Rasley et al., 2020) version 0.14.5 is used for SAIL-VL training. Flash-attention (Dao et al., 2022; Dao, 2023) implemented for 910B NPU (Huawei, 2023) is leveraged for fast attention computation.

We process training data sequences with a stream accumulator, which packs sequences in a micro batch into a long sequence for model training. This strategy speeds up SAIL-VL model training by approximately 40%.

Training resources. We conduct experiments with Huawei 910B x86 NPU (asc). To train the SAIL-VL-2B model, we allocate 90,053 NPU hours for pretraining and 10,992 NPU hours for SFT stages, resulting in a total of 101,045 NPU hours in model training. For the SAIL-VL-8B model, we use 26M samples in pretraining for efficiency, and the same data collections as the 2B model are used in SFT stages. The 8B model consumes 19,575 NPU hours to train, where 6,672 and 12,903 NPU hours are allocated in pretraining and SFT stages, respectively.

H.2 Experiment Results.

We show model training details for both SAIL-VL-2B and SAIL-VL-8B models in Table 10 and Table 11. We add InternVL2 (Team, 2024a), InterVL2.5 (Chen et al., 2024c), and Aquila (Gu et al., 2024) series models as supplementary baselines for evaluation.

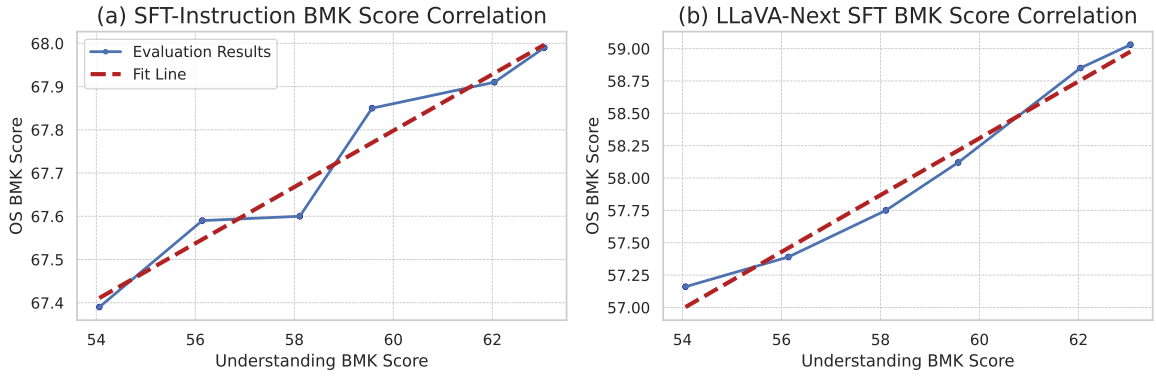

Figure 11: The correlation between SAIL-VL pretrained checkpoints' understanding performance and their performance on opensource benchmarks after SFT. "OS BMK Score" stands for average score on opensource benchmarks in our evaluations.

Benchmark	SAIL-VL	Qwen2-VL	InternVL2	InternVL2.5	InternVL2.5-MPO	Aquila	DeepSeekVL-2-Tiny
Overall Performance							
Opensource Average	**69.1**	64.4	62.9	66.5	67.7	66.5	67.0
General VQA	60.4	58.3	55.6	62.6	**63.1**	59.8	59.4
OCR VQA	**75.9**	72.5	68.2	68.7	71.1	71.7	74.4
Math&Knowledge	**79.0**	59.0	70.6	73.0	75.3	75.4	71.3
Hallucination	**66.2**	62.9	61.7	66.2	64.5	62.9	63.6
General VQA							
MMStar (2024b)	**55.1**	46.3	50.5	53.5	54.3	54.7	49.9
$\text{MMBench}_{\text{DEV}}$ (2024c)	72.4	68.8	70.3	73.1	72.5	**74.4**	68.3
MMMU_{VAL} (2024)	40.1	39.9	34.2	40.7	41.2	**44.1**	39.6
MME (2023)	1969	1923	1859	2090	**2123**	1808	1910
$\text{SEEDBench}_{\text{IMG}}$ (2023a)	**74.7**	72.0	70.9	73.4	73.2	73.9	72.5
RealWorldQA (2024)	63.8	60.9	56.7	60.9	60.7	64.1	**64.8**
MMVet (2024b)	46.1	51.2	40.4	61.7	**64.0**	42.7	52.8
OCR VQA							
$\text{AI2D}_{\text{TEST}}$ (2016)	**79.0**	72.3	74.2	75.0	75.3	75.0	74.6
$\text{DocVQA}_{\text{VAL}}$ (2021)	**89.2**	88.7	86.0	87.4	87.8	85.0	88.6
$\text{InfoVQA}_{\text{VAL}}$ (2022)	**67.2**	63.4	57.5	61.6	61.6	60.5	63.8
$\text{ChartQA}_{\text{TEST}}$ (2022)	81.0	70.6	71.7	73.3	70.9	76.6	**81.2**
$\text{TextVQA}_{\text{VAL}}$ (2019)	75.7	78.8	73.4	76.4	77.2	76.4	**80.5**
$\text{OCRVQA}_{\text{TEST}}$ (2019)	**58.5**	54.3	36.2	28.3	40.0	51.3	51.4
OCRBench (2024d)	806	794	786	789	**846**	772	808
Math&Knowledge							
$\text{MathVista}_{\text{MINI}}$ (2023)	**62.8**	45.0	46.8	51.1	55.3	59.4	54.5
$\text{ScienceQA}_{\text{VAL}}$ (2022)	95.3	73.0	94.4	94.9	95.3	91.4	88.1
Hallucination							
HallusionBench (2024)	**45.7**	38.3	38.2	42.5	39.2	42.1	38.4
POPE (2023)	86.7	87.6	85.3	**89.9**	89.8	83.6	88.8

Table 10: Complete evaluation results for SAIL-VL-2B and opensource VLMs of comparable sizes. Denotations are defined the same as Table 4.

Benchmark	SAIL-VL	Qwen2-VL	InternVL2	InternVL2.5	InternVL2.5-MPO	DeepSeekVL-2-Small
Overall Performance						
Opensource Average	**74.5**	73.0	70.0	73.2	74.3	72.7
General VQA	68.3	68.5	66.6	70.1	**71.2**	66.8
OCR VQA	**79.8**	79.6	72.6	75.0	76.3	79.0
Mah&Knowledge	**83.3**	71.0	78.4	81.5	83.2	79.0
Hallucination	68.7	67.5	64.5	69.5	**69.7**	65.3
General VQA						
MMStar (2024b)	64.2	58.3	61.6	62.5	**65.3**	57.7
$\text{MMBench}_{\text{DEV}}$ (2024c)	79.5	79.5	80.3	83.1	**83.3**	78.1
MMMU_{VAL} (2024)	48.2	50.9	47.6	52.4	**52.8**	47.6
MME (2023)	2244	2321	2215	**2339**	2321	2149
$\text{SEEDBench}_{\text{IMG}}$ (2023a)	75.5	75.3	75.4	**77.0**	76.9	76.8
RealWorldQA (2024)	**71.9**	69.7	64.7	69.9	70.2	70.2
MMVet (2024b)	58.3	62.6	57.7	62.1	**66.8**	60.3
OCR VQA						
$\text{AI2D}_{\text{TEST}}$ (2016)	83.7	82.9	83.7	**84.6**	84.1	82.0
$\text{DocVQA}_{\text{VAL}}$ (2021)	92.2	**93.7**	90.8	91.8	92.1	92.3
$\text{InfoVQA}_{\text{VAL}}$ (2022)	75.2	75.9	61.5	75.5	**76.2**	72.5
$\text{ChartQA}_{\text{TEST}}$ (2022)	84.6	81.6	82.0	82.9	77.6	84.6
$\text{TextVQA}_{\text{VAL}}$ (2019)	77.7	**83.8**	77.6	79.0	79.2	83.3
$\text{OCRVQA}_{\text{TEST}}$ (2019)	**61.4**	56.2	38.1	29.5	36.7	54.5
OCRBench (2024d)	835	833	746	819	**880**	834
Math&Knowledge						
$\text{MathVista}_{\text{MINI}}$ (2023)	68.4	57.3	59.4	65.4	**68.5**	61.8
$\text{ScienceQA}_{\text{VAL}}$ (2022)	**98.2**	84.6	97.4	97.6	97.9	96.2
Hallucination						
HallusionBench (2024)	**52.2**	48.5	44.6	50.1	50.3	41.2
POPE (2023)	85.2	86.5	84.4	88.8	89.1	**89.4**

Table 11: Complete evaluation results for SAIL-VL-8B and opensource VLMs of comparable sizes. Denotations are defined the same as Table 4.