AndesVL Technical Report: An Efficient Mobile-side Multimodal Large Language Model

AndesVL Team, OPPO AI Center

⌂ https://github.com/OPPO-Mente-Lab/AndesVL_Evaluation
🤗 https://huggingface.co/OPPOer

Abstract

In recent years, while cloud-based MLLMs such as QwenVL, InternVL, GPT-4o, Gemini, and Claude Sonnet have demonstrated outstanding performance with enormous model sizes reaching hundreds of billions of parameters, they significantly surpass the limitations in memory, power consumption, and computing capacity of edge devices such as mobile phones. This paper introduces AndesVL, a suite of mobile-side MLLMs with 0.6B to 4B parameters based on Qwen3's LLM and various visual encoders. We comprehensively outline the model architectures, training pipeline, and training data of AndesVL, which achieves first-tier performance across a wide range of open-source benchmarks, including fields such as text-rich image understanding, reasoning and math, multi-image comprehension, general VQA, hallucination mitigation, multilingual understanding, and GUI-related tasks when compared with state-of-the-art models of a similar scale. Furthermore, we introduce a 1+N LoRA architecture alongside a Quantization-Aware LoRA Fine-Tuning (QALFT) framework to facilitate efficient task adaptation and model compression during mobile-side deployment of AndesVL. Moreover, utilizing our cache eviction algorithm—OKV—along with customized speculative decoding and compression strategies, we achieve a 6.7x peak decoding speedup ratio, up to 30.9% memory reduction, and 1.8 bits-per-weight when deploying AndesVL-4B on MediaTek Dimensity 9500 chips. We release all models on https://huggingface.co/OPPOer.

1 Introduction

In recent years, the advent of large language models (LLMs) represented by ChatGPT [178], the Qwen series [13, 241, 242, 240], and the DeepSeek series [17, 126, 68] has ushered in a new era of artificial intelligence. These LLMs have demonstrated remarkable capabilities in natural language processing tasks, such as text generation, question answering, and language translation. Building upon the success of LLMs, multimodal large language models (MLLMs) have emerged, expanding the functionality of large models from pure text to multiple modalities. MLLMs incorporate modalities such as image, video, and even audio, enabling more diverse and comprehensive interactions.

The typical training paradigm of MLLMs involves leveraging a pre-trained LLM. By aligning the LLMs with visual encoders and engaging in continual pre-training and fine-tuning, an MLLM that can process multimodal inputs and generate text outputs is developed. For effective training, a substantial amount of multimodal data is necessary, in addition to extensive text datasets. This data encompasses image-text pairs, optical character recognition (OCR) data, and visual question-answering (VQA) data. These datasets provide the model with a wide range of multimodal capabilities, such as image captioning, OCR, chart question answering, visual semantic recognition, and visual reasoning.

On the cloud side, there are numerous outstanding MLLMs. Models such as the QwenVL series [14, 227, 16], the InternVL series [30, 29, 28, 274, 229], GPT-4o [177], Gemini [210, 189, 44, 43, 219] and Claude Sonnet [8] have demonstrated SOTA competence in multimodal tasks. Despite their groundbreaking performance, these models are generally oriented towards reaching maximum performance, involving parameter sizes running into hundreds of billions. Such large-scale parameter demands significantly exceed the processing capabilities
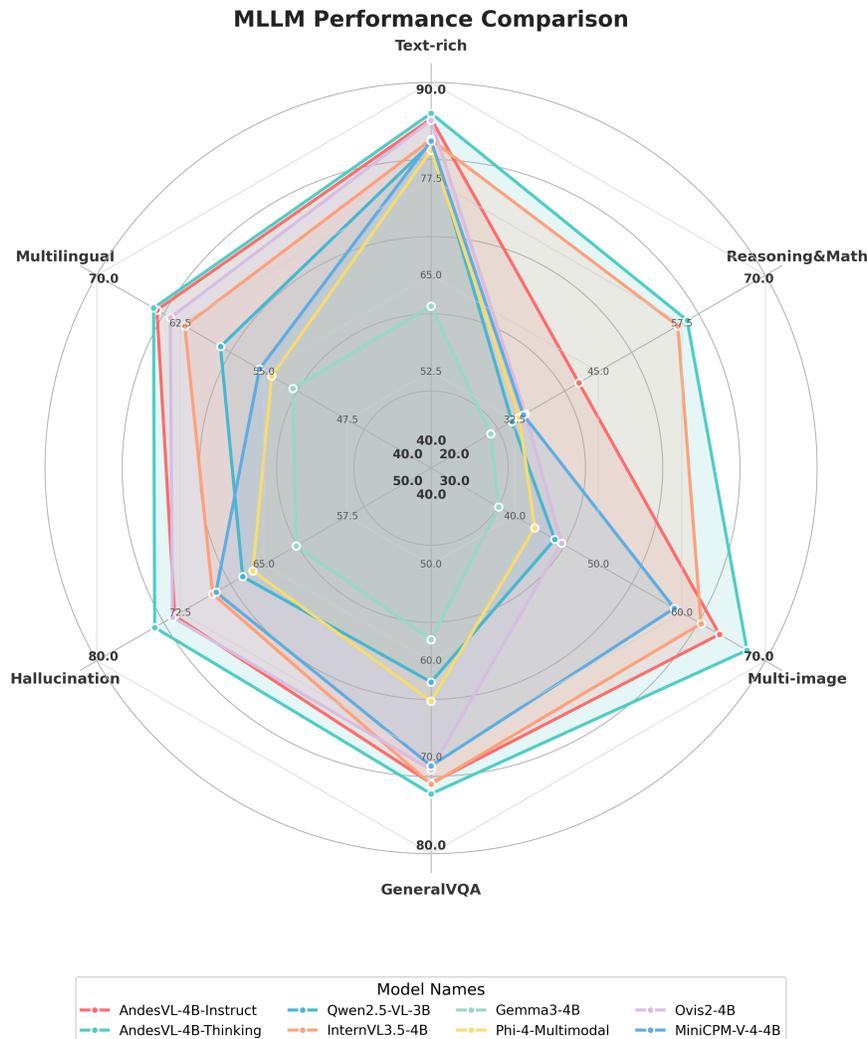
MLLM Performance Comparison

Figure 1: Overall performance comparisons over 6 domains (text-rich, reasoning and math, general VQA, multi-image, multilingual and hallucination) of different SOTA MLLMs with 4B parameters.

of mobile devices such as smartphones and tablets, particularly in terms of memory capacity, running speed, and computing power of chips. Consequently, MLLMs that typically run on mobile platforms are limited to approximately 4 billion parameters, as illustrated by Qwen2.5-VL-3B [16] and InternVL3.5-4B [229]. To maintain optimal functionality on mobile hardware, additional measures, such as quantization-aware training (QAT) and deployment optimization on the mobile side, are essential.

Currently, only a limited number of mobile-device manufacturers and internet companies have started exploring mobile-side MLLMs. For example, vivo has introduced BlueLM-V-3B [146] and BlueLM-2.5-3B [238], Meituan has launched the MobileVLM series [36, 37], Xiaomi has concentrated on mobile agents with the development of MobileVLM [233], and Apple has released the Ferret-UI series [249, 123] aimed at UI comprehension. Despite these initiatives, a thorough study explicating the training process, deployment strategies, and performance assessments on both general and mobile-specific benchmarks of mobile-side MLLMs is still absent.

In this paper, we introduce the AndesVL suite. By integrating Qwen3 [240] LLMs and various visual encoders, we have successfully developed mobile-side MLLMs with parameter sizes ranging from 0.6B to 4B. Our models focus on several key general-purpose capabilities, including knowledge acquisition, mathematical reasoning, handling text-rich content, dealing with hallucination issues, processing multi-image and multilingual inputs, and general VQA. We thoroughly introduce the model architectures, training pipeline, and data preparation strategies. Additionally, we have given special consideration to functions crucial for mobile-side MLLMs,

such as user interface (UI) understanding. To evaluate the performance of our models, we have developed mobile-specific benchmarks. Specifically, as inspired by Qwen3-4B-Instruct and Qwen3-4B-Thinking [240], we propose to train the instruct and thinking models of AndesVL separately to achieve the best instruct-following and reasoning abilities, respectively. Our floating-point models have achieved first-tier results among models of similar sizes across various benchmarks, as shown in Fig. 1, including 32 open-source benchmarks related to the domains mentioned above.

For practical application on mobile devices, we have also designed a 1 + N Low-Rank Adaptation (LoRA) [78] architecture to make the model adaptable to different tasks. Based on the AndesVL backbone model, downstream tasks can be clustered, and similar tasks can be fine-tuned using a single LoRA module to achieve optimal performance in specific domains. In addition to floating-point models within the 4B parameter range, to enable large models to run on the mobile side, the QAT and Quantization-Aware Lora Fine Tuning (QALFT) frameworks are necessary for model compression. With this pipeline, our mobile-side models have also demonstrated excellent results in various realistic applications. Additionally, we meticulously design a comprehensive mobile-side acceleration suite, with cache eviction, speculative decoding and sparsification, which achieve a block efficiency (BE) of up to 7.9, with about a 6.7x end-to-end decoding speedup over the baseline (with auto-regressive decoding and without compression optimization). Furthermore, we achieve a memory reduction of up to 30.9% and a weight compression of 1.8 bits-per-weight when deploying AndesVL-4B on MediaTek Dimensity 9500 chips. This work represents a significant step forward in the development and deployment of mobile-side MLLMs.

The structure of this article is as follows: Sec. 2 introduces the mobile-side MLLM and related work on mobile-side deployment. Sec. 3 focuses on the model architecture, training data, and training scheme of AndesVL. Sec. 4 introduces the mobile-side 1+N LoRA training architecture of AndesVL and the technical scheme for mobile-side deployment. Sec. 5 details the performance of AndesVL on public benchmarks and self-built mobile-side benchmarks, as well as its comparison with SOTA models. Sec. 6 presents the benchmark results and mobile-side performance of AndesVL after being deployed on mobile devices. Sec. 7 looks ahead to future directions. Sec. 8 summarizes the entire article.

The main contributions of this article can be summarized as follows:

- First, addressing the speed and performance trade-off for mobile implementations of MLLM, we introduce the AndesVL suite, which is a collection of MLLMs designed for efficient deployment on edge devices, with parameter scales ranging from 0.6B to 4B, demonstrating competitive performance with SOTA models with comparable parameters.

- Second, we offer separate models for Instruct and Thinking versions, making each ideal for tackling the challenges associated with high-efficiency tasks in understanding and generation, as well as applications in complex mathematical reasoning and planning.

- Third, we design a 1+N Lora training pipeline for mobile deployment, which enables efficient task clustering and adaptation. We further propose the QALFT framework to ensure flexible application of the 1+N Lora architecture on mobile devices.

- Finally, based on our mobile-side acceleration and compression strategies, *e.g.*, customized cache eviction, sparsification, and speculative decoding, AndesVL-4B can achieve a 6.7x peak decoding speedup ratio, a memory reduction of up to 30.9%, and 1.8 bits-per-weight on MediaTek Dimensity 9500 chips.

2 Related Works

2.1 Mobile-side MLLMs

Recent years have witnessed a proliferation of remarkable advances in MLLMs. Numerous remarkable MLLMs [14, 227, 15, 30, 29, 28, 274, 229, 177, 210, 189, 44, 43, 219, 8] have been introduced, primarily driven by the pursuit of exploring the upper bounds of model performance through scaling laws. This endeavor has resulted in models with astronomically large parameter counts, reaching hundreds of billions or even trillions. Nonetheless, this emphasis on large-scale models has left the development of mobile-side MLLMs relatively underexplored.

Among the efforts towards more mobile-friendly MLLMs, the Qwen series has made notable progress. Qwen2-VL [227] and Qwen2.5-VL [15] introduced model sizes of 2B and 3B, respectively, which are particularly suited for deployment on mobile devices. These model sizes effectively balance performance and the computational

limitations of mobile hardware. Similarly, the InternVL series [28, 274, 229] presented a range of model sizes—1B, 2B, and 4B—designed to fulfill various operational needs on mobile platforms.

In 2023, Meituan emerged as a pioneer in the mobile MLLM domain with the introduction of MobileVLM [36]. Built upon MobileLLaMA in a LLaVA-like [129] architecture, MobileVLM came in 1.7B and 3B model sizes. It achieved SOTA results in some benchmarks for models of similar sizes at that time. Meituan offered significant insights into the processing speeds on mobile and IoT platforms, reporting rates of 12.21 and 21.54 tokens per second, respectively. In 2024, the release of MobileVLM V2 [37] further advanced the field by exploring the data scaling law, improving training strategies, and optimizing the modality alignment design. These developments contributed to a comprehensive enhancement in the performance of the MobileVLM framework.

In 2024, the Apple MM series [154, 258] demonstrated that even relatively compact models, specifically those with 1B and 3B parameters, could achieve impressive performance through meticulous data curation and optimized training strategies. The Ferret UI series [249, 123] marked a significant step forward, as it was the first series extensively dedicated to improving the capabilities of screen UI understanding. It extended the capabilities of MLLMs to tasks such as referring and grounding on mobile UI screens and answering questions related to screen operations. However, Apple did not reveal the performance metrics for these models when deployed on mobile platforms.

Xiaomi's MobileVLM [233] also made important contributions by leveraging carefully constructed UI understanding and APP operation trajectory data. This enabled the model to expand its capabilities from understanding within a single UI (intra-UI) to understanding and operating across multiple UIs (inter-UI). Nevertheless, Xiaomi's 9.8B MobileVLM model was not successfully deployed on mobile devices.

Finally, vivo's BlueLM-V-3B [146] and BlueLM-2.5-3B [238] achieved mobile-side deployment of an MLLM through systematic optimizations in algorithms and hardware deployment. Specifically, BlueLM-V-3B achieved a running memory of 2.2G and a token throughput speed of 24.4 tokens/s on MediaTek Dimensity 9300 NPUs. This not only showcases its effectiveness but also provides practical performance metrics for mobile-side MLLMs.

Despite these efforts, there remains a gap in comprehensively documenting training processes, deployment solutions, and benchmark results for general and mobile-specific tasks of mobile-side MLLMs. Our work aims to fill this void by presenting the AndesVL suite, which offers a comprehensive approach to mobile-side MLLMs, including detailed training, deployment, and benchmarking aspects.

2.2 Mobile-Side Deployment of MLLM

The deployment of MLLMs on mobile devices presents unique challenges, including limited computational resources, diverse hardware architectures, and stringent energy constraints. To address these issues, various solutions [156, 61, 212, 82, 86, 114, 42, 10] have been proposed that take advantage of CPUs, GPUs, and NPUs.

CPU-based Deployment   In 2020, Alibaba developed the Mobile Neural Network (MNN) [86], an inference engine tailored for mobile applications. It introduces a "pre-inference" mechanism for runtime optimization, thorough kernel optimizations for optimal computation performance, and a back-end abstraction module that enables hybrid scheduling while maintaining a lightweight engine footprint on mobile CPUs.

In 2023, Georgi Gerganov [61] introduced llama.cpp, a lightweight, dependency-free C/C++ implementation designed for efficient LLM inference across diverse hardware platforms, including mobile CPUs. It includes support for several quantization levels (ranging from 1.5-bit to 8-bit), enabling reduced memory consumption and accelerated inference.

GPU-based Deployment   In 2024, a machine learning compiler and high-performance deployment engine for LLMs, MLC LLM [212], was developed, aiming to enable native deployment across various platforms, including mobile GPUs. It compiles models into optimized binaries compatible with platforms such as iOS, Android, and web browsers.

In addition, Li et al. [114] proposed Transformer-Lite, which focuses on the high-efficiency deployment of LLM on mobile phone GPUs. It introduced four optimization techniques: a symbolic expression-based approach for dynamic shape model inference, operator optimizations with execution priority settings, an FP4 quantization method termed M0E4 to reduce dequantization overhead, and a sub-tensor-based technique to eliminate the need for copying key-value (KV) cache after inference. These optimizations enable significant speedups in both prefill and decoding phases compared to existing CPU-based and GPU-based inference engines.
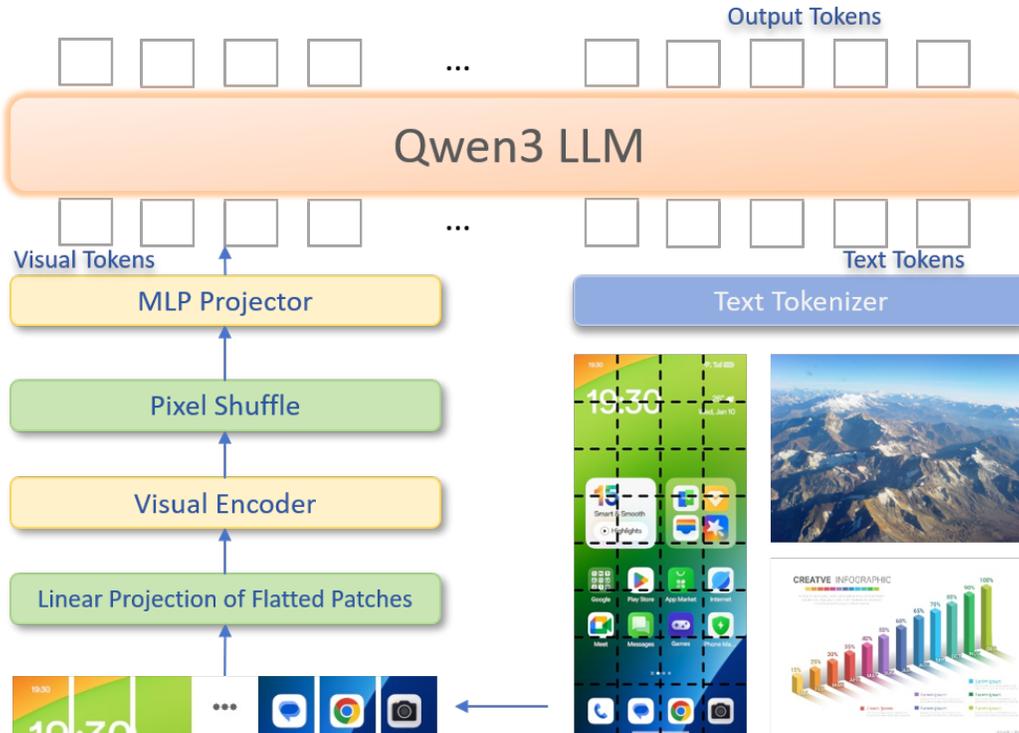
Figure 2: The overall architecture of AndesVL mainly includes a visual encoder, an MLP projector, and an LLM.

NPU-based Deployment Gemini Nano [42], developed by Google, is designed for on-device use cases, running within Android's AICore system service to leverage device hardware for low-latency inference. It is accessible through the AI Edge SDK, which allows developers to customize the inference and prompts. Gemini Nano models, such as Nano-1 (1.8B parameters) and Nano-2 (3.25B parameters), are distilled from larger Gemini models and optimized for edge devices such as smartphones.

Finally, Apple's On-Device Deployment utilizes the Core ML framework to optimize and deploy large language models on Apple silicon [10]. Techniques such as grouped-query attention (GQA) mechanisms, mixed 2-bit and 4-bit quantization, and efficient memory management strategies enable the deployment of models like Llama-3.1-8B-Instruct on devices such as the iPhone 15 Pro, achieving decoding speeds of approximately 30 tokens per second.

Despite notable progress in mobile-side deployment of MLLMs, several challenges persist. These include balancing model performance with resource constraints, ensuring cross-device compatibility, standardizing deployment processes, and establishing comprehensive evaluation frameworks for multimodal tasks. To address these issues, we introduce the AndesVL series, which offers a comprehensive suite of optimized deployment solutions tailored for mobile platforms. This includes detailed training methodologies, quantization techniques, compilation strategies, and hardware-specific optimizations. Our work aims to bridge existing gaps, providing a robust foundation for future research and practical applications in mobile-side MLLM deployment.

3   AndesVL

3.1   Model Architecture

AndesVL comprises models with parameters ranging from 0.6B to 4B parameters, with detailed architectures provided in Table 1. Following the paradigm of typical MLLMs [129, 26, 227], it consists of three fundamental components: a visual encoder, an MLP projector, and an LLM, as illustrated in Fig. 2.

As a general-purpose MLLM, AndesVL is designed to handle image inputs with arbitrary aspect ratios and resolutions. To achieve this, we avoid the image cropping methods employed in other works [128, 26, 239] and

Model Name	#Param (B)	Vision Encoder	Language Model
AndesVL-0.6B	0.695	SigLIP2-Base [217]	Qwen3-0.6B [240]
AndesVL-1B	0.927	AIMv2-Large [57]	Qwen3-0.6B [240]
AndesVL-2B	2.055	AIMv2-Large [57]	Qwen3-1.7B [240]
AndesVL-4B	4.360	AIMv2-Large [57]	Qwen3-4B [240]

Table 1: AndesVL model architectures of different sizes.

instead implement a Native Resolution ViT (NaViT) [45]-based strategy, allowing the visual encoder to process input of any resolution directly. This method is particularly beneficial for efficiently processing low-resolution images and ensures consistency between model inputs and the original data. The MLP projector includes two MLP layers used to align the ViT output with the LLM's embedding layer. To decrease the sequence length of the ViT output going into the LLM, a straightforward yet adaptable pixel shuffle operation is used to reduce the sequence length to a quarter of its original size. This operation combines and concatenates the data from adjacent 4×4 patches before passing them to the MLP. For the language model, AndesVL employs Qwen3 [240], utilizing the 0.6B, 1.7B, and 4B models from the Qwen3 series. To save memory for embedding parameters, we preserve the tied word embeddings configuration across all LLM variations.

3.2 Training Pipeline

In this paper, inspired by the recently released Qwen3-4B-Instruct and Thinking models [240], we propose to develop AndesVL in two distinct models: an instruction model (supporting only non-thinking mode) and a thinking model (supporting both non-thinking and thinking modes). Consequently, the training datasets and methodologies are distinct from one another, as will be thoroughly explained in the subsequent subsections.

Stages	Visual-Language alignment	Joint V-L pre-training	Multi-task pre-training
Main data type	Caption + OCR + VQA	Interleaved image-text + Pure text + VQA + Long CoT*	All multi-task data + Long CoT*
Trainable parameters	ViT + MLP	Full model	Full model
ViT sequence length	4,096 / 16,384	4,096	16,384
LLM sequence length	2,048 / 8,192	8,192	32,768
Trained tokens	100B / 100B	160B	160B

Table 2: Pre-training stages of AndesVL. * indicates data exclusively used for the Thinking models.

3.2.1 Pre-train

As illustrated in Table 2, the pre-training phase of AndesVL consists of three stages: vision-language alignment, joint vision-language pre-training, and multi-task pre-training. To improve training efficiency, we pack not only the ViT input but also the LLM input tokens. Our pre-training commences with the Qwen3-Instruct/Thinking versions of the language model. Throughout all pre-training stages, we incorporate a proportion of instruction-following data. This allows us to maintain the model's instruction-following capabilities and monitor its performance progression directly via instruction-based evaluation.

Vision-Language Alignment   Our primary visual encoder leverages AIMv2-Large [57], a compact 300M parameter model that offers superior power efficiency compared to larger alternatives such as Qwen2VL-ViT-675M [227], making it particularly well-suited for mobile deployment. To enhance the encoder's versatility across varying input resolutions, we integrate 2D Rotary Position Embeddings (2D-RoPE) [202], whose strong extrapolation capabilities enable our vision encoder to effectively process high-resolution inputs even when trained on lower resolutions. To maintain model performance, we preserve the original position embeddings with a length of 1024 and adapt them to different resolutions using bicubic interpolation.

We employ a two-stage training procedure for the visual encoder within our MLLM framework, keeping the LLM frozen while utilizing diverse training data from captions, OCR, and VQA tasks. The first stage processes 100B tokens with a ViT sequence length of 4,096, applying higher learning rates specifically to the randomly initialized MLP layers, while the second stage continues with an additional 100B tokens with a ViT sequence length of 16,384. For our 1B and 2B model variants, we streamline the training process by directly leveraging the vision encoder from our 4B model and performing alignment by training the MLP layer only. For our most compact 0.6B model variant, we adopt the SigLIP2-Base-Patch16-512 [217] model, which follows a similar

adaptation strategy that combines bicubic interpolation for position embeddings with 2D-RoPE and two-stage training.

Joint Vision-Language Pre-training   The second stage involves joint vision-language pre-training. After the visual encoder's output aligns well with the LLM's representations, we unfreeze the LLM parameters and conduct full-parameter pretraining using a relatively low learning rate. In this stage, we utilize a large volume of unsupervised interleaved image-text data, enabling the model to acquire extensive visual knowledge. During pre-training on this data, we compute loss only on text tokens, excluding image tokens from the calculation.

In unidirectional autoregressive transformers, inappropriate image positioning may prevent the model from learning multimodal knowledge effectively. For instance, images placed at the end of a sequence cannot contribute to learning even when encoded. To mitigate this issue, we employed a strategy where, with 50% probability, we maintained the original image positions. With the remaining 50% probability, we relocated all images in the data to precede all text content, replacing the images with their corresponding indices. Fig. 3 illustrates this transformation.

Original interleaved document:

```
The sunset over the Pacific Ocean was breathtaking.
<img>pacific_sunset.jpg</img> The vibrant colors painted
the sky in shades of orange and pink. Later that evening,
we hiked to the mountain viewpoint. <img>mountain_vista.jpg</img>
```

Transformed format:

```
<|image_0|> <img>pacific_sunset.jpg</img>
<|image_1|> <img>mountain_vista.jpg</img>
The sunset over the Pacific Ocean was breathtaking.
<|image_0|> The vibrant colors painted the sky in shades
of orange and pink. Later that evening, we hiked to the
mountain viewpoint. <|image_1|>
```

Figure 3: Image repositioning strategy for joint vision-language pre-training. Images are moved to the beginning of the sequence with 50% probability to ensure effective multimodal learning.

Since interleaved image-text data can be viewed as a multimodal extension of unlabeled text data, we also incorporate unlabeled text data from text pre-training. To maintain instruction-following capabilities, we include text instruction data in this stage as well. Furthermore, a certain proportion of multi-task pre-training data is added to enhance the model's overall multimodal abilities. For the Thinking version of the model, we additionally incorporate long CoT data, which will be detailed in Sec. 3.3.1.

Multi-task Pre-training   The final stage is multi-task pre-training. In this stage, our approach transitions from self-supervised learning with unsupervised data to supervised learning using annotated data, focusing solely on calculating the text token loss for the answer portions. Data types mainly consist of general VQA, captions, and OCR, alongside task-specific data like grounding/UI. For the Thinking model variant, we continue to incorporate long CoT data as in the previous stage, while increasing the proportion of multimodal types to enhance its step-by-step reasoning capabilities with visual inputs. Although we use 2D RoPE to allow model inference at high resolutions, we increased the ViT patch input from 4,096 to 16,384 to facilitate learning from data that require high resolution. To enhance the LLM's capabilities in long contexts, particularly its reasoning ability in Thinking mode, we expanded the LLM's sequence length from 8,192 to 32,768.

Consequently, by completing the three pre-training stages mentioned above, we developed the base versions for our Instruct and Thinking models, referred to as *AndesVL-Instruct-Base* and *AndesVL-Thinking-Base*, respectively, which are subsequently utilized for post-training.

3.2.2   Post-train

The AndesVL post-training process consists of two main stages: supervised fine-tuning (SFT) and reinforcement learning (RL). SFT is utilized for both instruction and thinking models. Notably, mixed preference optimization (MPO) [228] is adopted for refining the instruction models, while Group Relative Policy Optimization

(GRPO) [68] is employed for the thinking models. Following the application of SFT and MPO to AndesVL-Base, we derive the *AndesVL-Instruct* model. Conversely, the *AndesVL-Thinking* model is attained through the application of SFT and GRPO.

SFT   Supervised fine-tuning (SFT) of the pre-trained AndesVL model is conducted utilizing meticulously formatted instruction data. Recognizing the critical influence of data diversity and quality on the performance of downstream tasks, an extensive array of multimodal instruction data is compiled, covering a wide range of task areas. To improve the model's conversational abilities, the Chat-ML instruction data format is employed.

The instruction dataset is meticulously crafted to introduce the model to multiple input modalities, enabling the development of strong representational learning capabilities. Additionally, the dataset encompasses a diverse range of task objectives, such as image captioning, visual question answering, text summarization, and code generation. This deliberate diversification in data sources and task outlines is designed to enhance the model's generalization capacity and remain robust across various application scenarios. Compliance with the Chat-ML format supports seamless integration with contemporary dialogue-oriented systems, thus promoting coherent and informative conversation exchanges. This strategic SFT method is essential for unlocking the full potential of the AndesVL model, thereby facilitating its effective use in real-world scenarios.

MPO   Direct preference optimization (DPO) has emerged as the dominant approach for aligning LLMs with human preferences, as highlighted in [187], which can avoid complex on-policy RL pipelines and is suitable for training non-thinking models. Leveraging its effectiveness in language processing, recent research has extended the application of DPO to multimodal settings [113, 272]. Nonetheless, two challenges arise when implementing DPO in MLLM: the scarcity of comprehensive, high-quality multimodal reasoning preference datasets and DPO's inability to assess the absolute quality of individual responses. To address these issues, a novel approach known as Mixed Preference Optimization (MPO) was introduced by [228], which has shown enhancements across various multimodal reasoning evaluation sets.

We borrowed the MMPR dataset and MPO from [228]. During the training process, a joint loss consisting of preference loss $\mathcal{L}_p$, quality loss $\mathcal{L}_q$, and generation loss $\mathcal{L}_g$ was used, which can be formulated as

$$\mathcal{L} = w_p\mathcal{L}_p + w_q\mathcal{L}_q + w_g\mathcal{L}_g.$$

The preference loss $\mathcal{L}_p$ is formulated as

$$\mathcal{L}_p = -\log \sigma \left(\beta \log \frac{\pi_\theta \left(y_c \mid x\right)}{\pi_0 \left(y_c \mid x\right)} - \beta \log \frac{\pi_\theta \left(y_r \mid x\right)}{\pi_0 \left(y_r \mid x\right)} \right), \tag{1}$$

where β is the KL penalty coefficient, x, y_c, and y_r are user query, chosen response, and rejected response, respectively. The policy model π_θ is initialized from model π_0.

The quality loss $\mathcal{L}_q$ is formulated as

$$\mathcal{L}_q = -\log \sigma(\beta \log \frac{\pi_\theta \left(y_c \mid x\right)}{\pi_0 \left(y_c \mid x\right)} - \delta) - \log \sigma(-(\beta \log \frac{\pi_\theta \left(y_r \mid x\right)}{\pi_0 \left(y_r \mid x\right)} - \delta)), \tag{2}$$

where δ represents the reward shift, which is introduced by [89], calculated as the moving average of previous rewards to stabilize training.

The generation loss $\mathcal{L}_g$ is the standard cross-entropy loss:

$$\mathcal{L}_g = -\sum_{t=1}^{T} \log p_\theta(y_t \mid x, y_{<t}),$$

where p_θ represents the conditional probability distribution of the language model over tokens.

GRPO   In terms of training thinking model, subsequent to the SFT phase, our research transitions to on-policy RL training. Initially, a dataset comprising approximately 200k high-quality data pairs is collected from different sources, which will be detailed in Sec. 3.3.2. A difficulty score is subsequently assigned to each data sample, serving as a metric derived from the number of correct responses elicited across eight successive rollouts of the SFT version of AndesVL. Empirical observations indicate that data samples exhibiting either extreme difficulty or excessive simplicity do not meaningfully contribute to learning gains following reinforcement training. Consequently, we strategically select a subset of data with difficulty scores ranging from 1 to 4 for our

Task Type	Dataset Name
Caption	Emu2 [206], ShareGPT-4V [25], Laion-ZH [190], Wukong [64], Taisu [135]
OCR	DocMatrix [102], DocStruct [76, 77], Leopard-Instruct [83], Pixmo-doc [46], Anyword-3M [218], PIN-14M [225], In-house collected and synthesized (by Synthdog [225]) OCR data
Visual Grounding	Visual Genome [100], RefCOCOs [250, 125], LVIS [182], Flickr30k-Entities [182], Groma [149]
GUI and Agent	AITW [188], AITZ [260], AMEX [19], Android Control [117], Widget Caption [118], Rico [47], SeeClick [32], UIbert [12], Screen2Words [223], MultiUI [130], Aria-UI[244], OS-Atlas [234], Mind2Web [48], GUI-Odyssey [144], OmniAct [92], In-house AndesUI training set
VQA	Infinity-MM [65], MAmmoTH [70], LLaVA-OneVision [108], The Cauldron [105], VisualWebInstruct [84], PangeaInstruct [255]
Long CoT	OpenMathReasoning [160], OpenCodeReasoning [5], OpenThoughts [67], Nemotron [164], In-house multiModal long CoT data
Interleaved Image-Text	MMC4 [275], MINT [11], Multimodal Textbook [264], Wanjuan [72], OmniCorpus [116], In-house crawled data from Chinese websites and Apps
Pure Text	Fineweb-Edu-Chinese [253], Fineweb-Edu [138], FineMath [6], OpenCoder [79], Infinity-Instruct [175]

Table 3: The detailed lists of pre-training datasets.

training regimen. This refinement yields a final training dataset of 43.6K examples, including 15.3K pure text samples and 28.3K multimodal data instances.

Recent work on ReVisual-R1 [27] has demonstrated that subsequent text-only RL training, following a multimodal RL phase, further enhances multimodal reasoning capabilities. Our experiments with AndesVL similarly reveal that this two-stage RL training paradigm significantly improves the model's multimodal reasoning. Furthermore, we observed that training the model with RL in an "easy-to-hard" manner more effectively enhances model performance; thus, training samples are ordered according to their difficulty scores. Consequently, AndesVL also undergoes a two-stage training process, incorporating this easy-to-hard curriculum, utilizing the GRPO [194] algorithm. The empirical results unequivocally showcase a notable enhancement in AndesVL's domain-specific reasoning capabilities.

3.3 Training Data

3.3.1 Pre-train Data

In this section, we present in detail the data we utilized during the pre-training stage, including several open-source datasets and our in-house data, as illustrated in Table 3.

Image Caption Our image caption data comprises both Chinese and English languages. The Chinese image caption data primarily originates from Laion [190], Wukong [64], and Tasisu [135] datasets. To refine the quality of these descriptions, we utilized Qwen2.5-VL-72B [16] to generate re-captioned versions. During training, we randomly employed the original captions with a 50% probability and the re-captioned versions also with a 50% probability, culminating in a dataset of around 116 million samples. The English image caption data are primarily derived from the Infinity-MM [175] stage 1 subset, using Emu2 [206] for caption generation and consisting of approximately 10 million samples.

OCR OCR data plays a critical role in bridging visual and textual modalities within vision-language models (VLMs). Our OCR dataset is primarily derived from three sources: open-source data, synthetic data, and in-house collected data. For real-world textual images, we curated and refined widely used open-source datasets through our dedicated data engine. To further enhance data diversity, we also extracted text-rich images from the LAION-2B [190] dataset using PaddleOCR [110].

Synthetic data serves as another essential component in strengthening the model's OCR capabilities. Specifically, to improve recognition accuracy for Chinese characters, we generated a large-scale Chinese OCR dataset using SynthDog [98]. Additionally, we produced substantial amounts of non-semantic English OCR data to help mitigate the model's tendency toward hallucination. To further improve robustness, we applied extensive data augmentation techniques, including geometric transformations, noise injection, and style variations, ensuring the model generalizes effectively across diverse and challenging real-world scenarios.

Visual Grounding   We followed the bounding box structure utilized in Qwen2-VL [227] and prepared data for both single and multiple grounding scenarios. The grounding datasets were chosen from publicly available sources like Visual Genome [100], RefCOCOs [250, 125], Flickr30k-Entities [182], and Groma [149]. These datasets were screened and categorized into four classifications: object referring, region captioning, referenced entity recognition, and grounded image captioning. Inspired by Ferret [123] and Ferret-v2 [259], we ensured an equitable distribution of our grounding data across the Region-in-Text-out and Text-in-Region-Out formats.

GUI and Agent   We divided the GUI data into four categories, which are detailed caption, recognition, action, and element grounding. In the pre-training stage, the data were formatted in single-turn style. For the element grounding data, we kept the structure the same as our visual grounding data. During the data synthesis and reconstruction stages, we kept the balance between different task categories and platform domains.

Besides the publicly available GUI data, we built an in-house GUI dataset using ColorOS UI and application widgets, namely AndesUI. We gathered 90 apps in total, including 65 popular download apps from the OPPO Software Store, spanning a variety of categories frequently used by users, along with 25 ColorOS pre-installed apps. Annotators were directed to capture screenshots of different heterogeneous pages within each app, ensuring that each screenshot contained unique layouts and content. Ultimately, we collected a total of 10k screenshots from third-party apps and 2.5k from ColorOS pre-installed apps. Then, we aimed to annotate all the widgets within each screenshot. This included defining bounding boxes, identifying widget types, recording any available text on the widgets, and determining their clickability, among other details. On average, each interface produced 18 widgets. The training dataset resulted in a cumulative total of 227k widgets. Finally, we needed to construct both basic and advanced data. Basic data consists of grounding and referring datasets, whereas advanced data comprises overall descriptive data and natural question-answer pairs. As a result, the training set produced 227k referring data entries, 186k grounding data entries, 13k comprehensive description data, and 108k natural question-answer pairs. All the details of the AndesUI dataset are presented in Appendix B.

VQA   Our VQA dataset primarily originated from the open-source community, encompassing general VQA datasets, Infinity-MM [65], Llava One Vision [108], and The Cauldron [105]. Additionally, it included reasoning datasets such as MAmmoTH [70] and VisualWebInstruct [84], along with multilingual and multicultural datasets like PangeaInstruct [255].

Interleaved Image-Text   Interleaved image-text data serves as a natural extension of pure text pretraining data into scenarios encompassing image inputs. Unlike instruction data differentiating between single-image and multi-image contexts, interleaved image-text data is inherently multi-modal and simplifies to pure text corpora when there are no image inputs. This is similar to pretraining on purely textual data, which enables models to develop in-context learning abilities. We collected interleaved image-text data from the open-source community, which includes MMC4 [275], MINT [11], Multimodal Textbook [264], Wanjuan [72], and OmniCorpus [116]. Furthermore, to improve the model's understanding of Chinese language and culture, we created an in-house interleaved image-text dataset based on Chinese web content.

Pure-text   Pure text data plays a crucial role in maintaining the text capabilities of MLLMs. Our openly accessible pure text corpus comprises the Chinese FineWeb-Edu [253], the English FineWeb-Edu [138], the mathematical corpus FineMath [6], the code corpus derived from OpenCoder [79] annealing data. Besides, we also constructed a large quantity of in-house text pre-training corpora. Furthermore, we incorporated text instruction data from Infinity-Instruct [175].

Long COT data   The long CoT data construction pipeline is illustrated in Fig. 4. Our long CoT dataset was constructed from two distinct pipelines: one leveraged human annotations in combination with a DeepSeek-based Chain-of-Thought (CoT) data generation pipeline, while the other relied on distilling knowledge from existing CoT models. In the first pipeline, we began by collecting a diverse set of STEM (science, technology, engineering, and math) samples and common real-life images to serve as our visual inputs. Subsequently, human annotators developed pertinent questions based on these images and derived corresponding correct answers.
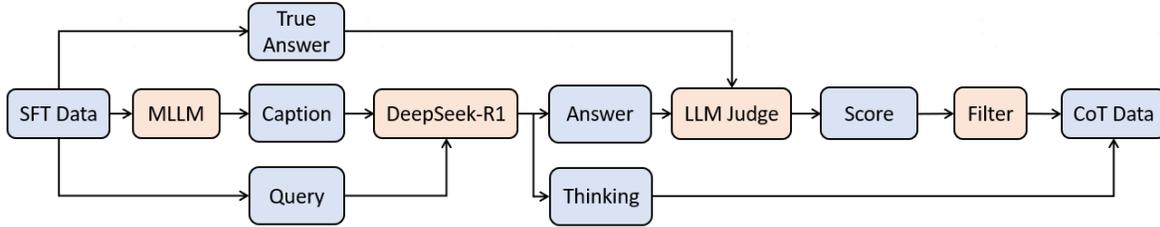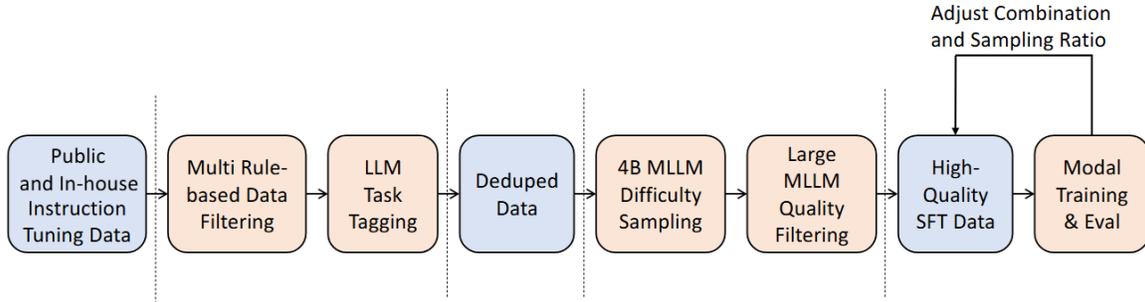
Figure 4: CoT data construction pipeline

Figure 5: SFT Data Filtering Pipeline

Concurrently, MLLMs, such as GPT-4o [177], were utilized to produce initial image descriptions. These descriptions were then manually refined, and query-related information was extracted to maintain alignment and relevance.

In the second pipeline, which focuses on distillation from existing CoT models, we employed a hybrid strategy that merges MLLMs with DeepSeek-R1 [68]. Specifically, MLLMs are used to generate descriptive captions for the input images. These captions, along with the associated queries, are input into DeepSeek, which generates detailed reasoning chains as output responses, thereby producing high-quality CoT examples.

3.3.2 Post-train Data

AndesVL undergoes two distinct post-training phases: SFT leveraging instruction data in a specific format and MPO utilizing sample & reject data pairs for AndesVL-Instruct or GRPO data for AndesVL-Thinking.

SFT Data   The SFT phase enhances the model's conversational capabilities and instruction-following abilities by using a large-scale, diverse dataset of instruction data obtained from publicly available resources and meticulously curated by closed-source strong models.

As illustrated in Fig. 5, we adopted a multi-stage data filtering process to further enhance the quality of these datasets. Initially, we utilized traditional rule-based single-modality filtering only on text and images, eliminating basic noise (*e.g.*, invalid or blurry images, incorrect instructions) and inappropriate data within each dataset. After this, we employed Qwen2.5-VL-72B [15] that clusters all datasets into different task categories. Then, these clustered image-text pairs were filtered on the measurement of quality and difficulty level by the LLM-as-a-judge [271] approach. We used GPT-4o [177] to measure quality according to factual accuracy, image-text correspondence, and hallucination levels. We employed previously trained checkpoints for difficulty filtering to generate multiple responses for the image-query pair. The image-text pair was considered unsuitable for the SFT training if most of the generated responses were judged to be the same as or above the level of the image-text pairs.

The final SFT dataset encompasses approximately 16 million entries, distributed between unimodal text data (10%) and multimodal data, incorporating interleaved image-text sequences. This data composition, characterized by a significant proportion of multimodal data complemented by a supplementary portion of pure text, enables the model to maintain robust performance even in purely text-based scenarios.

Task Type	Dataset Name
VQA	EATEN [69], PMC-VQA [265], OmniAlign-V [268], Dvqa [90], mm-localized-narratives [183], OCR-VQA[157], Plotqa[155], ShareGPT-4o [179], geoqa-plus[22], Figureqa [91], Tallyqa [4], LACR-I2I [165], Robut-wikisql [172], Mimic-cgd [104], clevr [87], textvqa [199], scienceqa [141], mpdocvqa [214], nlvr2 [203], ShareGPT4V [23], cocoqa [53], ScreenQa [75], raven [257], unigeo [21], docmatix [103], robut-wtq [270], mapqa [20], iconqa [143], chart2text [174], docreason25k [76], kvqa [99], scitsr [34], mm-LADD [166], tabmwp [142], KonIQ-10k [73], SVIT [267], sujet-finance-qa-vision [204], viet-sharegpt-4o-text-vqa [50], ChartQA [136], cambrian-10m [216], mm-aokvqa [191], ai2diagram [96], Clevr-CoGenT-TrainA-70K-Complex [87], laion-gpt4v [101], geometry3k [140], DocVQA [153], vistext [208], simchart9k [235], geomverse [95], spot-the-diff [207], robut-sqa [270], multihiertt [269], lrv-instruction [127], objects365 [193], mtvqa-train [209], iam [33], finqa [31], viquae [107], fsc147-controlnet [213], mm-tat-qa [273], infographic-vqa [63], vsr [262], orand-car-a [60], mm-tqa [170], mm-intergps [169], JourneyBench-Hallucination [88], mm-vqarad [171], mm-diagram-image-to-text [168], ctw[133], naf[41], LIME-DATA-ai2d-train [124], mmc-inst [56], COCO-Text [74], HME100K [221], st-vqa [220], fintabnet [9] , CoSyn-400K [46], PuzzleVQA [35]
OCR	anytext [218], CORD [180], invoice-to-json [1], arxiv-ocr [173], textocr [158], invoices-and-receipts-ocr-v2 [163], mall-receipt-extraction [18], invoices-and-receipts-ocr-v1 [162], ds-receipts-v2-train [161], dataset-receipt [80], invoices-donut-data-v1 [94], Vision-OCR-Financial-Reports [7], handwritten-text-ocr [215], nutritional-data-poie [93]
Pure Text Dialogue	sharegpt-gpt4 [198], ruozhiba [137], Ner-sentiment-analysis-sharegpt [246], chinese-ner-sft [184], few-shot-ner-sft [185], SystemChat [38]
Image Captioning	Detailed Caption [122], VizWiz [71]
Code Generation	WebSight [106], mm-datikz [167]
Function Calling	Function-Calling-Dataset-V1 [81], glaive-function-calling [62]
Markdown	Docomini[76], Mdocr Chinese markdown [225]
Math	cmm-math[131], MMR1-Math-RL-Data-v0 [159], Codegebragpt-multimodal [201], Geo170K [147], MathV360K [196], Multimath-300k [181], Unimer-math-ocr [222], Openr1-math-220k [176], MathInstruct [254], MetaMathQA [251]
In-house Data	Meticulously auto-generated, labeled, and curated instruction data

Table 4: The detailed lists of SFT datasets.

To facilitate the acquisition of more comprehensive and superior capabilities, the instruction data covers a wide range of task types, including VQA, OCR, image captioning, pure text dialogue, code generation, function calling, markdown format generation, pure text math, and multimodal math. We list the details of the SFT datasets in Table 4.

MPO Data The MPO dataset was derived from two distinct sources: one is constructed through our in-house data generation pipeline, and the other is the publicly available MMPR-v1.2 dataset [228] introduced by InternVL [274].

The MPO data construction pipeline, as illustrated in Fig. 6, began with the collection of SFT data from specific domains to act as input sources. Following this, sampling-based inference was conducted using AndesVL after SFT, yielding an n-best list of responses for each input. Subsequently, an MLLM, *e.g.*, GPT-4o [177], was leveraged to evaluate these candidate responses, selecting the most precise and coherent response as the *chosen response* and the least relevant or erroneous one as the *rejected response*, based on the input image, query, and reference answer.

Then a filtering strategy was applied to guarantee the quality of data. In particular, two similarity scores were calculated: (a) comparing the chosen and rejected responses and (b) comparing the rejected response with the ground-truth answer. Instances were discarded if (a) they fell below a designated threshold—indicating insufficient contrast—or if (b) they exceeded a specific level, implying that the rejected response is too similar to
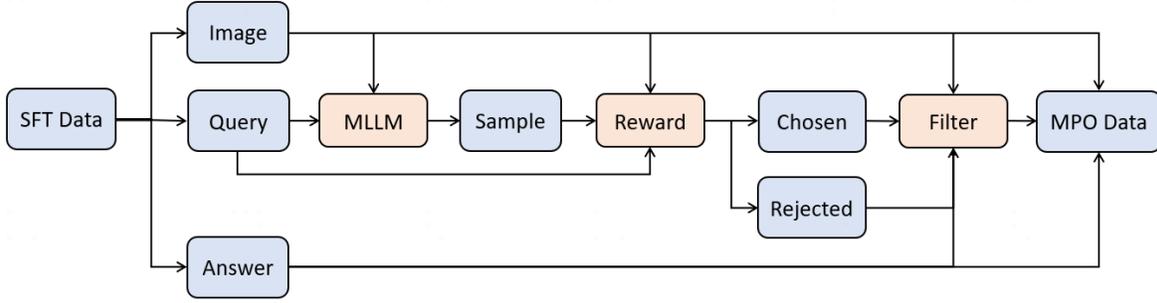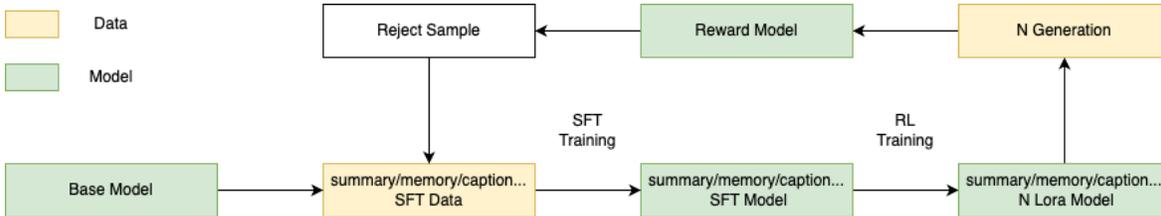
Figure 6: MPO data construction pipeline

Figure 7: Multi-scenario LoRA training based on AndesVL.

the correct answer, which could mislead the preference learning process. Finally, we obtained about 80k valid MPO data through this process.

GRPO Data The GRPO data primarily comprised tasks from STEM-related domains, encompassing both unimodal text and multimodal inputs. The data sources included publicly available datasets as well as annotated in-house data. Specifically, the GRPO dataset integrated the following: We-Math2.0 [186], MathV360K [197], KVQA [192], ChartQA [151], ThinkLite [231], STEM-500k [195], deepscaler [148] and Statics-5k. Among these, Stats-5K is an in-house annotated dataset specifically designed for computational tasks involving statistical charts in English-language contexts. To enhance data quality and training efficacy, we applied post-processing procedures, including difficulty grading and content categorization.

Difficulty Grading Difficulty grading refers to performing n rollouts on these data using the thinking model after SFT training and then categorizing the difficulty levels based on the number of correct answers obtained in the n rollouts. A higher number of correct responses corresponds to a lower difficulty level.

Content Categorization This involved first identifying the model's deficient capabilities, followed by employing an LLM to select and group data instances with semantically similar content.

Through these refinement strategies, we constructed a reinforcement learning dataset predominantly composed of mathematical and STEM tasks, amounting to approximately 43.6k samples.

4 Mobile-side Application of AndesVL

Based on the AndesVL model after both pre-training and post-training, we build a **1+N LoRA** [243] on-device AI framework. This architecture comprises a foundational model and multiple scenario-specific LoRA adapters for each scenario. Based on this framework, we further perform quantization and mobile-side acceleration and release multiple on-device AI applications on OPPO AI phones.

4.1 Multi-scenario LoRA Training

During multi-scenario deployment of AndesVL models, it is imperative to balance the generalization capacity of the model with its domain-specific adaptability. To address practical requirements during application, based on AndesVL, we further designed a dedicated multi-scenario LoRA training stage, structured as in Fig. 7.

In practical deployment scenarios, task-specific fine-tuning is often required while computational resources remain limited. It is infeasible to train separate large models for each individual scenario. Therefore, based on

Task Type	Reward Criteria
OCR tasks	Text detection accuracy and structural integrity
Caption generation	Semantic relevance and linguistic conciseness
Text summarization	Content coverage and instruction alignment

Table 5: Reward signals for different specific real-world tasks

our AndesVL described above, we trained multiple LoRA models for different scenarios while keeping the base model parameters frozen. This approach only requires fine-tuning a minimal number of parameters to adapt to various application scenarios. It significantly reduces training resource consumption, while effectively prevents catastrophic forgetting and enhances the model's generalization across multiple scenarios. The LoRA training for each scenario consist of two phases: SFT and RL.

The SFT data construction for LoRA fine-tuning was more scenario-specific and customized, where for each scenario we collected and annotated high-quality, strongly relevant dedicated data; designed data labels and task definitions that closely align with actual requirements; and implemented customized training loss functions tailored to specific scenarios. For example, in image caption generation tasks, to increase entity density, we designed an entity-weighted cross-entropy loss that assigns higher loss weights to entity words (*e.g.*, colors, quantities, object categories) in captions, thereby encouraging the model to focus more on generating these critical pieces of information. The entity-weighted cross-entropy loss enhances key information generation through differential weighting:

$$\mathcal{L}_{\text{entity}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \alpha_{i,t} \log P(w_{i,t}|x_i, w_{i,<t}, \theta), \quad (3)$$

where $\alpha_{i,t}$ denotes the weighting factor ($\alpha_{i,t} > 1$ for entity tokens, $= 1$ otherwise), N represents the batch size, T_i is the sequence length of the i-th sample, and $w_{i,t}$ indicates the t-th token in the i-th sample. The total training objective that combines entity-focused and fluency-preserving terms is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{entity}} + \lambda_2 \mathcal{L}_{\text{BLEU/ROUGE}}, \quad (4)$$

where λ_1 is the weight for entity preservation, λ_2 is the fluency control coefficient maintaining grammatical quality and $\mathcal{L}_{\text{BLEU/ROUGE}}$ denotes standard metric-based loss for text quality.

The RL phase in the multi-scenario LoRA stage is also highly customized. For example, in the captioning task, in addition to ensuring the accuracy and conciseness of the generated captions, it is also necessary to achieve a higher density of entity words (*i.e.*, including useful information such as color, quantity, etc.). This enables the model to output more informative content within the same token length, thereby facilitating improvements in downstream tasks such as album search. We define the **Entity Density Reward** as

$$R_{\text{entity}} = \frac{\text{Number of entity words in caption}}{\text{Total number of words in caption}}, \quad (5)$$

the **Key Information Reward**,

$$R_{\text{info}} = \beta_1 \cdot \mathbb{I}(\text{caption contains color}) + \beta_2 \cdot \mathbb{I}(\text{caption contains number}), \quad (6)$$

where $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition is satisfied and 0 otherwise), and β_1, β_2 are weighting coefficients. So finally, the **Total Reward** is

$$R_{\text{total}} = \lambda_1 R_{\text{entity}} + \lambda_2 R_{\text{info}} + \lambda_3 R_{\text{BLEU/ROUGE}}, \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3$ are weighting coefficients and λ_3 ensures the fluency and relevance of the caption. Various detailed examples of reward signals are presented in Table 5. This mechanism ensures consistent and high-quality outputs in diverse scenarios.

The multi-scenario LoRA training phase focuses on **deep customization** and **strong adaptation**, utilizing scenario-specific data along with customized loss and reward functions to significantly enhance model precision and practical utility in targeted application scenarios.

4.2 Quantization and Deployment

We have established an end-to-end quantization optimization pipeline, comprising a QAT framework for base models and a scenario-specific Quantization-Aware LoRA Fine-Tuning (QALFT) framework. This pipeline

leverages cloud-based computational resources and engineering investments, to maximally preserve AndesVL performance on edge devices, while simultaneously enhance on-device inference efficiency through fine-grained mixed-precision quantization.

4.2.1 Quantization-Aware Training for AndesVL

Although post-training quantization (PTQ) techniques have advanced rapidly, directly deploying models to mobile devices via PTQ still incurs significant performance degradation. Moreover, the inherent unpredictability of PTQ-induced accuracy loss imposes an additional burden on algorithm validation and testing.

To address these challenges, we have developed a robust and flexible Quantization-Aware Training (QAT) framework. It supports multiple quantization configurations: weights can be quantized to 2, 3, 4, or 8 bits, and activations to 8 or 16 bits. The framework also enables fine-grained mixed-precision combinations and includes automated precision assignment strategies to maintain model accuracy while maximizing inference efficiency. Furthermore, through close collaboration with silicon vendors, we have established a deterministic mapping mechanism that directly translates static-QAT models into hardware-compatible, edge-deployable quantized representations. This approach aims to fundamentally eliminate the performance uncertainty on edge devices that arises from PTQ.

4.2.2 QALFT

QAT effectively satisfies the accuracy requirements for deploying a single base model on edge devices. However, in multi-LoRA scenarios, the activation quantization encodings of the base model must jointly account for the activation ranges introduced by all LoRA adapters. Consequently, any update to a LoRA adapter necessitates re-quantizing both the base model and all associated LoRAs to maintain optimal performance across diverse use cases—an impractical requirement for edge deployment.

To overcome this limitation, we co-developed the Quantization-Aware LoRA Fine-Tuning (QALFT) framework in collaboration with MediaTek. QALFT begins by applying PTQ to a QAT-pretrained base model and permanently freezing its quantization encodings. Subsequent LoRA weights are then trained on top of this fixed, quantized backbone—following a paradigm analogous to QLoRA [49]. This design enables independent updates of scenario-specific LoRA modules without re-quantizing the base model, thereby eliminating quantization-induced performance degradation during deployment and significantly streamlining the iteration cycle for task-specific algorithms.

As illustrated in Fig. 8, QALFT employs a layered architectural design. Its core principle is the complete decoupling of three essential components: the floating-point base model, training data, and the QALFT trainer. This decoupling ensures that the training logic remains agnostic to and isolated from vendor-specific hardware infrastructure, thereby facilitating seamless and efficient deployment on MediaTek platforms.

4.3 Mobile-side Acceleration with Cache Eviction

The key-value cache (KV cache) plays a crucial role in enhancing the inference performance of LLMs. Nevertheless, as the input sequence length expands, the size of the KV cache also grows proportionally—this not only imposes significant pressure on memory resources but also undermines time efficiency. This issue is particularly pronounced for edge devices such as mobile devices: on these platforms, both performing inference on long text inputs and storing massive volumes of KV cache data are highly inefficient and impractical.

Surprisingly, the long text prompt itself is extremely sparse, which means that only a small number of tokens contribute most of the value. Therefore, we can perform an eviction operation on the KV cache.

Classic cache eviction solutions include streamingLLM [236], H2O [266], and snapKV [120]. The essence of these solutions lies in retaining the latest token and the previous key token based on observations, attention weights, etc. We designed a new solution called OKV that outperforms these solutions while maintaining the same compression rate and supports context lengths up to 128K.

4.4 Mobile-Side Acceleration with Speculative Decoding

Due to the sequential nature of auto-regressive LLMs, the decoding phase is expensive and slow. Speculative decoding has been proven to be an effective solution to this problem: EAGLE-2 [121] performs auto-regression at the feature level, reusing top-layer features from the target model in drafting to achieve better results than vanilla speculative decoding; HASS [261] proposes a training-time-testing method, which further improves
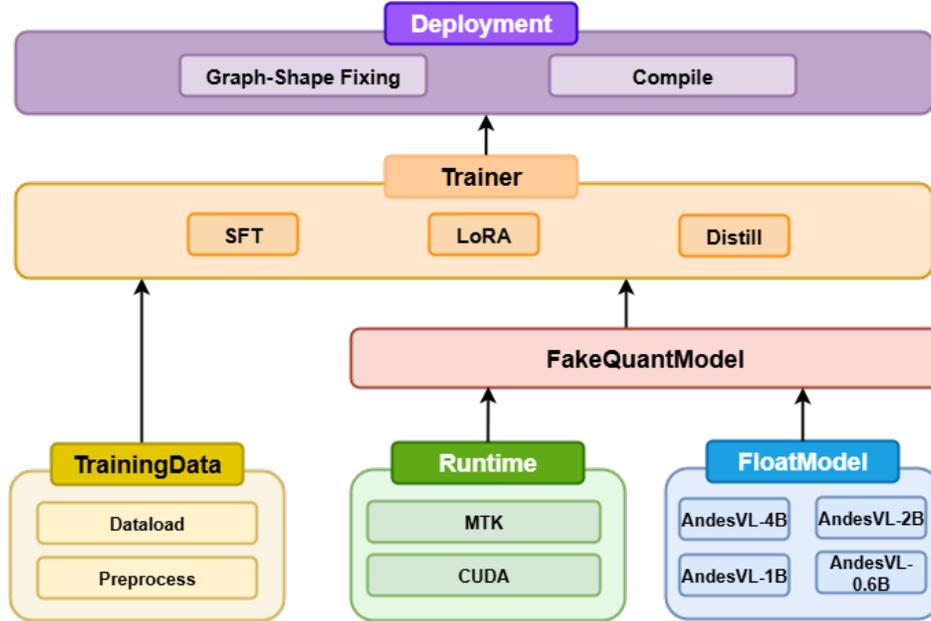
Figure 8: QALFT framework.

accept length by reusing features of the draft model in the training phase to maintain consistency in the inference phase.

Based on the characteristics of mobile-side devices, we made some customizations and adaptations to existing Eagle-like methods reusing top layer features, to fully utilize the storage and computation resources on devices. We experimentally evaluate the AndesVL model with speculative decoding on multiple specific tasks. The results show that our customized method achieves a **block efficiency (BE)** of up to **7.9**. Additionally, when combined with hardware compression and LLM sparsification, it obtains a **6.7x** peak speedup ratio over the baseline.

5 Experiments

In this section, we provide experimental results to demonstrate the comprehensive capabilities of AndesVL. The general multimodal capabilities of AndesVL are compared with those of SOTA MLLMs using widely adopted multimodal benchmarks. Following this, the domain-specific performances of AndesVL are detailed, covering text-rich image understanding (including OCR, chart, and document comprehension), reasoning and math, multi-image comprehension, general VQA, hallucination mitigation, multilingual capability, and GUI-related tasks. Most of the benchmarks are tested using VLMEvalKit [54].

5.1 Overall Performance

5.1.1 Benchmarks

We evaluated the performance of AndesVL in comparison to several advanced multimodal models of comparable model size, including Qwen2-VL [227], Qwen2.5-VL [16], InternVL3 [274], InternVL3.5 [229], Gemma3 [211], BlueLM-2.5-3B [238], Phi-3.5-Vision [2], Phi-4-Multimodal [3], MiniCPM-V [245], R-4B [85], Qianfan-VL [51], Ovis2 [145], SAIL-VL-1.5-2B [52], SAIL-VL2-2B [247], and SmolVLM2 [150]. For fair comparisons, these models are grouped by their parameter sizes in the following evaluations.

The diverse multimodal capabilities of the proposed AndesVL are assessed using 32 commonly adopted benchmarks, covering various multimodal tasks across six domains: reasoning and math, text-rich, multi-image, general VQA, hallucination and multilingual capability. Detailed benchmark information is presented in subsequent subsections.

Model	Text-rich	Reasoning &Math	Multi-image	General VQA	Hallucination	Multilingual	Overall
Phi-3.5-Vision [2]	65.8	24.0	36.8	55.4	63.9	47.0	48.8
Phi-4-Multimodal [3]	81.2	33.2	42.4	64.2	66.0	54.3	56.9
Gemma3-4B [211]	61.0	28.9	38.1	57.8	62.1	52.4	50.0
Qwen2.5-VL-3B [16]	82.1	32.1	44.8	62.2	66.9	58.9	57.8
Ovis2-4B [145]	85.1	34.1	45.6	71.3	73.2	63.4	62.1
MiniCPM-V-4-4B [245]	82.4	33.8	59.1	70.9	69.3	55.4	61.8
R-4B-RL [85]	-	57.1	-	-	-	-	-
InternVL3.5-4B [229]	82.6	56.9	62.3	72.8	69.6	62.1	67.7
AndesVL-4B-Instruct	85.3	42.1	64.5	72.7	73.0	64.6	67.0
AndesVL-4B-Thinking	**86.0**	**58.3**	**67.8**	**73.8**	**74.8**	**64.9**	**70.9**
Qwen2-VL-2B [227]	75.8	23.1	49.7	60.5	66.1	52.8	54.7
MiniCPM-V-2-2B [245]	60.5	20.8	50.5	53.5	63.6	40.2	48.2
SAIL-VL-1.5-2B [52]	82.1	29.5	55.6	68.4	70.5	57.3	60.6
Ovis2-2B [145]	81.8	29.5	59.7	67.2	70.3	58.0	61.1
InternVL3-2B [274]	78.3	31.6	54.3	**69.4**	67.9	57.4	59.8
InternVL3.5-2B [229]	79.6	**49.9**	56.6	68.3	70.5	58.0	63.8
AndesVL-2B-Instruct	**82.4**	33.8	56.5	66.1	70.9	**60.3**	61.7
AndesVL-2B-Thinking	81.3	45.7	**59.8**	68.3	**71.8**	59.4	**64.4**
Ovis2-1B [145]	77.3	24.3	53.0	59.5	65.4	52.4	55.3
InternVL3-1B [274]	71.2	23.8	47.8	61.2	65.4	47.9	52.9
InternVL3.5-1B [229]	73.5	32.8	52.2	59.9	65.4	49.1	55.5
AndesVL-1B-Instruct	76.7	27.4	52.1	60.7	67.0	53.3	56.2
AndesVL-1B-Thinking	**77.4**	**35.8**	**54.3**	**63.4**	**67.4**	**54.1**	**58.8**
SmolVLM2-0.5B [150]	55.5	18.4	42.0	43.6	54.4	26.1	40.0
AndesVL-0.6B-Instruct	**73.5**	26.0	51.5	55.3	65.7	**51.0**	53.8
AndesVL-0.6B-Thinking	73.3	**29.4**	**53.1**	**57.1**	**65.9**	49.7	**54.7**

Table 6: The overall comparison of AndesVL with existing MLLMs on 32 benchmarks, which are grouped into 6 domains. The best results are marked in **bold**.

5.1.2 Evaluation Results

Table 6 summarizes the overall performance of various existing MLLMs across 32 benchmarks spanning six different categories: Text-rich, Reasoning & Math, Multi-image, General VQA, Hallucination, and Multilingual. We compute the average scores, drawn from the models' original papers or the OpenCompass leaderboard [39], to represent their capabilities across specific domains and overall.

Our proposed AndesVL series substantially outperform existing models of similar sizes on multiple test sets, across all evaluated scales. These statistics highlight the effectiveness of our advanced training strategies and the quality of the training corpus utilized.

Specifically, across 32 benchmarks, the AndesVL-4B-Thinking model achieves an overall score of 70.9, outperforming the second-best model, InternVL3.5-4B [229], by a margin of 3.2 points. Across every multimodal task category, the AndesVL-4B-Thinking model secures a significant margin of 0.9 to 5.5 points, underscoring its universal superiority in diverse multimodal scenarios. AndesVL-4B-instruct also demonstrates remarkably strong performance across multiple vertical domains, especially on multilingual and multi-image tasks.

At the 2B scale, the AndesVL-2B-Thinking model achieves the highest overall score of 64.4. It exhibits a clear advantage in multi-image understanding and hallucination mitigation over existing models, even surpassing some 4B-scale models.

For even more compact and lightweight models, our proposed 1B and 0.6B models command a decisive advantage across all metrics, with their Thinking and Instruct versions occupying the top spots and suppressing other leading models in the literature. Notably, our 0.6B variants, the AndesVL-0.6B series, achieve a performance even comparable to existing 1B models, such as InternVL3.5-1B.

Above results underscore the model's proficiency in addressing a wide range of real-world tasks that require multimodal perception, understanding, knowledge, and reasoning. Moreover, the diversity in our models' sizes, combined with their strong performance, enables them suitable for deployment in a wide range of mobile scenarios, including those with highly limited computing resources.

Model	MMMU (val)	MMMU Pro	MathVista (mini)	MathVision	MathVerse (vision-only)	DynaMath (worst case)	WeMath	LogicVista	Overall
Qwen2.5-VL-3B [16]	51.2	30.9*	60.9*	18.8*	25.7*	11.0*	23.2*	35.1*	32.1
BlueLM-2.5-3B [238]	47.5	-	70.8	28.5	-	-	-	-	-
BlueLM-2.5-3B-thinking [238]	51.3	-	78.4	47.7	-	-	-	-	-
Qianfan-VL-3B [51]	46.4	-	-	-	-	-	-	-	-
Gemma3-4B [211]	47.3	24.5*	46.3	23.1*	23.2*	11.0*	23.1*	32.7*	28.9
Phi-3.5-Vision-4B [2]	44.6	23.6*	43.3	14.9*	19.7*	9.8*	11.2*	25.1*	24.0
Phi-4-Multimodal [3]	55.1	23.6*	62.4	19.7*	22.0*	13.0*	19.2*	35.6*	33.2
Ovis2-4B [145]	49.0	28.0*	69.6	21.1*	39.0*	12.6*	18.0*	35.1*	34.1
MiniCPM-V-4-4B [245]	51.2	33.4*	66.9	20.7	22.0*	14.2*	30.0*	32.0*	33.8
R-4B-RL [85]	68.1	46.5	78.0	47.8	64.9	**39.5**	52.8	**59.1**	57.1
InternVL3.5-4B [229]	66.6	**53.5***	77.1	**54.4**	61.7	35.7	50.1	56.4	56.9
AndesVL-4B-Instruct	58.0	37.6	73.3	27.1	34.3	21.2	33.7	41.6	40.9
AndesVL-4B-Thinking	**66.9**	51.4	**79.4**	51.0	**66.9**	35.5	**57.4**	57.7	**58.3**
Qwen2-VL-2B [227]	42.2	19.9*	48.0	17.3*	16.8*	4.0*	11.3*	25.5*	23.1
MiniCPM-V-2B [245]	38.2	20.8*	39.8	15.0*	16.8*	2.8*	6.3*	26.6*	20.8
SAIL-VL-1.5-2B [52]	46.7*	23.6*	67.3	18.0*	21.7*	8.6*	16.5*	33.8*	29.5
SAIL-VL2-2B [247]	47.7	-	71.1	23.4	-	10.2	22.7	36.2	-
SAIL-VL2-2B-Thinking [247]	-	-	68.5	27.5	-	20.2	38.8	47.0	-
Ovis2-2B [145]	45.6	23.8*	64.1	17.6*	30.7*	10.0*	10.4*	33.6*	29.5
InternVL3-2B [274]	43.2	26.9*	57.0	19.5*	21.8*	14.6	22.4	47.7	31.6
InternVL3.5-2B [229]	**59.0**	**42.6***	71.8	**42.8**	53.4	**31.5**	48.5	**49.4**	**49.9**
AndesVL-2B-Instruct	46.1	30.7	64.9	22.4	26.8	15.2	30.3	34.0	33.8
AndesVL-2B-Thinking	52.1	37.3	**73.3**	35.2	**54.8**	27.5	41.1	44.3	45.7
Ovis2-1B [145]	36.1	20.9*	59.4	16.0	23.9*	2.8*	9.6*	26.0*	24.3
InternVL3-1B [274]	43.4	20.1*	45.8	18.8	18.7	5.8	13.4	29.8	24.5
InternVL3.5-1B [229]	**44.2**	25.7*	59.3	**27.3**	37.8	**17.2**	21.5	29.3	32.8
AndesVL-1B-Instruct	43.1	24.4	53.8	18.1	18.5	10.2	21.0	30.2	27.4
AndesVL-1B-Thinking	44.0	**27.9**	66.4	23.5	**45.1**	11.6	**33.9**	**34.0**	**35.8**
SmolVLM2-0.5B [150]	34.1	14.7*	37.5	13.2*	14.0*	3.2*	7.2*	23.5*	18.4*
AndesVL-0.6B-Instruct	40.7	**24.9**	51.8	**19.2**	18.7	6.4	16.2	**29.8**	26.0
AndesVL-0.6B-Thinking	**43.3**	24.3	**54.9**	**19.2**	**34.0**	**7.0**	**22.8**	29.3	**29.4**

Table 7: Comparison of reasoning and mathematical performance. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or the OpenCompass leaderboard.

5.2 Reasoning and Math

5.2.1 Benchmarks

To evaluate AndesVL's multimodal reasoning and mathematical capabilities, we extensively evaluate the model on various benchmarks for mathematical reasoning as follows:

MMMU [256]: MMMU evaluates MLLMs on college-level tasks across six disciplines, testing expert-level reasoning and advanced perception in specific fields. The accuracy results achieved from the model's direct answer on its validation set are recorded.

MMMU Pro [256]: MMMU Pro evaluates the multimodal understanding and reasoning capabilities of the model from a wide range of academic disciplines. It is the upgraded version of the MMMU benchmark. The overall accuracy score of the direct answer is reported.

MathVista [139]: MathVista evaluates the mathematical reasoning ability, such as algebra, geometry, and statistics, of MLLMs with visual contexts. The accuracy scores on the testmini set are recorded.

MathVision [226]: MathVision is made up of math problems with visual contexts. The problems are sourced from real math competitions. The results on full set of the benchmark are reported.

MathVerse [263]: MathVerse evaluates a model's capability of solving visual diagram-based math problems. The performance of its vision-only set is reported.

DynaMath [276]: DynaMath consists of variant-generated questions for a seed question under various conditions. The worst-case accuracy is reported to reflect the model's reliability of MLLMs' reasoning abilities.

WeMath [186]: WeMath decomposes composite visual math problems into sub-problems to hierarchically assess inherent issues in MLLMs' reasoning, covering 67 knowledge concepts across 5 levels of granularity.

LogicVista [237]: LogicVista evaluates models across five logical reasoning tasks: spatial, deductive, inductive, numeric, and mechanical reasoning, leveraging a diverse dataset of visual multiple-choice questions.

5.2.2 Evaluation Results

As shown in Table 7, AndesVL-4B-Thinking achieves the highest overall score of 58.3 across various math and reasoning benchmarks among exiting models. Notably, AndesVL-4B exhibits considerable superiority over advanced models on the MathVista, MathVerse and WeMath benchmarks. With an overall score of 45.7, the AndesVL-2B-Thinking model ranks second, performing very close to the top score of 49.9 in literature. Furthermore, the AndesVL's 1B and 0.6B Thinking models deliver dominant performance within their respective size groups, achieving top ranks not only overall but also on most individual benchmarks.

These improvements over exiting models highlight the efficacy of our training strategy. Our approach enhances the visual-text joint reasoning ability by leveraging a large corpus of refined, long Chain-of-Thought (CoT) multimodal data in pre-training and through an intricately designed reinforcement learning process in post-training.

Collectively, these findings underscore AndesVLs' comprehensive capabilities in addressing multimodal mathematical problems, as well as reasoning challenges in scientific, engineering, and real-world contexts.

5.3 Text-rich Image Understanding

5.3.1 Benchmarks

In order to evaluate the OCR, chart, and document understanding capabilities of AndesVL, we perform assessments over a variety of text-rich datasets, including the following seven benchmarks.

AI2D [97]: AI2D consists of visual questions based on elementary school science diagrams. The results of its test set with and without mask settings are reported.

OCRBench [134]: OCRBench evaluates the overall OCR capabilities of MLLMs across five tasks: text recognition, scene text VQA, document VQA, key information extraction, and handwritten math expression recognition.

ChartQA [151]: ChartQA requires a model to comprehend charts and graphs visually. The average relaxed accuracy across both human and augmented test sets in ChartQA is taken as the evaluation metric.

TextVQA [200]: TextVQA evaluates a model's capability on visual reasoning with visual context from texts within images. The accuracy in the validation set is reported.

DocVQA [153]: DocVQA requires a model to read, comprehend, and retrieve texts within document images to answer related questions. Performance is reported on the test set using the ANLS text similarity metric.

InfoVQA [152]: InfoVQA consists of various complex infographics that combine text, graphics, and visual elements in creative layouts. The ANLS similarity score computed on the test set is reported.

SEEDBench-2-Plus [109]: SEEDBench-2-Plus evaluates a model's multimodal capability on text-rich visual tasks across charts, maps, and webs. The average accuracy on this dataset is reported.

5.3.2 Evaluation Results

Table 8 shows a detailed comparison of AndesVL with several existing promising MLLMs on OCR-related benchmarks. AndesVL demonstrates superior or competitive performance to them.

Among existing models, our AndesVL-4B-Thinking model claims the top rank with an overall score of 86.0, and it achieves the top results on four of eight benchmarks. Meanwhile, the AndesVL-4B-Instruct model also delivers strong and comparable performance on text-rich tasks with a score of 85.3. Most notably, on ChartQA, the AndesVL-4B-Thinking model scores 90.4, exceeding the previous best, InternVL3.5-4B (86.0), by 4.4 points. A similar marked advantage is observed on DocVQA. AndesVL's success on the ChartQA and DocVQ benchmarks, featuring long-text images and complex questions, directly illustrates its ability to not only recognize long texts accurately but also apply advanced reasoning to solve challenging, contextual problems effectively.

Moreover, the advantages over existing models on text-rich tasks persist down to our smaller-scale versions. To be specific, our proposed AndesVL-2B-Instruct, AndesVL-1B-Thinking, and AndesVL-0.6B-Instruct models all rank first in their respective model-size groups, with overall scores of 82.4, 77.4, and 73.5, respectively.

These outcomes demonstrate the effectiveness of our models' multimodal recognition and comprehension capabilities across a variety of text-rich tasks.

Model	AI2D (w M)	AI2D (w/o M)	ChartQA (test)	TextVQA (val)	DocVQA (test)	InfoVQA (test)	OCR Bench	SEED 2-Plus	Overall
Qwen2.5-VL-3B [16]	81.4	91.3*	84.2*	79.2*	93.0*	77.0*	82.6*	68.2*	82.1
BlueLM-2.5-3B [238]	83.0	-	-	-	-	-	82.6	-	-
BlueLM-2.5-3B-thinking [238]	82.6	-	-	-	-	-	84.0	-	-
Qianfan-VL-3B [51]	81.4	-	81.8	80.1	-	-	83.1	67.6	-
Gemma3-4B [211]	70.7	86.3*	33.7	57.7	70.0*	43.0*	66.0	60.7	61.0
Phi-3.5-Vision-4B [2]	77.8	87.6*	70.0*	65.1*	69.0*	35.0*	59.9	62.2	65.8
Phi-4-Multimodal [3]	82.3	91.7*	81.4	75.6	93.2	72.7	84.4	68.5	81.2
Ovis2-4B [145]	85.7	94.2*	84.2*	83.2*	94.0*	79.0*	91.1	69.3	85.1
MiniCPM-V-4-4B [245]	82.9	93.0*	84.4	80.8	93.0*	69.0*	89.4	67.0	82.4
R-4B-RL [85]	86.2	-	-	-	91.0	-	83.6	-	-
InternVL3.5-4B [229]	82.6	92.3	86.0	77.9	92.4	78.0*	82.2	69.4	82.6
AndesVL-4B-Instruct	84.5	**94.6**	87.8	81.6	**96.0**	81.0	86.1	70.9	85.3
AndesVL-4B-Thinking	84.9	94.1	**90.4**	82.1	95.4	**81.9**	87.0	**72.0**	**86.0**
Qwen2-VL-2B [227]	74.7	84.1*	72.5*	79.5*	90.0*	65.0*	79.7	61.2	75.8
MiniCPM-V-2B [245]	62.9	68.8*	55.6	73.2	71.0*	40.0*	60.5	51.9	60.5
SAIL-VL-1.5-2B [52]	**83.7**	**92.4***	78.4*	**82.0**	92.0*	72.0*	88.5	68.0*	82.1
SAIL-VL2-2B [247]	83.0	-	-	-	93.1	-	**89.5**	-	-
Ovis2-2B [145]	82.7	91.8*	81.3*	80.0*	92.0*	72.0*	87.3	67.4	81.8
InternVL3-2B [274]	78.7	87.4	80.2	77.0	88.0*	67.0*	83.5	64.6	78.3
InternVL3.5-2B [229]	78.8	89.1	80.7	76.5	89.4	70.8	83.6	68.0	79.6
AndesVL-2B-Instruct	80.1	89.9	**87.4**	79.9	94.2	74.2	84.6	**68.8**	**82.4**
AndesVL-2B-Thinking	77.8	89.3	86.6	80.0	93.9	72.9	82.9	67.1	81.3
Ovis2-1B [145]	**76.4**	85.3*	74.9*	**78.4***	89.0*	64.0*	**89.0**	61.4	77.3
InternVL3-1B [274]	69.4	78.3	75.3	74.1	81.9	53.7	79.0	58.2	71.2
InternVL3.5-1B [229]	69.3	81.8	77.7	71.5	85.6	60.5	79.5	62.3	73.5
AndesVL-1B-Instruct	71.5	83.8	80.4	77.0	**91.5**	65.3	78.9	64.8	76.7
AndesVL-1B-Thinking	74.4	**86.1**	**82.3**	76.2	91.4	**65.8**	77.7	**65.5**	**77.4**
SmolVLM2-0.5B [150]	57.3	59.5*	59.6	60.3	70.0*	29.0*	60.9	47.7	55.5
AndesVL-0.6B-Instruct	68.4	82.1	**78.2**	69.7	**89.5**	63.4	72.2	**64.3**	**73.5**
AndesVL-0.6B-Thinking	**68.8**	**82.9**	78.2	68.9	88.8	61.0	**73.8**	64.0	73.3

Table 8: Comparison of OCR, chart, and document understanding performance. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or the OpenCompass leaderboard.

5.4 Multi-Image Understanding

5.4.1 Benchmarks

To evaluate AndesVL's capabilities in perception and understanding of multi-image relation, we conducted assessments on various multi-image benchmarks.

BLINK [59]: BLINK contains visual questions on multiple images from 14 computer vision tasks. Over half of the questions involve multiple images. The accuracy result on the validation set is reported.

MMT-Bench [248]: MMT-Bench consists of multimodal tasks across recognition, reasoning, and planning, with many sub-tasks requiring multi-image understanding. The accuracy metric in the validation set is reported.

MuirBench [224]: MuirBench evaluates MLLMs' capabilities in multi-image understanding on 12 tasks and 10 types of multi-image relations. The accuracy score is reported.

Q-Bench [232]: Q-Bench assesses the abilities of MLLMs in low-level visual perception and understanding. The accuracy metric in the validation set is reported.

5.4.2 Evaluation Results

The detailed results presented in Table 9 indicate that AndesVL-4B-Thinking achieves superior outcomes across various multi-image benchmarks, culminating in a top overall score of 67.8, outperforming the previous best (InternVL3.5-4B, 62.3) by a margin of 5.5 points. It also scores the top on three out of four individual multi-image benchmarks. Moreover, as the model scale decreases, the models persist to demonstrate highly competitive accuracy, with the 0.6B variant attaining a score of 53.1.

This superiority suggests that the advanced pre-training strategies and enhanced training datasets employed in AndesVL significantly enhance its ability to capture and reason about inter-image relationships by concurrently understanding and analyzing the relationships among multiple images.

Model	BLINK	Q-Bench1 val	MMT val	MuirBench	Overall
Qwen2.5-VL-3B [16]	49.3*	30.3*	61.0*	38.7*	44.8
Qianfan-VL-3B [51]	50.0	73.5	62.2	-	-
Gemma3-4B [211]	43.9	14.2*	53.9	40.5*	38.1
Phi-3.5-Vision-4B [2]	58.3	3.7*	61.6	23.6*	36.8
Phi-4-Multimodal [3]	**61.3**	10.2*	60.4	37.6*	42.4
Ovis2-4B [145]	53.0	20.9*	65.2	43.4*	45.6
MiniCPM-V-4-4B [245]	54.0	76.5*	59.7	46.1	59.1
R-4B-RL [85]	56.3	-	-	-	-
InternVL3.5-4B [229]	58.1	73.8*	64.3	53.1	62.3
AndesVL-4B-Instruct	58.2	**77.8**	66.4	55.5	64.5
AndesVL-4B-Thinking	58.4	77.5	**66.5**	**68.8**	**67.8**
Qwen2-VL-2B [227]	45.2	72.8*	55.0	25.9*	49.7
MiniCPM-V-2B [245]	41.2	67.0*	53.5	40.1*	50.5
SAIL-VL-1.5-2B [52]	45.4*	75.9*	**61.7***	39.5*	55.6
Ovis2-2B [145]	**65.7**	**76.2***	55.0	41.9*	59.7
InternVL3-2B [274]	50.3	68.4*	59.5	38.8*	54.3
InternVL3.5-2B [229]	51.3	72.4*	58.5	44.0	56.6
AndesVL-2B-Instruct	48.1	73.8	58.8	45.5	56.5
AndesVL-2B-Thinking	48.6	74.6	58.5	**57.4**	**59.8**
Ovis2-1B [145]	44.0	71.3	54.7*	42.0*	53.0
InternVL3-1B [274]	42.9	63.4*	53.5*	31.2	47.8
InternVL3.5-1B [229]	44.0	68.5*	54.5	41.7	52.2
AndesVL-1B-Instruct	**44.7**	70.4	55.2	38.0	52.1
AndesVL-1B-Thinking	**44.7**	**72.4**	**57.0**	**43.2**	**54.3**
SmolVLM2-0.5B [150]	40.7	56.5	44.7	26.2*	42.0
AndesVL-0.6B-Instruct	**46.6**	69.2	52.0	38.0	51.5
AndesVL-0.6B-Thinking	46.0	**71.7**	**52.7**	**42.0**	**53.1**

Table 9: Comparison of multi-image understanding performance. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or OpenCompass leaderboard.

5.5 General VQA

5.5.1 Benchmarks

We evaluate AndesVL's general visual question-answer ability through a range of benchmarks, including real-world understanding and comprehensive benchmarks. These evaluations test the model's capabilities under complex realistic tasks and various comprehensive tasks. The following six benchmarks are included in these evaluations.

MME [58]: MME evaluates models' perception and cognitive abilities across 14 sub-tasks. The overall summarization score across all tasks is reported. Notice that the MME score is divided by 28 to calcute the overall average score.

MMBench v1.1 [132]: MMBench v1.1 evaluates the multimodal understanding capability of MLLMs. It consists of multimodal questions over 20 dimensions and supports English and Chinese versions. The average performance scores on both test sets are reported.

MMVet [252]: MMVet evaluates six core competencies for MLLMs: recognition, knowledge, spatial awareness, language generation, OCR, and mathematics, across 16 integrated tasks.

MMStar [24]: MMStar evaluates the multimodal capabilities of MLLMs, focusing on advanced perception, reasoning, math, and science&technology for visual and language understanding.

RealWorldQA [40]: RealWorldQA evaluates the spatial understanding capabilities of MLLMs under various real-world scenarios.

R-Bench [111]: R-Bench focuses on evaluating the robustness of MLLMs to distortion in the real world, which covers 33 distortion dimensions. The accuracy on the distortion set is reported.

Model	MME_sum	MMBench v1.1	MMVet	MMStar	RealWorldQA	R-Bench (dis)	Overall
Qwen2.5-VL-3B [16]	2,181*	51.2	60.0	56.3	66.3*	61.8*	62.2
BlueLM-2.5-3B [238]	-	82.1	66.7	64.5	-	-	-
BlueLM-2.5-3B-thinking [238]	-	78.3	65.1	66.3	-	-	-
Qianfan-VL-3B [51]	-	-	48.2	57.9	65.8	-	-
Gemma3-4B [211]	1,744	66.4	57.8	47.9	55.6	56.6*	57.8
Phi-3.5-Vision-4B [2]	1,838	67.4	43.2	47.5	53.6	55.4*	55.4
Phi-4-Multimodal [3]	1,962	77.2	51.9	58.9	64.1	62.8*	64.2
Ovis2-4B [145]	2,162	81.4	65.5	61.9	71.1*	70.5*	71.3
MiniCPM-V-4-4B [245]	2,298	79.7	68.0	62.8	68.5	64.7*	71.0
R-4B-RL [85]	-	**84.8**	73.1	**81.9**	69.1	-	-
InternVL3.5-4B [229]	2,272	79.3*	**76.6**	65.0	66.3	68.7	72.8
AndesVL-4B-Instruct	2,345	81.2	61.2	66.1	72.2	**71.7**	72.7
AndesVL-4B-Thinking	**2,412**	81.7	61.9	69.9	**73.2**	69.9	**73.8**
Qwen2-VL-2B [227]	1,899	72.2*	51.5	47.5	60.7	62.8*	60.5
MiniCPM-V-2B [245]	1,808	65.8*	41.0	39.1	55.8	64.7*	53.5
SAIL-VL-1.5-2B [52]	2,063	**78.5***	61.4	62.8	67.1	66.7*	68.4
SAIL-VL2-2B [247]	2,144	-	68.7	**64.1**	-	-	-
Ovis2-2B [145]	2,005	77.0*	67.9	56.7	66.0	64.2*	67.2
InternVL3-2B [274]	2,221	78.0*	62.2	60.7	64.3	**71.4***	**69.4**
InternVL3.5-2B [229]	2,123	75.3*	**71.7**	62.7	62.0	62.4	68.3
AndesVL-2B-Instruct	2,081	77.3	52.0	60.0	**67.8**	65.3	66.1
AndesVL-2B-Thinking	**2,326**	75.4	59.5	62.7	64.8	64.2	68.3
Ovis2-1B [145]	1,720	68.4*	50.0*	52.1	63.9	61.0*	59.5
InternVL3-1B [274]	1,935	68.2*	**59.5***	51.5	58.2	60.4	61.2
InternVL3.5-1B [229]	1,910	67.6*	56.5*	51.9	57.6	57.4	59.9
AndesVL-1B-Instruct	1,938	70.9	43.5	52.5	65.0	63.4	60.7
AndesVL-1B-Thinking	**2,006**	**73.1**	48.9	**57.9**	**65.1**	**64.0**	**63.4**
SmolVLM2-0.5B [150]	1,448	41.6*	29.9	38.2	52.7*	47.4*	43.6
AndesVL-0.6B-Instruct	1,866	65.3	**39.7**	44.3	58.6	57.2	55.3
AndesVL-0.6B-Thinking	**1,925**	**66.3**	36.1	**49.7**	**61.6**	**59.8**	**57.1**

Table 10: Comparison of general VQA performance. Notice that the MME score is divided by 28 to calculate the overall score. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or the OpenCompass leaderboard.

5.5.2 Evaluation Results

As illustrated in Table 10, the AndesVL series (4B, 1B, and 0.6B) achieve the top performance in their respective groups, while the 2B model also delivers a highly competitive result. A detailed breakdown reveals that while there remains room for improvement on the MMVet benchmark, the AndesVL series demonstrates exceptionally strong and robust performance on both MME and RealWorldQA. This suggests that our model extracts robust representations and displays a strong ability to comprehend real-world scenarios, enabling it to effectively tackle complex and dynamic tasks.

5.6 Hallucination Evaluation

5.6.1 Benchmarks

We evaluate AndesVL's hallucination alleviation ability through a range of widely used hallucination benchmarks. These evaluations test the model's capabilities under visual hallucination settings. The following three benchmarks are included in these evaluations.

HallusionBench [66]: HallusionBench mainly evaluates a model's capabilities under language hallucination and visual illusion settings. The average of its three metrics—aAcc, fAcc, and qAcc—is taken as the reported performance score.

CRPE [230]: CRPE quantitatively evaluates the object recognition and relation comprehension ability of MLLMs. The accuracy on the CRPE Relation subset is reported.

POPE [119]: POPE evaluates object hallucination tendencies in MLLMs. The overall average score is reported.

Model	Hallucination	CRPE (relation)	POPE (avg)	Overall
Qwen2.5-VL-3B [16]	46.6	64.9*	89.3*	66.9
BlueLM-2.5-3B [238]	53.7	-	-	-
BlueLM-2.5-3B-thinking [238]	57.3	-	-	-
Qianfan-VL-3B [51]	-	-	85.1	-
Gemma3-4B [211]	40.8	61.0*	84.6	62.1
Phi-3.5-Vision-4B [2]	40.5	68.5*	82.8	63.9
Phi-4-Multimodal [3]	40.5	72.0*	85.6	66.0
Ovis2-4B [145]	53.8	**77.0***	88.7	73.2
MiniCPM-V-4-4B [245]	50.8	74.6*	82.4	69.3
R-4B-RL [85]	58.9	-	-	-
InternVL3.5-4B [229]	44.8	75.0	88.9	69.6
AndesVL-4B-Instruct	54.7	75.8	88.5	73.0
AndesVL-4B-Thinking	**59.2**	75.5	**89.8**	**74.8**
Qwen2-VL-2B [227]	42.4	68.5*	87.3	66.1
MiniCPM-V-2B [245]	36.1	68.5*	86.3	63.6
SAIL-VL-1.5-2B [52]	49.8	73.9*	87.7*	70.5
SAIL-VL2-2B [247]	51.7	75.2	-	-
Ovis2-2B [145]	50.2	73.0*	87.8	70.3
InternVL3-2B [274]	42.5	71.5	89.6	67.9
InternVL3.5-2B [229]	48.6	**75.6**	87.2	70.5
AndesVL-2B-Instruct	**51.8**	73.0	87.9	70.9
AndesVL-2B-Thinking	51.4	74.1	**89.8**	**71.8**
Ovis2-1B [145]	45.2	63.2	87.7	65.4
InternVL3-1B [274]	41.4	64.0	**90.7**	65.4
InternVL3.5-1B [229]	41.0	68.4	86.8	65.4
AndesVL-1B-Instruct	43.2	68.7	89.2	67.0
AndesVL-1B-Thinking	**45.6**	**68.8**	87.7	**67.4**
SmolVLM2-0.5B [150]	27.7	52.9*	82.7	54.4
AndesVL-0.6B-Instruct	**45.3**	67.4	84.3	65.7
AndesVL-0.6B-Thinking	42.5	**68.3**	**86.8**	**65.9**

Table 11: Comparison of hallucination alleviation performance. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or the OpenCompass leaderboard.

5.6.2 Evaluation Results

As illustrated in Table 11, the AndesVL series achieve exceptionally high scores overall: the 4B, 2B, 1B and 0.6B models attaining 74.8, 71.8, 67.4, and 65.9 points, respectively. It maintains a substantial lead over other models of comparable size by a margin of 1.5 to 11.5 points - a lead that becomes even more pronounced with smaller model scales. This finding demonstrates that our architecture delivers superior hallucination alleviation capabilities while maintaining high accuracy, a key strength that persists even in the smallest 0.6B variants.

5.7 Multimodal Multilingual Understanding

5.7.1 Benchmarks

We evaluate AndesVL's multilingual understanding capabilities through the following three benchmarks:

MMMB [205]: MMMB assesses multilingual capabilities of MLLMs, comprising 6 languages, 15 categories, and 12,000 questions. The average score is reported.

Multilingual MMBench [205]: Multilingual MMBench extends the original MMBench [132] dataset to six languages, including English, Chinese, Portuguese, Arabic, Turkish, and Russian. The average score is reported.

MTVQA [209]: MTVQA evaluates the multilingual capability of MLLMs with human-annotated, text-rich images across 9 diverse languages. The average accuracy on the test set is reported.

Model	MMMB	Multilingual MMBench	MTVQA	Overall
Qwen2.5-VL-3B [16]	76.9*	74.9*	24.8	58.9
Qianfan-VL-3B [51]	-	-	26.5	-
Gemma3-4B [211]	69.6*	65.3*	22.1	52.4
Phi-3.5-Vision-4B [2]	61.3*	67.4*	12.3*	47.0
Phi-4-Multimodal [3]	74.5*	74.2*	14.3	54.3
Ovis2-4B [145]	79.7*	81.2*	29.4	63.4
MiniCPM-V-4-4B [245]	72.8*	70.7*	22.6*	55.4
InternVL3.5-4B [229]	80.2	76.4	29.6	62.1
AndesVL-4B-Instruct	**81.9**	80.8	**31.2***	64.6
AndesVL-4B-Thinking	81.7	**83.2**	29.9	**64.9**
Qwen2-VL-2B [227]	71.3*	66.3*	20.8	52.8
MiniCPM-V-2B [245]	60.0*	51.3*	9.3	40.2
SAIL-VL-1.5-2B [52]	76.0*	72.9*	22.9*	57.3
Ovis2-2B [145]	**76.6***	72.0*	25.6	58.0
InternVL3-2B [274]	73.6	71.9	26.7	57.4
InternVL3.5-2B [229]	74.6	70.9	28.5	58.0
AndesVL-2B-Instruct	76.5	**75.3**	**29.1**	**60.3**
AndesVL-2B-Thinking	76.5	75.0	26.7	59.4
Ovis2-1B [145]	70.8*	62.6*	23.7	52.4
InternVL3-1B [274]	63.2	58.2	22.2	47.9
InternVL3.5-1B [229]	66.0	58.5	22.9	49.1
AndesVL-1B-Instruct	**72.0**	63.0	**24.9**	53.3
AndesVL-1B-Thinking	71.3	**67.5**	23.6	**54.1**
SmolVLM2-0.5B [150]	46.8*	23.7*	7.7	26.1
AndesVL-0.6B-Instruct	**70.3**	**60.8**	**21.8**	**51.0**
AndesVL-0.6B-Thinking	69.4	58.4	21.3	49.7

Table 12: Comparison of multilingual performance. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers or the OpenCompass leaderboard.

5.7.2 Evaluation Results

As demonstrated in Table 12, both the Thinking and Instruct variants of AndesVL-4B demonstrate exceptional multilingual capabilities, achieving a leading score of 64.9, which surpasses the previous best model, Ovis2-4B [145], by 1.5 points. This advantage persists in the smaller-scale variants of AndesVL, with each one achieving top multilingual scores within their respective sub-groups.

The model's professional-grade multilingual capability provides a foundation for the cross-lingual transfer of its multimodal functions, a feature paramount for extending its global utility in mobile applications.

5.8 GUI Tasks

5.8.1 Benchmarks

In order to validate the capability of the AndesVL in terms of UI understanding, we conducted experiments on ScreenSpot [32], ScreenSpot-V2 [234], ScreenSpot-Pro [112], and our proposed AndesUI-Bench.

ScreenSpot [32]: ScreenSpot is a realistic GUI grounding benchmark that encompasses mobile, desktop, and web environments. It contains over 600 screenshots and 1200 instructions from iOS, Android, macOS, Windows, and webpages. It specifically includes both text-based elements and a variety of widgets and icons.

ScreenSpot-V2 [234]: ScreenSpot-V2 is an enhanced version of the ScreenSpot benchmark that addresses annotation errors and ambiguities in the original dataset. Specifically, it corrects spelling errors, clarifies ambiguous instructions, removes overly similar questions, and fixes mislabeled ground-truth bounding boxes. These improvements ensure a more accurate and reliable evaluation for GUI grounding tasks.

ScreenSpot-Pro [112]: ScreenSpot-Pro is a new benchmark designed to evaluate the grounding capabilities of MLLMs in high-resolution professional settings. It includes 1,581 unique instructions in high-resolution screenshots sourced from 23 applications across five industries (development, creative, CAD, scientific, and

Model	ScreenSpot	ScreenSpot_v2	ScreenSpot_Pro	Overall
Qwen2.5-VL-3B [16]	55.5*	80.9*	27.3*	54.6
OS-Atlas-4B [238]	70.1	71.9	-	-
InternVL3.5-4B [229]	83.6	85.1	18.1*	62.3
AndesVL-4B-Instruct	84.3	86.1	28.2	66.2
AndesVL-4B-Thinking	**85.2**	**87.4**	**32.5**	**68.4**
UI-TARS-2B [227]	**82.3**	**84.7**	**27.7**	**64.9**
InternVL3-2B [274]	45.1*	47.0*	1.0*	31.0
InternVL3.5-2B [229]	77.2*	79.6*	12.2*	56.3
AndesVL-2B-Instruct	74.6	76.3	20.9	57.3
AndesVL-2B-Thinking	67.2	70.2	19.6	52.4
InternVL3-1B [274]	31.3*	30.9*	0.6*	20.9
InternVL3.5-1B [229]	60.9*	61.9*	9.2*	44.0
AndesVL-1B-Instruct	71.8	73.2	**23.1**	56.0
AndesVL-1B-Thinking	**73.9**	**74.4**	20.9	**56.4**

Table 13: Comparison of UI understanding performance with other general and specific models on ScreenSpot testset. The best results are marked in **bold**. Data marked with * are from our evaluation, while others are from their original papers.

Model	Grounding	Referring	QA	Overall
Qwen2.5-VL-3B [16]	61.4	62.1	15.0	46.2
InternVL3.5-4B [229]	91.9	68.9	82.4	81.1
AndesVL-4B-Instruct	**95.1**	72.5	**82.6**	83.4
AndesVL-4B-Thinking	94.5	**73.4**	82.5	**83.5**

Table 14: Comparison of UI understanding performance on AndesUI-Bench testset. The best results are marked in **bold**. All results are from our evaluation.

office) and three operating systems (Linux, macOS, and Windows). The benchmark highlights the challenges of high-resolution displays, smaller target sizes, and complex environments in professional applications.

For the Screenspot, ScreenSpot-V2, and ScreenSpot-Pro datasets, the annotation format is bounding boxes. For each data instance, the model is required to output a specific coordinate; if the coordinate falls within the annotated bounding box, it is considered a correct prediction and contributes to the final accuracy.

AndesUI-Bench: The AndesUI-Bench was developed to evaluate the smartphone UI understanding capabilities of MLLMs. As mentioned in Appendix B, the AndesUI-Bench represents the test set of the AndesUI dataset. This dataset includes 9k referring entries, 7.6k grounding entries, 455 comprehensive description entries, and 1.2k complex question-answer entries.

5.8.2 Evaluation Results

In this study, we present a comprehensive analysis of various models' performance in UI understanding tasks. As illustrated in Tab. 13, AndesVL-4B surpasses other models of comparable size in accuracy, achieving a leading score of 68.4. While slightly trailing behind UI-TARS-2B, a specialized model in the GUI domain, the AndesVL 2B and 1B variants maintain highly competitive performance, demonstrating robust UI comprehension capabilities.

Tab. 14 shows the performance comparison between AndesVL and other leading open-source models on the AndesUI-Bench testset. AndesVL-4B delivers outstanding performance across all evaluation metrics, achieving the top score of 83.5. These results collectively demonstrate our model's substantial expertise and distinct competitive advantage in UI understanding and application.

5.9 Ablation Studies

In Table 15, we present ablation results for AndesVL-2B-Instruct-Base, AndesVL-2B-Instruct-SFT and AndesVL-2B-Instruct-MPO. We find that MPO enhances the mathematical reasoning and multimodal understanding capabilities of the Instruct model, with the MPO model achieving improvements of over 1.0 pp

Model	MMVet	MathVerse (vision-only)	RealWorldQA	OCRBench	HallusionBench	Overall
AndesVL-2B-Instruct-Base	48.3	22.8	65.1	82.4	49.3	53.6
AndesVL-2B-Instruct-SFT	51.0	25.9	66.9	83.9	49.4	55.4
AndesVL-2B-Instruct-MPO	52.0	26.8	67.8	84.6	51.8	56.6

Table 15: Comparison on several general benchmarks among AndesVL-2B-Instruct-Base, AndesVL-2B-Instruct-SFT and AndesVL-2B-Instruct-MPO.

Model	MathVista (mini)	MathVision	WeMath	MathVerse (vision-only)	MMMU	MMMU_Pro	Overall
AndesVL-2B-Thinking-Base	68.1	32.1	38.2	51.7	48.0	35.7	45.6
AndesVL-2B-Thinking-SFT	69.7	32.0	37.0	52.3	52.6	35.5	46.5
AndesVL-2B-Thinking-RL	73.3	35.2	41.1	54.8	52.1	37.3	49.0
AndesVL-4B-Thinking-Base	76.2	48.1	49.5	64.9	62.3	46.0	57.8
AndesVL-4B-Thinking-SFT	77.4	48.4	54.2	66.4	64.8	48.7	60.0
AndesVL-4B-Thinking-RL	79.4	51.0	57.4	66.9	66.9	51.4	62.2

Table 16: Comparison on several reasoning and math benchmarks among AndesVL-2B-Thinking-Base, AndesVL-2B-Thinking-SFT, AndesVL-2B-Thinking-RL, AndesVL-4B-Thinking-Base, AndesVL-4B-Thinking-SFT and AndesVL-4B-Thinking-RL.

and 0.9 pp on MMVet and MathVerse respectively. Furthermore, MPO improves the model's ability to resist hallucinations, yielding a 1.4 pp gain on HallusionBench. Notably, RealWorldQA and OCRBench show 0.9 pp and 0.7 pp improvement respectively. We thought that this is because MPO corrects errors made by the SFT model on instances it was originally capable of solving correctly.

In Table 16, we present the ablation studies for AndesVL-2B-Thinking-Base, AndesVL-2B-Thinking-SFT, AndesVL-2B-Thinking-RL, AndesVL-4B-Thinking-Base, AndesVL-4B-Thinking-SFT and AndesVL-4B-Thinking-RL. As shown, the model trained with RL exhibits significant improvements in mathematical reasoning. For instance, performance increases by about 2 pp on MathVista, MathVision and WeMath on both AndesVL-2B-Thinking and AndesVL-4B-Thinking. Moreover, the RL-enhanced model also demonstrates improved performance on complex multimodal understanding tasks, such as over 2 pp improvement on MMMU_Pro. These results indicate that RL significantly enhances the model's multimodal understanding and mathematical reasoning capabilities evenif the model has only 2B parameters.

Overall, both the Instruct and Thinking models exhibit improved performance after the SFT stage. For the Instruct model, MPO leads to significant gains in mathematical reasoning, multimodal understanding capabilities, OCR accuracy, and hallucination resistance. For the Thinking model, RL notably enhances its abilities in mathematical reasoning and complex multimodal understanding.

6 On-Device Performance

6.1 Results of Quantization-Aware Training

To evaluate the capabilities of our on-device models, we use OCR capabilities as a testbed and conduct experiments on multiple OCR-related benchmarks, including DocVQA [153], InfoVQA [152], TextVQA [200] and ChartQA [151].

As mentioned in Sec. 4.2.1, directly applying PTQ to floating-point models can significantly degrade model performance, and we introduced QAT to solve this. We compare the quantized and floating-point models based on Top-1 overlap across multiple OCR-related benchmarks. The experimental results are shown in Table 17. In Table 17, AndesVL-4B-Instruct-Base (PTQ) represents the model of AndesVL-4B-Instruct-Base post-trained on OCR data with PTQ, AndesVL-4B-Instruct-Base (QAT+PTQ) is the model of AndesVL-4B-Instruct-Base post-trained on OCR data with QAT and PTQ. The results demonstrate that QAT+PTQ can achieve 95% Top-1 overlap [115] between the quantized and floating-point models, and achieves significant improvement over PTQ alone.

Model	DocVQA (test)	InfoVQA (test)	TextVQA (val)	ChartQA (test)	Overall
AndesVL-4B-Instruct-Base (PTQ)	93.2	89.0	91.4	89.3	90.7
AndesVL-4B-Instruct-Base (QAT+PTQ)	95.4	95.2	97.5	95.1	95.8

Table 17: Top-1 overlap between AndesVL-4B-Instruct-Base (PTQ) and AndesVL-4B-Instruct-Base (QAT+PTQ) on 4 OCR benchmarks.

Eviction Ratio	Method	ROUGE-1	ROUGE-2	ROUGE-L
0% (Baseline)	AndesVL-4B-Instruct-Base	0.59	**0.33**	**0.42**
25%	SnapKV	0.55	0.30	0.39
	OKV	**0.60**	**0.33**	0.41
50%	SnapKV	0.50	0.25	0.36
	OKV	0.56	0.30	0.39

Table 18: ROUGE performance of the reproduced SnapKV and our OKV under 25% and 50% key-value cache eviction ratios on the call summary task.

6.2 Results of Cache Eviction

Our cache eviction strategy is tailored for tasks with long prompts. We use the call summary task, which is a popular and pioneering feature of OPPO AI phones and involves substantial input information redundancy, to verify its effectiveness. In this task, our proprietary OKV cache eviction algorithm results in a more than 10% improvement in Rouge-1 relative to SnapKV with 50% eviction ratios. In certain instances, it even outperformed the baseline with full KV caches. Comprehensive results presented in Table 18. All experiments are based on the same AndesVL-4B-Instruct-Base model and are carried out on one device. The baseline AndesVL-4B-Instruct-Base is supervised fine-tuned on the call summary task; SnapKV and OKV are applied to the model for inference respectively.

6.3 Results of Speculative Decoding

Our customized speculative decoding achieves significant decoding acceleration across multiple multimodal and text-only tasks. We combined it with our key breakthrough in LLM sparsification and MediaTek's hardware-aware compression, and show the final results in Table 19. In this table, the PTQ (baseline) represents the quantized version of the floating point AndesVL-4B-Instruct-Base, + Hardware-aware compression represents PTQ (baseline) with hardware compression, + Sparsification denotes PTQ (baseline) with hardware-aware compression and sparsification, and + Speculative decoding denotes PTQ (baseline) with speculative decoding, sparsification, and hardware-aware compression. The results show that we can achieve 6.7x peak decoding speedup ratio and 1.8 bits-per-weight under extreme sparsification and hardware-aware compression. Moreover, we achieved a memory reduction of up to 30.9% on the MediaTek Dimensity 9500 chips.

7 Future Directions

In the future, several promising directions can be explored to further enhance the capabilities of mobile-side MLLMs. First, designing more optimal visual encoder solutions holds great potential. By leveraging advanced

Compression & Acceleration Method	Peak Speedup	BPW
PTQ (baseline)	1.0×	3.0
+ Hardware-aware compression	1.1×	3.0
+ Sparsification	1.6×	1.8
+ Speculative decoding	6.7×	1.8

Table 19: Peak decoding speedup ratio and bits-per-weight (BPW) of AndesVL-4B-Instruct-Base under various compression and acceleration techniques on an edge device. The baseline is PTQ-only.

network architectures and novel feature extraction strategies, we aim to improve the efficiency and accuracy of visual information processing, enabling the model to better understand complex visual content on resource-constrained mobile-side devices.

Second, developing superior post-training schemes is crucial. Refining the post-training process can optimize the model performance in handling various multimodal tasks, reduce hallucinations, and enhance the consistency and reliability of generated outputs. This may involve exploring new types of training data, adjusting training objectives, and optimizing training algorithms to make the model more adaptable to real-world scenarios.

Third, implementing effective distillation schemes between large and small models can significantly improve the performance-to-resource ratio of mobile-side models. By transferring knowledge from large, high-performance cloud-based models to smaller mobile-side counterparts, we can boost the capabilities of the latter while maintaining low computational costs and memory requirements.

Finally, the development of a unified mobile-side model integrating text, image, and speech modalities (a three-mode integrated model) represents an exciting frontier. Such a model would enable seamless interaction with users across multiple modalities, providing more natural and intelligent user experiences. This will require in-depth research on multimodal fusion techniques, cross-modal representation learning, and efficient inference algorithms to ensure the model's effectiveness and efficiency on mobile-side devices. These research directions will not only drive the progress of mobile-side MLLMs but also expand their application scope in various fields.

8 Conclusion

This paper presents AndesVL, a suite of mobile-side MLLMs with parameter sizes ranging from 0.6B to 4B. By integrating Qwen3's LLM and various visual encoders, AndesVL achieves first-tier performance on multiple open-source benchmarks and the self-developed AndesUI benchmark, especially excelling in mobile UI understanding. The proposed 1+N LoRA architecture and Quantization-Aware LoRA Fine Tuning (QALFT) framework enable efficient task adaptation and model compression. By employing our proposed OKV, meticulously designed speculative decoding techniques and compression strategies, we can achieve 1.8 bits-per-weight, 6.7x peak decoding speed ratio and up to 30.9% memory reduction when deploying AndesVL-4B on MediaTek Dimenisity 9500 chips. This work bridges the gap between cloud-based MLLMs and edge devices, providing a practical solution for mobile-side MLLM and paving the way for future advancements in edge AI.

References

[1] Invoice-to-json: A document understanding and information extraction dataset, 2024. 12

[2] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 16, 17, 18, 20, 21, 22, 23, 24

[3] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 16, 17, 18, 20, 21, 22, 23, 24

[4] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084, 2019. 12

[5] Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. Opencodereasoning: Advancing data distillation for competitive coding. 2025. 9

[6] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. 9, 10

[7] Anas989898. Vision ocr financial reports 10k. https://huggingface.co/datasets/Anas989898/Vision-OCR-Financial-Reports-10k, 2024. 12

[8] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com, 2024. 1, 3

[9] apoidea. fintabnet. https://huggingface.co/datasets/apoidea/fintabnet-html, 2024. 12

[10] Apple. On device llama 3.1 with core ml. `https://machinelearning.apple.com/research/core-ml-on-device-llama?utm_source=chatgpt.com`, 2024. 4, 5

[11] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024. 9, 10

[12] Chongyang Bai, Xiaoxue Zang, Ying Xu, Srinivas Sunkara, Abhinav Rastogi, Jindong Chen, et al. Uibert: Learning generic multimodal representations for ui understanding. *arXiv preprint arXiv:2107.13731*, 2021. 9

[13] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[14] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3

[15] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 11

[16] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 9, 16, 17, 18, 20, 21, 22, 23, 24, 25

[17] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024. 1

[18] CC1984. mall receipt extraction dataset. `https://huggingface.co/datasets/CC1984/mall_receipt_extraction_dataset`, 2023. 12

[19] Yuxiang Chai, Siyuan Huang, Yazhe Niu, Han Xiao, Liang Liu, Dingyu Zhang, Peng Gao, Shuai Ren, and Hongsheng Li. Amex: Android multi-annotation expo dataset for mobile gui agents. *arXiv preprint arXiv:2407.17490*, 2024. 9

[20] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 12

[21] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 12

[22] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021. 12

[23] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 12

[24] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 21

[25] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 9

[26] Qiaoling Chen, Diandian Gu, Guoteng Wang, Xun Chen, YingTong Xiong, Ting Huang, Qinghao Hu, Xin Jin, Yonggang Wen, Tianwei Zhang, et al. Internevo: Efficient long-sequence large language model training via hybrid parallelism and redundant sharding. *arXiv preprint arXiv:2401.09149*, 2024. 5

[27] Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025. 9

[28] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 3, 4

[29] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1, 3

[30] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 3

[31] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021. 12

[32] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024. 9, 24

[33] Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. Iam: a comprehensive and large-scale dataset for integrated argument mining tasks. *arXiv preprint arXiv:2203.12257*, 2022. 12

[34] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 12

[35] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *arXiv preprint arXiv:2403.13315*, 2024. 12

[36] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 2, 4

[37] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 2, 4

[38] cognitive computations. Systemchat-2.0. https://huggingface.co/datasets/cognitivecomputations/SystemChat-2.0, 2024. 12

[39] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023. 17

[40] X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model. https://x.ai/blog/grok-1.5v, 2024. 21

[41] Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. Deep visual template-free form parsing. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 134–141. IEEE, 2019. 12

[42] Deepmind. Gemini-nano. https://deepmind.google/models/gemini/nano/, 2024. 4, 5

[43] Google Deepmind. Gemini 2.0 is now available to everyone. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/, 202. 1, 3

[44] Google Deepmind. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024. 1, 3

[45] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 6

[46] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 9, 12

[47] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 845–854, 2017. 9

[48] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[49] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. 15

[50] Khang T. Doan, Bao G. Huynh, Dung T. Hoang, Thuc D. Pham, Nhat H. Pham, Quan T. M. Nguyen, Bang Q. Vo, and Suong N. Hoang. Vintern-1b: An efficient multimodal large language model for vietnamese, 2024. 12

[51] Daxiang Dong, Mingming Zheng, Dong Xu, Bairong Zhuang, Wenyu Zhang, Chunhua Luo, Haoran Wang, Zijian Zhao, Jie Li, Yuxuan Li, et al. Qianfan-vl: Domain-enhanced universal vision-language models. *arXiv preprint arXiv:2509.18189*, 2025. 16, 18, 20, 21, 22, 23, 24

[52] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. *arXiv preprint arXiv:2501.05952*, 2025. 16, 17, 18, 20, 21, 22, 23, 24

[53] Tianjiao Du, Junming Cao, Qinyue Wu, Wei Li, Beijun Shen, and Yuting Chen. Cocoqa: Question answering for coding conventions over knowledge graphs. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1086–1089. IEEE, 2019. 12

[54] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 16

[55] Abhishek Dutta, Ankush Gupta, and Andrew Zisserman. Vgg image annotator (via). https://www.robots.ox.ac.uk/~vgg/software/via/. 44

[56] Felprot75. Mmc instructed dataset. https://huggingface.co/datasets/Felprot75/MMC_Instructed_Dataset, 2024. 12

[57] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9641–9654, 2025. 6

[58] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 21

[59] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 20

[60] Fzkuji. orand-car-a. https://github.com/Fzkuji/Labeled_ORAND-CAR-A, 2024. 12

[61] Ggerganov. llama.cpp - llm inference with minimal setup and state-of-the-art performance on a wide range of hardware. https:github.com/ggerganov/llama.cpp/, 2023. 4

[62] glaiveai. glaive-function-calling. https://huggingface.co/datasets/glaiveai/glaive-function-calling, 2023. 12

[63] Othón González-Chávez, Guillermo Ruiz, Daniela Moctezuma, and Tania Ramirez-delReal. Are metrics measuring what they should? an evaluation of image captioning task metrics. *Signal Processing: Image Communication*, 120:117071, 2024. 12

[64] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 9

[65] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*, 2024. 9, 10

[66] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023. 22

[67] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. 9

[68] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 8, 11

[69] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction. In *International Conference on Document Analysis and Recognition*, pages 254–259, 2019. 12

[70] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024. 9, 10

[71] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 12

[72] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*, 2023. 9, 10

[73] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 12

[74] howard hou. Coco-text. `https://huggingface.co/datasets/howard-hou/COCO-Text`, 2024. 12

[75] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*, 2022. 12

[76] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 9, 12

[77] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024. 9

[78] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3

[79] Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J. Yang, J. H. Liu, Chenchen Zhang, Linzheng Chai, Ruifeng Yuan, Zhaoxiang Zhang, Jie Fu, Qian Liu, Ge Zhang, Zili Wang, Yuan Qi, Yinghui Xu, and Wei Chu. Opencoder: The open cookbook for top-tier code large language models. 2024. 9, 10

[80] ilhamxx. dataset receipt. `https://huggingface.co/datasets/ilhamxx/dataset_receipt`, 2024. 12

[81] "Teknium" "interstellarninja". Hermes-function-calling-dataset-v1. `https://huggingface.co/NousResearch/hermes-function-calling-v1`. 12

[82] Venkatraman Iyer, Sungho Lee, Semun Lee, Juitem Joonwoo Kim, Hyunjun Kim, and Youngjae Shin. Automated backend allocation for multi-model, on-device ai inference. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(3):1–33, 2023. 4

[83] Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Meng Jiang, and Dong Yu. Leopard: A vision language model for text-rich multi-image tasks. *arXiv preprint arXiv:2410.01744*, 2024. 9

[84] Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. Visualwebinstruct: Scaling up multimodal instruction data through web search. *arXiv preprint arXiv:2503.10582*, 2025. 9, 10

[85] Jie Jiang, Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, and Houwen Peng. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. *arXiv preprint arXiv:2508.21113*, 2025. 16, 17, 18, 20, 21, 22, 23

[86] Xiaotang Jiang, Huan Wang, Yiliu Chen, Ziqi Wu, Lichuan Wang, Bin Zou, Yafeng Yang, Zongyang Cui, Yu Cai, Tianhang Yu, et al. Mnn: A universal and efficient inference engine. *Proceedings of Machine Learning and Systems*, 2:1–13, 2020. 4

[87] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 12

[88] JourneyBench. Journeybench hallucination. `https://huggingface.co/datasets/JourneyBench/JourneyBench_Hallucination`, 2024. 12

[89] Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024. 8

[90] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018. 12

[91] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 12

[92] Raghav Kapoor, Yash Parag Butala, Melisa Russak, Jing Yu Koh, Kiran Kamble, Waseem AlShikh, and Ruslan Salakhutdinov. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pages 161–178. Springer, 2025. 9

[93] kashindra mahato. nutritional-data-poie. `https://huggingface.co/datasets/kashindra-mahato/nutritional-data-poie-1`, 2024. 12

[94] katanaml org. invoices-donut-data-v1. `https://huggingface.co/datasets/katanaml-org/invoices-donut-data-v1`, 2023. 12

[95] Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023. 12

[96] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 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, 2016. 12

[97] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, pages 235–251, 2016. 19

[98] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 10

[99] Jin-Hwa Kim, Soohyun Lim, Jaesun Park, and Hansu Cho. Korean localization of visual question answering for blind people. In *SK T-Brain-AI for Social Good Workshop at NeurIPS*, volume 2, 2019. 12

[100] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. 9, 10

[101] laion. laion-gpt4v. `https://huggingface.co/datasets/laion/gpt4v-dataset`, 2024. 12

[102] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 9

[103] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024. 12

[104] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024. 12

[105] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 9, 10

[106] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024. 12

[107] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022. 12

[108] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 9, 10

[109] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 19

[110] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022. 9

[111] Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. R-bench: Are your large multimodal model robust to real-world corruptions? *arXiv preprint arXiv:2410.05474*, 2024. 21

[112] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025. 24

[113] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023. 8

[114] Luchang Li, Sheng Qian, Jie Lu, Lunxi Yuan, Rui Wang, and Qin Xie. Transformer-lite: High-efficiency deployment of large language models on mobile phone gpus. *arXiv preprint arXiv:2403.20041*, 2024. 4

[115] Minchong Li, Feng Zhou, and Xiaohui Song. Bild: Bi-directional logits difference loss for large language model distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1168–1182, 2025. 26

[116] Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024. 9, 10

[117] Wei Li, William E Bishop, Alice Li, Christopher Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the effects of data scale on ui control agents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 9

[118] Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. Widget captioning: Generating natural language description for mobile user interface elements. *arXiv preprint arXiv:2010.04295*, 2020. 9

[119] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 22

[120] Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024. 15

[121] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-2: Faster inference of language models with dynamic draft trees. *arXiv preprint arXiv:2406.16858*, 2024. 15

[122] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*, 2023. 12

[123] Zhangheng Li, Keen You, Haotian Zhang, Di Feng, Harsh Agrawal, Xiujun Li, Mohana Prasad Sathya Moorthy, Jeff Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui 2: Mastering universal user interface understanding across platforms. *arXiv preprint arXiv:2410.18967*, 2024. 2, 4, 10

[124] LIME-DATA. Lime-data-ai2d-train. https://huggingface.co/datasets/LIME-DATA/ai2d, 2024. 12

[125] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 9, 10

[126] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

[127] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, 2023. 12

[128] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, January 2024. 5

[129] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023. 4, 5

[130] Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding, 2024. 9

[131] Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*, 2024. 12

[132] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 21, 23

[133] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. 12

[134] Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 19

[135] Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717, 2022. 9

[136] Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*, 2024. 12

[137] LooksJuicy. ruozhiba. https://huggingface.co/datasets/LooksJuicy/ruozhiba, 2024. 12

[138] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. 9, 10

[139] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 18

[140] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 12

[141] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 12

[142] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 12

[143] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 12

[144] Quanfeng Lu, Wenqi Shao, Zitao Liu, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Gui odyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. *arXiv preprint arXiv:2406.08451*, 2024. 9

[145] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 16, 17, 18, 20, 21, 22, 23, 24

[146] Xudong Lu, Yinghao Chen, Cheng Chen, Hui Tan, Boheng Chen, Yina Xie, Rui Hu, Guanxin Tan, Renshou Wu, Yan Hu, et al. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*, 2024. 2, 4

[147] Luckyjhg. Geo170k. https://huggingface.co/datasets/Luckyjhg/Geo170K, 2024. 12

[148] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca30301` 2025. Notion Blog. 13

[149] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024. 9, 10

[150] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 16, 17, 18, 20, 21, 22, 23, 24

[151] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2263–2279, 2022. 13, 19, 26

[152] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 19, 26

[153] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. 12, 19, 26

[154] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 4

[155] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 12

[156] Microsoft. Accelerated edge machine learning. `https://onnxruntime.ai/`, 2023. 4

[157] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition*, pages 947–952, 2019. 12

[158] MiXaiLL76. Textocr ocr. `https://huggingface.co/datasets/MiXaiLL76/TextOCR_OCR`, 2025. 12

[159] MMR1. Mmr1-math-rl-data-v0. `https://huggingface.co/datasets/MMR1/MMR1-Math-RL-Data-v0`, 2025. 12

[160] Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. 9

[161] mychen76. ds receipts v2 train. `https://huggingface.co/datasets/mychen76/ds_receipts_v2_train`, 2023. 12

[162] mychen76. invoices and receipts ocr v1. `https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1`, 2023. 12

[163] mychen76. invoices and receipts ocr v2. `https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v2`, 2023. 12

[164] Dhruv Nathawani, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Ameya Sunil Mahabaleshwarkar, , Jian Zhang, and Jane Polak Scowcroft. Nemotron-Post-Training-Dataset-v1, 2025. 9

[165] nimapourjafar. Lacr i2i. `https://huggingface.co/datasets/nimapourjafar/mm_LACR_I2I`, 2024. 12

[166] nimapourjafar. Ladd. `https://huggingface.co/datasets/nimapourjafar/mm_LADD`, 2024. 12

[167] nimapourjafar. mm datikz. `https://huggingface.co/datasets/nimapourjafar/mm_datikz`, 2024. 12

[168] nimapourjafar. mm diagram image to text. `https://huggingface.co/datasets/nimapourjafar/mm_diagram_image_to_text`, 2024. 12

[169] nimapourjafar. mm intergps. `https://huggingface.co/datasets/nimapourjafar/mm_intergps`, 2024. 12

[170] nimapourjafar. mm tqa. `https://huggingface.co/datasets/nimapourjafar/mm_tqa`, 2024. 12

[171] nimapourjafar. mm vqarad. `https://huggingface.co/datasets/nimapourjafar/mm_vqarad`, 2024. 12

[172] nimapourjafar. robut-wikisql. `https://huggingface.co/datasets/nimapourjafar/mm_robut_wikisql`, 2024. 12

[173] nz. arxiv-ocr. `https://huggingface.co/datasets/nz/arxiv-ocr-v0.1-sft`, 2024. 12

[174] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. *arXiv preprint arXiv:2010.09142*, 2020. 12

[175] Beijing Academy of Artificial Intelligence (BAAI). Infinity instruct. *arXiv preprint arXiv:2406.XXXX*, 2024. 9, 10

[176] open r1. Openr1-math-220k. `https://huggingface.co/datasets/open-r1/OpenR1-Math-220k`, 2025. 12

[177] OpenAI. Gpt-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2025. 1, 3, 11, 12

[178] OpenAI. Introducing chatgpt. `https://openai.com/index/chatgpt/`, 2025. 1

[179] OpenGVLab. Sharegpt-4o. `https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o`, 2005. 12

[180] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. 2019. 12

[181] pengshuai rin. multimath-300k. `https://huggingface.co/datasets/pengshuai-rin/multimath-300k`, 2024. 12

[182] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2641–2649, 2015. 9, 10

[183] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 12

[184] qgyd2021. chinese ner sft. `https://huggingface.co/datasets/qgyd2021/chinese_ner_sft`, 2023. 12

[185] qgyd2021. few shot ner sft. `https://huggingface.co/datasets/qgyd2021/few_shot_ner_sft`, 2024. 12

[186] Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning, 2025. 13, 18

[187] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[188] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Androidinthewild: A large-scale dataset for android device control. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[189] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 3

[190] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 9

[191] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 12

[192] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884, 2019. 13

[193] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 12

[194] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 9

[195] Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models. *arXiv preprint arXiv:2402.17205*, 2024. 13

[196] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 12

[197] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 13

[198] shibing624. sharegpt gpt4. `https://huggingface.co/datasets/shibing624/sharegpt_gpt4`, 2023. 12

[199] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 12

[200] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 19, 26

[201] sr5434. Codegebragpt data. `https://huggingface.co/datasets/sr5434/CodegebraGPT_data`, 2024. 12

[202] Jianlin Su. Transformer upgrade path: 4. rotary position encoding for two-dimensional positions, May 2021. 6

[203] Alane Suhr and Yoav Artzi. Nlvr2 visual bias analysis. *arXiv preprint arXiv:1909.10411*, 2019. 12

[204] Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa, 2024. 12

[205] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*, 2024. 23

[206] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 9

[207] Lynette Yihui Tan. Spot the diff. `https://huggingface.co/datasets/Lancelot53/spot-the-diff`, 2013. 12

[208] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 12

[209] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024. 12, 23

[210] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 3

[211] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. 16, 17, 18, 20, 21, 22, 23, 24

[212] MLC team. Mlc-llm - universal llm deployment engine with ml compilation. `https://github.com/mlc-ai/mlc-llm/`, 2023-2024. 4

[213] Theonewhomadethings. fsc147-controlnet. `https://huggingface.co/datasets/Theonewhomadethings/fsc147-controlnet`, 2024. 12

[214] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022. 12

[215] toghrultahirov. handwritten text ocr. `https://huggingface.co/datasets/toghrultahirov/handwritten_text_ocr`, 2024. 12

[216] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 12

[217] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 6

[218] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. 2023. 9, 12

[219] v DeepMind. Gemini 2.5 pro. `https://deepmind.google/technologies/gemini/pro/`, 2025. 1, 3

[220] vikhyatk. st-vqa. `https://huggingface.co/datasets/vikhyatk/st-vqa`, 2024. 12

[221] VLM-Perception. Hme100k-400. `https://huggingface.co/datasets/VLM-Perception/HME100k-400`, 2025. 12

[222] Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. Unimernet: A universal network for real-world mathematical expression recognition, 2024. 12

[223] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021. 9

[224] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 20

[225] Junjie Wang, Yin Zhang, Yatai Ji, Yuxiang Zhang, Chunyang Jiang, Yubo Wang, Kang Zhu, Zekun Wang, Tiezhen Wang, Wenhao Huang, et al. Pin: A knowledge-intensive dataset for paired and interleaved multimodal documents. *arXiv preprint arXiv:2406.13923*, 2024. 9, 12

[226] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*, 2024. 18

[227] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 3, 5, 6, 10, 16, 17, 18, 20, 21, 22, 23, 24, 25, 44

[228] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024. 7, 8, 12

[229] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 2, 3, 4, 16, 17, 18, 20, 21, 22, 23, 24, 25

[230] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *European Conference on Computer Vision*, pages 471–490. Springer, 2024. 22

[231] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025. 13

[232] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024. 20

[233] Qinzhuo Wu, Weikai Xu, Wei Liu, Tao Tan, Jianfeng Liu, Ang Li, Jian Luan, Bin Wang, and Shuo Shang. Mobilevlm: A vision-language model for better intra-and inter-ui understanding. *arXiv preprint arXiv:2409.14818*, 2024. 2, 4

[234] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*, 2024. 9, 24

[235] Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Junchi Yan, and Yu Qiao. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*, 2023. 12

[236] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 15

[237] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. 18

[238] Baojiao Xiong, Boheng Chen, Chengzhi Wang, Daxiong Luo, Dongsheng Xu, Dongyang Liu, Fan Yang, Fangyuan Li, Fei Teng, Feng Wang, et al. Bluelm-2.5-3b technical report. *arXiv preprint arXiv:2507.05934*, 2025. 2, 4, 16, 18, 20, 22, 23, 25

[239] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 5

[240] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 2, 3, 6

[241] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[242] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, and Fei Huang. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[243] Jianxin Yang. Longqlora: Efficient and effective method to extend context length of large language models. *arXiv preprint arXiv:2311.04879*, 2023. 13

[244] Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. Aria-ui: Visual grounding for gui instructions, 2025. 9

[245] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 16, 17, 18, 20, 21, 22, 23, 24

[246] Yeenyi. ner sentiment analysis sharegpt. `https://huggingface.co/datasets/Yeenyi/ner_sentiment_analysis_sharegpt`, 2024. 12

[247] Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025. 16, 18, 20, 22, 23

[248] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. 20

[249] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms. In *European Conference on Computer Vision*, pages 240–255. Springer, 2024. 2, 4

[250] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85, 2016. 9, 10

[251] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023. 12

[252] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 21

[253] Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. Opencsg chinese corpus: A series of high-quality chinese datasets for llm training, 2025. 9, 10

[254] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023. 12

[255] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024. 9, 10

[256] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 18

[257] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 12

[258] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, et al. Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024. 4

[259] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 10

[260] Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*, 2024. 9

[261] Lefan Zhang, Xiaodan Wang, Yanhua Huang, and Ruiwen Xu. Learning harmonized representations for speculative sampling. *arXiv preprint arXiv:2408.15766*, 2024. 15

[262] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 115–130. Springer, 2021. 12

[263] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2025. 18

[264] Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2.5 years in class: A multimodal textbook for vision-language pretraining. *arXiv preprint arXiv:2501.00958*, 2025. 9, 10

[265] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 12

[266] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023. 15

[267] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 12

[268] Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*, 2025. 12

[269] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022. 12

[270] Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv preprint arXiv:2306.14321*, 2023. 12

[271] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. 11

[272] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 8

[273] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, August 2021. Association for Computational Linguistics. 12

[274] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 1, 3, 4, 12, 16, 17, 18, 20, 21, 22, 23, 24, 25

[275] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023. 9, 10

[276] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024. 18

Category	APP Names
Shopping	Alibaba, Dewu, JD, Pinduoduo, Taobao, Taote, Xianyu, Vipshop
Transportation	Baidu Maps, Amap, Tencent Maps, Hello, Didi, Traffic 12123, Railway 12306
Lifestyle Services	Meituan, Dazhong Dianping, Ele.me, Meituan Waimai, Ctrip, Qunar, SF Express
Automotive	Dongchedi, Autohome
Telecommunications	State Grid Online, China Telecom, China Unicom, China Mobile
Video	Tencent Video, iQIYI, Bilibili, Youku, Kuaishou, Douyin, Migu Video, Tencent Animation, Hongguo Short Drama
Social Media	Toutiao, Weibo, WeChat, Xiaohongshu, Douban, Zhihu, Baidu Tieba, Momo, Facebook, YouTube
Gaming	Xiaohonghe, League of Legends Mobile, Happy Match
Music	NetEase Cloud Music, Ximalaya
Fitness	Keep
Tools	Tianyancha, Quark, Cloud Flash Pay, Industrial and Commercial Bank of China, 58 City, Meitu Xiuxiu
OPPO Built-in Apps	Settings, Phone Migration, Xiaobu Assistant, Clock, Weather, Calendar, Notes, Calculator, Compass, Camera, Recorder, Album, Music, OPPO Video, Reader, Contacts, Dialer, Messages, Mini Games, Game Center, Wallet, Cloud Services, My OPPO, OPPO Store, Main App Store

Table 20: App List from Andes-UI Dataset Collection

Data Type	Training Set	Test Set
Total Screenshots	13002	455
Referring Data Count	226901	8642
Grounding Data Count	185968	7194
Overall descriptive data	13002	455
Natural Q&A Pairs	107688	1181

Table 21: AndesUI Dataset Statistics

A Contributor

Zhiwei Jin, Xiaohui Song, Nan Wang, Yafei Liu, Chao Li, Xin Li, Ruichen Wang, Zhihao Li, Qi Qi, Long Cheng, Dongze Hao, Quanlong Zheng, Yanhao Zhang, Haobo Ji, Jian Ma, Zhitong Zheng, Zhenyi Lin, Haolin Deng, Xin Zou, Xiaojie Yin, Ruilin Wang, Liankai Cai, Haijing Liu, Yuqing Qiu, Ke Chen, Zixian Li, Chi Xie, Huafei Li, Chenxing Li, Chuangchuang Wang, Kai Tang, Zhiguang Zhu, Kai Tang, Wenmei Gao, Rui Wang, Jun Wu, Chao Liu, Qin Xie, Chen Chen[1], Haonan Lu[2]

B AndesUI Dataset

In this section, we provide a comprehensive presentation of the AndesUI dataset construction pipeline, including the data collection process, human annotation, and data generation.

Selection of APPs. We collected a total of 90 APPs, comprising 65 popular download APPs from the OPPO Software Store, covering a wide range of categories commonly used by users, along with 25 ColorOS pre-installed APPs. These APPs are listed in Table 20.

Screenshot Data Collection. For each APP, we instructed annotators to capture screenshots of various diverse pages within the app, ensuring that each screenshot had distinct layouts and content. If two screenshots had similar layout structures but differed solely in text and images, they were classified as homogeneous interfaces. Our objective was to maximize diversity within the dataset while covering all typical interfaces of the app. Depending on the homogeneity degree, we collected between 1 and 10 screenshots for each heterogeneous page. For example, in the Xiaohongshu post interface, the display of different users' posts is similar enough to be

---

[1]chenchen4@oppo.com

[2]luhaonan@oppo.com

regarded as a homogeneous page; however, since some posts include images while others do not, we aimed to collect additional screenshots from this homogeneous interface.

Throughout the screenshot collection process, we focused on capturing various atypical scenarios, including network interruptions and pop-ups (encompassing advertisement, log-in, confirmation, and phone pop-ups). For the training dataset, we collected a total of 10,747 screenshots from third-party apps and 2,255 screenshots from system pre-installed apps. In the testing set, there were a total of 455 screenshots. These screenshots were heterogeneous to reduce duplicate and similar pages. All detailed statistics of the dataset is shown in Table 21

Annotation of Widgets. Our objective was to provide annotations for all widgets present within each screenshot. This included delineating bounding boxes, identifying widget types, recording any text on the widgets (when available), and indicating whether they are clickable, among other details. For this process, we employed the VIA-2.0.12 tool [55]. Annotating all widgets manually from scratch is a labor-intensive endeavor; hence, we initially used Qwen2-VL-72B [227] to generate preliminary annotations on each screenshot, converting these annotations into a JSON format compatible with VIA. Subsequent modifications and refinements were then carried out by annotators. On average, each interface resulted in 18 widgets. The training dataset contained a total of 226,901 widgets, while the testing dataset included 9,068 widgets. Examples of labeled widgets of screenshots are provided in Fig. 9.
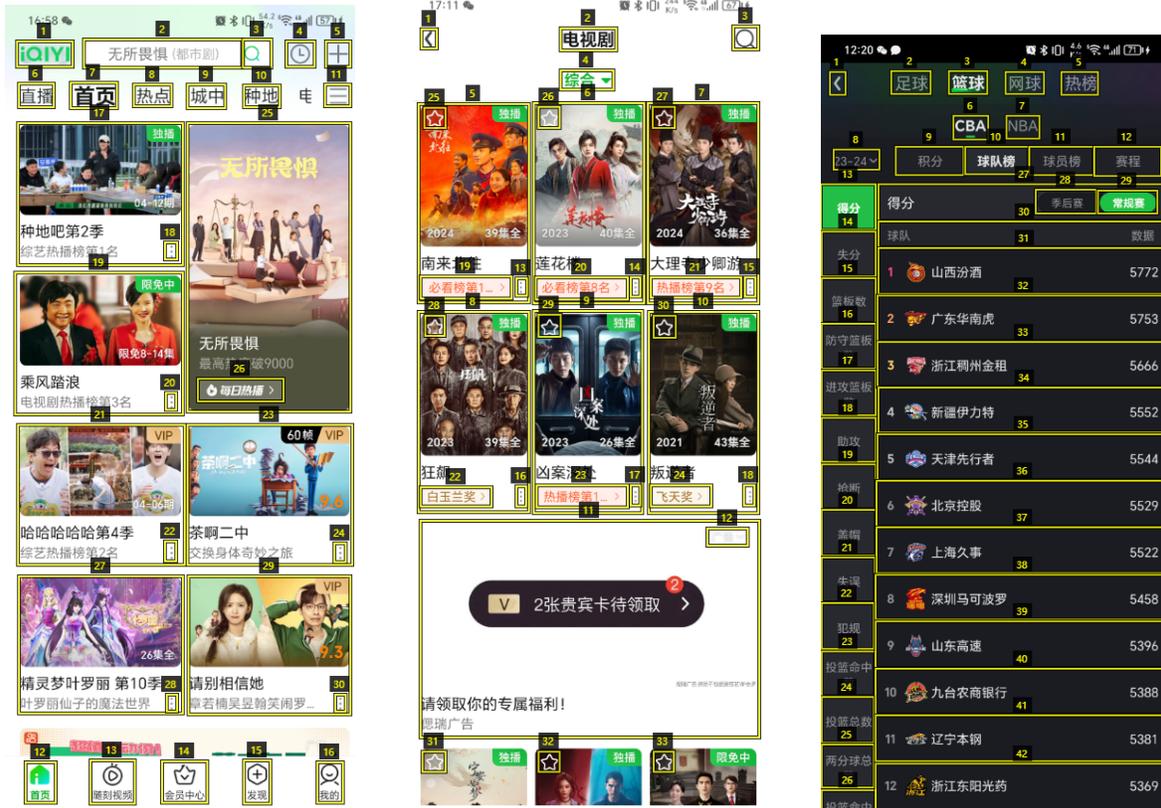

Figure 9: Examples of widget labels in the AndesUI dataset.

We needed to construct both basic and advanced data. Basic data includes grounding and referring data, while advanced data consists of comprehensive descriptive data and natural question-answer pairs. Basic data can be generated through programmatic means. In particular, for each widget, a single grounding data entry and a single referring data entry are generated. As an illustration, for a "send" widget with coordinates [3212, 1045, 3550, 2242], the associated grounding and referring data are

- Question: "Can you tell me the coordinates of the widget named 'send'?" Answer: "<|box_start|>(3212, 1047),(3550, 2242)<|box_end|>"

44

- Question: "What is the widget located within the bounding box <|box_start|>(3212, 1047),(3550, 2242)<|box_end|>?" Answer: "send"

Consequently, the training dataset produced 226,901 data entries for referring and 185,968 for grounding. The test dataset included 7,194 grounding entries and 8,642 referring entries. This discrepancy occurs because a single page can contain several widgets sharing the same name, leading to fewer grounding entries. The questions were randomly selected from a seed library of questions. Initially, GPT-4 was employed to create 100 different question formulations. The bounding box coordinates underwent normalization and were then scaled by a factor of 10,000.

Generation of Advanced Data. For comprehensive descriptive data, each screenshot was analyzed by aggregating the details of individual widgets. Subsequently, GPT-4 was utilized to generate a detailed description of the current page, including the theme, function, spatial arrangement of principal widgets, and a general usage guide for the interface.

For dataset creation involving natural question-answer pairs, we constructed several pairs for each screenshot. To achieve this, we initially utilized the information from each widget to instruct GPT-4 to formulate approximately ten question-answer pairs, emulating possible user inquiries during real-world application. The natural question-answer pairs were divided into four categories: descriptive questions, locating questions, interaction questions, and questions regarding natural scenes. They can also be classified by difficulty level: easy, medium, and hard. Questions classified as "easy" can be immediately answered, whereas "hard" questions might necessitate reasoning or multiple steps to resolve. Initially, GPT-4 was employed to generate several preliminary questions, which were then refined by annotators. Ultimately, we generated 107,688 natural question-answer pairs for the training set and constructed 1,181 pairs for the test set. Below is the system prompt used to guide GPT-4 in generating the natural question-answer pairs:

You are an AI visual assistant capable of analyzing mobile screens. You will receive a screenshot from the {app_info} app of the {page_description} page, along with a string representation of a widget dictionary. Each element in the dictionary is a dictionary that represents a UI widget, where the key is the widget number and the value contains information about the widget, including its bounding box coordinates, widget type, and widget description. The bounding box coordinates are represented as (x1, y1, x2, y2), with floating-point values ranging from 0 to 1. Based on the provided text and coordinates, please design several simulated question-and-answer dialogues that represent interactions between the user and the system. These dialogues should focus on the user's potential actions on the screen (rather than perceptions). The questions you create should be divided into three levels of difficulty: easy, medium, and hard. Easy questions can be answered directly from the widget dictionary. Medium and hard questions require some reasoning. The questions can also be categorized into four content types: descriptive questions, locating questions, interaction questions, and natural scene questions. Here are four examples for reference; they may not be related to this image, and please do not restrict yourself to these few questions. Examples of descriptive questions: Can you describe the function of widget_23? / How many video list items are there in this screenshot? Examples of locating questions: How do I access the creation page from the current page? / How can I view detailed information about the fourth video? Examples of interaction questions: Can widget_2 be swiped? Examples of natural scene questions: How can I search for the latest movies? / How do I share the second video on social media? For non-descriptive questions, you do not need to specify the type information of the widget in your responses. When creating dialogues involving specific widgets, please strictly refer to the widget number (e.g., "widget_3") rather than using the bounding box coordinates. This is necessary for me to use this data for function calls, so clear reference is required. Your response format should be: [{"User":"...","System":"...","Difficulty Level":"","Content Category":""},...]

C   Qualitative Examples

In this section, a series of qualitative examples are illustrated for various capabilities of the AndesVL model by presenting responses generated from AndesVL-4B. Fig. 22 shows strong image understanding; Fig. 23, 24 and 25 demonstrate multi-image perception ability; Fig. 26 shows basic reasoning capability of AndesVL; and Fig. 27 and 28 illustrate document understanding and OCR capabilities.

Examples demonstrating AndesVL-4B-Thinking's visual reasoning capability are provided in Fig. 29, 30 and 31. Fig. 29 shows a case of AndesVL-4B-Thinking solving a geometry problem. Fig. 31 illustrates the model's capability of handling reasoning tasks related to charts.

UI understanding and reasoning cases of AndesVL-4B are shown in Fig. 32 and 33.

	Prompt: 详细描述一下这张图片的内容。 **Model Response**: 这张图片展示了一个夜晚的场景。在图片的右侧，有一只猫，它正坐在一根树枝上，猫的轮廓清晰地映衬在背景的明亮光线中。这只猫的姿势看起来很放松，它正面对着树枝的另一端。背景中，有一个巨大的满月，月光呈现出橙红色调，与夜晚的黑色背景形成鲜明对比。月球的表面细节不是很清晰，但可以看到一些暗淡的云状纹理。整个场景给人一种宁静而神秘的感觉。

Table 22: A dense captioning example of AndesVL.

	Prompt: What are these attractions? Please give a short description of the attractions and output in Markdown format. **Model Response**: - **Left**: The Pyramids of Giza, located in Egypt, are the largest and oldest of the three pyramids in the Giza pyramid complex. They were built as tombs for the Pharaohs of Ancient Egypt. - **Right**: The Great Wall of China is an ancient series of walls and fortifications that was built to protect China from invasions. It stretches for over 13,000 miles and is one of the most famous landmarks in the world.

Table 23: A knowledge question answering example of AndesVL.
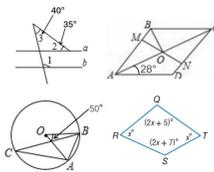
	Prompt: Here is one math problem with its ground-truth answer, the solving of which requires interpreting the paired diagram. Please select the correct paired math diagram from the four given images. Question: As shown in the figure, the straight line a is parallel to b, then the degree of angle 1 is (). # Choices: - A:75° - B:105° - C:140° - D:145° Answer the question first and give a brief explanation of your choice. **Model Response**: B: 105° The given angles are 40° and 35°. Since line a is parallel to line b, angle 2 is equal to 40°. Therefore, angle 1 is the sum of angles 2 and 35°, which is 40° + 35° = 75°. However, angle 1 is an exterior angle to the triangle formed by angles 35° and 40°, so the exterior angle theorem tells us that angle 1 is equal to 180° - 75° = 105°.

Table 24: An Example of Multi-Image Understanding of AndesVL.

	Prompt: A point is circled on the first image, labeled with REF. We change the camera position or lighting and shoot the second image. You are given multiple red-circled points on the second image, choices of "A, B, C, D" are drawn beside each circle. Which point on the second image corresponds to the point in the first image? # Select from the following options first and briefly explain the reason: - A: Point A - B: Point B - C: Point C - D: Point D **Model Response**: B: Point B The first image's REF point is in the center of the ski lift area, and Point B in the second image is in the same relative position, maintaining the same perspective and orientation.

Table 25: An Example of Multi-Image Understanding of AndesVL.
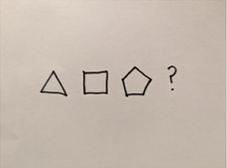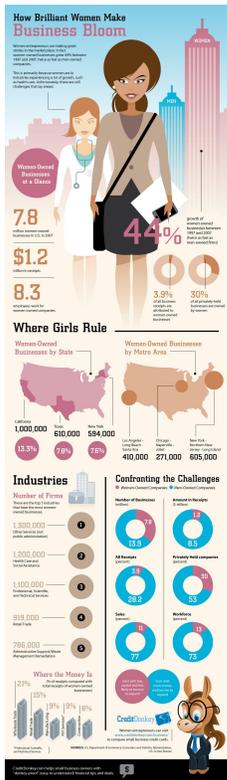
	Prompt: Look at this sequence of three shapes. What shape should come as the fourth shape? Explain your reasoning with detailed descriptions of the first shapes. **Model Response**: The sequence consists of a triangle, a square, and a pentagon. The fourth shape in the sequence should be a hexagon. The reasoning is as follows: Each shape in the sequence has one more side than the previous shape. The triangle has three sides, the square has four sides, and the pentagon has five sides. Therefore, the next shape should have six sides, which is a hexagon.

Table 26: A visual reasoning example of AndesVL.

Prompt: Make a detailed summary of the image.
Model Response:
This image is an infographic titled "How Brilliant Women Make Business Bloom."
It highlights the growth and impact of women-owned businesses in the United States. Here's a detailed summary of the content:
1. Title and Introduction:
   · The infographic begins with the title "How Brilliant Women Make Business Bloom."
   · A brief introduction mentions that women entrepreneurs are making significant strides in the marketplace. It states that women-owned businesses grew by 44% between 1997 and 2007 ...
2. Where Girls Rule:
   · Women-Owned Businesses by State:
      · California: 1,000,000 businesses (13.3% of total)
      ...
   · Women-Owned Businesses by Metro Area:
      · Los Angeles - Long Beach - Santa Ana: 410,000 businesses.
      ...
3. Industries:
   The infographic lists the top 5 industries with the most women-owned businesses:
   1. Other Services (not public administration): 1,300,000 businesses.
   ...
4. Confronting the Challenges:
...
Additional Information:
   · The infographic also includes a note on financial tips and deals, mentioning CreditDonkey.com, which helps small business owners with "donkey-proof" financial advice.
   ...
Visual Elements:
...

Prompt: Which state has the second highest number of women-owned businesses? Answer the question with detailed explanation.
Model Response:
According to the text, Texas has 610,000 women-owned businesses, which is the second highest number after California. The text also provides percentages for the number of women-owned businesses in different industries and metro areas. However, the question specifically asks for the state with the second highest number of women-owned businesses, which is Texas.

Table 27: Demonstration of AndesVL's ability of accurately extracting and comprehensing texts and visual elements from an image, understanding the user's query, and retrieving the answer from the document.

	Prompt: Parse the image into a markdown table.
	Model Response:
	``markdown
	|10%氯化钾注射液（哈森）（省集采2)(高警示药品)|10ml|1 支|0.1900|0.19|
	|5%葡萄糖注射液(威高)|250ml|1 袋|4.2600|4.26|
	|一次性使用无菌注射器|1.2mm(1.2*38TWLB)100支*6盒/箱|1 支|0.64|0.64|
	|碳酸氢钠注射液（哈森）（省集采2）|10ml:0.5g|1 支|0.4000|0.40|
	|一次性使用无菌注射器|1.2mm(1.2*38TWLB)100支*6盒/箱|2 支|0.64|1.28|
	|浓氯化钠注射液(高警示药品)|10ml:1g|1 支|0.7040|0.70|
	|静脉输液（2瓶以上每瓶加收）|1|2 次|2.00|2.00|
	|一次性使用输液器(超低密度聚乙烯)|15支*20包/箱|1 支|5.00|5.00|
	|5%葡萄糖注射液(威高)|250ml|1 袋|4.2600|4.26|
	|静脉输液|4|1 组|4.00|4.00|
	|0.9%氯化钠注射液(威高)|100ml|1 瓶|4.8900|4.89|
	|注射用奥美拉唑钠(美斯汀)（国集采7）|40mg*1瓶/瓶|1 瓶|0.8400|0.84|
	|浓氯化钠注射液(高警示药品)|10ml:1g|1 支|0.7040|0.70|
	|维生素B6注射液(省集采1)|2ml:0.1g|1 支|0.4300|0.43|
	|合计：|29.59|
	``

Table 28: An example of extracting multilingual texts from a long receipt.

Prompt: What is the volume of this azure right square pyramid?
Model Response:
<think>
Got it, let's see. The problem is to find the volume of a right square pyramid.
First, I need to recall the formula for the volume of a pyramid. The volume V of a pyramid is $\frac{1}{3} \times$ base area $\times$ height.
First, identify the base. The base is a square, so the base area is side length squared.
From the diagram, the side length of the square base is 8. The height h of the pyramid is given as 7.
So, base area $= 8 \times 8 = 64$. Then, volume $= \frac{1}{3} \times 64 \times 7$.
...
So the answer is 149.33 (or 149.333... but let's present it as a floating number).
</think>
To find the volume of a right square pyramid, we use the formula:
$V = \frac{1}{3} \times$ base area $\times$ height
Step 1: Determine the base area
The base of the pyramid is a square with side length 8.
The area of a square is side^2, so:
Base area $= 8 \times 8 = 64$
Step 2: Identify the height of the pyramid
From the diagram, the height h of the pyramid is 7.
Step 3: Calculate the volume
Substitute the base area and height into the volume formula:
$V = \frac{1}{3} \times 64 \times 7$
First, multiply $64 \times 7 = 448$. Then divide by 3:
$V = \frac{448}{3} \approx 149.333\ldots$
Step 4: Present the result as a floating number
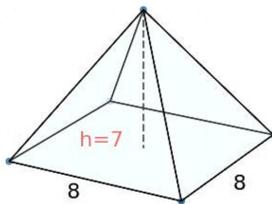Rounding to two decimal places, the volume is 149.33.
the answer is 149.33

Table 29: An example of solving a geometry problem.

$e - f = 8$ $g * h = 3$ $(e - f) / g = ?$ $e - f + g = 11$	**Prompt**: What is the value of (e - f) / g? **Model Response**: \<think\> Got it, let's try to solve this problem. First, we need to find the value of (e - f)/g. Let's list out the given equations: 1. e - f = 8 2. g * h = 3 3. e - f + g = 11 First, let's find the value of g. From equation 3: e - f + g = 11. We know from equation 1 that e - f = 8. So substitute that into equation 3: 8 + g = 11. Solving for g, we get g = 11 - 8 = 3. So g is 3. Now, we need to find (e - f)/g. We know e - f is 8 (from equation 1) and g is 3. So substitute those values in: 8 / 3. Let's compute that. 8 divided by 3 is approximately 2.666..., which as a floating number is 2.666666... (repeating). But we need to present it as a floating number. Let's check if we need to round or keep the exact value. Since it's a math problem, 8/3 is approximately 2.666..., so we can write it as 2.6667 (if we round to four decimal places) or just 2.666666... but let's confirm. Wait, let's make sure we didn't miss anything. Let's check the other equations. ... he problem says "convert the final result as a floating number". So 8 divided by 3 is approximately 2.666666..., so we can write it as 2.666666... or 2.6667. Let's go with 2.666666... (six decimal places) or 2.6667 (four decimal places). Let's pick 2.666666... as the floating number. \</think\> To solve for the value of (e - f)/g, we analyze the given equations step by step: Step 1: Find the value of g We know: e - f = 8 (from the first equation). e - f + g = 11 (from the fourth equation). Substitute e - f = 8 into e - f + g = 11: 8 + g = 11 Solve for g: g = 11 - 8 = 3 Step 2: Calculate (e - f)/g ... the answer is 2.666666

Table 30: An example of solving an algebra problem.

	Prompt: Find the value of $z * (z^*)$. **Model Response**: <think> Got it, let's try to solve this problem. First, we need to find the value of z multiplied by its complex conjugate z^*. Let's start by identifying the complex number z from the graph. Looking at the graph, the point (-2, 3i) corresponds to the complex number z = -2 + 3i. ... Alternatively, using the formula (a + bi)(a - bi) = a² + b². Here, a = -2, b = 3. So a² = (-2)² = 4, b² = 3² = 9. Then 4 + 9 = 13. So the value of $z * (z^*)$ is 13. </think> To solve for $z * (z^*)$, follow these steps: Step 1: Identify the complex number z from the graph The graph shows a point at $(-2, 3i)$. For a complex number $z = a + bi$, the real part a is the x - coordinate, and the imaginary part b is the y - coordinate. Thus, $z = -2 + 3i$. Step 2: Find the complex conjugate z^* The complex conjugate of $z = a + bi$ is $z^* = a - bi$. For $z = -2 + 3i$, the conjugate is $z^* = -2 - 3i$. Step 3: Calculate $z * z^*$ For any complex number $z = a + bi$, the product $z * z^*$ simplifies to $a^2 + b^2$ (derived from multiplying $(a + bi)(a - bi)$). For $z = -2 + 3i$: - Real part $a = -2$, so $a^2 = (-2)^2 = 4$. - Imaginary part $b = 3$, so $b^2 = 3^2 = 9$. Adding these: $a^2 + b^2 = 4 + 9 = 13$. The answer is 13.

Table 31: An example of solving an analytic geometry problem.

Prompt: 当前页面有弹窗吗？
Model Response: 是的，当前页面有一个弹窗，
控件(1741,7204),(8398,7858)是一个文本按钮, 描述为"立即抢购"。
Prompt: 如何关闭当前页面的弹窗？
Model Response: 您可以点击控件(4509,8167),(5565,8533)，
它是一个图标（Icon），描述为'关闭'。
Prompt: 我想找平板电脑，应该怎么办？
Model Response: 您可以点击控件(472,5638),(2000,6450)，
它是一个文本按钮，描述为"平板电脑"。
Prompt: 我想查看我的订单，应该怎么办？
Model Response: 您可以点击控件(8241,9371),(9250,9988)，
它是一个多标签（Multi-Tab），描述为"我的"，进入后可以查看您的订单。

Table 32: An example of Mobile UI understanding and reasoning.

Prompt: 这是什么页面?
Model Response: 这是音乐APP中的播放页面，用于控制和播放音乐。
Prompt: 现在正在播放什么音乐?
Model Response: 正在播放的音乐是'别怕我伤心'。
Prompt: 如何切换下一曲?
Model Response: 点击控件(6376,8752),(7342,9218)。

Table 33: An example of Mobile UI understanding and reasoning.