NVIDIA Nemotron Nano V2 VL

NVIDIA

Abstract. We introduce Nemotron Nano V2 VL, the latest model of the Nemotron vision-language series designed for strong real-world document understanding, long video comprehension, and reasoning tasks. Nemotron Nano V2 VL delivers significant improvements over our previous model, Llama-3.1-Nemotron-Nano-VL-8B, across all vision and text domains through major enhancements in model architecture, datasets, and training recipes. Nemotron Nano V2 VL builds on Nemotron Nano V2, a hybrid Mamba-Transformer LLM, and innovative token reduction techniques to achieve higher inference throughput in long document and video scenarios. We are releasing model checkpoints in BF16, FP8, and FP4 formats and sharing large parts of our datasets, recipes and training code.

Model (BF16) | Model (FP8) | Model (NVFP4-QAD) | Dataset

1. Introduction

We introduce Nemotron Nano V2 VL, an efficient 12B vision–language model that achieves leading accuracy on OCRBench v2 (Fu et al., 2024a) private data leaderboard[1], along with strong performance in reasoning, document understanding, long-video comprehension, visual question answering, and STEM reasoning. Nemotron Nano V2 VL delivers substantial improvements over our previous model, Llama-3.1-Nemotron-Nano-VL-8B[2], driven by enhancements in model architecture, dataset composition, and training methodology. These gains stem from the inclusion of higher-quality reasoning data, expanded OCR datasets, and additional long-context datasets.

In addition to overall benchmark improvements, we extend the model's context length from 16K to 128K, enabling better handling of long videos and complex reasoning tasks. Consistent with our prior open-source efforts, we release the model weights, along with substantial portions of the training datasets, recipes, and codebase, to support continued research and development.

Nemotron Nano V2 VL builds on top of the Nemotron Nano V2 (NVIDIA et al., 2025) 12B reasoning LLM and RADIOv2.5 vision encoder (Heinrich et al., 2025). In addition, Nemotron Nano V2 VL adopts the multimodal fusion architecture, training recipe as well as data strategy similar to Eagle 2 and 2.5 (Li et al., 2025b; Chen et al., 2025a). We use the open-source Megatron (Shoeybi et al., 2019) framework to train the model in FP8 precision using Supervised Finetuning (SFT) across several vision and text domains, followed by additional SFT stages to further improve video understanding, long context performance, and recover text-only capabilities to achieve competitive results across many vision and text benchmarks.

Compared to Llama-3.1-Nemotron-Nano-VL-8B, the hybrid Mamba-Transformer architecture of the LLM offers 35% higher throughput in long multi-page document understanding scenarios. Additionally, we employ Efficient Video Sampling (EVS) (Bagrov et al., 2025) to accelerate throughput in video understanding use cases by 2x or more with minimal or no impact on accuracy.

Furthermore, Nemotron Nano V2 VL supports both reasoning-on and reasoning-off modes, with the former enabling extended reasoning for tasks that require more complex problem-solving. This

[1] https://99franklin.github.io/ocrbench_v2/
[2] https://huggingface.co/nvidia/Llama-3.1-Nemotron-Nano-VL-8B-V1

design enables a balanced trade-off between computational efficiency and task performance.

We are releasing our model weights on HuggingFace in BF16, FP8 and FP4 formats:

- Nemotron-Nano-12B-v2-VL: The final model weights after our multi-stage training recipe.
- Nemotron-Nano-12B-v2-VL-FP8: Quantized model weights in FP8 format
- Nemotron-Nano-12B-v2-VL-NVFP4-QAD Quantized model weights in FP4 format using Quantization-aware Distillation (QAD)

Additionally, we release a large portion of our SFT dataset and tooling:

- Nemotron VLM Dataset V2: A collection of over 8 million training samples.
- NVPDFTex: A custom LaTeX compiler toolchain to generate annotated OCR ground truth.

The remainder of this report is organized as follows: Section 2 describes the model architecture and input processing pipeline; Section 3 outlines the training recipe, datasets, and hyperparameters; and Section 4 presents comprehensive evaluations of the model across both multimodal and pure-text tasks.

2. Model Architecture

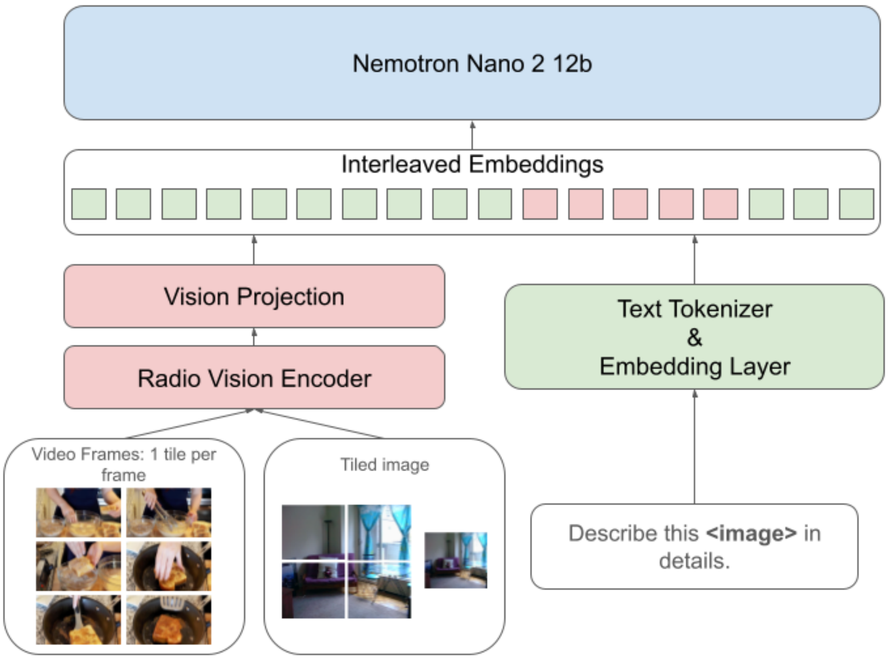


Figure 1 | Visualization of our VLM architecture. For images, we extract a dynamic number of tiles based on the image aspect ratio. For videos, we uniformly extract frames. Tiles and frames are resized to 512×512 pixels, and go through the RADIO vision encoder and an MLP connector. The image and text embeddings are interleaved, and fed to the Nemotron-Nano-12B-V2 LLM.

As illustrated in Figure 1, Nemotron Nano V2 VL consists of three modules: a vision encoder, an MLP projector, and a language model. We initialize the vision encoder using the c-RADIOv2-VLM-H version of the RADIOv2 vision encoder (Heinrich et al., 2025) and the language model with Nemotron-Nano-12B-V2 (NVIDIA et al., 2025).

Inspired by InternVL (Chen et al., 2024d), LLaVA-1.5 (Liu et al., 2024b) and Eagle (Li et al., 2025b; Chen et al., 2025a), we adopt a tiling strategy to handle varying image resolutions. First, each image is resized following the aspect ratio matching strategy employed by InternVL (Chen et al., 2024d) so that its width and height are multiples of s. Then it is divided into non-overlapping tiles of size $s \times s$. In this work, we set $s = 512$. With a patch size of 16, this results in 1024 visual tokens per tile. For scalability, we employ pixel shuffle with 2x downsampling to reduce the token count further to 256. During training, we set the maximum number of tiles to 12. Additionally, we use a single-tile thumbnail of the image to capture global image context. For video inputs, we limit each input frame to a single tile.

3. Training Recipe & Datasets

To preserve the base language model's text comprehension and reasoning abilities while improving its visual understanding, we adopt a multi-stage training approach as detailed below.

3.1. Stage 0

In this stage, we aim to warm up the MLP connector to establish cross-modal alignment between the language and vision domains. To this end, we freeze the vision encoder and language model weights and train only the MLP connector on a diverse multimodal subset of the Stage 1 SFT dataset (see Section 3.2), consisting of approximately 2.2 million samples (up to 36 billion tokens) spanning multiple tasks, including captioning, visual question answering, visual grounding, OCR, and document extraction.

3.2. SFT Stage 1: 16K context length

In this and all subsequent stages, we unfreeze all model components for training. In SFT Stage 1, the maximum sequence length is set to 16,384 tokens. This stage is trained on approximately 32.5 million samples (about 112.5 billion tokens). To preserve the text comprehension capabilities of the Nemotron-Nano-V2 (NVIDIA et al., 2025) LLM backbone, we incorporate a subset of the text reasoning data used in its Stage 1 SFT training, comprising approximately 6.5 million samples (around 40 billion tokens) spanning diverse domains and tasks such as mathematics, science, code, multilingual understanding, multi-turn dialogue, tool-use, and safety. In addition, we include multimodal datasets totaling 26 million samples (approximately 72 billion tokens) drawn from various tasks and sources, including:

1. **Image Captioning**: OpenImages (Kuznetsova et al., 2020), TextCaps (Sidorov et al., 2020), TextVQA (Singh et al., 2019), PixMo-cap (Deitke et al., 2024).
2. **Video Captioning**: Localized Narratives (Pont-Tuset et al., 2020), YouCook2 (Zhou et al., 2017), VaTeX (Wang et al., 2020a).
3. **General Visual QA**: TextVQA (Singh et al., 2019), VQAv2 (Goyal et al., 2017), OK-VQA (Marino et al., 2019), GQA (Hudson & Manning, 2019), CLEVR (Johnson et al., 2016), CLEVR-Math (Lindström & Abraham, 2022), TallyQA (Acharya et al., 2018), Dolly-15K (Conover et al., 2023), ScreenQA (Hsiao et al., 2025), VizWiz (Gurari et al., 2018), MapQA (Chang et al., 2022), ScienceQA (Lu et al., 2022), PMC-VQA (Zhang et al., 2024b), MetaMathQA (Yu et al., 2024), UniGeo (Chen et al., 2022), CMM-Math (Liu et al., 2024c), Geo-170K (Gao et al., 2025), VisualWebInstruct (Jia et al., 2025), LRV-Instruction (Liu et al., 2024a), OCR-VQA (Mishra et al., 2019), EST-VQA (Wang et al., 2020b), ST-VQA (Biten et al., 2019), PixMo-AskModelAnything (Deitke et al., 2024), ALLaVA-4v (Chen et al., 2024a), SLAKE

(Liu et al., 2021), VQA-RAD (Lau et al., 2018), DreamSim (Fu et al., 2023), Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), NLVR2 (Suhr et al., 2019).
4. **Video QA**: CLEVRER (Yi et al., 2020), Perception Test (Pătrăucean et al., 2023), ALFRED (Shridhar et al., 2020), NextQA (Xiao et al., 2021), VCG+112K (Maaz et al., 2024).
5. **Visual Grounding**: RefCOCO (Kazemzadeh et al., 2014).
6. **OCR, Table & Document Extraction**: SynthDog-en (Kim et al., 2022), SynthTabNet (Nassar et al., 2022), DocLayNet (Pfitzmann et al., 2022), WebSight (Laurençon et al., 2024b), TabRecSet (Yang et al., 2023), FinTabNet (Zheng et al., 2020), PubTables-1M (Smock et al., 2021), TextOCR (Singh et al., 2021), HierText (Long et al., 2022), FUNSD (Jaume et al., 2019), CASIA-HWDB2 (Liu et al., 2011), RCTW-17 (Shi et al., 2018), ReCTS-19 (Liu et al., 2019), human-annotated CommonCrawl [3] samples, synthetically generated tables, arXiv paper annotations generated using the NVPDFTex[4] pipeline and translated to several other languages using mBART-large-50 (Tang et al., 2020), and multilingual Wikimedia[5] dumps.
7. **Document, Chart, Table and GUI QA**: ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2021a), AI2D (Kembhavi et al., 2016), DocVQA (Mathew et al., 2021b), FigureQA (Kahou et al., 2018), ECD-10K (Yang et al., 2025b), ArXivQA (Li et al., 2024b), PlotQA (Methani et al., 2020), PixMo-Docs (Deitke et al., 2024), TabMWP (Lu et al., 2023), SlideVQA (Tanaka et al., 2023), Docmatix (Laurençon et al., 2024a), DocReason25K (Hu et al., 2024), UniChart (Masry et al., 2023), SimChart9K (Xia et al., 2024), MMTab (Zheng et al., 2024), VisText (Tang et al., 2023), ScreenQA (Hsiao et al., 2025), WaveUI-25K (AgentSea, 2024), as well as synthetic QA labels generated for FinTabNet (Zheng et al., 2020), HierText (Long et al., 2022) and CommonCrawl PDF samples transcribed using Nemo Retriever Parse[6] (Karmanov et al., 2025).
8. **Visual Grounding**: Visual7W (Zhu et al., 2016), OpenImages (Kuznetsova et al., 2020).
9. **Function Calling**: Glaive function calling (AI, 2023), xLAM-60K (Liu et al., 2024f).

We augment the corpus with both human-annotated reasoning traces and model-generated traces produced by Qwen2.5-VL-32B-Instruct (Qwen et al., 2025), GLM-4.1V, and GLM-4.5V (Hong et al., 2025) across multiple datasets to reinforce the extended reasoning ability of our model in reasoning-on mode. Additionally, for datasets lacking explicit QA labels, we generate synthetic question–answer pairs from existing OCR extractions or captions using LLMs from the Qwen2.5 (Qwen et al., 2025) and Qwen3 (Yang et al., 2025a) families. A large portion of the training data in this stage is released at Nemotron VLM Dataset V2.

3.3. SFT Stage 2: 49K context extension

In this stage, we extend the model's context length to 49,152 tokens to enhance its capability for multi-image and video understanding. The model is trained on approximately 11M samples (around 55B tokens), including a subset of the Stage 1 dataset. We experimented with varying proportions of Stage 1 data and found that a 25% reuse ratio offers a good balance between training efficiency and maintaining accuracy across text, vision, multi-frame and video benchmarks.

In addition to the reused Stage 1 subset, we curate video and multi-image datasets comprising approximately 1.4 million samples (around 17 billion tokens), covering a diverse range of tasks across several data sources, including: (1) **Video Classification**: Kinetics (Carreira & Zisserman, 2018); (2) **Dense Video Captioning**: YouCook2 (Zhou et al., 2017), HiREST (Zala et al., 2023),

[3] https://commoncrawl.org/
[4] https://github.com/NVIDIA-NeMo/Curator/tree/experimental/experimental/nvpdftex
[5] https://dumps.wikimedia.org/
[6] https://build.nvidia.com/nvidia/nemoretriever-parse

ActivityNet (Heilbron et al., 2015); (3) **Video Captioning**: EgoExoLearn (Huang et al., 2025b); (4) **Temporal Action Localization**: Breakfast Actions (Kuehne et al., 2014), Perception Test (Pătrăucean et al., 2023), HiREST (Zala et al., 2023), HACS Segment (Zhao et al., 2019), FineAction (Liu et al., 2022), Ego4D-MQ (Grauman et al., 2022), ActivityNet (Heilbron et al., 2015); (5) **Video Temporal Grounding**: YouCook2 (Zhou et al., 2017), QuerYD (Oncescu et al., 2021), MedVidQA (Gupta et al., 2022), Ego4D-NLQ (Grauman et al., 2022), DiDeMo (Hendricks et al., 2017); (6) **General Video QA**: LLaVA-Video-178K (Zhang et al., 2025c), Ego4D (Grauman et al., 2022), TVQA (Lei et al., 2019), Perception Test (Pătrăucean et al., 2023), NextQA (Xiao et al., 2021), EgoExoLearn (Huang et al., 2025b), CLEVRER (Yi et al., 2020), and relabeling of the following datasets with Qwen2.5-VL-72B-Instruct (Bai et al., 2025) into MCQ and open-ended questions: TAPOS (Shao et al., 2020), HC-STVG (Tang et al., 2021), EgoProceL (Bansal et al., 2022), CrossTask (Zhukov et al., 2019); (7) **Multi-page QA**: Synthetic multi-page QA data constructed from CommonCrawl PDF documents using Nemo Retriever Parse extractions; and (8) **Multi-image captions**: Mementos (Wang et al., 2024d).

We convert all the non-QA data into QA formats. For video classification, temporal action localization and temporal grounding data, we use template questions to generate QA pairs. For video captioning and multi-page OCR captions, we use existing LLM models from the Qwen2.5 family (Qwen et al., 2025) to synthesize both the questions and answers given the captions.

3.4. SFT stage 3: 49K text recovery

After SFT Stages 1 and 2, we observe a substantial drop in the LiveCodeBench score compared to the LLM backbone, despite including the text reasoning data from Nemotron Nano 2 (NVIDIA et al., 2025). To recover this loss, we introduce an additional SFT Stage 3 trained with a maximum sequence length of 49,152 using only code reasoning data totaling 1M samples or 15B tokens.

3.5. SFT stage 4: 300K context extension

Finally, we extend the model's context further and incorporate long-context data from the Stage 3 SFT stage of Nemotron Nano 2 (NVIDIA et al., 2025), accounting for around 74K samples or 12B tokens. The samples in this data are 160K tokens long on average, and we train with a maximum sequence length of 311,296 to accommodate the longest samples. We find that this stage helps improve the accuracy of the model in long-context benchmarks such as RULER (Hsieh et al., 2024).

3.6. Training details

At all stages, the model is trained with FP8 precision following a recipe similar to (NVIDIA et al., 2025) to accelerate training. The same configuration is applied to the LLM, vision encoder, and vision projection MLP, with the first and last layers of the LLM and the transformer blocks of the vision encoder kept in BF16. We did not observe any training instabilities with this setup, and both the training loss curve and benchmark scores closely match those of a full-BF16 model. Additionally, experiments keeping either the vision encoder or vision projection in BF16 did not yield a significant difference.

For video inputs to the model, we extract 2 frames per second, with a maximum of 128 frames for each video. If a video is longer than 64 seconds, we uniformly sample 128 frames instead.

As text-only, image and video samples vary significantly in sequence length, we find that sequence packing reduces training time by minimizing the number of padding tokens required for batching. We perform sequence packing online during training using a buffer containing several thousand

	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
Global Batch Size	1024	128			
Total Training iterations	2158	-	-	-	-
Linear Warmup Fraction	0.1				
Learning Rate	2×10^{-4}	2×10^{-5}			
Weight Decay	0.01	0.05			
Max Length	16384		49152		311296

Table 1 | Training hyperparameters for all stages

	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
# GPU nodes	32	64			
Training time (in hrs)	6	30	15	3.5	5.5

Table 2 | Hardware resources across training stages

samples. We employ the balance-aware data packing strategy described in (Li et al., 2025b).

To mitigate any bias towards shorter or longer sequences during training, we employ loss square-averaging, similar to InternVL (Chen et al., 2025b). We use the AdamW optimizer with β_1 and β_2 set to 0.9 and 0.999, respectively, and a cosine annealing schedule with a linear warmup. See Table 1 for an overview of the training hyperparameters.

Long Context Extension. For SFT stages 2, 3, and 4, we employ context parallelism in the LLM (Megatron Core, 2025). Context parallelism partitions the LLM input along the sequence dimension, mitigating out-of-memory issues at longer sequence lengths. We use 2-way and 8-way context parallelism for SFT stages 2–3 and stage 4, respectively. When using N-way context parallelism in the LLM, N replicas of the vision encoder and vision projection modules are instantiated. To efficiently utilize these replicas and further reduce GPU memory usage, we follow the approach of (Chen et al., 2024c) and split the vision encoder and projection inputs into N shards along the batch dimension. The N vision projection outputs are then gathered before being passed to the LLM.

Infrastructure. We train the model with the Megatron (Shoeybi et al., 2019) framework[7] using Transformer Engine[8] and the Megatron Energon[9] dataloader on NVIDIA H100 GPUs. Our training code is for large parts open-source. The training resources are summarized in Table 2.

Figure 2 | Overview of the training stages for the VLM along with the model context length and number of training tokens.

[7] https://github.com/NVIDIA/Megatron-LM
[8] https://github.com/NVIDIA/TransformerEngine
[9] https://github.com/NVIDIA/Megatron-Energon

4. Experiments

To comprehensively assess our model's capabilities, we evaluate it across a broad suite of multimodal benchmarks covering diverse tasks. In Section 4.1, we compare its performance with our previous-generation multimodal model, Llama-3.1-Nemotron-Nano-VL-8B, and with state-of-the-art multimodal models of similar scale. In Section 4.2, we demonstrate the importance of our multi-stage training strategy in preserving the text reasoning abilities of the LLM backbone. We also investigate reasoning budget control and present the model's behavior under different budget thresholds in Section 4.3. In Section 4.4, we investigate the efficiency gains achieved by applying Efficient Video Sampling (EVS) (Bagrov et al., 2025) to video inputs. Finally, we examine alternative approaches for processing image inputs in Section 4.5.

4.1. Multimodal Evaluations

We conduct a comprehensive evaluation of our model on 45 benchmarks over seven broad categories:

1. **General VQA**: MMBench V1.1 (Liu et al., 2024d), MMStar (Chen et al., 2024b), BLINK (Fu et al., 2024b), MUIRBench (Wang et al., 2024a), HallusionBench (Guan et al., 2024), ZeroBench (Roberts et al., 2025), CRPE (Wang et al., 2024c), POPE (Li et al., 2023), MME-RealWorld (Zhang et al., 2025b), RealWorldQA[10], MMT-Bench (Ying et al., 2024), R-Bench (Guo et al., 2025), WildVision (Lu et al., 2024b).
2. **STEM Reasoning**: MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), MathVista-Mini (Lu et al., 2024a), MathVision (Wang et al., 2024b), MathVerse-Mini (Zhang et al., 2024a), DynaMath (Zou et al., 2025), LogicVista (Xiao et al., 2024), WeMath (Qiao et al., 2024)
3. **Document Understanding, OCR & Charts**: MMLongBench-Doc (Ma et al., 2024), OCRBench (Liu et al., 2024e), OCRBench-V2 (Fu et al., 2024a), ChartQA (Masry et al., 2022), RDTableBench (AI, 2025), AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2021a), OCR-Reasoning (Huang et al., 2025a), VCR (Zhang et al., 2025a), SEED-Bench-2-Plus (Li et al., 2024a), CharXiv (Wang et al., 2024f)
4. **Visual Grounding & Spatial Reasoning**: TreeBench (Wang et al., 2025a), CV-Bench (Tong et al., 2024)
5. **GUI Understanding**: ScreenSpot (Cheng et al., 2024), ScreenSpot-v2 (Wu et al., 2024b), ScreenSpot Pro (Li et al., 2025a)
6. **Video Understanding**: LongVideoBench (Wu et al., 2024a), MLVU (Zhou et al., 2025), Video-MME (Fu et al., 2025)
7. **Multimodal Multilingual Understanding**: MTVQA (Tang et al., 2025), MMMB (Sun et al., 2025), Multilingual MMBench (Sun et al., 2025)

We use the VLMEvalKit (Duan et al., 2025)[11] framework for our evaluations with a vLLM (Kwon et al., 2023)[12] backend inference server. For the reasoning-off mode, we employ greedy decoding and cap the maximum number of generated tokens at 1,024 for all benchmarks except RDTableBench[13] where we use a limit of 16,384 tokens. For reasoning-on evaluations, we set the temperature to 0.6, top-p to 0.95, and the maximum output length to 16,384 tokens.

We present reasoning-off evaluation results on a subset of the benchmarks after each training stage in

[10] https://x.ai/news/grok-1.5v
[11] https://github.com/open-compass/VLMEvalKit
[12] https://github.com/vllm-project/vllm
[13] https://huggingface.co/datasets/reducto/rd-tablebench/tree/main

Task	Benchmark	Nemotron Nano V2 VL		InternVL3.5		GLM-4.5V		Qwen3-VL	
Size		12B	12B	14B	14B	106B (A12B)	106B (A12B)	8B	8B
Mode		Reasoning off	Reasoning on	Non-Thinking	Thinking	Non-Thinking	Thinking	Instruct	Thinking
General	MMBench V1.1 Dev (EN/ZH)	83.0/80.2	82.6/76.3	83.0*/82.3*	-	-	-	85.0/-	87.5/-
	MMStar	65.9	71.7	70.4	-	73.4	75.3	70.9	75.3
	BLINK (Val)	57.6	56.7	57.6	-	63.7	65.3	69.1	64.7
	MUIRBench	33.2	44.2	58.0	-	71.1	75.3	64.4	76.8
	HallusionBench	72.3	73.1	54.0	-	59.1	65.4	61.1	65.4
	ZeroBench (sub)	15.0	18.3	10.8*	-	21.9	23.4	22.8	18.6*
	CRPE (Relation)	71.3	64.3	76.6	-	-	-	71.7*	52.3*
	POPE	88.8	87.0	87.7	-	-	-	88.2*	86.2*
	MME-RealWorld (EN)	62.2	-	63.2	-	-	-	-	-
	RealWorldQA	76.2	72.2	70.5	-	-	-	71.5	73.5
	MMT-Bench (Val)	65.5	63.8	66.9*	-	-	-	-	-
	R-Bench (dis)	75.2	69.7	70.9	-	-	-	70.1*	69.3*
	WildVision (win rate)	16.0	20.2	73.0	-	-	-	75.0*	76.4*
STEM Reasoning	MMMU (val)	55.3	67.8	62.78*	73.3	68.4	75.4	69.6	74.1
	MMMU Pro	14.5	28.0	-	-	59.8	65.2	55.6	60.4
	MathVista-Mini	69.0	75.5	73.3*	80.5	78.2	84.6	77.2	81.4
	MathVision	31.5	53.6	-	59.9	52.5	65.6	53.9	62.7
	MathVerse-Mini (Vision-only)	34.3	58.2	41.75*	62.8	-	-	38.2*	68.7*
	Dynamath (worst case)	14.8	32.3	-	38.7	44.1	53.9	38.7*	-
	LogicVista	38.7	58.4	-	60.2	54.8	62.4	58.8*	59.3*
	WeMath	31.5	39.3	-	58.7	58.9	68.8	57.1*	-
Document Understanding, OCR & Charts	MMLongBench-Doc	32.1	-	-	-	41.1	44.7	47.9	48.0
	OCRBench	85.6	83.5	83.6	-	87.2	86.5	89.6	81.9
	OCRBenchV2 (EN/ZH)	62.0/44.2	54.8/39.8	-	-	-	-	65.4/61.2	63.9/59.2
	ChartQA (Test)	89.8	84.9	86.5	-	-	-	83.4*	-
	RDTableBench	67.8	57.4	-	-	-	-	82.4*	43.9*
	AI2D (Test)	87.2	84.7	85.1	-	-	-	85.7	84.9
	TextVQA (Val)	85.4	76.1	77.8	-	-	-	82.9*	-
	DocVQA (Test)	94.7	93.2	93.4	-	-	-	96.1	95.3
	InfoVQA (Test)	79.4	80.4	78.3	-	-	-	83.1	86.0
	OCR-Reasoning	21.0	33.9	-	-	-	-	39.3*	48.4*
	VCR-EN-Easy (EM/Jaccard)	71.8/80.0	-	93.4/97.7	-	-	-	91.8*/95.1*	87.4*/94.4*
	SEED-Bench-2-Plus	71.3	70.0	70.7	-	-	-	71.4*	69.9*
	CharXiv (RQ/DQ)	41.7/76.5	41.3/77.2	47.9/76.6	-	-	-	46.4/83.0	53.0/85.9
Visual Grounding & Spatial Reasoning	TreeBench	38.5	42.5	42.0*	-	47.9	50.1	26.9*	25.4*
	CV-Bench	81.0	78.3	80.1*	-	86.5	87.3	73.1*	79.5*
GUI	ScreenSpot	39.4	40.1	87.5	-	-	-	94.4	93.6
	ScreenSpot-v2	41.7	42.8	88.6	-	-	-	-	-
	ScreenSpot-Pro	4.8	5.5	-	-	-	-	54.6	46.6
Video Understanding	LongVideoBench	63.6	57.0	62.7	-	-	-	-	-
	MLVU (M-Avg)	73.6	-	72.1	-	-	-	-	-
	VideoMME (w/o sub)	66.0	63.0	67.9	-	74.3	74.6	71.4	71.8

Table 3 | Comparison of Nemotron Nano V2 VL with existing open-source multimodal models. All results marked with * were calculated in VLMEvalKit.

Table 4, along with a comparison to our previous generation multimodal model, Llama-3.1-Nemotron-Nano-VL-8B. We also compare our model against state-of-the-art open-source multimodal models of similar scale including InternVL3.5 (14B) (Wang et al., 2025b), GLM-4.5V (106B-A12B) (Hong et al., 2025) and Qwen3-VL (8B) (Yang et al., 2025a)[14] in Tables 3 and 5. Unless otherwise noted, we report the evaluation scores directly from the respective model reports. For benchmarks not covered therein, we independently evaluate the models using VLMEvalKit whenever possible.

Benchmark	Stage 0	SFT Stage 1: 16K context length	SFT Stage 2: 49K context extension	SFT stage 3: 49K text recovery	SFT stage 4: 300K context extension	Llama-3.1-Nemotron Nano-VL-8B
AI2D (Test)	67.6	87.1	87.1	**87.3**	87.2	85.0
ChartQA (Test)	70.9	89.9	90.0	**90.2**	89.8	86.3
DocVQA (Val)	79.1	**94.4**	94.3	94.2	**94.4**	91.2
InfoVQA (Val)	51.3	**80.2**	79.2	79.2	79.2	77.4
LongVideoBench	45.1	59.4	**63.6**	63.1	**63.6**	-
MMLongBench-DOC	10.75	29.2	32.0	30.8	**32.1**	-
MMMU (Val)	49.0	54.8	55.0	54.3	**55.3**	48.2
MathVista-Mini	53.1	67.7	**69.3**	69.2	69.0	-
OCRBench	61.4	84.8	85.3	85.4	**85.6**	83.9
OCRBench-V2 (CN)	18.3	42.9	43.4	43.4	**44.2**	37.9
OCRBench-V2 (EN)	38.3	**62.5**	**62.5**	61.8	62.0	60.1
TextVQA (Val)	76.7	**86.0**	85.0	85.2	85.4	-
Video-MME	36.3	57.6	65.8	65.4	**66.0**	54.7

Table 4 | Vision benchmarks for our previous-generation VLM, Llama-3.1-Nemotron-Nano-VL-8B, and after each training stage of Nemotron Nano V2 VL with reasoning-off.

Evaluation across training stages. Table 4 highlights the effectiveness of our long-context extension training introduced in SFT stages 2 and 4. We observe substantial gains on Video-MME (Fu et al., 2025) and MMLongBench-Doc (Ma et al., 2024) following SFT stage 2, and these improvements are retained in the final model after stage 4 training.

Comparison with Llama-3.1-Nemotron-Nano-VL-8B. As shown in Table 4, we observe consistent improvements across all benchmarks compared to Llama-3.1-Nemotron-Nano-VL-8B. We attribute these gains to a combination of factors, including an enhanced LLM backbone, expanded and higher-quality training datasets, and an improved training recipe.

4.2. Pure Text Evaluations

We conduct all pure-text evaluations using the NeMo-Skills[15] framework with a maximum output length of 32,768 tokens, temperature set to 0.6, and top-p of 0.95. We report Pass@1 average of 16 runs for AIME-2025; an average of 4 runs for MATH-500 (Lightman et al., 2023), GPQA-Diamond (Rein et al., 2023), LiveCodeBench (07/24 - 12/24) (Jain et al., 2024), IFEval (Zhou et al., 2023); and score of 1 run for SciCode (Tian et al., 2024) and RULER (Hsieh et al., 2024). For MMLU-Pro (Wang et al., 2024e), we report the accuracy on a 1000-sample subset.

Evaluation across training stages. We report the model's text evaluation scores after each training stage in Table 6. Comparing the results between stage 0 (where the LLM backbone remains

[14] https://github.com/QwenLM/Qwen3-VL
[15] https://github.com/NVIDIA-NeMo/Skills

Task	Benchmark	Nemotron Nano V2 VL Reasoning-off	InternVL3.5 Non-Thinking	GLM-4.5V	Qwen3-VL Instruct
Size		12B	14B	106B (A12B)	8B
MTVQA	Avg	24.3	34.2	22.0†	32.2*
MMMB	en	86.1	85.1	87.1†	85.4*
	zh	84.1	84.1	86.9†	82.5*
	pt	83.5	82.7	84.8†	81.5*
	ar	83.4	80.3	84.5†	80.3*
	tr	77.4	79.4	84.6†	78.7*
	ru	83.9	83.5	84.3†	81.9*
Multilingual MMBench	en	84.9	84.0	89.1†	85.5*
	zh	82.5	83.7	89.3†	85.6*
	pt	82.5	80.0	86.9†	82.5*
	ar	81.5	77.8	83.7†	79.0*
	tr	75.7	77.0	84.0†	79.0*
	ru	82.2	77.0	87.2†	81.3*

Table 5 | Comparison of Nemotron Nano V2 VL with SOTA multimodal models on multimodal multilingual benchmarks. All scores marked by * were reproduced by us using VLMEvalKit, and those marked with † were obtained from the InternVL3.5 (Wang et al., 2025b) technical report.

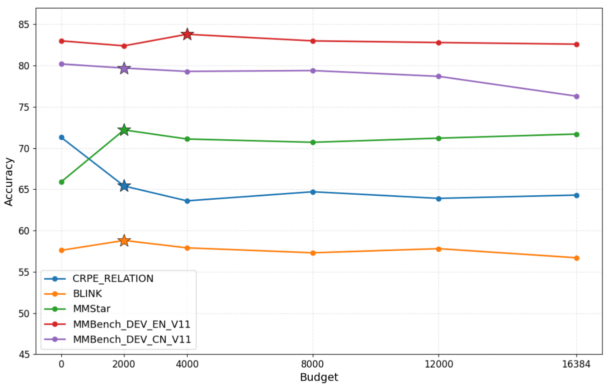
Benchmark	Stage 0: Pretraining	SFT Stage 1: 16K context length	SFT Stage 2: 49K context extension	SFT stage 3: 49K text recovery	SFT stage 4: 300K context extension
`MATH-500`	97.7	96.8	97.3	97.6	96.9
`AIME-25`	75.9	68.0	72.7	72.7	71.3
`GPQA`	65.0	60.9	63.0	60.6	64.1
`LiveCodeBench`	70.0	50.9	55.0	69.8	69.4
`IFEval_prompt_strict`	84.2	77.5	77.3	76.5	78.2
`IFEval_instruction_strict`	89.3	84.1	83.9	83.4	84.7
`SciCode_problem_accuracy`	7.5	5.0	6.9	5.0	6.9
`SciCode_subtask_accuracy`	22.3	17.5	14.9	17.2	17.6
`MMLU-Pro-1000`	77.8	75.2	75.8	76.7	77.1
`RULER`	77.9	8.8	17.4	21.5	72.1

Table 6 | Text benchmarks with reasoning on for the different stages. Our goal was to add vision capabilities with minimal impact to the text reasoning capabilities of the underlying LLM Nemotron-Nano-12B-V2. The text reasoning benchmarks of Stage 0 corresponds to the benchmarks results of the underlying LLM since the LLM is frozen during Stage 0. After Stage 1, we see a significant drop in text reasoning benhchmarks (eg. LiveCodeBench goes from 70 to 50.87) and long context benchmark (RULER goes from 77.91 to 8.80). The video long context extension (Stage 2) helps recover a little bit of the benchmark score, but we still observe very low LiveCodeBench (55.00) and RULER (17.39). We apply a code reasoning recovery stage (Stage 3), and a long context extension stage (Stage 4), to recover both benchmarks (LiveCodeBench of 69.44 and RULER of 72.12).

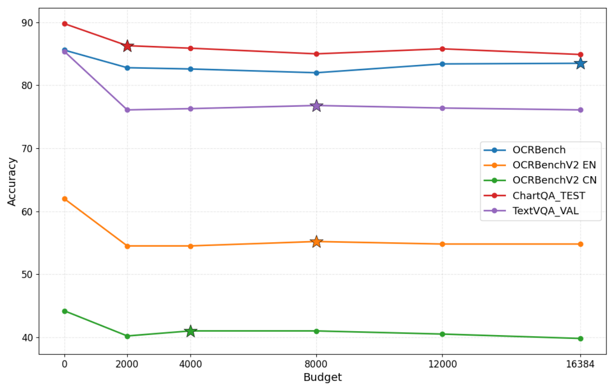
frozen) and SFT stage 1, we observe a significant drop in the LiveCodeBench score - from 70.0 to 50.87. To address this degradation without compromising vision performance, we explored several mitigation strategies that were unsuccessful, including augmenting the SFT stage 1 dataset with additional code reasoning examples and disabling loss scaling. Ultimately, we introduced an additional SFT stage 3 focused exclusively on code reasoning, which helped restore the LiveCodeBench score without affecting other benchmarks. As shown in Table 4, the vision benchmarks remain stable between SFT stage 2 and stage 3, and the final model largely preserves the text reasoning capabilities of the original LLM backbone across most tasks.

Furthermore, we see a significant improvement in the RULER score across SFT stages 2-4 following an initial drop after stage 1. These results further validate the effectiveness of our long-context training strategy.

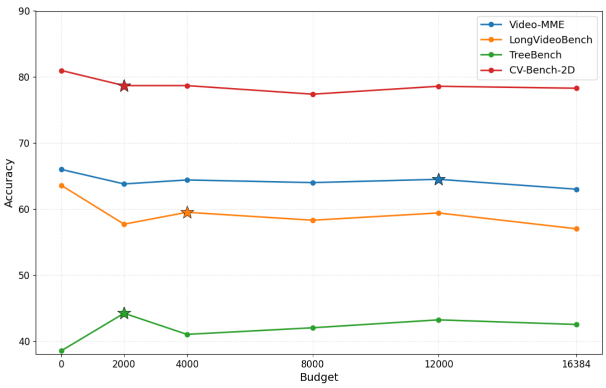
4.3. Reasoning Budget Control



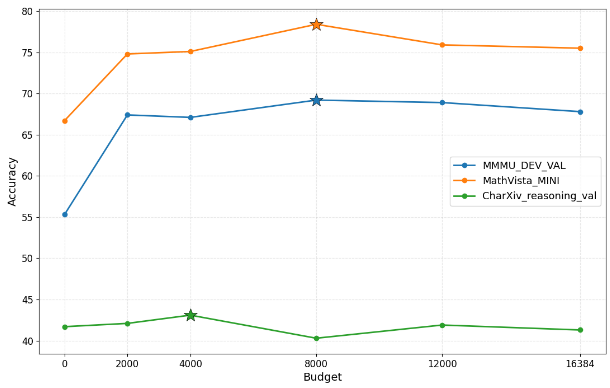
(a) General VQA



(b) Document, OCR & Chart Understanding



(c) Spatial & Video Understanding



(d) STEM & Chart Reasoning

Figure 3 | Effect of reasoning budget control at 2K, 4K, 8K, and 12K tokens across multiple tasks. A budget value of 0 corresponds to reasoning-off evaluations, while 16,384 denotes unrestricted reasoning-on evaluations with a maximum generation length of 16,384 tokens. The highest reasoning-on score (including the unrestricted case) is indicated with a ★ for each task.

Following Nemotron-Nano-V2 (NVIDIA et al., 2025), we experiment with varying reasoning budgets during inference. We evaluate the model's behavior under budgets of 2K, 4K, 8K, and 12K tokens, each with a 500-token grace period, as shown in Figure 3.

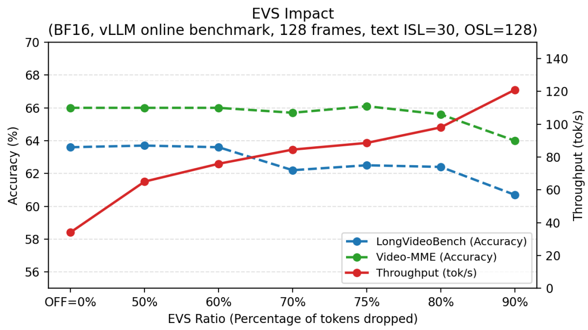
Our experiments indicate that tuning the reasoning budget can improve reasoning-on mode accuracy across multiple tasks, even in cases where the unrestricted reasoning-on score is lower than the reasoning-off score. These gains with budget control may arise from the early termination of malformed reasoning traces with repetition loops on out-of-distribution tasks, as well as the truncation of overly verbose reasoning chains for problems requiring minimal or straightforward reasoning.

4.4. Inference with Efficient Video Sampling

Building upon **Efficient Video Sampling (EVS)** (Bagrov et al., 2025), we integrate it directly into our video-processing pipeline. EVS reduces the number of visual tokens by identifying and pruning *temporally static patches* – spatial regions that remain nearly unchanged between consecutive frames, while preserving positional identity and semantic consistency. This enables our model to process substantially longer videos with lower latency and memory consumption, without requiring architectural changes or retraining.

We evaluate the model on two video benchmarks: **Video-MME** and **LongVideoBench**. Figure 4 presents EVS ablations for both BF16 and FP8 precision. The results demonstrate that EVS maintains strong performance on long-context video understanding tasks while significantly improving efficiency: as the EVS ratio increases, time-to-first-token (TTFT) decreases and throughput rises, with only a minor impact on accuracy.

EVS	LongVideoBench	Video-MME	TTFT (ms)	Throughput (tok/s)
OFF	63.6	66.0	4131	34
50%	63.7	66.0	2699	65
60%	63.6	66.0	2458	75
70%	62.2	65.7	2184	84
75%	62.5	66.1	2072	88
80%	62.4	65.6	1990	98
90%	60.7	64.0	1654	120



EVS	LongVideoBench	Video-MME	TTFT (ms)	Throughput (tok/s)
OFF	64.2	66.4	3436	51
50%	63.7	66.5	2384	80
60%	63.4	66.2	2223	85
70%	62.8	65.8	2000	95
75%	62.3	65.7	1840	103
80%	62.8	65.1	1717	103
90%	60.4	64.0	1567	132

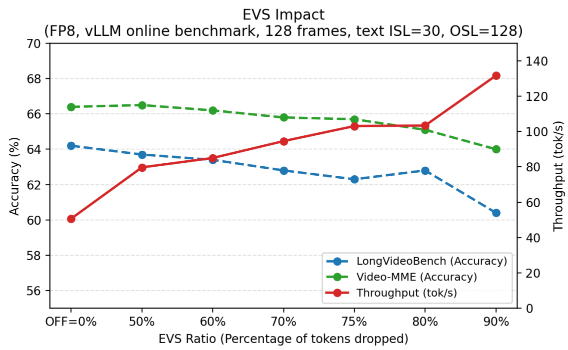


Figure 4 | EVS Ablation (RTX 6000 PRO SE, vLLM online benchmark, 128 frames, text ISL=30, OSL=128): top row shows BF16 results, bottom row shows FP8 results; left shows numeric tables (accuracy, time-TTFT, throughput), right shows corresponding visualizations.

4.5. Image Processing Ablations

We investigate an alternative approach to tiling for our input image processing pipeline. Instead of splitting the input image into non-overlapping tiles, we pass the entire image at native resolution to the vision encoder and MLP projector followed by convolutional token reduction for 4× sequence compression.

For each image processing variant, we train the model through Stages 0 and 1 of our full training recipe, excluding the text-only reasoning datasets from Stage 1. Table 7 compares these image processing strategies across a subset of multimodal benchmarks.

As shown in Columns 1 and 2, the native-resolution approach achieves comparable or even superior accuracy on several benchmarks. However, we observe a notable performance drop on OCRBench and OCRBench-V2 (English). Upon analysis, we found that the tiling algorithm occasionally applies large rescaling factors to smaller images to preserve aspect ratio. To isolate this effect, we conduct an additional experiment where we resize each image to match the size and aspect ratio that the tiling algorithm would have selected, without applying tiling. The results of this configuration are shown in Column 3 of Table 7, where we recover the performance drop on OCRBench. The gap on OCRBench-V2 (English), however, persists. We plan to further investigate strategies to address this gap in future work.

Benchmark	Tiling	Native Resolution	Native Resolution with tiling-size matching
AI2D (Test)	87.1	86.4	**87.8**
ChartQA (Test)	89.8	88.4	**90.3**
DocVQA (Val)	94.5	**95.1**	94.9
InfoVQA (Val)	80.2	**80.7**	79.0
MMMU (Val)	**56.8**	56.0	56.2
MathVista-Mini	69.7	**71.3**	71.2
OCRBench	84.5	82.8	**85.3**
OCRBench-V2 (CN)	40.5	**45.3**	42.7
OCRBench-V2 (EN)	**61.4**	57.6	57.6
TextVQA (Val)	85.6	84.6	**86.2**
Average	75.0	74.8	**75.1**

Table 7 | We compare our image tiling strategy with: 1) Native resolution image inputs to the vision encoder followed by convolutional token reduction 2) Native resolution image inputs with tiling-size matching where we resize the image to the size and aspect ratio the dynamic tiling algorithm would have picked, and then feed the resized image to the vision encoder followed by convolutional token merging. We ran the benchmarks on the last 10 saved checkpoints and, for each strategy, select the one with the highest average benchmark score.

4.6. Quantization

The model is trained with Transformer Engine's delayed-scaling FP8, using dynamic scaling factors computed from a running history of activation maxima across training iterations. By contrast, most downstream inference stacks (e.g. vLLM, TensorRT-LLM) assume static quantization with scaling factors fixed during a one-time calibration. To bridge this train–serve gap, we produce deployable checkpoints via post-training quantization (PTQ) and quantization-aware distillation (QAD) using software libraries TensorRT-Model-Optimizer and Megatron-LM. Unless otherwise noted, we quantize the language backbone - i.e., all linear layers, including both weights and activations - and retain the embedding layers, KV cache or others in higher precision. QAD is performed in BF16 to simulate lower-precision behavior. For FP8 we adopt the E4M3 format; we additionally provide NVFP4 variants from QAD.

PTQ We calibrate on 1,024 samples drawn from the training set and compute per-tensor static scales for weights and activations using the absolute amax aggregated over the calibration dataset.

The resulting FP8/NVFP4 checkpoints are directly usable in standard inference stacks such as vLLM.

QAD Because NVFP4 PTQ yields a small accuracy drop, we further improve quality by distilling a PTQ student from a BF16 teacher using a logit-matching loss (e.g. KL divergence) applied to the final model outputs only. Hyperparameters mirror SFT Stage 1, except we use a learning rate of 2×10^{-6}.

Table 8 summarizes the accuracy comparison across different quantization precisions and strategies. All evaluations are conducted within the vLLM framework under BF16, FP8, and NVFP4 runtime configurations. Minor discrepancies relative to previously reported baselines are attributable to variations in evaluation framework implementations.

Precision	AI2D	ChartQA	OCRBench	DocVQA-val	OCRBenchV2 English
BF16	87.21	89.68	854	94.22	61.74
FP8-PTQ	87.56	89.44	854	94.32	61.83
NVFP4-PTQ	86.37	88.84	863	92.38	60.88
NVFP4-QAD	87.14	89.96	851	93.95	61.94

Table 8 | Accuracy comparison across different precisions and tasks using vLLM backend.

5. Conclusion

In this work, we introduced Nemotron Nano V2 VL, an efficient 12B vision–language model built upon the Nemotron-Nano-V2 LLM. The model demonstrates substantial improvements in multimodal and text understanding, as well as reasoning capabilities, compared to its predecessor, Llama-3.1-Nemotron-Nano-VL-8B. We employed a multi-stage training strategy to enhance visual understanding while preserving the text comprehension abilities of the original backbone. We presented a comprehensive evaluation of the model across diverse tasks and modalities, along with investigations into efficient video sampling at inference and alternative image processing pipelines. Finally, we have open-sourced the model weights in BF16, FP8, and FP4 formats, along with a significant portion of our SFT dataset and associated tooling (see Section 1).

6. Contributors

Core Model Training and Data Study

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Guo Chen, Karan Sapra, Zhiding Yu, Adi Renduchintala, Charles Wang, Peter Jin.

Systems, Data and Infrastructure Arushi Goel, Mike Ranzinger, Lukas Voegtle, Philipp Fischer, Timo Roman, Wei Ping, Boxin Wang, Zhuolin Yang, Nayeon Lee, Shaokun Zhang, Fuxiao Liu, Zhiqi Li, Di Zhang, Greg Heinrich, Hongxu (Danny) Yin, Song Han, Pavlo Molchanov, Parth Mannan, Yao Xu, Jane Polak Scowcroft, Tom Balough, Subhashree Radhakrishnan, Paris Zhang, Sean Cha, Ratnesh Kumar, Zaid Pervaiz Bhat, Jian Zhang, Darragh Hanley, Pritam Biswas, Jesse Oliver, Kevin Vasques, Roger Waleffe, Duncan Riach, Oluwatobi Olabiyi, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Pritam Gundecha, Khanh Nguyen.

Inference and Optimization Alexandre Milesi, Eugene Khvedchenia, Ran Zilberstein, Ofri Masad, Natan Bagrov, Nave Assaf, Tomer Asida, Daniel Afrimi, Amit Zuker, Netanel Haber, Zhiyu Cheng, Jingyu (Justin) Xin, Di (Allan) Wu, Nik Spirin, Maryam Moosaei, Roman Ageev, Vanshil Atul Shah, Yuting Wu, Daniel Korzekwa, Unnikrishnan Kizhakkemadam Sreekumar, Wanli Jiang.

Evaluation, Safety, Product and Legal Padmavathy Subramanian, Alejandra Rico, Sandip Bhaskar, Saeid Motiian, Kedi Wu, Annie Surla, Chia-Chih Chen, Hayden Wolff, Matthew Feinberg, Melissa Corpuz, Marek Wawrzos, Eileen Long, Aastha Jhunjhunwala, Paul Hendricks, Farzan Memarian, Benika Hall, Xin-Yu Wang, David Mosallanezhad, Soumye Singhal, Luis Vega, Katherine Cheung, Krzysztof Pawelec, Michael Evans, Katherine Luna, Jie Lou, Erick Galinkin, Akshay Hazare, Kaustubh Purandare, Ann Guan, Anna Warno, Chen Cui, Yoshi Suhara, Shibani Likhite, Seph Mard, Meredith Price, Laya Sleiman, Saori Kaji.

Leadership Udi Karpas, Kari Briski, Joey Conway, Michael Lightstone, Jan Kautz, Mohammad Shoeybi, Mostofa Patwary, Jonathen Cohen, Oleksii Kuchaiev, Andrew Tao, Bryan Catanzaro.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering complex counting questions, 2018. URL https://arxiv.org/abs/1810.12440.

AgentSea. Waveui-25k. https://huggingface.co/datasets/agentsea/wave-ui-25k, 2024.

Glaive AI. Glaive function calling v2 dataset. https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2, 2023.

Reducto AI. RDTableBench: A benchmark for table-to-SQL / table reasoning for LLMs. https://huggingface.co/datasets/reducto/rd-tablebench, 2025.

Natan Bagrov, Eugene Khvedchenia, Borys Tymchenko, Shay Aharon, Lior Kadoch, Tomer Keren, Ofri Masad, Yonatan Geifman, Ran Zilberstein, Tuomas Rintamaki, Matthieu Le, and Andrew Tao. Efficient video sampling: Pruning temporally redundant tokens for faster VLM inference, 2025. URL https://arxiv.org/abs/2510.14624.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Siddhant Bansal, Chetan Arora, and C. V. Jawahar. My view is the best view: Procedure learning from egocentric videos, 2022. URL https://arxiv.org/abs/2207.10883.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering, 2019. URL https://arxiv.org/abs/1905.13648.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset, 2018. URL https://arxiv.org/abs/1705.07750.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps, 2022. URL https://arxiv.org/abs/2211.08545.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing GPT4V-synthesized data for lite vision-language models, 2024a. URL https://arxiv.org/abs/2402.11684.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Tuomas Rintamaki, Tyler Poon, Max Ehrlich, Tuomas Rintamaki, Tyler Poon, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. Eagle 2.5: Boosting long-context post-training for frontier vision-language models, 2025a. URL https://arxiv.org/abs/2504.15271.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression, 2022. URL https://arxiv.org/abs/2212.02746.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024b. URL https://arxiv.org/abs/2403.20330.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024c. URL https://arxiv.org/abs/2408.10188.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024d. URL https://arxiv.org/abs/2404.16821.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024. URL https://arxiv.org/abs/2401.10935.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Xiaozhe Li, Chaoyou Fu, Junbo Cui, Jixuan Chen, Enxin Song, Song Mao, Shengyuan Ding, Tianhao Liang, Zicheng Zhang, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2025. URL https://arxiv.org/abs/2407.11691.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024a. URL https://arxiv.org/abs/2501.00321.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. URL https://arxiv.org/abs/2306.09344.

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*, 2024b.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-llava: Solving geometric problem with multi-modal large language model, 2025. URL https://arxiv.org/abs/2312.11370.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL https://arxiv.org/abs/2310.14566.

Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, Bolin Ni, Guo-Wei Yang, Yongming Rao, Houwen Peng, Han Hu, Gordon Wetzstein, and Shi min Hu. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation, 2025. URL https://arxiv.org/abs/2505.02018.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering, 2022. URL https://arxiv.org/abs/2201.12888.

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, 2018. URL https://arxiv.org/abs/1802.08218.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

Greg Heinrich, Mike Ranzinger, Hongxu, Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models, 2025. URL https://arxiv.org/abs/2412.07679.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language, 2017. URL https://arxiv.org/abs/1708.01641.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots, 2025. URL https://arxiv.org/abs/2209.08199.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL https://arxiv.org/abs/2404.06654.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding, 2024. URL https://arxiv.org/abs/2403.12895.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. Ocr-reasoning benchmark: Unveiling the true capabilities of mllms in complex text-rich image reasoning, 2025a. URL https://arxiv.org/abs/2505.17163.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, and Yu Qiao. Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world, 2025b. URL https://arxiv.org/abs/2403.16182.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents, 2019. URL https://arxiv.org/abs/1905.13538.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images, 2018. URL https://arxiv.org/abs/1808.10584.

Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. Visualwebinstruct: Scaling up multimodal instruction data through web search, 2025. URL https://arxiv.org/abs/2503.10582.

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, 2016. URL https://arxiv.org/abs/1612.06890.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018. URL https://arxiv.org/abs/1710.07300.

Ilia Karmanov, Amala Sanjay Deshmukh, Lukas Voegtle, Philipp Fischer, Kateryna Chumachenko, Timo Roman, Jarno Seppänen, Jupinder Parmar, Joseph Jennings, Andrew Tao, and Karan Sapra. éclair – extracting content and layout with integrated reading order for documents, 2025. URL https://arxiv.org/abs/2502.04223.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

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 (ECCV)*, 2022.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 780–787, 2014. doi: 10.1109/CVPR.2014.105.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024a. URL https://arxiv.org/abs/2408.12637.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024b. URL https://arxiv.org/abs/2403.09029.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering, 2019. URL https://arxiv.org/abs/1809.01696.

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, 2024a. URL https://arxiv.org/abs/2404.16790.

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, 2025a. URL https://arxiv.org/abs/2504.07981.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models, 2024b. URL https://arxiv.org/abs/2403.00231.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2305.10355.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models, 2025b. URL https://arxiv.org/abs/2501.14818.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning, 2022. URL https://arxiv.org/abs/2208.05358.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021. URL https://arxiv.org/abs/2102.09542.

Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *2011 international conference on document analysis and recognition*, pp. 37–41. IEEE, 2011.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024a. URL https://arxiv.org/abs/2306.14565.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024b. URL https://arxiv.org/abs/2310.03744.

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*, 2024c.

Xi Liu, Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, Xiang Bai, Baoguang Shi, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar 2019 robust reading challenge on reading chinese text on signboard, 2019. URL https://arxiv.org/abs/1912.09641.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization, 2022. URL https://arxiv.org/abs/2105.11107.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024d. URL https://arxiv.org/abs/2307.06281.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024e. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, and Caiming Xiong. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets, 2024f. URL https://arxiv.org/abs/2406.18518.

Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis, 2022. URL https://arxiv.org/abs/2203.15143.

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.

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, 2023. URL https://arxiv.org/abs/2209.14610.

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. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255, 2024b.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations, 2024. URL https://arxiv.org/abs/2407.01523.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arxiv*, 2024. URL https://arxiv.org/abs/2406.09418.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL https://arxiv.org/abs/1906.00067.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning, 2023.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. Infographicvqa, 2021a. URL https://arxiv.org/abs/2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021b. URL https://arxiv.org/abs/2007.00398.

Megatron Core. context_ parallel package, September 2025. URL https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots, 2020. URL https://arxiv.org/abs/1909.00997.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers, 2022. URL https://arxiv.org/abs/2203.01017.

NVIDIA, :, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, Ann Guan, Anna Shors, Anubhav Mandarwal, Arham Mehta, Arun Venkatesan, Ashton Sharabiani, Ashwath Aithal, Ashwin Poojary, Ayush Dattagupta, Balaram Buddharaju, Banghua Zhu, Barnaby Simkin, Bilal Kartal, Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Brian Yu, Bryan Catanzaro, Charles Wang, Charlie Truong, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christian Munley, Christopher Parisien, Dan Su, Daniel Afrimi, Daniel Korzekwa, Daniel Rohrer, Daria Gitman, David Mosallanezhad, Deepak Narayanan, Dima Rekesh, Dina Yared, Dmytro Pykhtar, Dong Ahn, Duncan Riach, Eileen Long, Elliott Ning, Eric Chung, Erick Galinkin, Evelina Bakhturina, Gargi Prasad, Gerald Shen, Haifeng Qian, Haim Elisha, Harsh Sharma, Hayley Ross, Helen Ngo, Herman Sahota, Hexin Wang, Hoo Chang Shin, Hua Huang, Iain Cunningham, Igor Gitman, Ivan Moshkov, Jaehun Jung, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jian Zhang, Jiaqi Zeng, Jimmy Zhang, Jinze Xue, Jocelyn Huang, Joey Conway, John Kamalu, Jonathan Cohen, Joseph Jennings, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kezhi Kong, Krzysztof Pawelec, Kumar Anik, Kunlun Li, Kushan Ahmadian, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Luis Vega, Maer Rodrigues de Melo, Makesh Narsimhan Sreedhar, Marcin Chochowski, Mark Cai, Markus Kliegl, Marta Stepniewska-Dziubinska, Matvei Novikov, Mehrzad Samadi, Meredith Price, Meriem Boubdir, Michael Boone, Michael Evans, Michal Bien, Michal Zawalski, Miguel Martinez, Mike Chrzanowski, Mohammad Shoeybi, Mostofa Patwary, Namit Dhameja, Nave Assaf, Negar Habibi, Nidhi Bhatia, Nikki Pope, Nima Tajbakhsh, Nirmal Kumar Juluru, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Oluwatobi Olabiyi, Pablo Ribalta, Padmavathy Subramanian, Parth Chadha, Pavlo Molchanov, Peter Dykas, Peter Jin, Piotr Bialecki, Piotr Januszewski, Pradeep Thalasta, Prashant Gaikwad, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Rabeeh Karimi Mahabadi, Rajen Patel, Ran El-Yaniv, Ranjit Rajan, Ria Cheruvu, Rima Shahbazyan, Ritika Borkar, Ritu Gala, Roger Waleffe, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Sahil Jain, Samuel Kriman, Sanjeev Satheesh, Saori Kaji, Sarah Yurick, Saurav Muralidharan, Sean Narenthiran, Seonmyeong Bak, Sepehr Sameni, Seungju Han, Shanmugam Ramasamy, Shaona Ghosh, Sharath Turuvekere

Sreenivas, Shelby Thomas, Shizhe Diao, Shreya Gopal, Shrimai Prabhumoye, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Siddhartha Jain, Somshubra Majumdar, Soumye Singhal, Stefania Alborghetti, Syeda Nahida Akter, Terry Kong, Tim Moon, Tomasz Hliwiak, Tomer Asida, Tony Wang, Tugrul Konuk, Twinkle Vashishth, Tyler Poon, Udi Karpas, Vahid Noroozi, Venkat Srinivasan, Vijay Korthikanti, Vikram Fugro, Vineeth Kalluru, Vitaly Kurin, Vitaly Lavrukhin, Wasi Uddin Ahmad, Wei Du, Wonmin Byeon, Ximing Lu, Xin Dong, Yashaswi Karnati, Yejin Choi, Yian Zhang, Ying Lin, Yonggan Fu, Yoshi Suhara, Zhen Dong, Zhiyu Li, Zhongbo Zhu, and Zijia Chen. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations, 2021. URL https://arxiv.org/abs/2011.11071.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 3743–3751. ACM, August 2022. doi: 10.1145/3534678.3539043. URL http://dx.doi.org/10.1145/3534678.3539043.

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

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. URL https://arxiv.org/abs/2305.13786.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning?, 2024. URL https://arxiv.org/abs/2407.01284.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra- and inter-action understanding via temporal action parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17), 2018. URL https://arxiv.org/abs/1708.09585.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. URL https://arxiv.org/abs/2003.12462.

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

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8802–8812, 2021.

Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents, 2021. URL https://arxiv.org/abs/2110.00061.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. URL https://arxiv.org/abs/1811.00491.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Parrot: Multilingual visual instruction tuning, 2025. URL https://arxiv.org/abs/2406.02539.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images, 2023. URL https://arxiv.org/abs/2301.04883.

Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. URL http://vis.csail.mit.edu/pubs/vistext.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yangfan He, Kuan Lu, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2025. URL https://arxiv.org/abs/2405.11985.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers, 2021. URL https://arxiv.org/abs/2011.05049.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. Scicode: A research coding benchmark curated by scientists, 2024. URL https://arxiv.org/abs/2407.13168.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024a. URL https://arxiv.org/abs/2406.09411.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology, 2025a. URL https://arxiv.org/abs/2507.07999.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=QWTCcxMpPA.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world, 2024c. URL https://arxiv.org/abs/2402.19474.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, 2020a. URL https://arxiv.org/abs/1904.03493.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10126–10135, 2020b.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences, 2024d. URL https://arxiv.org/abs/2401.10529.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024e. URL https://arxiv.org/abs/2406.01574.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024f. URL https://arxiv.org/abs/2406.18521.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024a. URL https://arxiv.org/abs/2407.15754.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. Os-atlas: A foundation action model for generalist gui agents, 2024b. URL https://arxiv.org/abs/2410.23218.

Renqiu Xia, Haoyang Peng, Hancheng Ye, Mingsheng Li, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, Junchi Yan, and Bo Zhang. Structchart: On the schema, metric, and augmentation for visual chart understanding, 2024. URL https://arxiv.org/abs/2309.11268.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. URL https://arxiv.org/abs/2105.08276.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. URL https://arxiv.org/abs/2407.04973.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Fan Yang, Lei Hu, Xinwu Liu, Shuangping Huang, and Zhenghui Gu. A large-scale dataset for end-to-end table recognition in the wild. *Scientific Data*, 10(1):110, 2023.

Yuwei Yang, Zeyu Zhang, Yunzhong Hou, Zhuowan Li, Gaowen Liu, Ali Payani, Yuan-Sen Ting, and Liang Zheng. Effective training data synthesis for improving mllm chart understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. URL https://arxiv.org/abs/1910.01442.

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, 2024. URL https://arxiv.org/abs/2404.16006.

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, 2024. URL https://arxiv.org/abs/2309.12284.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL https://arxiv.org/abs/2409.02813.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yasher Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning, 2023. URL https://arxiv.org/abs/2303.16406.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024a. URL https://arxiv.org/abs/2403.14624.

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. Vcr: A task for pixel-level complex reasoning in vision language models via restoring occluded text, 2025a. URL https://arxiv.org/abs/2406.06462.

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, 2024b. URL https://arxiv.org/abs/2305.10415.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans?, 2025b. URL https://arxiv.org/abs/2408.13257.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025c. URL https://arxiv.org/abs/2410.02713.

Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization, 2019. URL https://arxiv.org/abs/1712.09374.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding, 2024. URL https://arxiv.org/abs/2406.08100.

Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context, 2020. URL https://arxiv.org/abs/2005.00589.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. URL https://arxiv.org/abs/2406.04264.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos, 2017. URL https://arxiv.org/abs/1703.09788.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016. URL https://arxiv.org/abs/1511.03416.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, 2019. URL https://arxiv.org/abs/1903.08225.

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, 2025. URL https://arxiv.org/abs/2411.00836.