Nemotron 3 Nano Omni: Efficient and Open Multimodal Intelligence

NVIDIA

Abstract. We introduce Nemotron 3 Nano Omni, the latest model in the Nemotron multimodal series and the first to natively support audio inputs alongside text, images, and video. Nemotron 3 Nano Omni delivers consistent accuracy improvements over its predecessor, Nemotron Nano V2 VL, across all modalities, enabled by advances in architecture, training data and recipes. In particular, Nemotron 3 delivers leading results in real-world document understanding, long audio-video comprehension, and agentic computer use. Built on the highly efficient Nemotron 3 Nano 30B-A3B backbone, Nemotron 3 Nano Omni further incorporates innovative multimodal token-reduction techniques to deliver substantially lower inference latency and higher throughput than other models of similar size. We are releasing model checkpoints in BF16, FP8, and FP4 formats, along with portions of the training data and codebase to facilitate further research and development.

Model-BF16 | Model-FP8 | Model-FP4 | Dataset | Megatron-Bridge | NeMo-RL | Example Data Pipeline

1. Introduction

In this work, we present Nemotron 3 Nano Omni, an efficient omni-modal model built on the Nemotron 3 Nano 30B-A3B (NVIDIA et al., 2025a) language model backbone, augmented with the C-RADIOv4-H[1] (Ranzinger et al., 2026; Heinrich et al., 2025) vision encoder and the Parakeet-TDT-0.6B-v2[2](Xu et al., 2023; Rekesh et al., 2023; Sekoyan et al., 2025) audio encoder. Nemotron 3 Nano Omni extends the Nemotron multimodal family with native audio support and improved reasoning capability across all supported modalities. It is particularly effective in practical multimodal settings, including real-world document understanding, long audio-video comprehension, and agentic computer use. In addition, Nemotron 3 Nano Omni incorporates innovative multimodal token-reduction techniques that substantially reduce inference latency and increase throughput, enabling efficient deployment without sacrificing model quality.

Compared to our previous release, Nemotron Nano V2 VL (NVIDIA et al., 2025c), Nemotron 3 Nano Omni introduces several key design choices and architectural advances:

1. **Improved LLM Backbone.** We replace the dense Nemotron Nano V2 12B hybrid backbone with the Nemotron 3 Nano 30B-A3B Mixture-of-Experts (MoE) hybrid backbone, enabling more efficient processing of long multimodal sequences and higher inference throughput.
2. **Native Audio Support.** We extend the model to natively support audio inputs in addition to text, images, and video.
3. **Dynamic Image Resolution.** We replace the tiling-based image processing approach with a dynamic resolution strategy that better preserves native aspect ratios.
4. **Temporal Video Compression.** We introduce Conv3D-based temporal compression for video, achieving a 2× reduction in temporal tokens.
5. **Extended Context Length.** We increase the maximum context length from 128K to 256K tokens, improving performance on long-context multimodal reasoning tasks.

[1] https://huggingface.co/nvidia/C-RADIOv4-H
[2] https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2

Training an omni-modal MoE model introduces challenges in modality alignment, training stability, and data balancing across heterogeneous sources. To preserve the strong text reasoning capabilities of the base LLM while improving multimodal performacne, we adopt a multi-stage training strategy that progressively introduces new modalities and scales context length. This staged approach mitigates catastrophic forgetting and stabilizes cross-modal alignment during training.

Driven by these technical improvements, Nemotron 3 Nano Omni achieves substantial gains over Nemotron Nano V2 VL across a wide range of tasks. In particular, it attains leading results in document understanding, audio-visual reasoning, and audio benchmarks, ranking at or near the top of leaderboards such as OCRBench-V2 (Liu et al., 2024), MMLongBench-DOC (Ma et al., 2024), VoiceBench (Chen et al., 2024a), WorldSense (Hong et al., 2026), and DailyOmni (Zhou et al., 2026).

These improvements also translate into higher inference efficiency and lower latency. On NVIDIA B200, Nemotron 3 Nano Omni achieves 3× higher single-stream output token throughput than Qwen3-Omni (Xu et al., 2025) and 9× higher output token throughput per GPU at a fixed interactivity target. Compared with Nemotron Nano V2 VL, Nemotron 3 Nano Omni provides 3× higher throughput at the same interactivity target and 2× higher single-stream output token throughput. Nemotron 3 Nano Omni ranks as the most cost-efficient open video understanding model on MediaPerf.

Along with this report, we are releasing the model checkpoints on HuggingFace:

- Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16
- Nemotron-3-Nano-Omni-30B-A3B-Reasoning-FP8
- Nemotron-3-Nano-Omni-30B-A3B-Reasoning-NVFP4

We are also releasing part of our training datasets, pipelines, and code:

- Nemotron-Image-Training-v3: A collection of ∼6.9 million training samples.
- Examples of data generation pipelines
- Training code on Megatron-Bridge
- NeMo RL guide for Nemotron 3 Nano Omni

The remainder of this paper is organized as follows. Section 2 describes the model architecture. Section 3 details the training pipeline and datasets. Section 4 presents evaluation results across all modalities.

2. Model Architecture

Our model follows an encoder-projector-decoder design, combining the Nemotron 3 Nano 30B-A3B (NVIDIA et al., 2025a) language model with modality-specific encoders for vision and audio, connected via MLP projectors. An overview of the architecture is shown in Figure 1. The vision encoder is based on C-RADIOv4-H (Ranzinger et al., 2026; Heinrich et al., 2025), while the audio encoder is initialized with Parakeet-TDT-0.6B-v2 (Xu et al., 2023; Rekesh et al., 2023; Sekoyan et al., 2025).

To handle varying image resolutions, we replace the tiling strategy used in Nemotron Nano V2 VL (NVIDIA et al., 2025c) with dynamic resolution processing that preserves the native aspect ratio. Each image is decomposed into a variable number of 16 × 16 patches, with the total number of visual tokens per image constrained between 1,024 and 13,312. This equates to an image size of 512 × 512 and 1840 × 1840, respectively, for square images. Prior to projection, we apply a pixel shuffle operation with 4× downsampling to reduce the token count presented to the language model. For video frames, we use a Conv3D patch embedder that compresses every two frames into one, leading to a 2× reduction in the total number of tokens for video inputs.

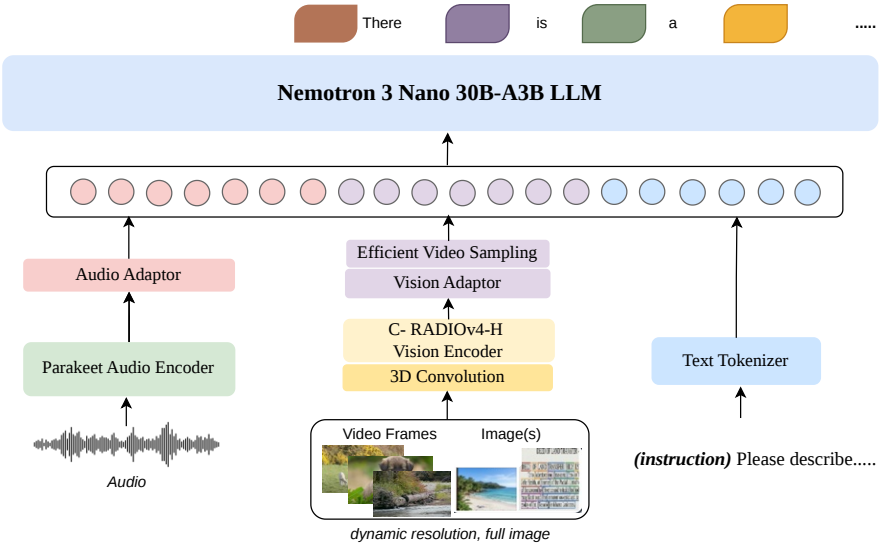

Figure 1 | Nemotron 3 Nano Omni architecture. For encoding images and videos we use dynamic resolution. Additionally, videos use Conv3D and optionally Efficient Video Sampling for higher throughput. Audio inputs are encoded using Parakeet v2 audio encoder. Visual, audio, and text tokens are concatenated and fed to the LLM.

Audio inputs are resampled to 16 kHz mono and encoded using the Parakeet-TDT-0.6B-v2 FastConformer encoder. We first compute log-mel spectrogram features with a 10 ms hop size, followed by three stride-2 convolutional subsampling layers, resulting in an overall $\sim 8\times$ temporal downsampling. This yields approximately 12.5 tokens per second of audio (i.e., $\sim$80 ms per token). Audio streams are segmented into 30-second clips (corresponding to $\sim$375 tokens per clip) with the last clip accounting for the remainder. Streams shorter than 30 seconds are not padded. We train the model to handle inputs ranging from 0.5 second to 20 minutes, but the model context length can accommodate audio input of over 5 hours.

For multimodal inputs containing both visual and audio streams (e.g., videos with audio), modality tokens are interleaved in temporal order during sequence construction to enable joint temporal reasoning across modalities.

3. Training Recipe & Datasets

Training an omni-modal reasoning model with heterogeneous encoders requires careful orchestration. To this end, we adopt a staged training strategy that first performs supervised fine-tuning (SFT) to progressively align modalities, improve multi-modal instruction-following and extend context capacity, followed by reinforcement learning (RL) to further refine reasoning and safety. Figure 2 illustrates the overall progression of these stages.

3.1. SFT

Our SFT pipeline is split into seven stages that progressively introduce new modalities and increase context length. This curriculum is designed to promote stable cross-modal alignment and mitigate catastrophic forgetting while improving multi-modal understanding. Detailed descriptions of each stage are provided in Sections 3.1.1–3.1.7, and an overview is provided in Table 1.

3.1.1. Stage 0: Vision projector warmup

We begin by training only the vision MLP projector to align the vision and language modalities with a maximum context length of 16384, while keeping all other components frozen. This stage

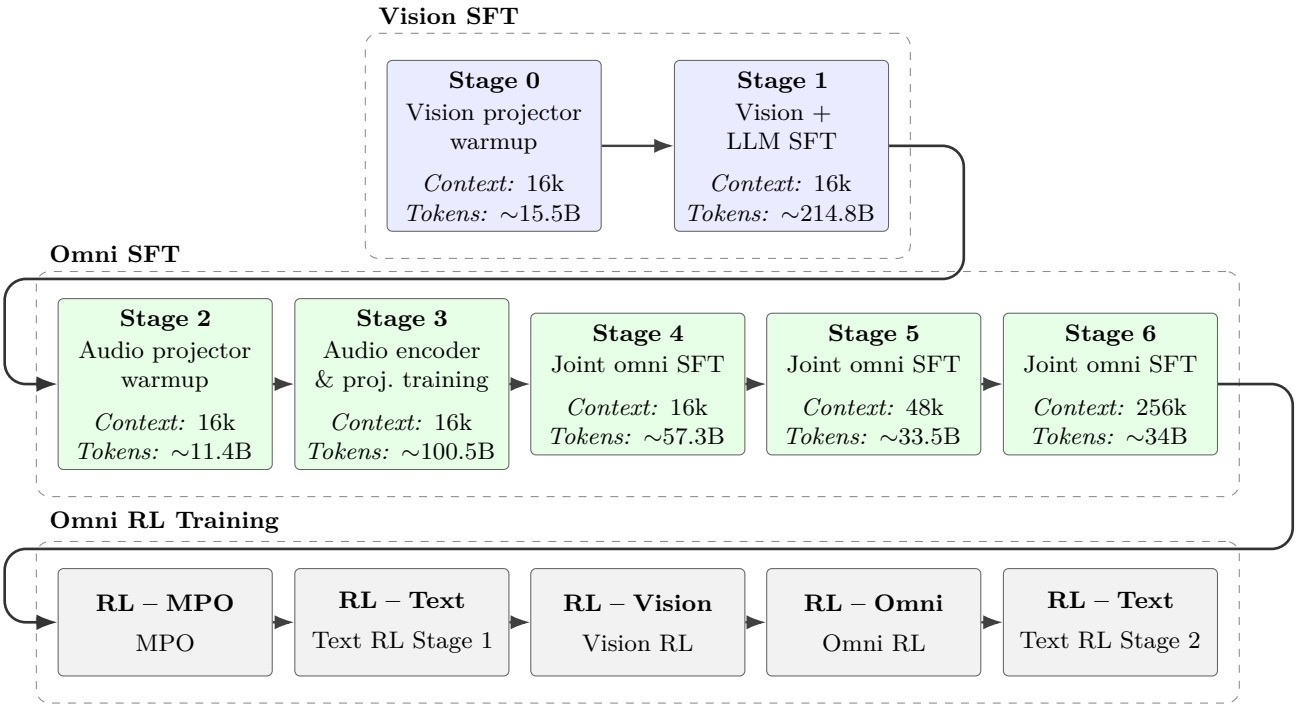

Figure 2 | Staged training recipe for the v3 omni-modal model. The pipeline first performs vision SFT, then joint omni SFT while progressively extending context length, followed by omni-modal RL training.

uses approximately 9.35 million vision–text samples (∼ 15.5B tokens), including a portion of the Stage 1 dataset (see Section 3.1.2) and covering a diverse set of tasks, including image captioning, visual grounding, OCR, document understanding, GUI understanding, and general visual question answering.

3.1.2. Stage 1: Vision SFT 16k

After training the vision projector, we unfreeze both the language model and the vision encoder for joint vision–language fine-tuning. During this stage, the model develops its core vision-language capabilities. The training data builds upon the SFT Stage 1 dataset used in Nemotron Nano V2 VL (NVIDIA et al., 2025c), with several key enhancements.

First, we replace its text-only subset with a portion of the SFT dataset from Nemotron 3 Nano 30B-A3B, resulting in higher-quality text reasoning samples. Second, we improve label quality by re-annotating noisy subsets using models from the Qwen3-VL series (Yang et al., 2025). Third, we enhance the availability and quality of reasoning traces by incorporating both human-annotated and model-generated chains of thought, leveraging models from the Qwen3-VL (Yang et al., 2025), Qwen3.5 (Qwen Team, 2026), and Kimi-K2.5 (Kimi Team et al., 2026) families.

Finally, we expand coverage across domains, including GUI understanding, visual grounding, charts, tables, document understanding, and video understanding, as well as across multiple languages. This is achieved through a combination of publicly available datasets, as well as internally curated data, including human annotation. To increase domain coverage, we additionally develop fully-synthetic data pipelines ensuring broad representation across domains, question types, and visual diversity. Guided by the gaps identified in the training blend, we source relevant data and generate synthetic question-answer pairs at scale using frontier open-source models such as Qwen3-VL (Yang et al., 2025), Qwen 3.5 (Qwen Team, 2026), GPT-OSS (OpenAI, 2025), Nemotron-Parse (Chumachenko

	Number of Samples	Number of Tokens	Primary Data Domains
Stage 0	9.35M	15.5B	Captioning, OCR, document, VQA
Stage 1	86.3M	214.8B	Comprehensive vision-language SFT
Stage 2	59.2M	11.4B	ASR (Granary)
Stage 3	242.0M	100.5B	ASR, sound, music, speech understanding
Stage 4	30.5M	57.3B	Vision, video, audio, text, omni, safety
Stage 5	6.08M	33.5B	Long video, omni, reasoning
Stage 6	623K	34.0B	Ultra-long documents, long-context text
Total (all stages)	**434.1M**	**466.9B**	

Table 1 | Approximate values for the total number of samples and tokens (including masked tokens from the prompt) in the training datasets across the SFT stages. This includes any sample repetitions.

Dataset type	Number of samples	% of total tokens	Number of tokens
ASR	113.8M	22.7%	22.8B
Sound understanding	61.0M	24.4%	24.5B
Music understanding	19.8M	43.3%	43.5B
Speech understanding	47.5M	9.6%	9.6B
Total	**242.0M**		**100.5B**

Table 2 | Dataset composition for the audio pretraining stage.

et al., 2025), and DeepSeek-OCR (Wei et al., 2025). For each domain, we generate question-answer pairs from images, videos, or OCR extracted from images using domain-specific instructions. This is followed by distillation of reasoning traces and strict filtering of the resulting samples to ensure data correctness, usefulness, and overall quality.

The resulting dataset comprises approximately 86.3M samples ($\sim$ 214.8B tokens), including sample repetitions.

3.1.3. Stage 2: Audio projector warmup

Analogous to Stage 0 for vision, this stage warms up the audio MLP projector (Chen et al., 2024b) while keeping the LLM, vision encoder, and Parakeet-TDT audio encoder all frozen.

The training data consists of the Granary v1.1 ASR dataset (Koluguri et al., 2025), comprising approximately 59.2M samples ($\sim$11.4B tokens) of diverse automatic speech recognition data across varied acoustic conditions, speaking styles, and languages.

3.1.4. Stage 3: Audio projector & encoder

Building on Stage 2, this stage unfreezes the Parakeet-TDT audio encoder while keeping the LLM backbone and vision encoder frozen. The audio encoder and its associated projector are jointly trained on an expanded audio corpus.

As shown in Table 2, this stage is trained using a mixture of ASR data along with sound, music, and speech understanding. Audio samples are paired with captions, multiple-choice questions, and open-ended questions, with a subset further augmented with reasoning traces. Our synthetic data generation pipeline leverages open models like Qwen3-Omni-30B-A3B to produce captions and specialized music tools to produce metadata. These outputs are then used to generate QA pairs via GPT-OSS-120B.

Dataset type	Number of samples	% of total tokens	Number of tokens
Vision data	14.6M	53.4%	30.6B
Text data	948K	6.1%	3.5B
Text safety data	14K	0.02%	10.4M
Image safety data	9K	0.02%	10.0M
Short video data	1.3M	11.0%	6.3B
Short video reasoning data	388K	4.2%	2.4B
Short video omni data	251K	2.8%	1.6B
ASR data	2.9M	1.1%	640M
Audio reasoning data	765K	4.4%	2.5B
Audio data	9.3M	16.9%	9.7B
Total	**30.5M**		**57.3B**

Table 3 | Dataset composition for Stage 4: Joint Omni SFT at 16k context length.

Category	Number of samples	% of total tokens	Number of Tokens
ASR	650K	0.4%	0.12B
Audio	2.84M	11.3%	3.80B
Vision	1.17M	9.8%	3.28B
Text	101K	7.2%	2.42B
Safety	45K	0.1%	0.04B
Video (short)	25K	0.6%	0.21B
Video (medium)	96K	5.8%	1.95B
Video (long)	74K	3.3%	1.11B
Video reasoning	167K	10.2%	3.42B
Omni (short)	6K	0.3%	0.09B
Omni (medium+long)	710K	39.1%	13.10B
Omni reasoning	198K	11.8%	3.94B
Total	**6.08M**	**100%**	**~33.5B**

Table 4 | Stage 5 (Omni SFT 48k) data composition by category.

3.1.5. Stage 4: Omni SFT 16k

This is the first stage that jointly trains on all modalities. All model parameters, including the LLM backbone, are trainable. The data mixture combines vision SFT, text instruction following, safety, video understanding, omni (audio+video) QA and captioning, ASR, and audio reasoning data.

Table 3 summarizes the data composition. The dominant sources are the vision dataset (30.6B tokens), the audio dataset (9.7B tokens), and the short video data (6.3B tokens). The omni-modal data used in this stage is a blend of audio-visual captions, open-ended QA and MCQ style QA. Videos less than 2 minutes length are used as source media for this data. The question-answer pairs and captions are synthetically generated by first extracting audio-visual metadata from videos and then using that metadata for question-answer generation and summarization using open-source models Qwen3-Omni-30B-A3B and GPT-OSS-120B. The audio reasoning dataset comprises speech-to-text conversations synthesized by converting text SFT user turns into spoken form and generating LLM responses to a curated subset of ASR prompts.

3.1.6. Stage 5: Omni SFT 48k

This stage extends the context length to 49,152 tokens with all model parameters trainable. The data mixture is rebalanced to emphasize longer sequences, with reduced sampling of short-context data and increased weight on medium and long video, omni, and reasoning data.

Table 4 shows the per-category breakdown. Compared to Stage 4, this stage has a much higher

Dataset type	Number of samples	% of total tokens	Number of tokens
Long Context Vision	508K	90.9%	30.9B
Text	63K	7.3%	2.5B
Long Context Text	2.2K	1.5%	506M
Vision	50K	0.3%	106M
Total	**623K**	**100%**	**34.0B**

Table 5 | Dataset composition for the ultra-long context Stage 6.

proportion of long-context data: medium and long video, omni data with joint audio-video understanding, and reasoning traces. Short video and omni data are downsampled, while medium/long omni data and reasoning data receive the bulk of the training budget.

For the 48k SFT stage, omni-modal data comprising reasoning and non-reasoning single-turn QA is synthesized from diverse domains and categories. The pipeline segments videos into 20-second clips, extracts audio-visual metadata using multimodal models such as Qwen3-Omni-30B-A3B, and generates QA pairs and reasoning traces via open-source reasoning models such as GPT-OSS-120B

3.1.7. Stage 6: Omni SFT 256k

This stage extends the context length to 262,144 and is intended to significantly increase the model's long context capabilities. The data for this stage consists of ∼34.0B tokens across long-context text-only and vision domains such as long-context reasoning and long document understanding (see Table 5). It particularly improves the model's ability to analyze documents spanning 10 to 100+ pages, including reasoning over text, charts, and complex tables. We assemble a diverse collection of long-form documents, including academic papers, financial reports, and presentations, and leverage vision-language models to generate synthetic question-answer pairs and reasoning traces at the page, multi-page, and full-document levels. To support long-document understanding, we release runnable data pipeline recipes[3] using NeMo Data Designer (The NeMo Data Designer Team, 2025).

The audio encoder and projector are frozen during this stage to focus model capacity on long-context text and document understanding.

3.1.8. Training Details

All SFT stages are trained using the Megatron framework (Shoeybi et al., 2019)[4] , together with Transformer Engine[5] and the Megatron Energon[6] dataloader. Training is conducted on 32–128 nodes of NVIDIA H100 GPUs, depending on the stage.

We employ 2-way tensor parallelism (TP), 32-way expert parallelism (EP), and sequence parallelism to efficiently scale training. All stages are trained in BF16 mixed precision and use online sequence packing with a balanced greedy knapsack algorithm to maximize GPU utilization.

To fit long sequences in GPU memory, we use selective activation recomputation for the LLM backbone (recomputing core attention, MLP, LayerNorm, and MoE activations) and full block-level recomputation for all 32 vision encoder layers. Sound model activations are recomputed starting from Stage 4. Vision projection and sound projection recomputation are enabled from Stage 5 onward to support the increased memory requirements of longer sequences. Additionally, context parallelism is introduced in later stages, with 2-way and 16-way CP in Stages 5 and 6, respectively, to accommodate increasingly long sequence lengths.

[3] https://github.com/NVIDIA-NeMo/DataDesigner/tree/main/docs/assets/recipes/vlm_long_doc
[4] https://github.com/NVIDIA/Megatron-LM
[5] https://github.com/NVIDIA/TransformerEngine
[6] https://github.com/NVIDIA/Megatron-Energon

	Stage 0	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
Context Length	16K					48k	256k
Max Video Frames	–	64	–	–	64	256	256
Global BS	128	256	512			256	128
CP	-					2	16
LR	10^{-3}	5×10^{-5}	10^{-3}	2.5×10^{-5}	10^{-5}	10^{-6}	
Minimum LR	10^{-5}	0	10^{-5}	0	10^{-7}	0	
Linear Warmup Fraction	0.1					0.01	0.1
Weight Decay	0.01	0.05	0.01	0.05			
Trainable Modules	Vision Projector	All except audio	Audio Projectior	Audio Encoder & Projector	All		All except audio
# GPU Nodes	32	64		128	64		

Table 6 | Summary of the SFT training hyperparameters. All stages use the AdamW optimizer (β_1=0.9, β_2=0.999), a cosine LR decay, BF16 precision, TP=2 and EP=32

The vision encoder's CPE layers are kept in eval mode in stages 1, 4 and 5 to stabilize training. For videos, we sample up to 64 frames in Stages 1 and 4, and up to 256 frames in Stages 5 and 6. We also employ video augmentation that randomly selects the target number of patches per video frame from $\{256, 512, 768, 1024\}$. This allows us to reduce the image resolution at inference time, while scaling up the number of frames, to improve temporal information without increasing the number of tokens. 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. Table 6 summarizes the training hyperparameters for SFT stages.

3.2. Reinforcement Learning

After SFT, we apply multiple rounds of reinforcement learning to further improve instruction following, reasoning, and safety-alignment for text, image, and video modalities. We design a curriculum learning pipeline for post-training: (1) Preference Optimization, (2) Text-RL-stage-1, (3) Image-RL, (4) Omni-RL, and (5) Text-RL-stage-2.

3.2.1. Preference Optimization

To align our model using both preference-level and quality-level supervision, we adopt Mixed Preference Optimization (MPO) (Wang et al., 2024a), which combines a preference loss and a quality loss during the offline reinforcement learning stage. Specifically, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) as the preference loss and Binary Classifier Optimization (BCO) (Wang et al., 2024a) as the quality loss. To construct the training data, we apply rejection sampling to generate candidate responses in the vision domain and assign binary labels based on outcome correctness, yielding positive samples for accepted responses and negative samples for rejected ones.

3.2.2. Text-RL

During text-only RL, we only train the LM parameters of the model via multi-environment RLVR/RLHF for improving general capabilities. We reuse the RL data and infrastructure from the post-training of Nemotron 3 Nano and Super (NVIDIA et al., 2025b, 2026). As part of our staged multi-modal training, during text-only RL stages we additionally freeze the LM input token embedding parameters to mitigate representational drift between multi-modal stages.

3.2.3. Image RL

ImageRL is the first stage of our multimodal RL pipeline. We employ outcome-based RL on visual reasoning tasks, which can be divided into the following categories

- *Chart, document, and text-rich image reasoning*: numerical, comparative, and trend reasoning over plots, tables, diagrams, infographics, and natural images containing text (∼28K).
- *STEM and mathematical problems*: geometry, algebra, functions, and counting, in both English and Chinese (∼19K).
- *Game and puzzle reasoning*: rule-based reasoning over rendered game-board states (∼12K).
- *Visual question answering*: open-ended and multiple-choice questions covering spatial relations, attribute recognition, and yes/no judgements (∼8K).
- *Visual grounding*: click-coordinate prediction on desktop, mobile, and web screenshots (∼7K).

During training, each prompt is graded by a $[0, 1]$ scalar that linearly combines an outcome score. The outcome score comes from one of four rule-based verifiers, chosen per prompt: *string-match* for free-form text answers, *mathruler* for symbolic equivalence on numeric and algebraic answers, *multiple-choice* for selected-letter answers, and *gui-coordinate* for click-target predictions, where the reward decays smoothly with distance from the target. The format score rewards a single `<think>` reasoning block followed by a single `\boxed` answer, with partial credit when the policy emits extra reasoning or boxed entries. This keeps correct answers from being zeroed out by the surface format errors common in VLM checkpoints after SFT, while still discouraging verbose multi-answer outputs.

To ensure an informative learning signal, we apply pass-rate filtering using 8 rollouts per prompt from the initial policy checkpoint, retaining only prompts whose empirical pass rate is below 0.8; prompts that are trivially solvable at initialization are discarded. The filtering is based on the same verifiers that are used during training. We additionally include a small set of unanswerable or image-text-mismatched prompts to train the policy to abstain when visual evidence is insufficient.

The resulting corpus and verifier suite are inherited by OmniRL as the image component of its mixed-modality training mixture.

3.2.4. Omni-RL

Understanding is inherently challenging, and extending it to multiple modalities further increases complexity due to the need for sophisticated cross-modal reasoning. Prior advances in text reasoning have demonstrated the effectiveness of structured reasoning in improving model performance (Wei et al., 2022; Shao et al., 2024). More recently, omni-modal reasoning has also been shown to be beneficial to omni and video tasks (Ye et al., 2026). Motivated by these findings, we develop a unified reinforcement learning training stage aimed at enhancing the model's capacity for coherent reasoning across images, videos, and audio modalities.

To make omni RL training possible, we curate a diverse, omni-modal training corpus of approximately 120K prompts spanning 113 sub-datasets across four modality groups: image, video, audio, and text-only reasoning. The dataset is constructed by aggregating and filtering data from multiple sources: **(1)** *Omni RL data* (∼17.6K samples): synthetic data generated from video content with accompanying audio, covering diverse visual understanding and temporal reasoning tasks; **(2)** *Video RL data* (∼8.5K): video-only question–answer pairs targeting spatial, temporal, and causal reasoning. **(3)** *Image RL data* (∼32K): a large-scale image understanding set drawing from OCR (∼10.5K), chart analysis (∼8.9K), game-related visual QA (∼11.9K), GUI grounding (∼7.1K), and additional curated domains; **(4)** *Audio RL data* (∼4.2K) and *ASR* (∼3.8K): audio question-answering and automatic speech recognition tasks at various utterance lengths. We incorporate an ASR verifier to stabilize the speech recognition capability of our model. The reward is 1 - WER, where WER is computed after text normalization.

To ensure balanced difficulty and effective learning signals, we apply pass-rate filtering based on the initial policy checkpoint. We retain only prompts on which the base model achieves a pass rate between 0.1 and 0.9 (with stricter 0.3–0.7 bands for AudioQA), thereby excluding prompts that are either trivially solvable or entirely intractable for the current policy. The verification pipeline supports five task types: multiple-choice (34%), string matching (31%), mathematical rule-based verification (26%), GUI coordinate grounding (6%), and ASR evaluation (3%). We additionally include a small set of unanswerable or mismatched samples (∼4K) to train the model to appropriately abstain when evidence is insufficient.

3.2.5. RL Training Details

Training is conducted on NVIDIA B200 and H100 GPU cluster using a Ray-based distributed training framework built on NeMo-RL (NVIDIA, 2025). The global batch size is set to 4,096 with 16 rollouts for each prompt and a micro-batch size of 1. We apply an adapted version of Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025; Shao et al., 2024) as the RL training algorithm.

We use a multimodal deduplication strategy during the generation phase to leverage a unique multimodal tensor with rollouts associated with each prompt. We leverage tensor, expert, and context parallelism during training. All experiments are run with the AdamW optimizer with β_1 and β_2 set to 0.9 and 0.999, respectively, and a linear warmup.

4. Experiments

In Sections 4.1–4.4, we conduct a comprehensive evaluation of the model's ability to reason over vision, audio, and text inputs, and present the corresponding results. In Section 4.6, we analyze the efficiency gains achieved through Efficient Video Sampling (EVS) for video inputs. In Sections 4.7 and 4.8, we examine the impact of quantization on model accuracy and efficiency.

For the vision and audio evaluations in Sections 4.1–4.3, we use the VLMEvalKit (Duan et al., 2025)[7] framework with a vLLM (Kwon et al., 2023)[8] backend. Text evaluations are conducted using the NeMo-Skills[9] framework.

4.1. Visual Evaluations

We conduct a comprehensive evaluation of our model on the following broad categories:

1. **STEM Reasoning**: MMMU (Yue et al., 2024), MathVista-Mini (Lu et al., 2024)
2. **Document Understanding, OCR & Charts**: MMLongBench-Doc (Ma et al., 2024), OCRBench(Liu et al., 2024), OCRBench-V2 (Fu et al., 2024), ChartQA (Masry et al., 2022), 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., 2025), CharXiv (Wang et al., 2024b)
3. **Visual Grounding & Spatial Reasoning**: TreeBench (Wang et al., 2025), CV-Bench (Tong et al., 2024), RefCOCO(Kazemzadeh et al., 2014)
4. **GUI Understanding**: ScreenSpot (Cheng et al., 2024), ScreenSpot-v2 (Wu et al., 2024), ScreenSpot Pro (Li et al., 2025), OSWorld (Xie et al., 2024)
5. **Video Understanding**: Video-MME (Fu et al., 2025)

As shown in Table 7, we observe significant improvements compared to Nemotron Nano V2 VL across all benchmarks and even outperforms Qwen3-Omni on several categories.

[7] https://github.com/open-compass/VLMEvalKit
[8] https://github.com/vllm-project/vllm
[9] https://github.com/NVIDIA-NeMo/Skills

Task	Benchmark	Nemotron 3 Nano Omni		Nemotron Nano V2 VL		Qwen3-Omni	Qwen3.5-Omni
Open-Source		✓		✓		✓	✗
Size		30B-A3B		12B		30B-A3B	Flash
Mode		Reasoning off	Reasoning on	Reasoning off	Reasoning on		
STEM Reasoning	MMMU (val)	55.2	70.8	55.3	67.8	75.6	**76.9**
	MathVista-Mini	71.9	82.8	69.0	75.5	80.0	**82.9**
Document Understanding, OCR & Charts	MMLongBench-Doc	46.1	**57.5**	32.1	38.0	49.5	53.6
	OCRBench	88.3	86.6	85.6	83.5	86.0	**89.1**
	OCRBenchV2 (EN/ZH)	65.8/52.0	**67.0/52.7**	62.0/44.2	54.8/39.8	-	-
	ChartQA (Test)	89.9	**90.3**	89.8	84.9	89.5	-
	DocVQA (Test)	93.3	**95.6**	94.7	93.2	95.3	-
	AI2D (Test)	88.5	88.5	87.2	84.7	86.62	**89.0**
	TextVQA (Val)	85.1	81.0	**85.4**	76.1	81.7	-
	InfoVQA (Test)	83.6	**86.8**	79.4	80.4	83.31	-
	OCR-Reasoning	22.2	**54.14**	21.0	33.9	49.9	-
	CharXiv (RQ/DQ)	49.1/81.9	63.6/**88.9**	41.7/76.5	41.3/77.2	61.1/-	**64.4**/-
Visual Grounding & Spatial Reasoning	TreeBench	43.7	**51.6**	38.5	42.5	-	-
	CV-Bench	**84.2**	84.0	81.0	78.3	-	-
	RefCOCO	80.6	90.5	-	-	-	**92.6**
GUI	ScreenSpot	**90.3**	89.3	39.4	42.5	-	-
	ScreenSpot-v2	**93.4**	92.8	41.7	42.8	-	-
	ScreenSpot-Pro	59.3	57.8	4.8	5.5	**59.7**	-
	OSWorld	-	**47.4**	-	11.1	29.0	-
Video Understanding	VideoMME (w/o sub)	70.8	72.2	66.0	63.0	70.5	**77.0**

Table 7 | Comparison of Nemotron 3 Nano Omni with our previous release, Nemotron Nano V2 VL, as well as other state-of-the-art omni-modal models.

4.2. Audio Evaluations

We evaluate our model across three broad categories:

1. **Automatic Speech Recognition (ASR):** We use the OpenASR leaderboard (Srivastav et al., 2026), and report word error rate on its English subset, including AMI, Earnings22, GigaSpeech, LibriSpeech, SPGISpeech, TED-LIUM and VoxPopuli. For long-form ASR, we additionally evaluate on TED-LIUM Longform (Fox et al., 2024), which tests transcription quality and long-context consistency on continuous speech.
2. **Audio Understanding:** We evaluate on MMAU (Sakshi et al., 2024), a benchmark of ∼10k audio clips with QA pairs spanning speech, environmental sounds, and music, covering 27 skills in information extraction and multi-step reasoning.
3. **Voice Interaction & Reasoning:** We use VoiceBench (Chen et al., 2024a), which assesses LLM-based voice assistants on realistic spoken interactions, evaluating knowledge, instruction following, and safety across diverse speakers and environments.

As shown in Table 8, Nemotron 3 Nano Omni outperforms Qwen family models on ASR and VoiceBench benchmarks.

4.3. Audio-Visual Evaluations

We evaluate our model on audio-visual perception and reasoning using two complementary benchmarks:

1. **DailyOmni** (Zhou et al., 2026): an audio-visual QA benchmark for cross-modal reasoning in daily scenarios, with 684 videos (segmented into 30 and 60 secoond clips) and 1,197 multiple-choice questions across six tasks, testing temporal alignment, event understanding, causal reasoning, and cross-modal consistency.
2. **WorldSense** (Hong et al., 2026): a large-scale omni-modal benchmark with 1,662 long-context videos and 3,172 multiple-choice questions across 26 tasks, evaluating long-range dependencies,

Task	Benchmark	Subtask	Nemotron 3 Nano Omni	Qwen3-Omni	Qwen3.5-Omni
Open-Source			✓	✓	✗
Size			30B-A3B	30B-A3B	Flash
ASR	OpenASR (Reasoning off)	AMI	**11.09**	12.52	–
		Earnings22	**11.27**	12.3	–
		GigaSpeech	9.66	**8.49**	–
		LibriSpeech (clean)	1.57	1.52	**1.3**
		LibriSpeech (other)	2.96	3.22	**2.4**
		SPGISpeech	**1.98**	3.69	–
		TED-LIUM	3.44	**2.38**	–
		VoxPopuli	**5.6**	8.26	–
		OpenASR Avg	**5.95**	6.55	–
Long-form ASR	TED-LIUM (Reasoning off)	–	3.11	**2.4**	–
Audio Understanding	MMAU (Reasoning off)	Music	74.2	–	–
		Audio	76.9	–	–
		Speech	72.8	–	–
		MMAU Avg	74.6	77.5	**80.4**
Voice Interaction	VoiceBench (Reasoning on)	IFEval	**88.7**	80.6	–
		BBH	**91.1**	88.9	–
		AdvBench	**100**	97.2	–
		AlpacaEval	95.0	**96.4**	–
		CommonEval	**91.3**	90.5	–
		WildVoice	**91.7**	90.5	–
		OpenBookQA	93.0	**94.3**	–
		MMSU	82.3	**83.0**	–
		SD-QA	71.4	**78.1**	–
		VoiceBench Avg	**89.4**	88.8	87.8

Table 8 | Comparison of Nemotron 3 Nano Omni with other state-of-the-art open source models on diverse audio and speech tasks, ASR (OpenASR), long-form ASR (TED-LIUM), MMAU, and VoiceBench. ASR tasks are measured by word error rate (lower is better). For MMAU and VoiceBench, higher is better. ASR and MMAU use non-reasoning settings, while VoiceBench uses reasoning.

sound grounding, temporal reasoning, and complex cross-modal inference.

As shown in Table 9, Nemotron 3 Nano Omni outperforms Qwen3-Omni on both reasoning on and off modes.

Benchmark	Nemotron 3 Nano Omni		Qwen3-Omni		Qwen3.5-Omni
Open-Source	✓		✓		✗
Size	30B-A3B		30B-A3B		Flash
Mode	Reasoning off	Reasoning on	Instruct	Thinking	
DailyOmni	74.5	74.1	71.9	73.6	**81.8**
WorldSense	55.2	55.4	54	-	**57.8**

Table 9 | Comparison of Nemotron 3 Nano Omni with other state-of-the-art open source models on Video+Audio (Omni) benchmarks, measured by accuracy (higher is better).

4.4. Text-only evaluations

We conduct all pure-text evaluations with a maximum output length of 131,072 tokens, temperature set to 1.0, and top-p of 1.0. We report Pass@1 average of 8 runs for AIME-2025; an average of 4 runs for GPQA-Diamond (Rein et al., 2023); and score of 1 run for SciCode (Tian et al., 2024), LiveCodeBench v5 (07/24 - 05/25) (Jain et al., 2024), IFBench (Zhou et al., 2023), and TauBench V2 (telecom). We additionally include MMLU-Pro to assess general academic and knowledge-intensive reasoning.

Table 10 shows the evaluation on a selected text benchmarks compared to the Nemotron 3 Nano

30B-A3B LLM that is used as the LLM backbone. The goal of the Omni model is maintain text benchmarks of the LLM while adding vision and audio understanding capabilities.

Benchmark	Nemotron 3 Nano Omni	Nemotron 3 Nano 30B-A3B	Qwen3-Omni
Open-Source	✓	✓	✓
Size	30B-A3B	30B-A3B	30B-A3B
MMLU-Pro	77.3	**78.3**	61.6
GPQA (no tools)	72.2	73.0	**73.1**
LiveCodeBench	63.2	**68.3**	-
AIME25 (no tools)	82.1	**89.1**	73.7
IFBench (prompt)	**74.2**	71.5	-
AA-LCR	**41.0**	35.9	-
TauBench V2 (Telecom)	**42.7**	42.2	-
SciCode	32.0	**33.3**	-

Table 10 | Comparison of Nemotron 3 Nano Omni, Nemotron 3 Nano LLM, and Qwen3-Omni across selected text-only benchmarks.

4.5. Reasoning budget control

We study the effect of inference-time reasoning budgets by evaluating model performance under two settings: (1) a base configuration with a maximum sequence length of 16,384 tokens, and (2) a reasoning-enabled configuration with a 13K reasoning budget, a 1,024-token grace period, and a maximum sequence length of 16,384 tokens (Table 11).

Our results suggest that reasoning budget adjustment yields accuracy gains on select benchmarks under reasoning-on mode, with no degradation observed on the remaining ones. 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.

Benchmark	MathVista-Mini	MMLongBench-Doc	DocVQA (Val)	Charxiv(RQ)	RefCOCO	VideoMME
w.o. reasoning budget	80.3	54.5	**95.3**	61.8	90.4	67.5
w. reasoning budget	**82.8**	**56.8**	95.2	**64**	**90.6**	**70.3**

Table 11 | Effect of reasoning budget across several key benchmarks.

4.6. Conv3D and Efficient Video Sampling (EVS)

Nemotron 3 Nano Omni reduces the cost of long video inputs through two stacked mechanisms on the vision side. Conv3D is an architecture change applied during both training and inference: every $T = 2$ consecutive frames are fused into a single "tubelet" before the first ViT block. This halves the number of vision tokens flowing through the ViT and the LLM, cutting both ViT prefill cost and LLM-side prefill, attention compute, and KV-cache footprint. EVS (Efficient Video Sampling)(Bagrov et al., 2025) is a runtime-only feature that drops video tokens after the ViT blocks and the vision adapter, immediately before they reach the LLM. For each spatial position (h, w) it computes the cosine dissimilarity between consecutive tubelets and keeps the globally most-dissimilar tokens up to a budget set by the pruning rate q; the entire first tubelet is pinned to maximum dissimilarity so it is always retained as an anchor. The two mechanisms compose multiplicatively: Conv3D halves the number of tokens in the temporal dimension, EVS prunes the remaining tokens in the spatial dimension.

Table 12 compares the four combinations of Conv3D and EVS for both BF16 and NVFP4 checkpoints, with EVS fixed at $q = 0.5$. Each accuracy column reports per-benchmark scores at 128 and 256 sampled frames, with reasoning off. Accuracy is averaged across three runs with identical settings.

TTFT is averaged across five concurrency-1 `aiperf` runs against a synthetic 512-frame, 512 × 512 video at 30 fps.

Configuration	DailyOmni		LongVideoBench		Video-MME		WorldSense		Avg		TTFT
	128f	256f	128f	256f	128f	256f	128f	256f	128f	256f	(ms)
BF16	74.74	74.77	66.23	67.90	69.13	69.70	54.80	54.50	66.23	66.72	7969
BF16 + EVS	74.46	74.38	65.70	67.80	69.80	70.10	54.87	55.40	66.21	66.92	6452
BF16 + Conv3D	74.41	74.24	66.30	67.20	68.70	70.70	54.83	54.43	66.06	66.64	5984
BF16 + Conv3D + EVS	73.74	73.54	65.70	66.60	68.60	70.70	55.07	54.43	65.78	66.32	5313
NVFP4	71.71	71.68	66.07	66.93	69.23	69.97	53.27	52.45	65.07	65.26	6885
NVFP4 + EVS	71.65	71.76	65.50	67.30	69.80	70.93	52.90	52.60	64.96	65.65	5977
NVFP4 + Conv3D	70.37	70.84	65.90	66.43	68.70	70.30	52.63	52.27	64.40	64.96	5635
NVFP4 + Conv3D + EVS	70.76	70.65	64.97	66.50	68.47	70.17	52.50	52.70	64.17	65.00	5083

Table 12 | Per-benchmark accuracy (128 frames / 256 frames) and TTFT at concurrency 1 across Conv3D and EVS combinations, with EVS rate $q = 0.5$, reasoning off.

Both mechanisms significantly reduce TTFT on BF16: Conv3D alone drops it from 7969 ms to 5984 ms (−25%), EVS alone drops it to 6452 ms (−19%), and stacking them yields 5313 ms (−33% versus the baseline) at a cost of about half a point of average accuracy. The same ordering holds on NVFP4. Underlying these gains is a substantial reduction in the number of input tokens for the LLM: a 512-frame video produces ∼141k input tokens without either mechanism, drops to ∼75k with Conv3D enabled (−47%), and drops further to ∼42k with Conv3D combined with EVS at $q = 0.5$ (−70% versus the baseline).

Table 13 sweeps the EVS pruning rate q on BF16 with Conv3D enabled. Per-benchmark accuracy is essentially flat through $q = 0.7$, slightly reduces at $q = 0.8$, and drops noticeably beyond, with LongVideoBench being the most sensitive benchmark to aggressive pruning. TTFT improves monotonically through the range, for a ∼14% reduction at $q = 0.7$ versus no-EVS.

EVS	DailyOmni		LongVideoBench		Video-MME		WorldSense		Avg		TTFT
q	128f	256f	128f	256f	128f	256f	128f	256f	128f	256f	(ms)
none	74.41	74.24	66.30	67.20	68.70	70.70	54.83	54.43	66.06	66.64	5984
0.5	73.74	73.54	65.70	66.60	68.60	70.70	55.07	54.43	65.78	66.32	5313
0.6	74.41	74.44	65.10	66.50	68.90	70.90	54.57	54.40	65.74	66.56	5173
0.7	73.82	73.77	65.00	65.40	68.70	70.30	54.10	53.80	65.41	65.82	5124
0.8	73.38	73.24	64.30	64.80	67.80	70.10	54.00	53.73	64.87	65.47	5182
0.9	71.54	71.71	59.80	61.00	67.10	68.30	52.97	52.87	62.85	63.47	4883
0.95	69.31	69.31	55.60	57.90	64.60	66.30	51.67	51.93	60.29	61.36	4804

Table 13 | BF16 with Conv3D enabled, varying EVS pruning rate q, reasoning off. Same column structure as Table 12.

4.7. Quantization

Inspired by the quantization recipe from Nemotron 3 Super, we pursued a mixed-precision strategy for FP4: routed MoE experts are quantized to NVFP4 (FP4 E2M1 values with per-block FP8 E4M3 scales over groups of 16 elements and an additional per-tensor FP32 global scale), while the Mamba `in_proj` / `out_proj`, shared experts, and attention `o_proj` are quantized to FP8 (per-tensor E4M3 values with a per-tensor FP32 scale). All remaining language-model layers are left in BF16, as are the vision and audio encoders and their MLP projectors. For the KV cache we use FP8, while the Mamba SSM state cache is kept at FP32 at serving time. This gives a model-weight footprint of 4.98 effective bits per weight (20.9 GB vs the 61.5 GB BF16 reference). For FP8 we quantize every linear layer in the language model to per-tensor E4M3 (with a per-tensor FP32 scale), with the exception of the MoE router and `lm_head`, and pair it with an FP8 KV cache. The vision and audio encoders and their MLP projectors are excluded entirely. This yields ∼8.5 bpw (32.8 GB). We evaluated the

quantized models across 25 text, image, video, and audio benchmarks and found a median accuracy drop of less than 1% vs BF16 for both FP8 and NVFP4.

Benchmark	**BF16**	**FP8**	**NVFP4**
Size (GB)	61.5	32.8	20.9
Effective bpw	16.00	8.5	4.98
MathVista-Mini	71.90	71.05	71.30
CharXiv (RQ)	49.10	48.05	47.95
OCR-Reasoning	22.20	23.43	22.78
MMLongBench-Doc	46.10	45.84	45.78
OCRBenchV2 (EN)	65.80	65.63	65.77
OCRBenchV2 (ZH)	52.00	50.24	50.39
CV-Bench	84.20	85.62	85.27
VideoMME	70.80	69.40	69.60
DailyOmni	74.50	74.06	74.23
WorldSense-AVLM	55.20	54.40	54.60
MMAU	74.62	74.56	74.34
TedLium-Longform (WER↓)	3.11	3.12	3.04
HF-ASR avg, 8 short-form (WER↓)	5.95	5.97	5.95
Mean (11 non-ASR)	**60.58**	**60.21**	**60.18**
Median (11 non-ASR)	**65.80**	**65.63**	**65.77**
Δ vs BF16 (mean)	—	−0.37	−0.40

Table 14 | Accuracy of Nemotron 3 Nano Omni at BF16, FP8, and NVFP4 across the eval suite.

4.8. Inference efficiency

NVFP4. Compared to BF16 precision, NVFP4 on NVIDIA B200 provides up to 7.5× the output token throughput at iso-interactivity (18200 tok/s versus 2400 tok/s, at 150 tok/s/user) on an single-image reasoning usecase.

Low-latency single-stream inference. Nemotron 3 Nano Omni delivers strong single-stream inference performance on NVIDIA B200, reaching more than 500 output tokens/s at a concurrency of 1. This low-latency generation rate is sustained at longer sequence lengths and with larger multimodal inputs, such as long videos or multi-document workloads, due to the hybrid architecture. This is approximately 2.4–2.9× faster than Qwen3-Omni, which reaches 175–210 output tokens/s depending on input size and sequence length, and 2× faster than Nemotron Nano V2 VL, which reaches 250 output tokens/s.

For a multi-document workload, Nemotron 3 Nano Omni achieves a time-to-first-token (TTFT) of approximately 1.3 s, compared to more than 2.5 s for Qwen3-Omni.

High-throughput serving. At maximum concurrency on a single NVIDIA B200, Nemotron 3 Nano Omni reaches 5000 output tokens/s on a multi-document workload. At an iso-interactivity target of 50 output tokens/s per user, the deployment provides 9× higher output throughput than Qwen3-Omni on long-video workloads and 7.5× higher output throughput on multi-document workloads. Compared with Nemotron Nano V2 VL, Nemotron 3 Nano Omni provides 3× higher throughput at the same interactivity target.

Experimental setup. All measurements use a single NVIDIA B200 GPU and vLLM nightly as of 2026-04-19 with EVS 50%. Nemotron 3 Nano Omni is evaluated in NVFP4, Qwen3-Omni with dynamic FP8 quantization, and Nemotron Nano V2 VL in NVFP4. Text ISL=50 and OSL=8000. The multi-document workload contains 32 images at 1024×1536 resolution. The long-video workload contains 512 frames at 512×512 resolution.

5. Conclusion

We introduced Nemotron 3 Nano Omni, an efficient omni-modal model that extends the Nemotron multimodal family with native audio support and consistently stronger reasoning across text, images, video, and audio. Built on the Nemotron 3 Nano 30B-A3B MoE hybrid backbone and augmented with the C-RADIOv4-H vision encoder and the Parakeet-TDT audio encoder, the model combines dynamic image resolution, Conv3D-based temporal video compression, and a 256K context length to process long, heterogeneous multimodal inputs with high accuracy. We use a multi-stage training recipe that progressively introduces new modalities and extends context to enable robust cross-modal alignment while preserving the text reasoning ability of the base LLM.

Across a broad evaluation suite, Nemotron 3 Nano Omni delivers consistent gains over Nemotron Nano V2 VL and achieves leading or competitive results on document understanding (OCRBench-V2, MMLongBench-Doc, ChartQA, CharXiv), agentic GUI use (ScreenSpot, ScreenSpot-Pro, OSWorld), long audio-video comprehension (WorldSense, DailyOmni), and voice interaction (VoiceBench), while retaining the text reasoning performance of the Nemotron 3 Nano 30B-A3B backbone. Combined with innovative multimodal token-reduction techniques, these capabilities translate into substantially lower inference latency and several-fold higher throughput than comparably sized models. We release model checkpoints in BF16, FP8, and FP4 formats alongside a large portion of our training data and code, with the goal of enabling the community to further advance efficient omni-modal modeling.

6. Contributors

Core Model Development

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, Collin McCarthy, Shaokun Zhang, Fuxiao Liu, Hanrong Ye, Yi Dong, Mingjie Liu, Yifan Peng, Piotr Zelasko, Zhehuai Chen, Nithin Rao Koluguri, Nune Tadevosyan, Lilit Grigoryan, Ehsan Hosseini Asl, Pritam Biswas, Leili Tavabi, Yuanhang Su, Zhiding Yu, Peter Jin, Alexandre Milesi, Netanel Haber

Data Generation and Curation

Yao Xu, Sarah Amiraslani, Nabin Mulepati, Eric Tramel, Jaehun Jung, Ximing Lu, Brandon Cui, Jin Xu, Zhiqi Li, Shihao Wang, Yuanguo Kuang, Shaokun Zhang, Huck Yang, Boyi Li, Hongxu (Danny) Yin, Song Han, Pavlo Molchanov, Adi Renduchintala, Charles Wang,, David Mosallanezhad, Soumye Singhal, Luis Vega, Katherine Cheung, Sreyan Ghosh, Yian Zhang, Alexander Bukharin, Venkat Srinivasan, Johnny Greco, Andre Manoel, Maarten Van Segbroeck, Suseella Panguliri, Rohit Watve, Divyanshu Kakwani, Shubham Pachori, Jeffrey Glick, Radha Sri-Tharan, Aileen Zaman, Khanh Nguyen, Shi Chen, Jiaheng Fang, Qing Miao, Wenfei Zhou, Yu Wang, Zaid Pervaiz Bhat, Varun Praveen, Arihant Jain, Ramanathan Arunachalam, Tomasz Kornuta, Ashton Sharabiani, Amy Shen, Wei Huang

Systems, Data and Infrastructure

Yi-Fu Wu, Ali Roshan Ghias, Huiying Li, Brian Yu, Nima Tajbakhsh, Chen Cui, Wenwen Gao, Li Ding, Terry Kong, Manoj Kilaru, Anahita Bhiwandiwalla, Marek Wawrzos, Daniel Korzekwa, Pablo Ribalta, Grzegorz Chlebus, Besmira Nushi, Ewa Dobrowolska, Maciej Jakub Mikulski, Kunal Dhawan, Steve Huang, Jagadeesh Balam, Yongqiang Wang, Nikolay Karpov, Valentin Mendelev, George Zelenfroynd, Meline Mkrtchyan, Qing Miao, Omri Almog, Bhavesh Pawar, Rameshwar Shivbhakta, Sudeep Sabnis, Ashrton Sharabiani, Negar Habibi, Geethapriya Venkataramani, Pamela Peng, Prerit Rodney, Serge Panev, Richard Mazzarese, Nicky Liu, Michael Fukuyama, Andrii Skliar,

Roger Waleffe, Duncan Riach, Yunheng Zou, Jian Hu, Hao Zhang, Binfeng Xu, Yuhao Yang, Zuhair Ahmed

Inference and Optimization

Alexandre Milesi, Carlo del Mundo, Chad Voegele, Zhiyu Cheng, Nave Assaf, Andrii Skliar, Daniel Afrimi, Natan Bagrov, Ran Zilberstein, Ofri Masad, Eugene Khvedchenia, Natan Bagrov, Borys Tymchenko, Tomer Asida, Daniel Afrimi, Parth Mannan, Victor Cui

Safety

Michael Evans, Katherine Luna, Jie Lou, Pinky Xu, Guyue Huang, Negar Habibi, Michael Boone, Pradeep Thalasta, Adeola Adesoba, Dina Yared, Christopher Parisien, Leon Derczynski, Shaona Ghosh, Wes Feely, Micah Schaffer, Radha Sri-Tharan, Jeffrey Glick, Barnaby Simkin, George Zelenfroynd, Tomasz Grzegorzek, Rishabh Garg

Evaluation, Product and Legal

Aastha Jhunjhunwala, Sergei Kolchenko, Farzan Memarian, Haran Kumar, Shiv Kumar, Isabel Hulseman, Anjali Shah, Kari Briski, Padmavathy Subramanian, Joey Conway, Udi Karpas, Jane Polak Scowcroft, Annie Surla, Shilpa Ammireddy, Ellie Evans, Jesse Oliver, Tom Balough, Chia-Chih Chen, Sandip Bhaskar, Alejandra Rico, Bardiya Sadeghi, Seph Mard, Katherine Cheung, Meredith Price, Laya Sleiman, Saori Kaji, Wesley Helmholz, Wendy Quan

Leadership

Michael Lightstone, Jonathan Cohen, Jian Zhang, Oleksii Kuchaiev, Boris Ginsburg, Jan Kautz, Eileen Long, Mohammad Shoeybi, Mostofa Patwary, Oluwatobi Olabiyi, Andrew Tao, Bryan Catanzaro, Udi Karpas

References

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.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024a. URL https://arxiv.org/abs/2410.17196.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13521–13525. IEEE, 2024b.

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.

Kateryna Chumachenko, Amala Sanjay Deshmukh, Jarno Seppanen, Ilia Karmanov, Chia-Chih Chen, Lukas Voegtle, Philipp Fischer, Marek Wawrzos, Saeid Motiian, Roman Ageev, Kedi Wu, Alexandre Milesi, Maryam Moosaei, Krzysztof Pawelec, Padmavathy Subramanian, Mehrzad Samadi, Xin Yu, Celina Dear, Sarah Stoddard, Jenna Diamond, Jesse Oliver, Leanna Chraghchian, Patrick Skelly, Tom Balough, Yao Xu, Jane Polak Scowcroft, Daniel Korzekwa, Darragh Hanley,

Sandip Bhaskar, Timo Roman, Karan Sapra, Andrew Tao, and Bryan Catanzaro. Nvidia nemotron parse 1.1, 2025. URL https://arxiv.org/abs/2511.20478.

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.

Jennifer Drexler Fox, Desh Raj, Natalie Delworth, Quinn McNamara, Corey Miller, and Migüel Jetté. Updated corpora and benchmarks for long-form speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13246–13250. IEEE, 2024.

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, 2024. URL https://arxiv.org/abs/2501.00321.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22487–22497, June 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2026. URL https://arxiv.org/abs/2502.04326.

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, 2025. URL https://arxiv.org/abs/2505.17163.

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.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen,

Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, Yifan Peng, Sara Papi, Marco Gaido, Alessio Brutti, and Boris Ginsburg. Granary: Speech recognition and translation dataset in 25 european languages, 2025. URL https://arxiv.org/abs/2505.13404.

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.

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

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 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

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)*, 2024.

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.

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.

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.

NVIDIA. Nemo rl: A scalable and efficient post-training library. https://github.com/NVIDIA-NeMo/RL, 2025. GitHub repository.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Amy Shen, Anahita Bhiwandiwalla, Andrew Tao, Anjulie Agrusa, Ankur Verma, Ann Guan, Anubhav Mandarwal, Arham Mehta, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asma Kuriparambil Thekkumpate, Ayush Dattagupta, Banghua Zhu, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bita Darvish Rouhani, Boris Ginsburg, Brandon Norick, Brandon Soubasis, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Carlo del Mundo, Chantal Hwang, Charles Wang, Cheng-Ping Hsieh, Chenghao Zhang, Chenhan Yu, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christopher Parisien, Collin Neale, Cyril Meurillon, Damon Mosk-Aoyama, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Lo, Daniel Rohrer, Daniel Serebrenik, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, Deepak Narayanan, Dhruv Nathawani, Dima Rekesh, Dina Yared, Divyanshu Kakwani, Dong Ahn, Duncan Riach, Dusan Stosic, Edgar Minasyan, Edward Lin, Eileen Long, Eileen Peters Long, Elad Segal, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper,

Eric Tramel, Erick Galinkin, Erik Pounds, Evan Briones, Evelina Bakhturina, Evgeny Tsykunov, Faisal Ladhak, Fay Wang, Fei Jia, Felipe Soares, Feng Chen, Ferenc Galko, Frank Sun, Frankie Siino, Gal Hubara Agam, Ganesh Ajjanagadde, Gantavya Bhatt, Gargi Prasad, George Armstrong, Gerald Shen, Gorkem Batmaz, Grigor Nalbandyan, Haifeng Qian, Harsh Sharma, Hayley Ross, Helen Ngo, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huizi Mao, Huy C Nguyen, Huy Q Nguyen, Iain Cunningham, Ido Galil, Ido Shahaf, Igor Gitman, Ilya Loshchilov, Itamar Schen, Itay Levy, Ivan Moshkov, Izik Golan, Izzy Putterman, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jian Zhang, Jiaqi Zeng, Jie Lou, Jimmy Zhang, Jinhang Choi, Jining Huang, Joey Conway, Joey Guman, John Kamalu, Johnny Greco, Jonathan Cohen, Joseph Jennings, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kai Xu, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khushi Bhardwaj, Kirthi Shankar, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Li Ding, Lizzie Wei, Lucas Liebenwein, Luis Vega, Maanu Grover, Maarten Van Segbroeck, Maer Rodrigues de Melo, Mahdi Nazemi, Makesh Narsimhan Sreedhar, Manoj Kilaru, Maor Ashkenazi, Marc Romeijn, Marcin Chochowski, Mark Cai, Markus Kliegl, Maryam Moosaei, Matt Kulka, Matvei Novikov, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Meredith Price, Michael Andersch, Michael Boone, Michael Evans, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Minseok Lee, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Najeeb Nabwani, Natalie Hereth, Nave Assaf, Negar Habibi, Neta Zmora, Netanel Haber, Nicola Sessions, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nir Ailon, Nirmal Juluru, Nishant Sharma, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Puny, Oren Tropp, Ouye Xie, Parth Chadha, Pasha Shamis, Paul Gibbons, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pinky Xu, Piotr Januszewski, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Qing Miao, Qiyu Wan, Rabeeh Karimi Mahabadi, Rachit Garg, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Rich Harang, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Hesse, Roger Waleffe, Rohit Watve, Roi Koren, Ruoxi Zhang, Russell Hewett, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Sadegh Mahdavi, Sahil Modi, Samuel Kriman, Sangkug Lim, Sanjay Kariyappa, Sanjeev Satheesh, Saori Kaji, Satish Pasumarthi, Saurav Muralidharan, Sean Narentharen, Sean Narenthiran, Seonmyeong Bak, Sergey Kashirsky, Seth Poulos, Shahar Mor, Shanmugam Ramasamy, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shiqing Fan, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Simeng Sun, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Stas Sergienko, Stefania Alborghetti, Stephen Ge, Sugam Dipak Devare, Sumeet Kumar Barua, Suseella Panguluri, Suyog Gupta, Sweta Priyadarshi, Syeda Nahida Akter, Tan Bui, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tijmen Blankevoort, Tim Moon, Tom Balough, Tomer Asida, Tomer Bar Natan, Tomer Ronen, Tugrul Konuk, Twinkle Vashishth, Udi Karpas, Ushnish De, Vahid Noorozi, Vahid Noroozi, Venkat Srinivasan, Venmugil Elango, Victor Cui, Vijay Korthikanti, Vinay Rao, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wenfei Zhou, Will Jennings, William Zhang, Wojciech Prazuch, Xiaowei Ren, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Yigong Qin, Ying Lin, Yonatan Geifman, Yonggan Fu, Yoshi Subara, Yoshi Suhara, Yubo Gao, Zach Moshe, Zhen Dong, Zhongbo Zhu, Zihan Liu, Zijia Chen, and Zijie Yan. Nvidia nemotron 3: Efficient and open intelligence, 2025a. URL https://arxiv.org/abs/2512.20856.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa

Liu, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Amy Shen, Anahita Bhiwandiwalla, Andrew Tao, Ann Guan, Anubhav Mandarwal, Arham Mehta, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asma Kuriparambil Thekkumpate, Ayush Dattagupta, Banghua Zhu, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bita Darvish Rouhani, Boris Ginsburg, Brandon Norick, Brandon Soubasis, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Carlo del Mundo, Chantal Hwang, Charles Wang, Cheng-Ping Hsieh, Chenghao Zhang, Chenhan Yu, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christopher Parisien, Collin Neale, Damon Mosk-Aoyama, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Rohrer, Daniel Serebrenik, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, Deepak Narayanan, Dhruv Nathawani, Dima Rekesh, Dina Yared, Divyanshu Kakwani, Dong Ahn, Duncan Riach, Dusan Stosic, Edgar Minasyan, Edward Lin, Eileen Long, Eileen Peters Long, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Tramel, Erick Galinkin, Erik Pounds, Evan Briones, Evelina Bakhturina, Faisal Ladhak, Fay Wang, Fei Jia, Felipe Soares, Feng Chen, Ferenc Galko, Frankie Siino, Gal Hubara Agam, Ganesh Ajjanagadde, Gantavya Bhatt, Gargi Prasad, George Armstrong, Gerald Shen, Gorkem Batmaz, Grigor Nalbandyan, Haifeng Qian, Harsh Sharma, Hayley Ross, Helen Ngo, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huizi Mao, Huy C Nguyen, Huy Q Nguyen, Iain Cunningham, Ido Shahaf, Igor Gitman, Ilya Loshchilov, Ivan Moshkov, Izzy Putterman, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jian Zhang, Jiaqi Zeng, Jie Lou, Jimmy Zhang, Jining Huang, Joey Conway, Joey Guman, John Kamalu, Johnny Greco, Jonathan Cohen, Joseph Jennings, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kai Xu, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khushi Bhardwaj, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Li Ding, Lucas Liebenwein, Luis Vega, Maanu Grover, Maarten Van Segbroeck, Maer Rodrigues de Melo, Makesh Narsimhan Sreedhar, Manoj Kilaru, Maor Ashkenazi, Marc Romeijn, Mark Cai, Markus Kliegl, Maryam Moosaei, Matvei Novikov, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Meredith Price, Michael Boone, Michael Evans, Miguel Martinez, Mike Chrzanowski, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Natalie Hereth, Nave Assaf, Negar Habibi, Neta Zmora, Netanel Haber, Nicola Sessions, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nirmal Juluru, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Ouye Xie, Parth Chadha, Pasha Shamis, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pinky Xu, Piotr Januszewski, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Qing Miao, Rabeeh Karimi Mahabadi, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Rich Harang, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Roger Waleffe, Rohit Watve, Roi Koren, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Sadegh Mahdavi, Sahil Modi, Samuel Kriman, Sanjay Kariyappa, Sanjeev Satheesh, Saori Kaji, Satish Pasumarthi, Sean Narentharen, Sean Narenthiran, Seonmyeong Bak, Sergey Kashirsky, Seth Poulos, Shahar Mor, Shanmugam Ramasamy, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shiqing Fan, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Simeng Sun, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Stefania Alborghetti, Stephen Ge, Sugam Dipak Devare, Sumeet Kumar Barua, Suseella Panguluri, Suyog Gupta, Sweta Priyadarshi, Syeda Nahida Akter, Tan Bui, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tijmen Blankevoort, Tom Balough, Tomer Asida, Tomer Bar Natan, Tugrul Konuk, Twinkle Vashishth, Udi Karpas, Ushnish De, Vahid Noorozi, Vahid Noroozi, Venkat Srinivasan, Venmugil Elango, Vijay Korthikanti, Vitaly Kurin, Vitaly Lavrukhin, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wenfei Zhou, Will Jennings, William Zhang, Wojciech Prazuch, Xiaowei Ren, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Ying Lin, Yonatan Geifman, Yonggan

Fu, Yoshi Subara, Yoshi Suhara, Yubo Gao, Zach Moshe, Zhen Dong, Zihan Liu, Zijia Chen, and Zijie Yan. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2025b. URL https://arxiv.org/abs/2512.20848.

NVIDIA, :, 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, 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 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, Alexandre Milesi, Eugene Khvedchenia, Ran Zilberstein, Ofri Masad, Natan Bagrov, Nave Assaf, Tomer Asida, Daniel Afrimi, Amit Zuker, Netanel Haber, Zhiyu Cheng, Jingyu Xin, Di Wu, Nik Spirin, Maryam Moosaei, Roman Ageev, Vanshil Atul Shah, Yuting Wu, Daniel Korzekwa, Unnikrishnan Kizhakkemadam Sreekumar, Wanli Jiang, 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, Udi Karpas, Kari Briski, Joey Conway, Michael Lightstone, Jan Kautz, Mohammad Shoeybi, Mostofa Patwary, Jonathen Cohen, Oleksii Kuchaiev, Andrew Tao, and Bryan Catanzaro. Nvidia nemotron nano v2 vl, 2025c. URL https://arxiv.org/abs/2511.03929.

NVIDIA, :, Aakshita Chandiramani, Aaron Blakeman, Abdullahi Olaoye, Abhibha Gupta, Abhilash Somasamudramath, Abhinav Khattar, Adeola Adesoba, Adi Renduchintala, Adil Asif, Aditya Agrawal, Aditya Vavre, Ahmad Kiswani, Aishwarya Padmakumar, Ajay Hotchandani, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Gronskiy, Alex Kondratenko, Alex Neefus, Alex Steiner, Alex Yang, Alexander Bukharin, Alexander Young, Ali Hatamizadeh, Ali Taghibakhshi, Alina Galiautdinova, Alisa Liu, Alok Kumar, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Anahita Bhiwandiwalla, Ananth Subramaniam, Andrew Tao, Anjaney Shrivastava, Anjulie Agrusa, Ankur Srivastava, Ankur Verma, Ann Guan, Anna Shors, Annamalai Chockalingam, Anubhav Mandarwal, Aparnaa Ramani, Arham Mehta, Arti Jain, Arun Venkatesan, Asha Anoosheh, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asli Sabanci Demiroz, Asma Kuriparambil Thekkumpate, Atefeh Sohrabizadeh, Avinash Kaur, Ayush Dattagupta, Barath Subramaniam Anandan, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Benjamin Chislett, Besmira Nushi, Bilal Kartal, Bill Thiede, Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Buvaneswari Mani, Carlo del Mundo, Chankyu Lee, Chanran Kim, Chantal Hwang, Chao Ni, Charles Wang, Charlie Truong, Cheng-Ping Hsieh, Chenhan Yu, Chenjie Luo, Cherie Wang, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Chris Holguin, Chris Wing, Christian Munley, Christopher Parisien, Chuck Desai, Chunyang Sheng, Collin Neale, Cyril Meurillon, Dakshi Kumar, Dan Gil, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Burkhardt Eliuth Triana, Daniel Egert, Daniel Fatade, Daniel Lo, Daniel Rohrer, Daniel Serebrenik, Daniil Sorokin, Daria Gitman, Daria Levy, Darko Stosic, David Edelsohn, David Messina, David Mosallanezhad, David Tamok, Deena Donia, Deepak Narayanan, Devin O'Kelly, Dheeraj Peri, Dhruv Nathawani, Di Wu, Dima Rekesh, Dina Yared, Divyanshu Kakwani,

Dmitry Konyagin Brandon Tuttle, Dong Ahn, Dongfu Jiang, Dorrin Poorkay, Douglas O'Flaherty, Duncan Riach, Dusan Stosic, Dustin Van Stee, Edgar Minasyan, Edward Lin, Eileen Peters Long, Elad Segal, Elena Lantz, Elena Lewis, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Pham-Hung, Eric W. Tramel, Erick Galinkin, Erik Pounds, Esti Etrog, Evan Briones, Evan Wu, Evelina Bakhturina, Evgeny Tsykunov, Ewa Dobrowolska, Farshad Saberi Movahed, Farzan Memarian, Fay Wang, Fei Jia, Felipe Soares, Felipe Vieira Frujeri, Feng Chen, Fengguang Lin, Ferenc Galko, Fortuna Zhang, Frankie Siino, Frida Hou, Gantavya Bhatt, Gargi Prasad, Geethapriya Venkataramani, Geetika Gupta, George Armstrong, Gerald Shen, Giulio Borghesi, Gordana Neskovic, Gorkem Batmaz, Grace Lam, Grace Wu, Greg Pauloski, Greyson Davis, Grigor Nalbandyan, Guoming Zhang, Guy Farber, Guyue Huang, Haifeng Qian, Haran Kumar Shiv Kumar, Harry Kim, Harsh Sharma, Hayate Iso, Hayley Ross, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huy Nguyen, Iain Cunningham, Ido Galil, Ido Shahaf, Igino Padovani, Igor Gitman, Igor Shovkun, Ikroop Dhillon, Ilya Loshchilov, Ingrid Kelly, Itamar Schen, Itay Levy, Ivan Moshkov, Izik Golan, Izzy Putterman, Jain Tu, Jan Baczek, Jan Kautz, Jane Polak Scowcroft, Janica Rosenberg, Jared Casper, Jarrod Pflum, Jason Grant, Jason Sewall, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jiacheng Xu, Jiafan Zhu, Jialin Song, Jian Zhang, Jiaqi Zeng, Jie Lou, Jill Milton, Jim Chow, Jimmy Zhang, Jinhang Choi, Jining Huang, Jocelyn Huang, Joel Caruso, Joey Conway, Joey Guman, Johan Jatko, John Kamalu, Johnny Greco, Jonathan Cohen, Jonathan Raiman, Joseph Jennings, Joyjit Daw, Juan Yu, Julio Tapia, Junkeun Yi, Jupinder Parmar, Jyothi Achar, Kari Briski, Kartik Mattoo, Katherine Cheung, Katherine Luna, Keith Wyss, Kevin Shih, Kezhi Kong, Khanh Nguyen, Khushi Bhardwaj, Kirill Buryak, Kirthi Shankar Sivamani, Konstantinos Krommydas, Kris Murphy, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Laikh Tewari, Laya Sleiman, Leo Du, Leon Derczynski, Li Ding, Lilach Ilan, Lingjie Wu, Lizzie Wei, Luis Vega, Lun Su, Maarten Van Segbroeck, Maer Rodrigues de Melo, Magaret Zhang, Mahan Fathi, Makesh Narsimhan Sreedhar, Makesh Sreedhar, Makesh Tarun Chandran, Manuel Reyes Gomez, Maor Ashkenazi, Marc Cuevas, Marc Romeijn, Margaret Zhang, Mark Cai, Mark Gabel, Markus Kliegl, Martyna Patelka, Maryam Moosaei, Matthew Varacalli, Matvei Novikov, Mauricio Ferrato, Mehrzad Samadi, Melissa Corpuz, Meng Xin, Mengdi Wang, Mengru Wang, Meredith Price, Micah Schaffer, Michael Andersch, Michael Boone, Michael Evans, Michael Z Wang, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Mike Hollinger, Mingyuan Ma, Minseok Lee, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Nader Khalil, Najeeb Nabwani, Nancy Agarwal, Nanthini Balasubramaniam, Narimane Hennouni, Narsi Kodukula, Natalie Hereth, Nathaniel Pinckney, Nave Assaf, Negar Habibi, Nestor Qin, Neta Zmora, Netanel Haber, Nick Reamaroon, Nickson Quak, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nir Ailon, Nirmal Juluru, Nirmalya De, Nowel Pitt, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Almog, Omri Puny, Oren Tropp, Otavio Padovani, Ouye Xie, Parth Chadha, Pasha Shamis, Paul Gibbons, Pavlo Molchanov, Peter Belcak, Peter Jin, Pinky Xu, Piotr Januszewski, Pooya Jannaty, Prachi Shevate, Pradeep Thalasta, Pranav Prashant Thombre, Prasoon Varshney, Prerana Gambhir, Pritam Gundecha, Przemek Tredak, Qing Miao, Qiyu Wan, Quan Tran Minh, Rabeeh Karimi Mahabadi, Rachel Oberman, Rachit Garg, Rahul Kandu, Raina Zhong, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Renee Yao, Renjie Pi, Richard Mazzarese, Richard Wang, Rick Izzo, Ridhima Singla, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Clark, Robert Hesse, Roger Waleffe, Rohit Varma Kalidindi, Rohit Watve, Roi Koren, Ron Fan, Ruchika Kharwar, Ruisi Cai, Ruoxi Zhang, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Ryota Egashira, Sadegh Mahdavi, Sagar Singh Ashutosh Joshi, Sahil Modi, Samuel Kriman, Sandeep Pombra, Sanjay Kariyappa, Sanjeev Satheesh, Santiago Pombo, Saori Kaji, Satish Pasumarthi, Saurav Mishra, Saurav Muralidharan, Scott Hara, Sean Narenthiran, Sebastian Rogawski, Seonjin Na, Seonmyeong Bak, Sepehr Sameni, Seth Poulos, Shahar Mor,

Shantanu Acharya, Shaona Ghosh Adam Lord, Sharath Turuvekere Sreenivas, Shaun Kotek, Shaya Gharghabi, Shelby Thomas, Sheng-Chieh Lin, Shibani Likhite, Shiqing Fan, Shiyang Chen, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuo Zhang, Shuoyang Ding, Shyam Renjith, Shyamala Prayaga, Siddhartha Jain, Simeng Sun, Sirisha Rella, Sirshak Das, Smita Ithape, Sneha Harishchandra S, Somshubra Majumdar, Soumye Singhal, Sri Harsha Singudasu, Sriharsha Niverty, Stas Sergienko, Stefana Gloginic, Stefania Alborghetti, Stephen Ge, Stephen McCullough, Sugam Dipak Devare, Suguna Varshini Velury, Sukrit Rao, Sumeet Kumar Barua, Sunny Gai, Suseella Panguluri, Sushil Koundinyan, Swathi Patnam, Sweta Priyadarshi, Swetha Bhendigeri, Syeda Nahida Akter, Sylendran Arunagiri, Tailling Yuan, Talor Abramovich, Tan Bui, Tan Yu, Terry Kong, Thanh Do, Thomas Gburek, Thorgane Marques, Tiffany Moore, Tijmen Blankevoort, Tim Moon, Timothy Ma, Tiyasa Mitra, Tomasz Grzegorzek, Tomer Asida, Tomer Bar Natan, Tomer Keren, Tomer Ronen, Traian Rebedea, Trenton Starkey, Tugrul Konuk, Twinkle Vashishth, Tyler Condensa, Udi Karpas, Ushnish De, Vahid Noorozi, Vahid Noroozi, Vanshil Atul Shah, Veena Vaidyanathan, Venkat Srinivasan, Venmugil Elango, Victor Cui, Vijay Korthikanti, Vikas Mehta, Virginia Adams, Virginia Wu, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Wan Seo, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wei-Ming Chen, Wendy Quan, Wenliang Dai, Wenwen Gao, Will Jennings, William Zhang, Xiaowei Ren, Xiaowen Xin, Xin Li, Yang Yu, Yangyi Chen, Yaniv Galron, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Ying Lin, Yonatan Geifman, Yonggan Fu, Yoshi Suhara, Youngeun Kwon, Yuan Zhang, Yuki Huang, Zach Moshe, Zhilin Wang, Zhiyu Cheng, Zhongbo Zhu, Zhuolin Yang, Zihan Liu, Zijia Chen, Zijie Yan, and Zuhair Ahmed. Nemotron 3 super: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning, 2026. URL https://arxiv.org/abs/2604.12374.

OpenAI. Introducing gpt-oss, 2025. URL https://openai.com/index/introducing-gpt-oss/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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, 2023. URL https://arxiv.org/abs/2305.18290.

Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report), 2026. URL https://arxiv.org/abs/2601.17237.

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.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. Fast conformer with linearly scalable attention for efficient speech recognition, 2023. URL https://arxiv.org/abs/2305.05084.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast, 2025. URL https://arxiv.org/abs/2509.14128.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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.

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.

Vaibhav Srivastav, Steven Zheng, Eric Bezzam, Eustache Le Bihan, Nithin Rao Koluguri, Piotr Żelasko, Somshubra Majumdar, Adel Moumen, and Sanchit Gandhi. Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2026. URL https://arxiv.org/abs/2510.06961.

NVIDIA The NeMo Data Designer Team. Nemo data designer: A framework for generating synthetic data from scratch or based on your own seed data. https://github.com/NVIDIA-NeMo/DataDesigner, 2025. GitHub Repository.

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.

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, 2025. URL https://arxiv.org/abs/2507.07999.

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, 2024a. URL https://arxiv.org/abs/2411.10442.

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, 2024b. URL https://arxiv.org/abs/2406.18521.

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression, 2025. URL https://arxiv.org/abs/2510.18234.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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, 2024. URL https://arxiv.org/abs/2410.23218.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. Efficient sequence transduction by jointly predicting tokens and durations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38462–38484. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23g.html.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

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, 2025. URL https://arxiv.org/abs/2505.09388.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *ICLR*, 2026.

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

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.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026. URL https://arxiv.org/abs/2505.17862.