# Model Overview

## Description:

The Nemovision-4B-Instruct model uses the latest NVIDIA recipe for distilling, pruning and quantizing to make it small enough to be performant on a broad range of RTX GPUs with the accuracy developers need. This is a model for generating responses for roleplaying, retrieval augmented generation, and function calling with vision understanding and reasoning capabilities. VRAM usage has been minimized to approximately 3.5 GB, providing fast Time to First Token. This model is ready for commercial use.

## License/Terms of Use

The use of this model is governed by the NVIDIA AI Foundation Models Community License Agreement

## Model Architecture:

**Architecture Type:** Transformer
**Network Architecture**

- Vision Encoder: openai/clip-vit-large-patch14
- Language Encoder: Nemotron-Mini-4B-Instruct

## Input

**Input Type(s):** Image(s), Text
**Input Format(s):** Red, Green, Blue (RGB), String
**Input Parameters:** 2D, 1D
**Other Properties Related to Input:** The model has a maximum of 4096 input tokens.

## Output

**Output Type(s):** Text
**Output Format(s):** String
**Output Parameters:** 1D
**Other Properties Related to Input:** The model has a maximum of 4096 input tokens. Maximum output for both versions can be set apart from input.

# Prompt Format:

**Single Turn**

```
<extra_id_0>System
{system prompt}

<extra_id_1>User
<image>
{prompt}
<extra_id_1>Assistant\n
<extra_id_0>System
{system prompt}

<extra_id_1>User
{prompt}
<extra_id_1>Assistant\n
```

**Multi-image**

```
<extra_id_0>System
{system prompt}

<extra_id_1>User
<image>
<image>
<image>
{prompt}
<extra_id_1>Assistant\n
```

**Multi-Turn or Few-shot**

```
<extra_id_0>System
{system prompt}

<tool> ... </tool>
<context> ... </context>
```

```
<extra_id_1>User
{prompt}
<extra_id_1>Assistant
<toolcall> ... </toolcall>
<extra_id_1>Tool
{tool response}
<extra_id_1>Assistant\n
```

## Software Integration:

**Runtime(s):** AI Inference Manager (NVAIM) Version 1.0.0
**Supported Hardware Microarchitecture Compatibility:** GPU supporting DirectX 11/12 and Vulkan 1.2 or higher
**[Preferred/Supported] Operating System(s):**

- Windows

## Software Integration: (Cloud)

**[Preferred/Supported] Operating System(s):**

- Linux

# Training & Evaluation:

## Training Dataset:

NV-Pretraining and NV-VILA-SFT data were used. Additionally, the following datasets were used:

- OASST1
- OASST2
- Localized Narratives
- TextCaps
- TextVQA
- RefCOCO
- VQAv2

- GQA
- SynthDoG-en
- A-OKVQ
- WIT
- CLEVR
- CLEVR-X
- CLEVR-Math

**Data Collection Method by dataset:**

- Hybrid: Automated, Human

**Labeling Method by dataset:**

- Hybrid: Automated, Human

**Properties:**

NV-Pretraining data was collected from 5M subsampled NV-CLIP dataset. Stage 3 NV-SFT data has 3.47M unique images and 3.78M annotations on images that only have commercial license. Trained on commercial text dataset.

## Evaluation Dataset:

**Data Collection Method by dataset**

- Hybrid: Automated, Human

**Labeling Method by dataset**

- Human

**Properties:**

A collection of different benchmarks, including academic VQA benchmarks and recent benchmarks specifically proposed for language understanding and reasoning, instruction-following, and function calling LMMs.

- VQAv2
- GQA
- ScienceQA Image
- Text VQA
- POPE
- MMBench
- SEED-Bench

- MMMU
- IF-Eval
- BFCL

| Benchmark | VQAv2 | GQA | SQA Image | Text VQA | POPE (Popular) | MMB ench-en | SEED | SEED Image | MMMU val (beam 5) |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 73.92 | 53.47 | 69.81 | 57.03 | 87.13 | 59.96 | 58.89 | 66.18 | 36.8 |

Berkeley Function Calling

| Benchmark | Simple | Multiple Functions | Parallel Functions | Parallel Multiple | Relevance |
|---|---|---|---|---|---|
| Accuracy | 85.25 | 90 | 77.5 | 76.5 | 17.08 |

Instruction Following Eval

| Benchmark | Prompt Level Accuracy | Instruction Level Accuracy |
|---|---|---|
| Accuracy | 46.95 | 57.79 |

# Inference:

**Test Hardware:**

- H100
- A100
- A10g
- L40s

**Supported Hardware Platform(s):** L40s, A10g, A100, H100

# Ethical Considerations:

NVIDIA believes Trustworthy AI is a shared responsibility and we have established policies and practices to enable development for a wide array of AI applications. When downloaded or used in accordance with our terms of service, developers

should work with their internal model team to ensure this model meets requirements for the relevant industry and use case and addresses unforeseen product misuse.

Please report security vulnerabilities or NVIDIA AI Concerns here.

## Model Card ++ Safety

| Field | Response |
|---|---|
| Generatable or reverse engineerable personally-identifiable information (PII)? | None |
| Was consent obtained for any personal data used? | Not Applicable |
| Personal data used to create this model? | Datasets used for fine-tuning did not introduce any personal data that did not exist in the base model. |
| How often is dataset reviewed? | Before Release |
| Is a mechanism in place to honor data subject right of access or deletion of personal data? | Not Applicable |
| If personal data is collected for the development of the model, was it collected directly by NVIDIA? | Not Applicable |
| If personal data is collected for the development of the model by NVIDIA, do you maintain or have access to disclosures made to data subjects? | Not Applicable |

| Field | Response |
|---|---|
| If personal data is collected for the development of this AI model, was it minimized to only what was required? | Not Applicable |
| Is there provenance for all datasets used in training? | Yes |
| Does data labeling (annotation, metadata) comply with privacy laws? | Yes |
| Is data compliant with data subject requests for data correction or removal, if such a request was made? | Not Applicable |

## Model Card ++ Privacy

| Field | Response |
|---|---|
| Intended Application(s) & Domain(s): | Visual Question Answering. |
| Model Type: | Vision Language Model |
| Intended Users: | The primary intended users of the model are practitioners and researchers in computer vision, natural language processing, machine learning, and artificial intelligence. |
| Output: | Text |
| Describe how the model works: | Chat based on image content. Chat for roleplaying, retrieval augmented generation, and function calling. |

| Field | Response |
|---|---|
| Technical Limitations: | This model may not perform well on domain specific images. |
| Known Risk: | The Model may produce output that is biased, toxic, or incorrect responses. Therefore, the model may amplify those biases and return toxic responses especially when prompted with toxic prompts. The Model may also generate answers that may be inaccurate, omit key information, or include irrelevant or redundant text, producing socially unacceptable or undesirable text, even if the prompt itself does not include anything explicitly offensive. |
| Verified to have met prescribed NVIDIA standards: | Yes |
| Performance Metrics: | Visual Question Answering (VQA), GQA, MMBench-en, MMMU, SQA-Image, etc. Text based evaluations include Automatic Evaluation - Commonsense Reasoning & Language Understanding, Automatic Evaluation - Open LLM Leaderboard, Instruction Following, Berkeley Function Calling, and ChatRAG-Bench. |
| Potential Known Risks: | None Known |
| Licensing: | NVIDIA AI Foundation Models Community License |

## Model Card ++ Explainability

| Field | Response |
| --- | --- |
| Participation considerations from adversely impacted groups protected classes in model design and testing: | None |
| Measures taken to mitigate against unwanted bias: | None |