NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

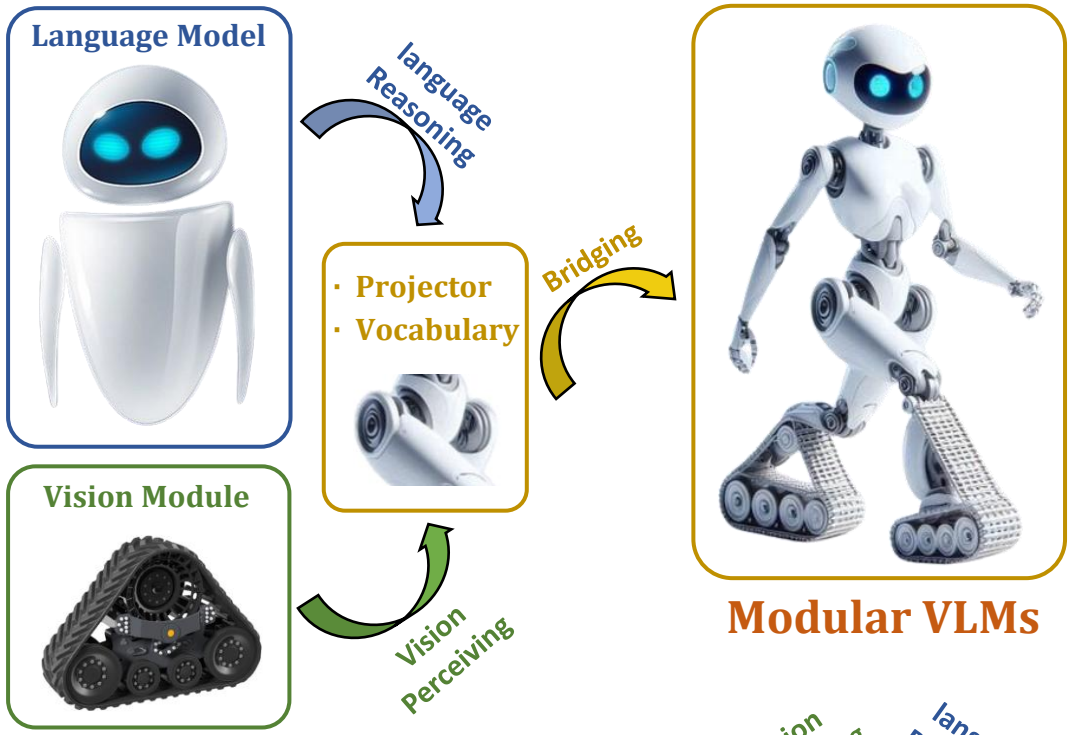
# NEO series:
## Building Native Multimodal Models End to End

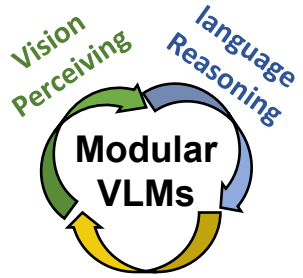
Ziwei Liu 刘子纬

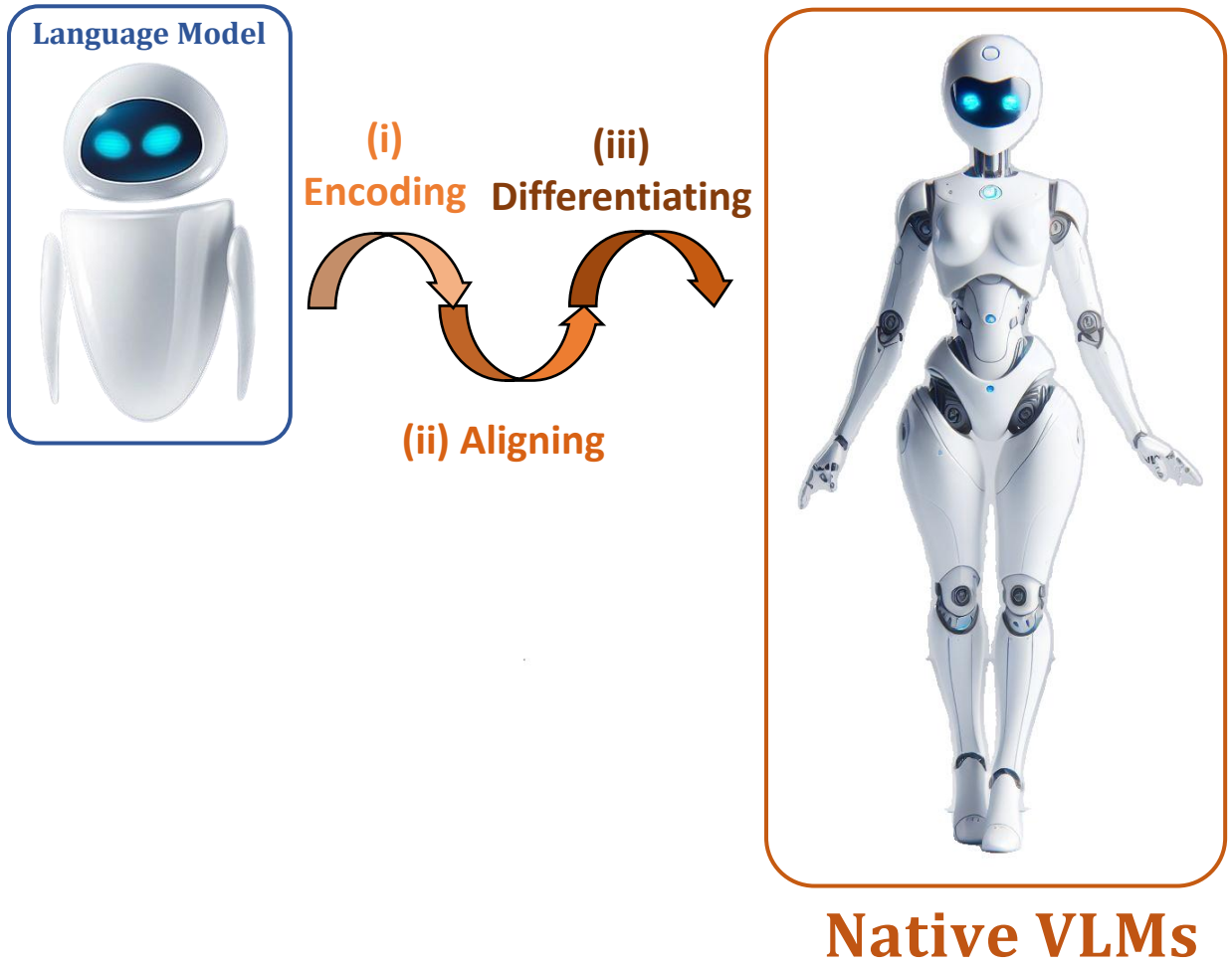
Nanyang Technological University

https://liuziwei7.github.io

# Modular vs. Native Vision-Language Models

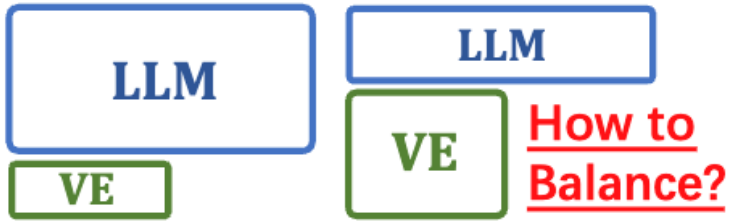


# Advantages： Native Vision-Language Models

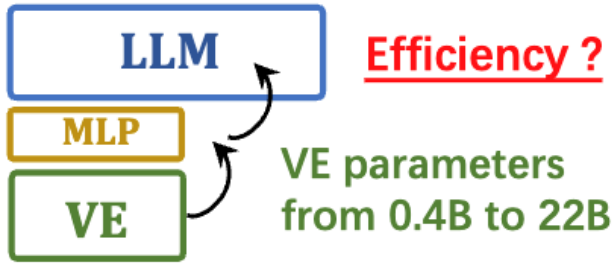
- Remove **Pre-training Scaling Law Bottleneck** !



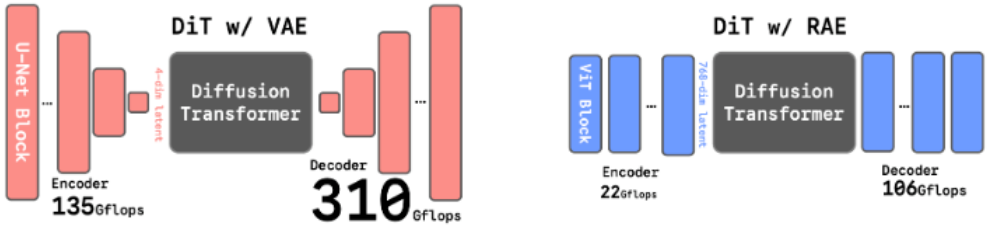

-- NaViL
-- NMMs

- Largely **Simplify Infrastructure Configuration** !

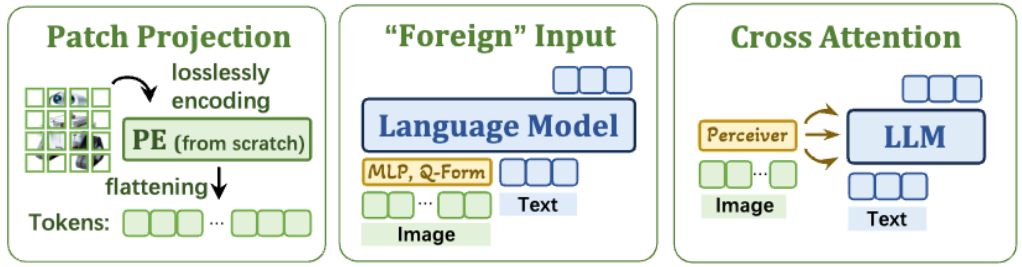
  -- Pre-Training Stages (scale, modality)
  -- Post-Training Stages (context length)
  -- Deployment Stages



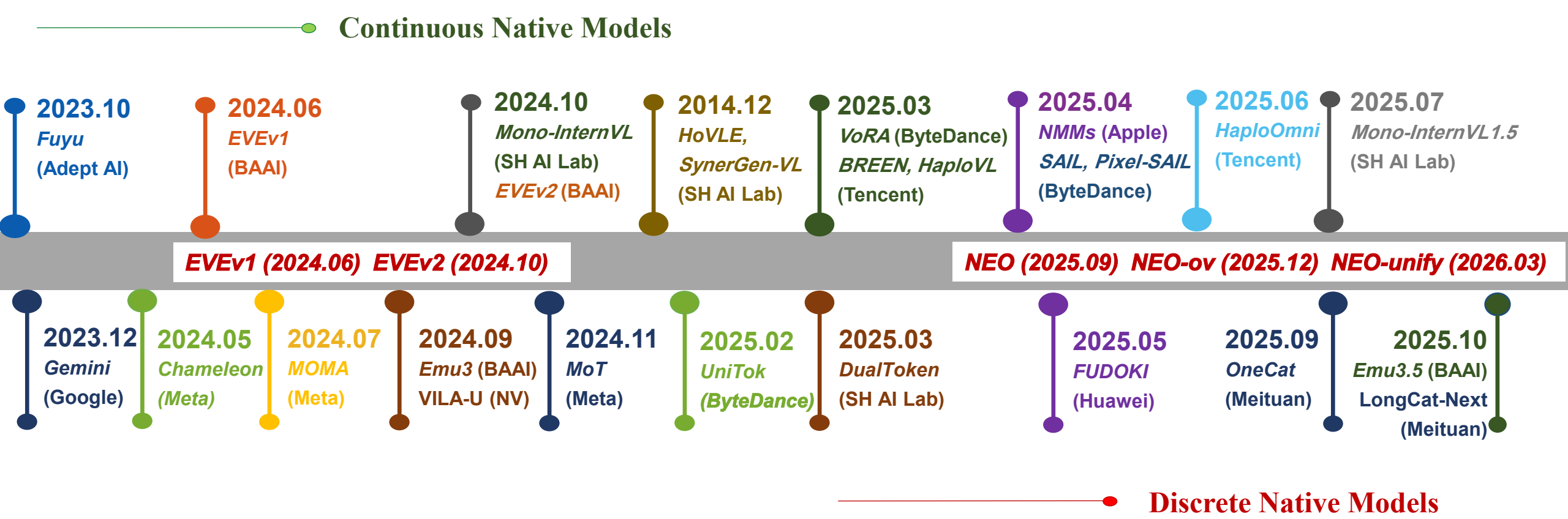

- Understanding, Reconstruction, Generation **Pixel and Semantic Representations** !




- **Lossless Input** and **Early-Fusion Interaction** for Potential Zero-shot Scenarios !



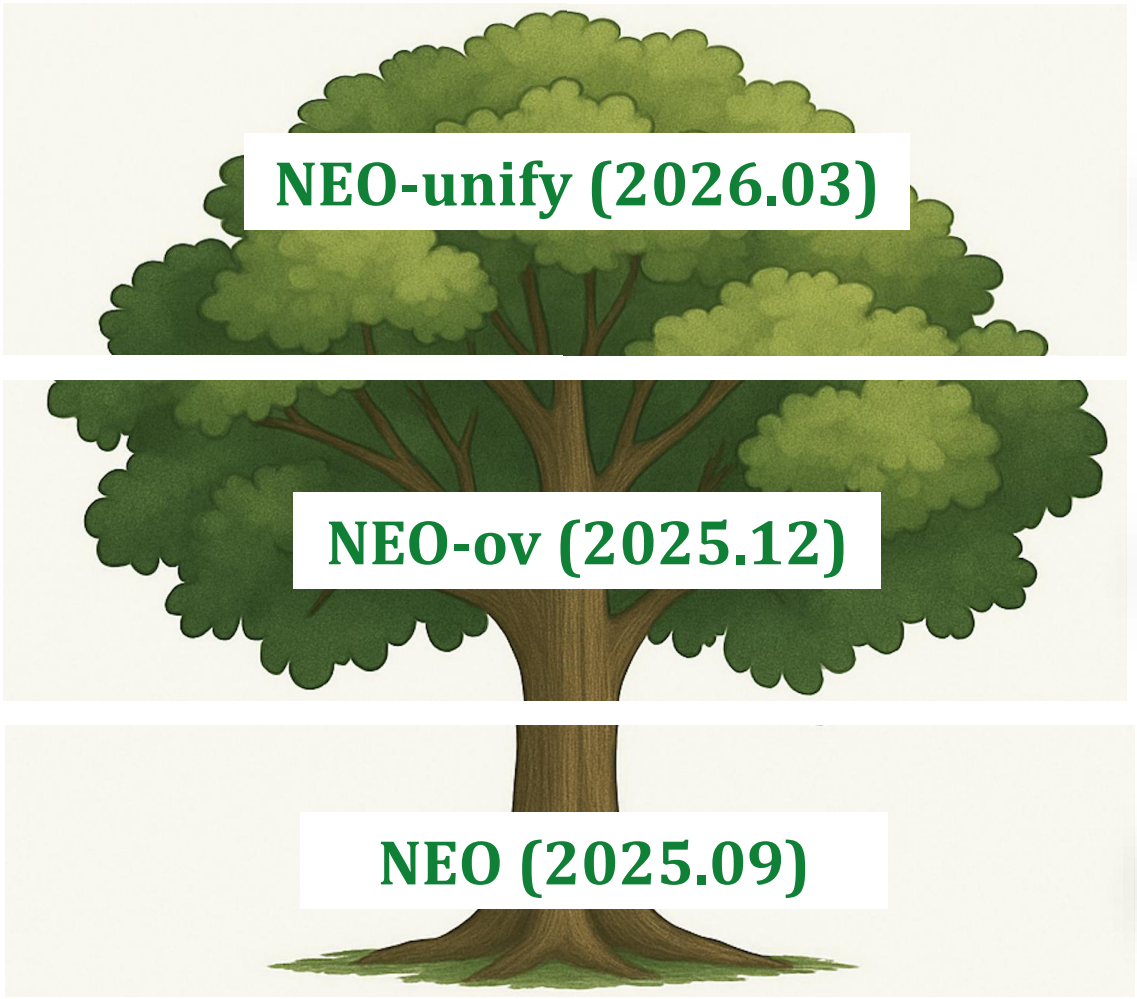
# Background： Native Vision-Language Models




**Research Focus:**

(1) Training Data and Stages (2) Model Architecture (3) Task Applications **(4) Scaling law !**

# Outline： Native Vision-Language Models




## Native VLMs for Unified Models

**Between Pixels and Words -- Towards Native Multimodal Unified Models at Scale**

Tokenizer-free VLMs, Asymmetry - Lossless, MoT Backbone, Multi-Modality Reasoning

## Native VLMs for One-Vision Models

**From Pixels to Words: Towards Native One-Vision Models at Scale**

Dense Models for Single-Image, Multi-Image, 3D spatial, Video Understanding

## Native VLMs for Image Understanding

**From Pixels to Words: Towards Native Vision-Language Primitives at Scale**

Native Vision-Language Primitive, Training Recipe, Vision-Language Conflict

# Native VLMs for Image Understanding

## From Pixels to Words: Towards Native Vision-Language Primitives at Scale

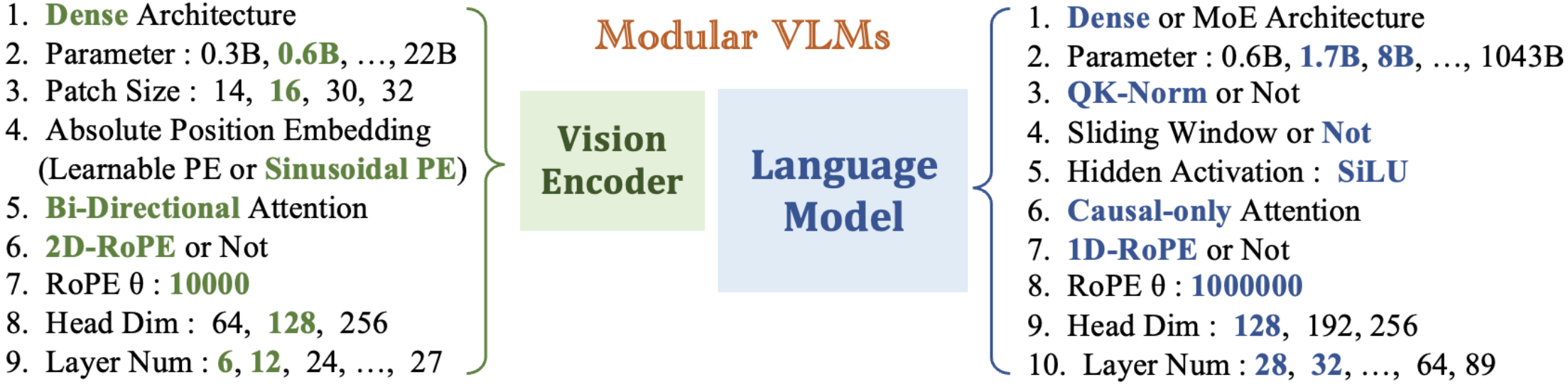
Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaohua Wang, Hanming Deng, Lewei Lu, Dahua Lin, Ziwei Liu

(**NEO**, ICLR 2026)

# Challenge

**Question 1:**

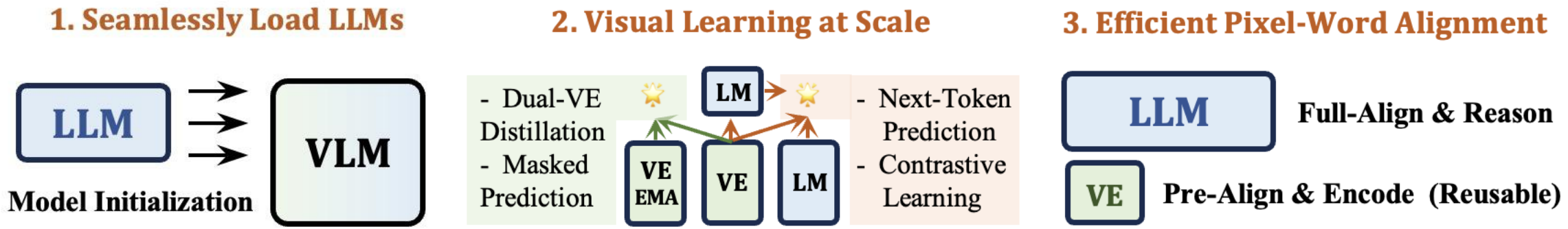
- What **constraints** set native VLMs apart from modular ones, and to what extent can these barriers be **overcome**?



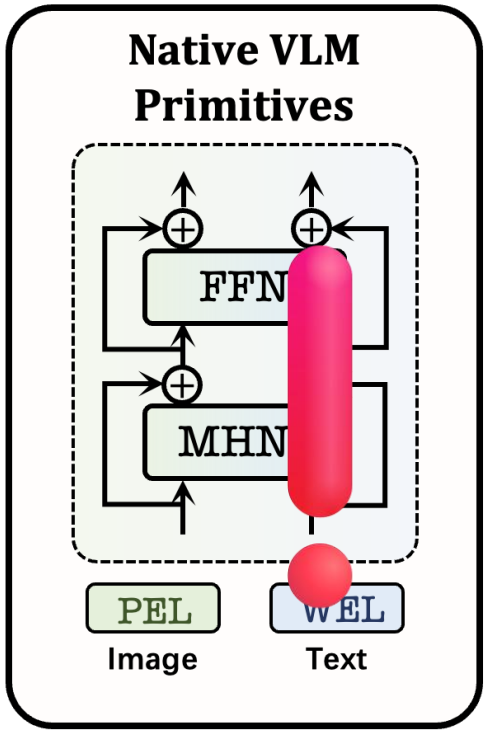
# Challenge

**Question 2:**

- How to make research in native VLMs more **accessible** and **democratized**, thereby accelerating progress in the field?



# Key Idea

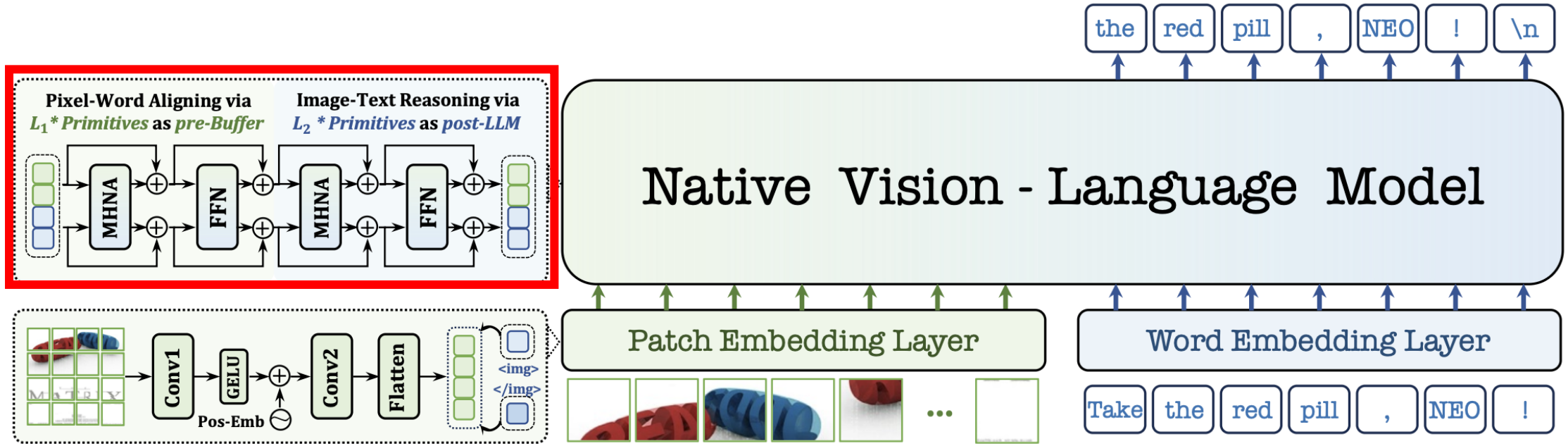



**From first principles, one base VLM block should :**

- Integrate the **strengths** of formerly **separate VE and LLM blocks**
- Inherently embody various multi-modal properties that support unified vision-language **encoding**, **aligning**, and **reasoning**

**Native VLM Primitive**

# Methodology

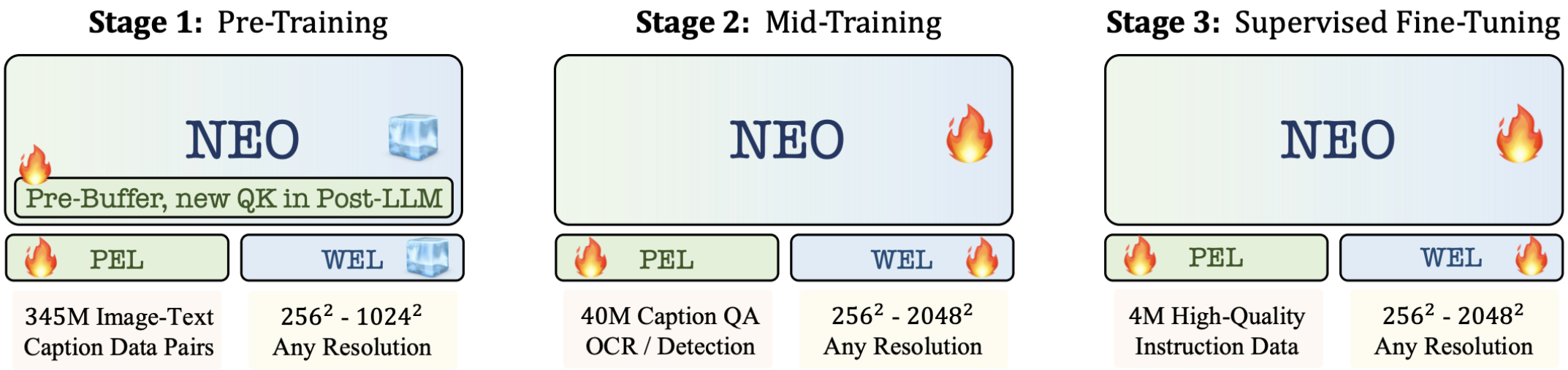



**Modality-agnostic Encoding / Reduce Disturbance towards LLM / Reusable for Future VLMs**

- Modality-shared pre-Buffer maps vision and language into a unified representation space.
- Post-LLM absorbs strong language proficiency and powerful reasoning capabilities of pre-trained LLMs.

# Methodology

## (-) Training Recipe




**End-to-End Training Procedure / Quite Efficient with Limited Data**

- Using entire **390M** image-text samples for developing visual perception from scratch
- Text-only : Image-Text (pre-training, mid-training) with **3 : 7** for mitigating vision-language conflicts inside one dense model.

# Main Results

## (NEO-2B) General Vision-Language Benchmarks

| Model | LLM | # Data | MMMU | MMB | MMVet | MMStar | SEED-I | POPE | HallB |
|---|---|---|---|---|---|---|---|---|---|
| ▼ *Modular Vision-Language Models (2B)* | | | | | | | | | |
| Qwen2-VL | Qwen2-1.5B | – / – / – | 41.1 | 74.9 | 49.5 | 48.0 | – | – | 41.7 |
| InternVL2.5 | InternLM2.5-1.8B | >6B / 100M / 16M | 43.6 | 74.7 | 60.8 | 53.7 | – | **90.6** | 42.6 |
| InternVL3† | Qwen2.5-1.5B | >6B / 100M / 22M | **48.6** | **81.1** | **62.2** | **60.7** | – | 89.6 | 42.5 |
| Qwen2.5-VL† | **Qwen2.5-3B** | – / – / – | 51.2 | 79.1 | 61.8 | 55.9 | – | – | 46.3 |
| Encoder-Based | Qwen3-1.7B | >6B / 40M / 4M | 47.1 | 75.8 | 37.4 | 52.7 | **73.6** | 87.0 | **44.4** |
| ▼ *Native Vision-Language Models (2B)* | | | | | | | | | |
| Mono-InternVL | InternLM2-1.8B | 1.2B / 143M / 7M | 33.7 | 65.5 | 40.1 | – | 67.4 | – | 34.8 |
| Mono-InternVL-1.5 | InternLM2-1.8B | 400M / 150M / 7M | 39.1 | 64.0 | **54.0** | – | 66.9 | – | 32.5 |
| HoVLE | InternLM2-1.8B | 550M / 50M / 7M | 32.2 | 73.3 | 43.8 | – | 70.9 | 87.4 | 38.4 |
| OneCAT | Qwen2.5-1.5B | 436M / 70M / 13M | 39.0 | 72.4 | 42.4 | – | 70.9 | – | – |
| NEO | Qwen3-1.7B | 345M / 40M / 4M | **48.6** | **76.0** | 49.6 | **54.2** | **74.2** | **87.5** | **43.1** |

- **NEO approaches** the top-tier modular VLMs, e.g., **InternVL3**.
- **NEO outperforms** the best native VLM competitors, from **EVE** series to **SAIL**.

# Main Results

## (NEO-2B) Visual Question Answering Benchmarks

| Model | Input | RoPE | Backbone | AI2D | DocVQA | ChartQA | InfoVQA | TextVQA | OCRBench |
|---|---|---|---|---|---|---|---|---|---|
| ▼ *Modular Vision-Language Models (2B)* | | | | | | | | | |
| Qwen2-VL | Any Res. | M-RoPE | Dense | 74.7 | **90.1** | 73.5 | 65.5 | **79.7** | 80.9 |
| InternVL2.5 | Tile-wise | 1D-RoPE | Dense | 74.9 | 88.7 | 79.2 | 60.9 | 74.3 | 80.4 |
| InternVL3† | Tile-wise | 1D-RoPE | Dense | **78.7** | 88.3 | **80.2** | **66.1** | 77.0 | **83.5** |
| Qwen2.5-VL† | Any Res. | M-RoPE | Dense | 81.6 | 93.9 | 84.0 | 77.1 | 79.3 | 79.7 |
| Encoder-Based | Tile-wise | 1D-RoPE | Dense | 77.4 | 89.9 | 78.4 | 65.9 | 73.3 | **83.5** |
| ▼ *Native Vision-Language Models (2B)* | | | | | | | | | |
| Mono-InternVL | Tile-wise. | 1D-RoPE | MoE | 68.6 | 80.0 | 73.7 | 43.0 | 72.6 | 76.7 |
| Mono-InternVL-1.5 | Tile-wise. | 1D-RoPE | DaC | 67.4 | 81.7 | 72.2 | 47.9 | 73.7 | **80.1** |
| HoVLE | Tile-wise. | 1D-RoPE | Dense | 73.0 | 86.1 | 78.6 | 55.7 | 70.9 | 74.0 |
| OneCAT | Any Res. | M-RoPE | Dense | 72.4 | 87.1 | 76.2 | 56.3 | 67.0 | – |
| NEO | Any Res. | Native-RoPE | Dense | **80.1** | **89.9** | **81.2** | **63.2** | **74.0** | 77.1 |

# Main Results

## (NEO-8B) General Vision-Language Benchmarks

| Model | LLM | # Data | MMMU | MMB | MMVet | MMStar | SEED-I | POPE | HallB |
|---|---|---|---|---|---|---|---|---|---|
| ▼ *Modular Vision-Language Models (8B)* | | | | | | | | | |
| Qwen2-VL | Qwen2-7B | – / – / – | 54.1 | 83 | 62.0 | 60.7 | – | 88.1 | 50.6 |
| InternVL2.5 | InternLM2.5-7B | >6B / 50M / 4M | 56.0 | **84.6** | 62.8 | 64.4 | – | 90.6 | 50.1 |
| Qwen2.5-VL† | Qwen2.5-7B | – / – / – | 55.0 | 83.5 | 67.1 | 63.9 | – | 86.4 | **52.9** |
| InternVL3† | Qwen2.5-7B | >6B / 100M / 22M | **62.7** | 83.4 | **81.3** | **68.2** | – | **91.1** | 49.9 |
| Encoder-Based | Qwen3-8B | >6B / 40M / 4M | 54.1 | 84 | 60.0 | 63.5 | **76.2** | 87.8 | 51.4 |
| ▼ *Native Vision-Language Models (8B)* | | | | | | | | | |
| Fuyu | Persimmon-8B | – / – / – | 27.9 | 10.7 | 21.4 | – | 59.3 | 84.0 | – |
| Chameleon | from scratch | 1.4B / 0M / 1.8M | 25.4 | 31.1 | 8.3 | – | 30.6 | 19.4 | 17.1 |
| EVE | Vicuna-7B | 33M / 0M / 1.8M | 32.6 | 52.3 | 25.7 | – | 64.6 | 85.0 | 26.4 |
| SOLO | Mistral-7B | 44M / 0M / 2M | – | 67.7 | 30.4 | – | 64.4 | 78.6 | – |
| Emu3 | from scratch | – / – / – | 31.6 | 58.5 | 37.2 | – | 68.2 | 85.2 | – |
| EVEv2 | Qwen2.5-7B | 77M / 15M / 7M | 39.3 | 66.3 | 45.0 | – | 71.4 | 87.6 | – |
| BREEN | Qwen2.5-7B | 13M / 0M / 4M | 42.7 | 71.4 | 38.9 | 51.2 | – | – | 37.0 |
| VoRA | Qwen2.5-7B | 30M / 0M / 0.6M | 32.0 | 61.3 | 33.7 | – | 68.9 | 85.5 | – |
| SAIL | Mistral-7B | 512M / 86M / 6M | – | 70.1 | 46.3 | 53.1 | 72.9 | 85.8 | **54.2** |
| NEO | Qwen3-8B | 345M / 40M / 4M | **54.6** | **82.1** | **53.6** | **62.4** | **76.3** | **88.4** | 46.4 |

# Main Results

## (NEO-8B) Visual Question Answering Benchmarks

| Model | Input | RoPE | Backbone | AI2D | DocVQA | ChartQA | InfoVQA | TextVQA | OCRBench |
|---|---|---|---|---|---|---|---|---|---|
| ▼ *Modular Vision-Language Models (8B)* | | | | | | | | | |
| Qwen2-VL | Any Res. | M-RoPE | Dense | 83.0 | 94.5 | 83 | 76.5 | 84.3 | 86.6 |
| InternVL2.5 | Tile-wise | 1D-RoPE | Dense | 84.5 | 93.0 | 84.8 | 77.6 | 79.1 | 82.2 |
| Qwen2.5-VL† | Any Res. | M-RoPE | Dense | 83.9 | **95.7** | **87.3** | **82.6** | **84.9** | 86.4 |
| InternVL3† | Tile-wise | 1D-RoPE | Dense | **85.2** | 92.7 | 86.6 | 76.8 | 80.2 | **88** |
| Encoder-Based | Tile-wise | 1D-RoPE | Dense | 82.9 | 92.1 | 83.5 | 75 | 77.1 | 85.3 |
| ▼ *Native Vision-Language Models (8B)* | | | | | | | | | |
| Fuyu | Any Res. | 1D-RoPE | Dense | 64.5 | – | – | – | – | 36.6 |
| Chameleon | Fix Res. | 1D-RoPE | Dense | 46.0 | 1.5 | 2.9 | 5.0 | 4.8 | 0.7 |
| EVE | Any Rat. | 1D-RoPE | Dense | 61.0 | 53.0 | 59.1 | 25.0 | 56.8 | 39.8 |
| SOLO | Any Res. | 1D-RoPE | Dense | 61.4 | – | – | – | – | 12.6 |
| Emu3 | Fix Res. | 1D-RoPE | Dense | 70 | 76.3 | 68.6 | 43.8 | 64.7 | 68.7 |
| EVEv2 | Any Rat. | 1D-RoPE | DaC | 74.8 | – | 73.9 | – | 71.1 | 70.2 |
| BREEN | Any Res. | 1D-RoPE | MoE | 76.4 | – | – | – | 65.7 | – |
| VoRA | Any Res. | 1D-RoPE | Dense | 61.1 | – | – | – | 58.7 | – |
| SAIL | Any Res. | M-RoPE | Dense | 76.7 | – | – | – | **77.1** | **78.3** |
| NEO | Any Res. | Native-RoPE | Dense | **83.1** | **88.6** | **82.1** | **60.9** | 75.0 | 77.7 |

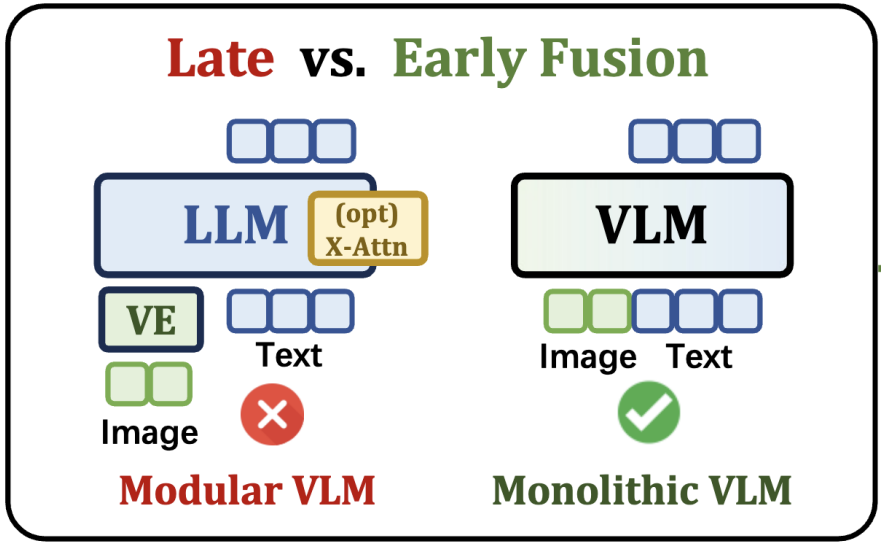
# Native VLMs for One-Vision Understanding

## From Pixels to Words -- Towards Native One-Vision Models at Scale

Haiwen Diao, Jiahao Wang, Penghao Wu, Yuhao Dong, Yuwei Niu, Yue Zhu, Zhongang Cai, Weichen Fan, Linjun Dai, Silei Wu, Xuanyu Zheng, Mingxuan Li, Yuanhan Zhang, Bo Li, Hanming Deng, Huchuan Lu, Quan Wang, Lei Yang, Lewei Lu, Dahua Lin, Ziwei Liu

(**NEO-OneVision**, Arxiv: 2026)

# Challenge

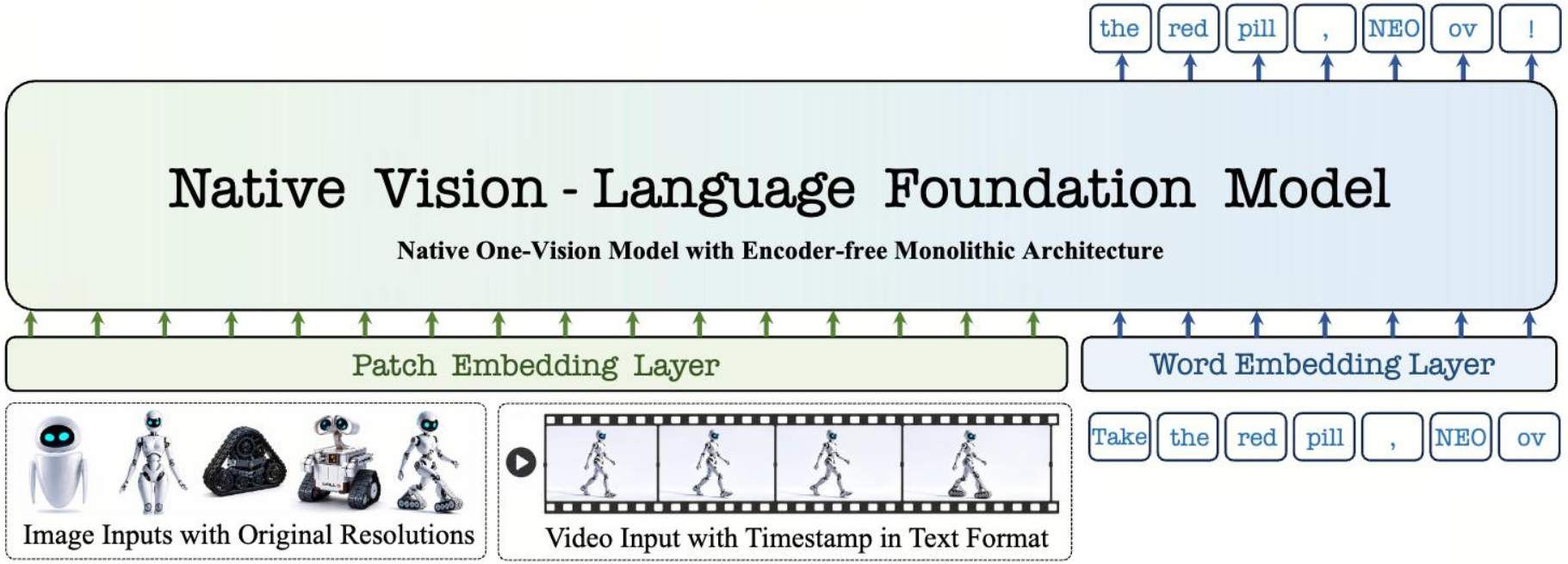



**Question：**

- Can native VLMs generalize across single-image, multi-image, **video**, and **3D spatial** scenarios?
- What advantages of our native VLMs, especially **early-fusion** for **pixel-pixel**， **pixel-word**?
- **Stronger** and **comprehensive** baseline over Qwen3-VL for subsequent **RL community**?

# Methodology

## (-) Model Architecture

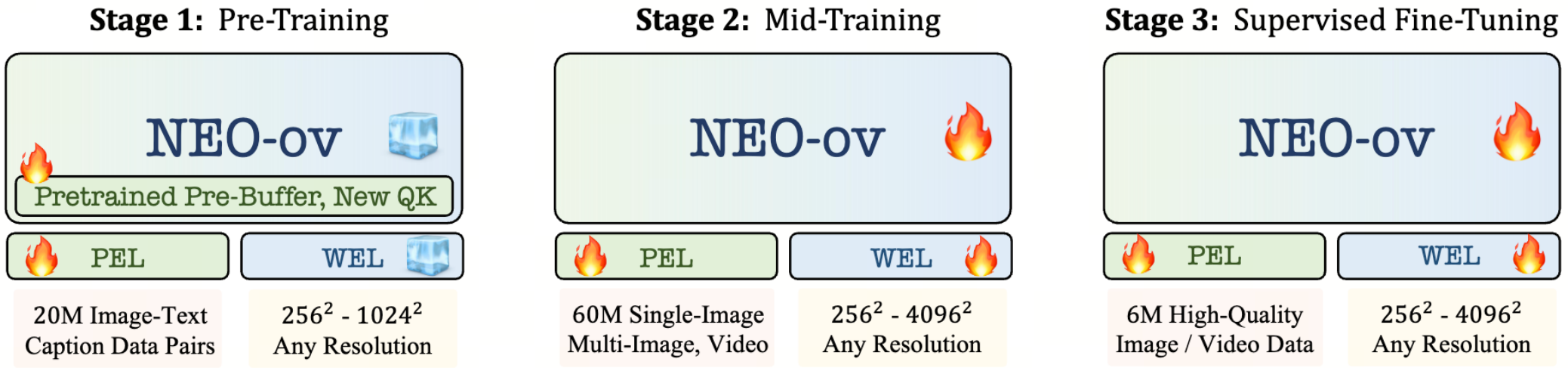



**Efficient Native One-Vision VLMs construction**

- Pre-trained pre-Buffer adopted from NEO for **native visual inputs**, singe-image, multi-image, video
- Initialize new Post-LLM from existing state-of-the-art LLM series for **strong reasoning capability**

# Methodology

## (-) Training Recipe




**Continuous Training Procedure / Context length, Resolution, Data Ratio do count !**

- Context length rises from 16K to **36K.** (training **128 frame**), Image Resolution varies from 256*256 to **4096*4096.**
- Text-only : Image-Text : Multi-Image, Video-Text with **2: 4 : 1 : 1** for harmonizing various application scenarios.

# Main Results: Image Understanding

| Model | General VQA Understanding | | | | | | OCR Recognization | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MMMU | MMB | RWQA | MMStar | SEED-I | HallB | AI2D | DocVQA | ChartQA | TextVQA | OCRBench |
| ▼ *Modular Vision-Language Models (Instruct-2B)* | | | | | | | | | | | |
| Qwen2-VL | 41.1 | 74.9 | 62.6 | 48.0 | – | 41.7 | 74.7 | 90.1 | 73.5 | **79.7** | 80.9 |
| InternVL3 | 48.6 | **81.1** | **64.3** | 60.7 | – | 42.5 | 78.7 | 88.3 | 80.2 | 77.0 | 83.5 |
| InternVL3.5 | 53.0 | 78.2 | 62.0 | **62.7** | **75.3** | 48.6 | **78.8** | 89.4 | **80.7** | 76.5 | 83.6 |
| Qwen3-VL | **53.4** | 78.4 | 63.9 | 58.3 | – | **51.4** | 76.9 | **93.3** | 79.1 | – | **85.8** |
| ▼ *Native Vision-Language Models (Instruct-2B)* | | | | | | | | | | | |
| Mono-VL | 33.7 | 65.5 | – | – | 67.4 | 34.8 | 68.6 | 80.0 | 73.7 | 72.6 | 76.7 |
| Mono-VL1.5 | 39.1 | 64.0 | – | – | 66.9 | 32.5 | 67.4 | 81.7 | 72.2 | 73.7 | 80.1 |
| HoVLE | 32.2 | 73.3 | – | – | 70.9 | 38.4 | 73.0 | 86.1 | 78.6 | 70.9 | 74.0 |
| OneCAT | 39.0 | 72.4 | – | – | 70.9 | – | 72.4 | 87.1 | 76.2 | 67.0 | – |
| NEO | 48.6 | 76.0 | 63.1 | 54.2 | 74.2 | 43.1 | 80.1 | 89.9 | 81.2 | 74.0 | 77.1 |
| **NEO-ov** | **54.7** | **80.0** | **64.4** | **58.6** | **76.2** | **54.5** | **81.4** | **91.2** | **83.1** | **77.3** | **81.2** |
| ▼ *Modular Vision-Language Models (Instruct-8B)* | | | | | | | | | | | |
| Qwen2.5-VL | 55.0 | 83.5 | 68.5 | 63.9 | – | 52.9 | 83.9 | 95.7 | 87.3 | **84.9** | 86.4 |
| InternVL3 | 62.7 | 83.4 | 70.8 | 68.2 | – | 49.9 | 85.2 | 92.7 | 86.6 | 80.2 | 88.0 |
| InternVL3.5 | 68.1 | 82.7 | 67.5 | 69.3 | **77.1** | 54.5 | 84.0 | 92.3 | 86.7 | 78.2 | 84.0 |
| Qwen3-VL | **69.6** | **84.5** | **71.5** | **70.9** | – | **61.1** | **85.7** | **96.1** | **89.6** | – | **89.6** |
| ▼ *Native Vision-Language Models (Instruct-8B)* | | | | | | | | | | | |
| Fuyu | 27.9 | 10.7 | 43.7 | – | 59.3 | – | 64.5 | – | – | – | 36.6 |
| EVE | 32.6 | 52.3 | – | – | 64.6 | 26.4 | 61.0 | 53.0 | 59.1 | 56.8 | 39.8 |
| SOLO | – | 67.7 | 44.7 | – | 64.4 | – | 61.4 | – | – | – | 12.6 |
| EVEv2 | 39.3 | 66.3 | 62.4 | – | 71.4 | – | 74.8 | – | 73.9 | 71.1 | 70.2 |
| BREEN | 42.7 | 71.4 | – | 51.2 | – | 37.0 | 76.4 | – | – | 65.7 | – |
| VoRA | 32.0 | 61.3 | 60.1 | – | 68.9 | – | 61.1 | – | – | 58.7 | – |
| SAIL | – | 70.1 | 63.9 | 53.1 | 72.9 | 54.2 | 76.7 | – | – | 77.1 | 78.3 |
| NEO | 54.6 | 82.1 | 67.3 | 62.4 | 76.3 | 46.4 | 83.1 | 88.6 | 82.1 | 75.0 | 77.7 |
| **NEO-ov** | **68.1** | **85.1** | **67.8** | **67.3** | **76.6** | **59.8** | **85.4** | **91.9** | **86.2** | **78.5** | **81.6** |

- NEO-ov **remarkably improves the capability** of previous NEO baseline
- NEO-ov **largely bridges the gap** to the top-tier modular VLMs, e.g., **Qwen3-VL.**
- NEO-ov **outperforms** the best native VLM counterparts, from **EVE** to **NEO**.

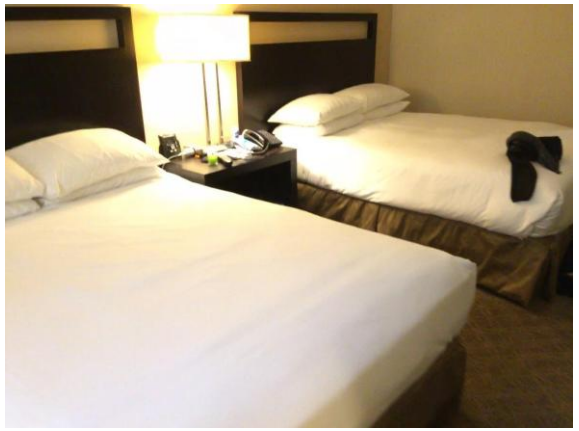
# Main Results: Multiple Image / Video Understanding

| Model | Multi-Image | | Video Understanding | | | | | |
|---|---|---|---|---|---|---|---|---|
| | BLINK | MUIRBENCH | VideoMME | MVBench | LVBench | MLVU | LongVideoBench | VideoMMMU |
| ▼ *Modular Vision-Language Models (Instruct-2B)* | | | | | | | | |
| VideoLLaMA3 | 44.2 | – | 59.6 | 65.5 | 41.6 | 65.4 | 57.1 | – |
| InternVL3.5 | 51.3 | 44.0 | 58.4 | **65.9** | 37.6 | 64.4 | **57.4** | **42.7** |
| Qwen3-VL | **53.8** | **47.4** | **61.9** | 61.7 | **47.4** | **68.3** | 55.6 | 41.9 |
| ▼ *Native Vision-Language Models (Instruct-2B)* | | | | | | | | |
| ELVA | – | – | 41.8 | 43.5 | – | 47.6 | – | – |
| **NEO-ov** | **53.9** | **56.8** | **60.4** | **65.7** | **43.3** | **64.8** | **56.8** | **42.3** |
| ▼ *Modular Vision-Language Models (Instruct-8B)* | | | | | | | | |
| LLaVA-Video | – | – | 63.3 | 58.6 | 44.2 | 70.8 | 58.2 | – |
| VideoLLaMA3 | 56.7 | – | 66.2 | 69.7 | 45.3 | 73.0 | 59.8 | – |
| InternVL3.5 | 59.5 | 55.8 | 66.0 | **72.1** | 45.9 | 70.2 | 62.1 | 54.9 |
| Qwen3-VL | **69.1** | **64.4** | **71.4** | 68.7 | **58.0** | **78.1** | **63.6** | **65.3** |
| ▼ *Native Vision-Language Models (Instruct-8B)* | | | | | | | | |
| Fuyu | – | – | 28.7 | 31.6 | – | 31.1 | – | – |
| EVE | – | – | 29.3 | 34.9 | – | 36.8 | – | – |
| ELVA | – | – | 47.1 | 51.2 | – | 51.8 | – | – |
| **NEO-ov** | **62.8** | **58.2** | **67.4** | **70.7** | **46.4** | **69.3** | **63.5** | **51.6** |

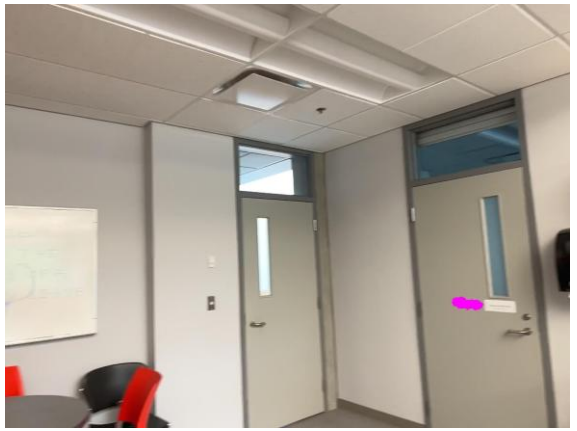
# Main Results: Spatial Intelligence

| Model | VSI-Bench | MMSI | Mindcube | ViewSpatial | SITE | 3DSR | EmbSpatial | SPAR | Omni-Spatial |
|---|---|---|---|---|---|---|---|---|---|
| ▼ *Spatial-specialist Models (Instruct-2B)* | | | | | | | | | |
| Cambrian-S (3B) | 56.1 | 27.0 | 38.4 | 41.0 | 31.0 | 41.4 | **63.5** | 33.0 | **41.9** |
| Sensenova-SI | **63.7** | **34.2** | **41.8** | **52.7** | **36.8** | **50.5** | 62.8 | **38.0** | 26.4 |
| ▼ *General-purpose Models (Instruct-2B)* | | | | | | | | | |
| InternVL3.5 | 53.8 | 25.6 | 42.1 | 37.9 | 34.8 | 31.4 | 61.5 | 32.4 | **44.4** |
| Qwen3-VL | 53.9 | 27.8 | 34.2 | 36.7 | 35.8 | 47.6 | **69.2** | 34.1 | 36.3 |
| **NEO-ov** | **58.4** | **33.6** | **77.2** | **52.8** | **38.4** | **52.9** | 63.8 | **41.2** | 43.1 |
| ▼ *Spatial-specialist Models (Instruct-8B)* | | | | | | | | | |
| Cambrian-S | 67.5 | 25.8 | 39.6 | 40.9 | 33.0 | 45.0 | 72.8 | 37.9 | **41.9** |
| Sensenova-SI | 68.8 | **43.3** | **85.7** | **54.7** | 47.7 | **55.5** | 72.0 | 45.8 | 33.0 |
| GeoThinker | **72.6** | 30.9 | 83.0 | 45.9 | **55.9** | 51.9 | **78.8** | **68.2** | 40.1 |
| ▼ *General-purpose Models (Instruct-8B)* | | | | | | | | | |
| InternVL3.5 | 56.3 | 29.1 | 40.4 | 40.0 | **54.4** | 35.3 | 75.7 | 38.2 | **47.8** |
| Qwen3-VL | 59.4 | 31.2 | 29.6 | 41.9 | 45.4 | 52.9 | 77.8 | 40.3 | 47.0 |
| **NEO-ov** | **64.8** | **41.3** | **90.0** | **55.2** | 54.3 | **61.7** | **78.8** | **48.8** | 45.0 |

# Visualized Results : Spatial Intelligence



**User:** Can you describe the position of the lamp relative to the nightstand?
**Response:** B. above



**User:** Which object has a farther proximity to the red point box: water kettle (green point) or chair (blue point)? Calculate or judge based on the 3D center points of these objects.
**Response:** A. chair (blue point)


**User:** While driving, what should you be more cautious of now?
**Response:** B. 2

**User:** Based on these two views showing the same scene: in which direction did I move from the first view to the second view?
**Response:** A. Diagonally forward and left

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

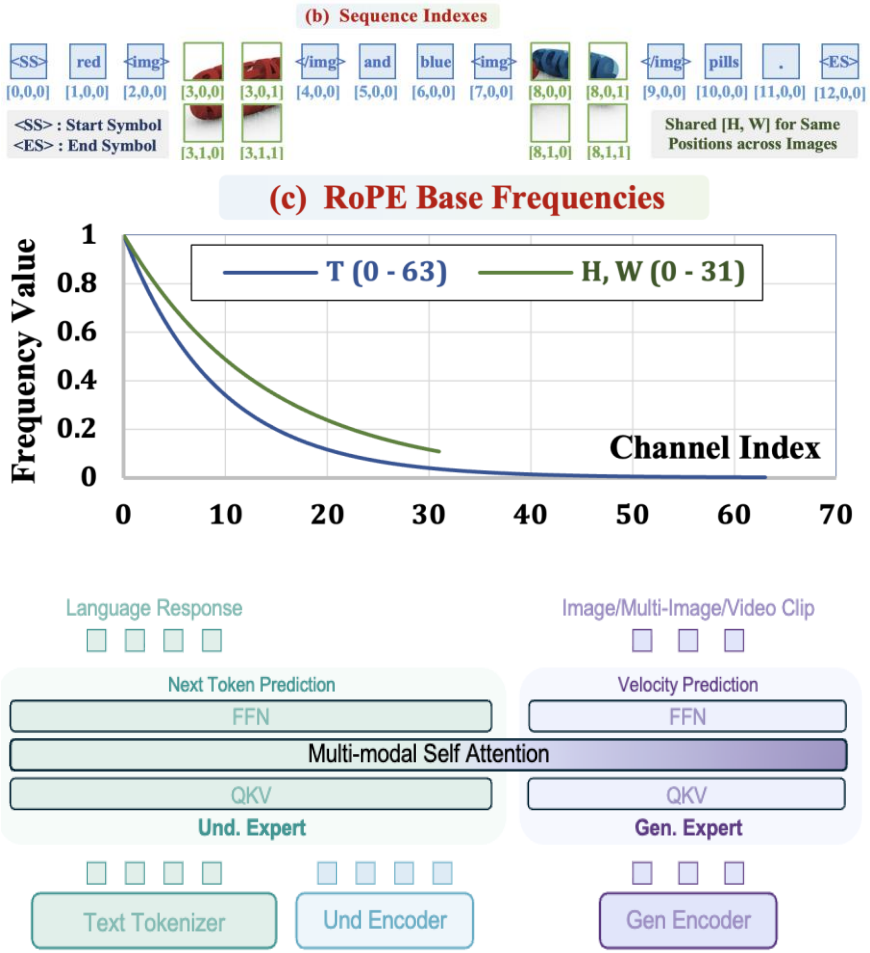
# Native VLMs for Understanding and Generation

## Between Pixels and Words -- Towards Native Multimodal Unified Models at Scale

S-Lab NTU, SenseTime Research

**NEO-unify & SenseNova-U1**, https://huggingface.co/blog/sensenova/neo-unify

# Challenge




## Question:

- Is NEO suitable for unification models?

  **Pixel and Semantic emerging from model itself !**

- What will the architecture of NEO-unify start?

  **MoE or Dense Models. First understand then generation !**

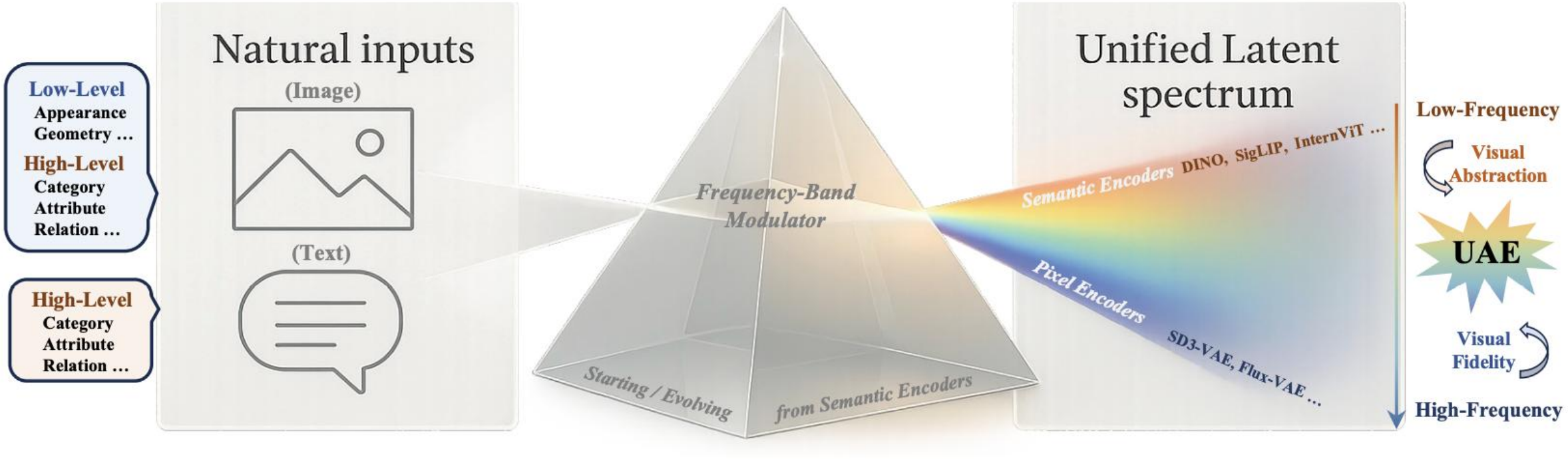
- Why do we need to build up NEO-unify?

  **End pixel-semantic argument ! Multi-modality reasoning !**

  **Embodied AI – World Model - Omni-modal Reasoning**

# Background

## (-) Representation Analyses




**The Prism Hypothesis: Harmonizing Semantic and Pixel Representations via Unified Autoencoding**
Weichen Fan, Haiwen Diao, Quan Wang, Dahua Lin, Ziwei Liu https://arxiv.org/pdf/2512.19693

# Background

## (-) Representation Analyses

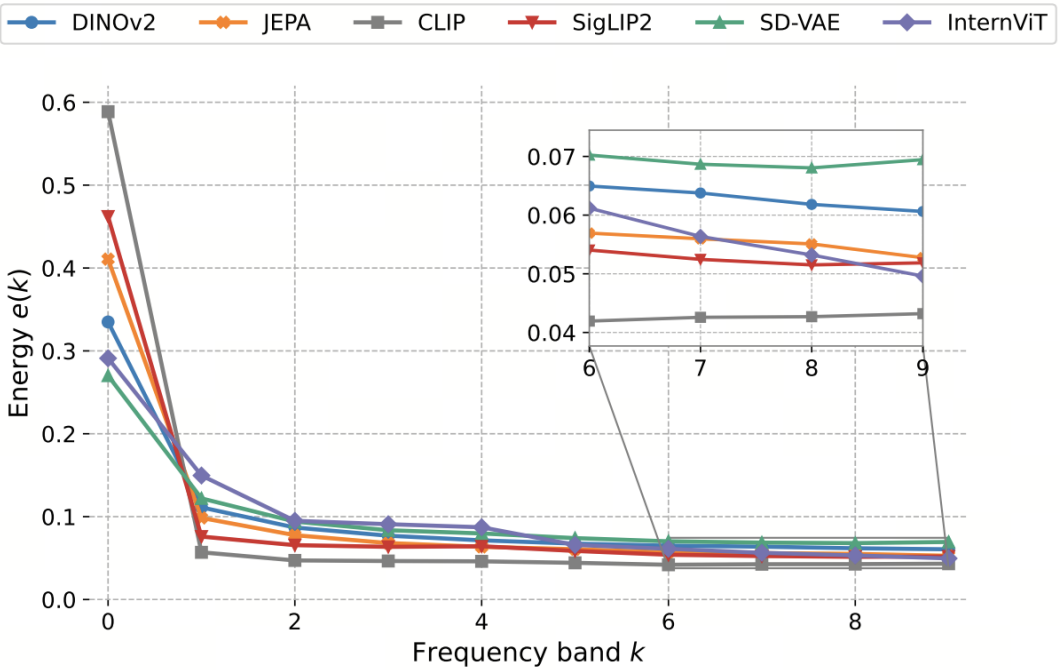



Figure 2. **Frequency energy distribution.** Normalized energy $e(k)$ across frequency bands for diverse tokenizers. DINOv2 and CLIP focus on low-frequency (semantic) content, while SD-VAE retains more high-frequency energy, capturing finer details.

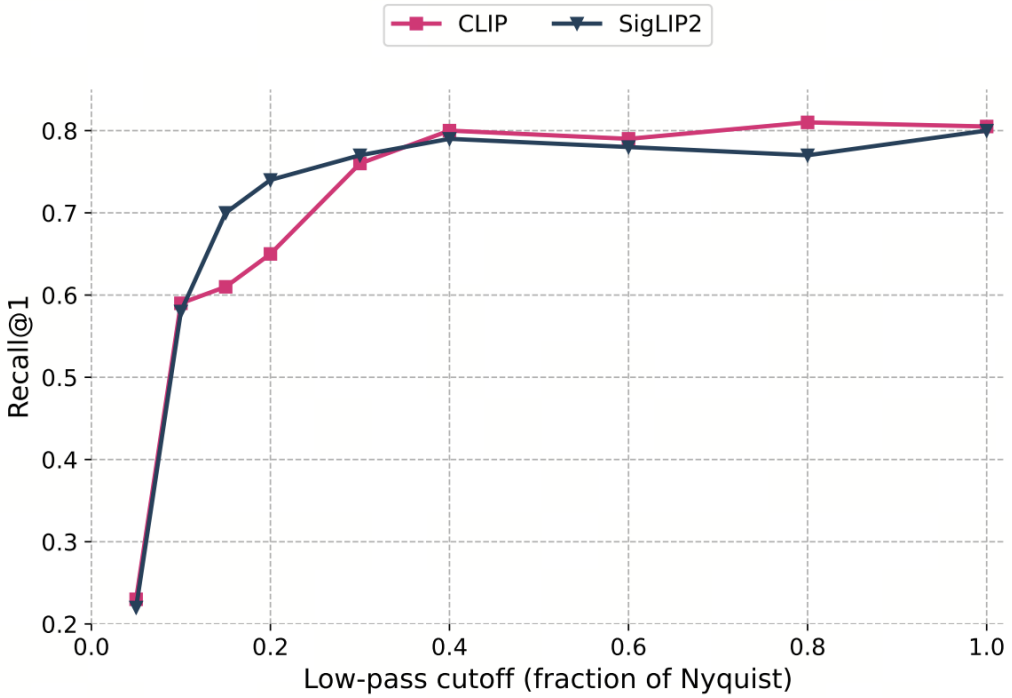
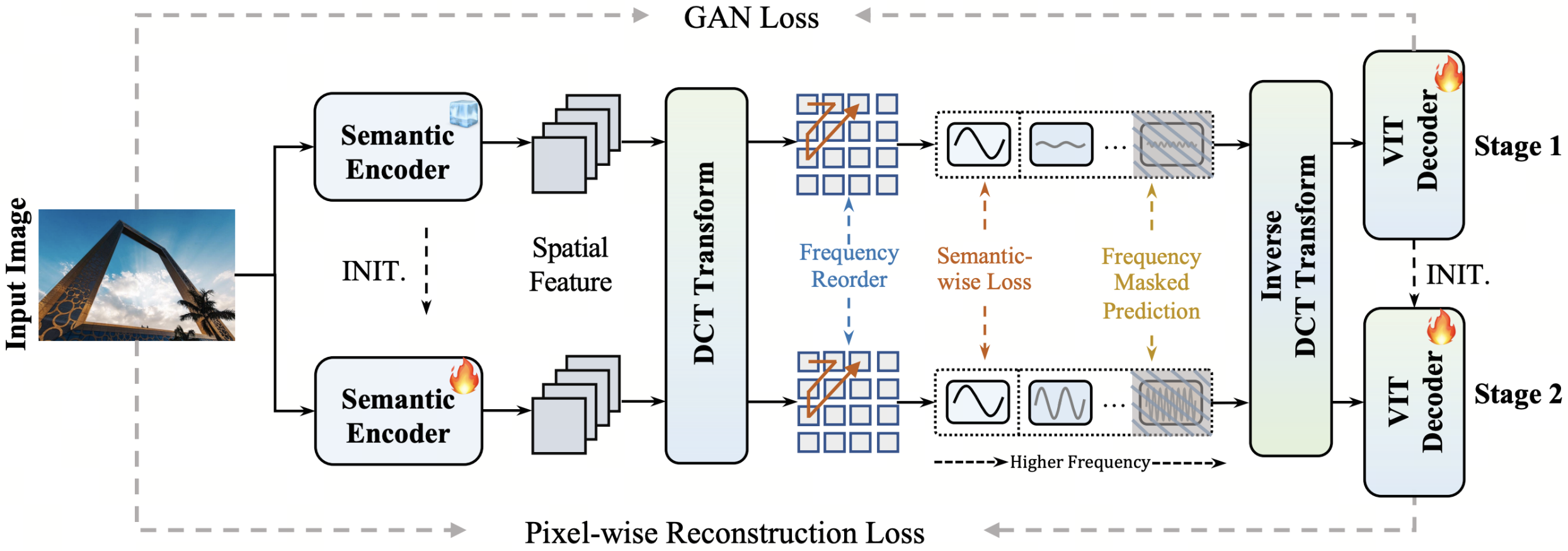



Figure 3. **Retrieval results via frequency filtering.** Text–Image retrieval remains stable under low-pass filtering but degrades sharply under high-pass filtering, confirming that semantic alignment primarily resides in low-frequency components.

# Background

## (-) Representation Analyses



# Background

## (-) Representation Analyses

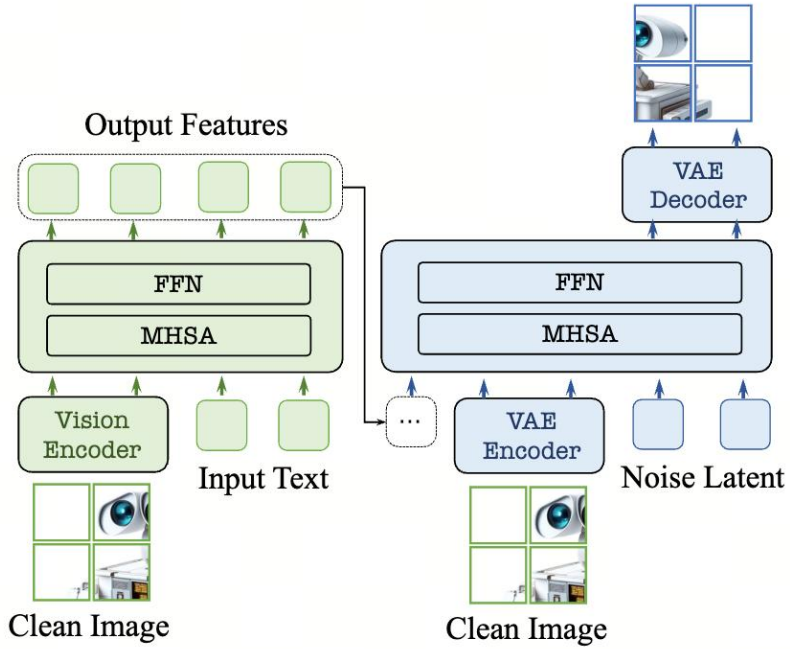
| Method | Type | Ratio | Linear prob (Acc)↑ | ImageNet-1K | | | MS-COCO 2017 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | PSNR↑ | SSIM↑ | rFID↓ | PSNR↑ | SSIM↑ | rFID↓ |
| SD-VAE-1.x [33] | Continuous | 8 | – | 23.54 | 0.68 | 1.22 | 23.21 | 0.69 | 5.94 |
| SD-VAE [33] | Continuous | 8 | – | 25.68 | 0.72 | 0.75 | 25.43 | 0.73 | 0.76 |
| SD-VAE-2.x [33] | Continuous | 8 | – | 23.54 | 0.68 | 1.22 | 26.62 | 0.77 | 4.26 |
| SD-VAE-XL [28] | Continuous | 8 | – | 27.37 | 0.78 | 0.67 | 27.08 | 0.80 | 3.93 |
| SD-VAE-3 [20] | Continuous | 8 | – | 31.29 | 0.87 | 0.20 | 31.18 | 0.89 | 1.67 |
| FLUX-VAE [18] | Continuous | 8 | – | **32.74** | **0.92** | **0.18** | **32.32** | **0.93** | **1.35** |
| VA-VAE | Continuous | 16 | – | 27.96 | 0.79 | 0.28 | 27.50 | 0.81 | 2.71 |
| SVG (DINOv3) [34] | Continuous | 16 | – | 24.25 | 0.67 | 0.78 | - | - | - |
| RAE (DINOv2-B) [50] | Continuous | 14 | – | 18.05 | 0.5 | 2.04 | 18.36 | 0.47 | 6.01 |
| UniFlow (DINOv2-L) [49] | Continuous | 14 | – | 32.32 | 0.91 | 0.17 | 30.66 | 0.94 | 2.81 |
| **UAE (Siglip2)** | Continuous | 16 | **80.1 (81.2)** | **31.00** | **0.91** | **0.43** | **30.20** | **0.89** | **2.91** |
| **UAE (DINOv2-B)** | Continuous | 14 | **83.0 (83.0)** | **32.17** | **0.92** | **0.35** | **31.19** | **0.91** | **2.01** |
| **UAE (CLIP-L)** | Continuous | 14 | **77.2 (79.4)** | **36.58** | **0.96** | **0.04** | **36.25** | **0.97** | **0.41** |

**Strong image fidelity** together with competitive **semantic representation quality**.

# Background

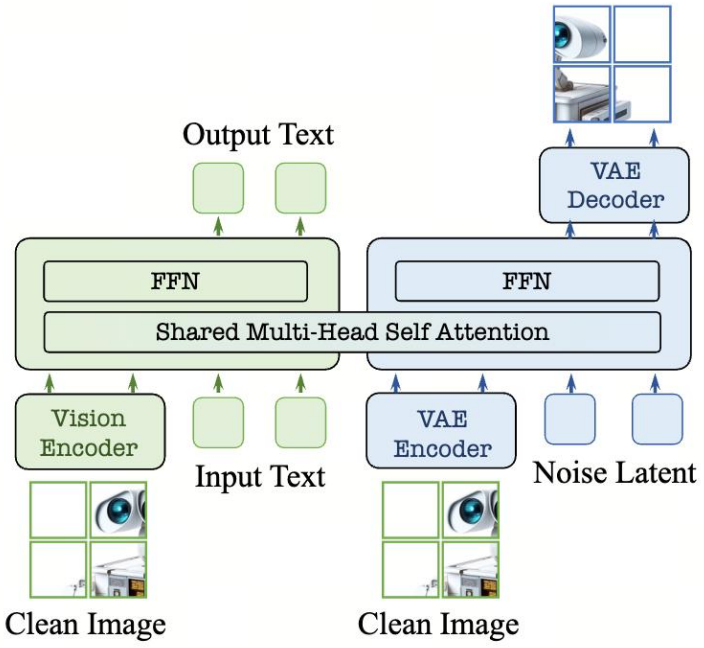
## (-) Architecture Analyses

**(1) Sequential Architecture with Vision Encoder, Projector, VAE-like Encoder and Decoder**



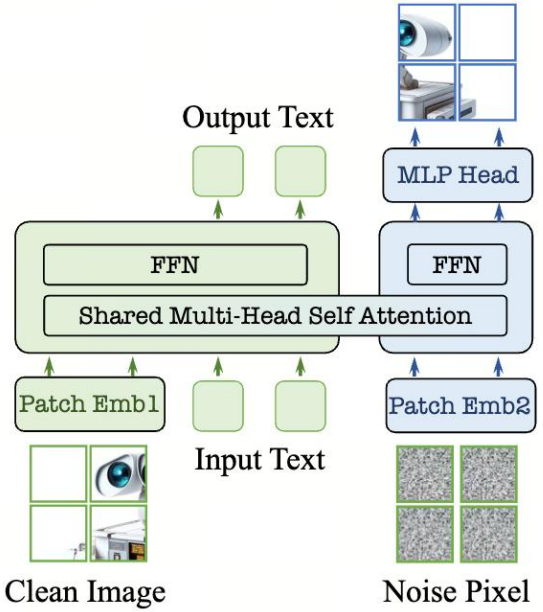

Related Works: Omnigen2, Qwen-Image, etc.

**(2) Parallel Architecture with Vision Encoder, Projector, VAE-like Encoder and Decoder**




Related Works: MoT, LMFusion, Bagel, etc.

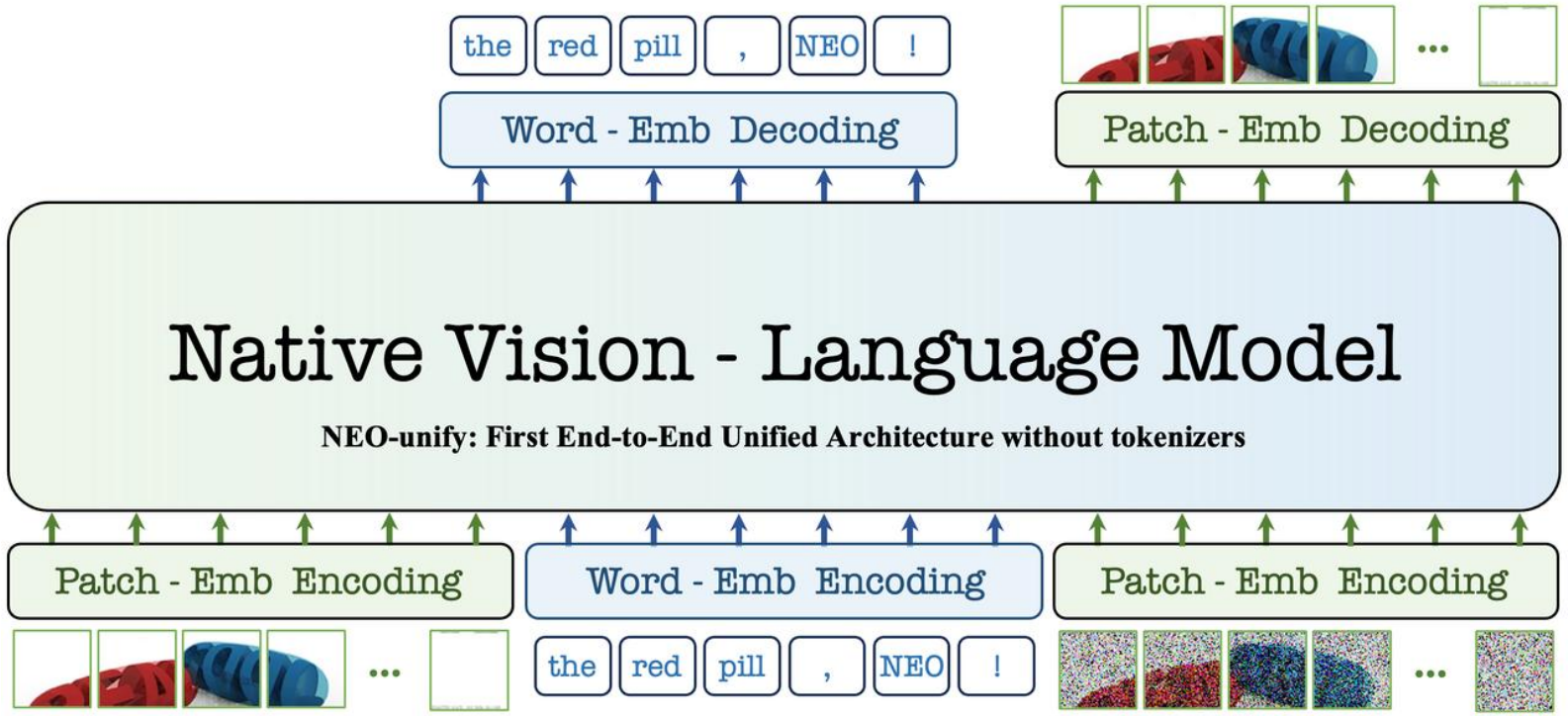
**(3) Parallel Native Architecture with Lightweight Encoding / Decoding Layers**




NEO-unify, SenseNova-U1

# Methodology

## (-) Model Architecture




### Native Unified VLMs

- **Near-lossless visual interface** for input and output via patch embedding
- **Native Mixture-of-Transformer (MoT)** for synergizing understanding and generation as native architecture
- **Unified learning** with autoregressive cross-entropy modeling for texts and pixel flow matching for vision.

# Main Results: Image Understanding

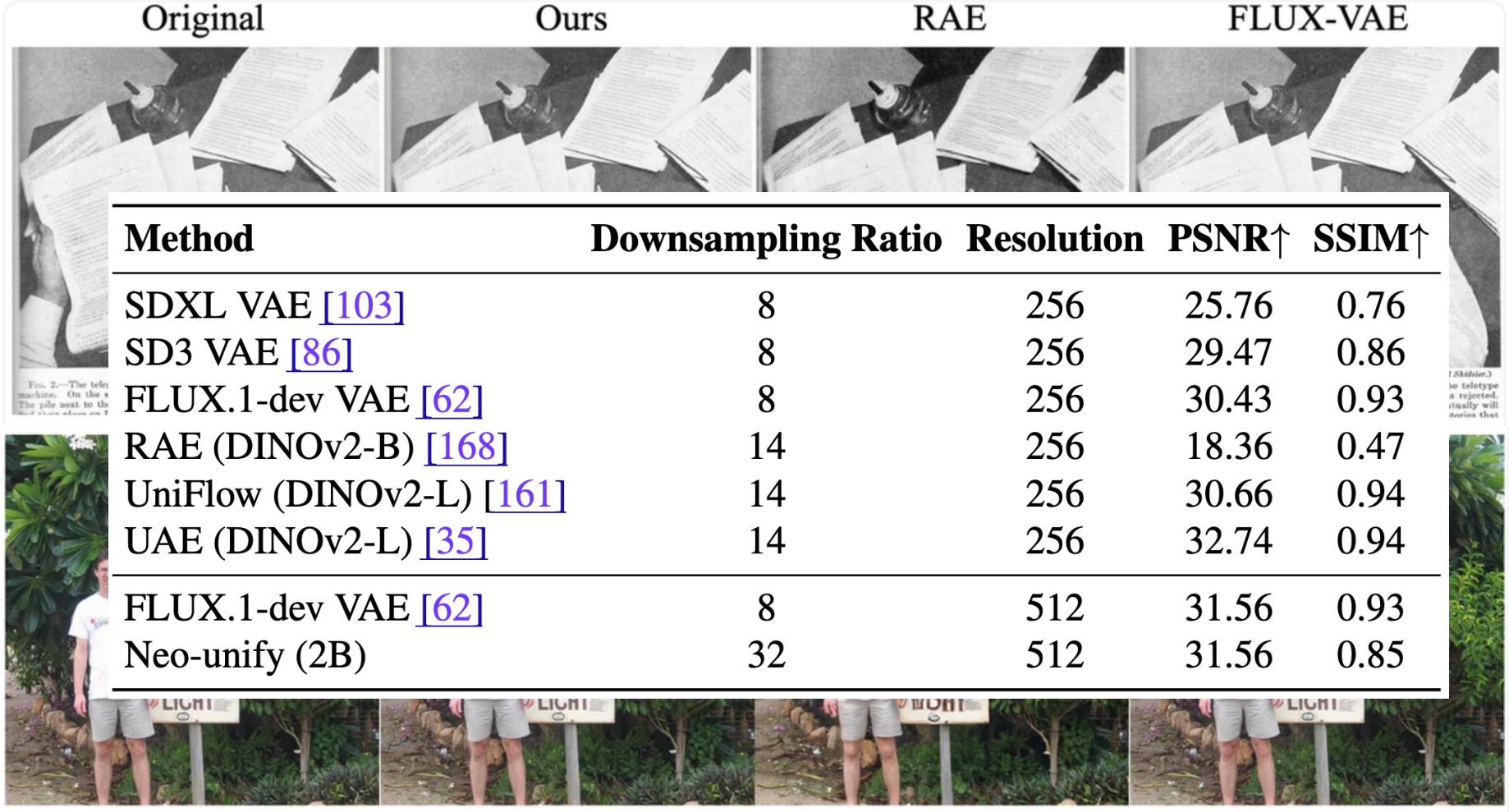
| Model | MMMU | MMB | RWQA | MMStar | SEED | HallB | AI2D | DocVQA | ChartQA | TextVQA | OCRBench |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ▼ *Vision-Language Models (2B)* | | | | | | | | | | | |
| Qwen2-VL | 41.1 | 74.9 | 62.6 | 48.0 | – | 41.7 | 74.7 | 90.1 | 73.5 | **79.7** | 80.9 |
| InternVL3 | 48.6 | **81.1** | **64.3** | 60.7 | – | 42.5 | 78.7 | 88.3 | 80.2 | 77.0 | 83.5 |
| InternVL3.5 | 53.0 | 78.2 | 62.0 | **62.7** | – | 48.6 | **78.8** | 89.4 | **80.7** | 76.5 | 83.6 |
| Qwen3-VL | **53.4** | 78.4 | 63.9 | 58.3 | – | **51.4** | 76.9 | **93.3** | 79.1 | – | **85.8** |
| Mono-VL | 33.7 | 65.5 | – | – | – | 34.8 | 68.6 | 80.0 | 73.7 | 72.6 | 76.7 |
| Mono-VL1.5 | 39.1 | 64.0 | – | – | – | 32.5 | 67.4 | 81.7 | 72.2 | 73.7 | 80.1 |
| HoVLE | 32.2 | 73.3 | – | – | – | 38.4 | 73.0 | 86.1 | 78.6 | 70.9 | 74.0 |
| NEO | 48.6 | 76.0 | 63.1 | 54.2 | – | 43.1 | 80.1 | 89.9 | 81.2 | 74.0 | 77.1 |
| Janus-Pro | 36.3 | 75.5 | – | – | 68.3 | – | – | – | – | – | – |
| Show-o2 | 37.1 | 67.4 | – | 43.4 | 65.6 | – | 69.0 | – | – | – | – |
| OneCAT | 39.0 | 72.4 | – | – | 70.9 | – | 72.4 | 87.1 | 76.2 | 67.0 | – |
| BAGEL | 43.2 | **79.2** | – | – | – | – | – | – | – | – | – |
| **NEO-unify** | **53.8** | 78.2 | **63.5** | **59.5** | **75.0** | **61.0** | **81.1** | **91.2** | **80.4** | **77.4** | **81.0** |
| ▼ *Vision-Language Models (8B)* | | | | | | | | | | | |
| Qwen2.5-VL | 55.0 | 83.5 | 68.5 | 63.9 | – | 52.9 | 83.9 | 95.7 | 87.3 | **84.9** | 86.4 |
| InternVL3 | 62.7 | 83.4 | 70.8 | 68.2 | – | 49.9 | 85.2 | 92.7 | 86.6 | 80.2 | 88.0 |
| InternVL3.5 | 68.1 | 82.7 | 67.5 | 69.3 | – | 54.5 | 84.0 | 92.3 | 86.7 | 78.2 | 84.0 |
| Qwen3-VL | **69.6** | **84.5** | **71.5** | **70.9** | – | **61.1** | **85.7** | **96.1** | **89.6** | – | **89.6** |
| Fuyu | 27.9 | 10.7 | 43.7 | – | – | – | 64.5 | – | – | – | 36.6 |
| EVE | 32.6 | 52.3 | – | – | – | 26.4 | 61.0 | 53.0 | 59.1 | 56.8 | 39.8 |
| EVEv2 | 39.3 | 66.3 | 62.4 | – | – | – | 74.8 | – | 73.9 | 71.1 | 70.2 |
| NEO | 54.6 | 82.1 | 67.3 | 62.4 | – | 46.4 | 83.1 | 88.6 | 82.1 | 75.0 | 77.7 |
| Janus-Pro | 41.0 | 79.2 | – | – | 72.1 | – | – | – | – | – | – |
| Show-o2 | 48.9 | 79.3 | – | 56.6 | 69.8 | – | 78.6 | – | – | – | – |
| Mogao | 44.2 | 75.0 | – | – | 74.6 | – | – | – | – | – | – |
| BAGEL | 55.3 | 85.0 | – | – | – | – | – | – | – | – | – |
| **NEO-unify** | **68.9** | **85.1** | **67.5** | **65.5** | **75.8** | **67.2** | **85.8** | **91.6** | **84.9** | **78.6** | **81.5** |

# Main Results: Image Generation

| Model | GenEval | DPG-Bench | WISE | LongText-en | LongText-zh |
|---|---|---|---|---|---|
| GLM-Image | – | 84.78 | – | 0.952 | 0.979 |
| Z-Image | 84 | 88.14 | – | 0.935 | 0.936 |
| Qwen-Image | **87** | 88.32 | **62** | 0.943 | 0.946 |
| Seedream4.5 | – | **88.63** | – | **0.989** | **0.987** |
| ▼ *Vision-Language Models (2B)* | | | | | |
| Janus-Pro | 73 | 82.63 | – | – | – |
| Show-o2 | 73 | 85.02 | – | – | – |
| OneCat | 85 | 81.72 | – | – | – |
| **NEO-unify** | 84 [87] | 86.54 | 41 (59) | 0.748 | 0.495 |
| ▼ *Vision-Language Models (8B)* | | | | | |
| Janus-Pro | 80 | 84.19 | 35 | – | – |
| Show-o2 | 76 | 86.14 | – | – | – |
| OmniGen2 | 80 [86] | 83.57 | – | 0.561 | 0.059 |
| BAGEL | 82 [88] | 85.07 | 52 (70) | 0.373 | 0.310 |
| Mogao | – [89] | 84.33 | – | – | – |
| **NEO-unify** | **85** [90] | **86.71** | 47 (**72**) | **0.914** | **0.755** |

# Key Insights

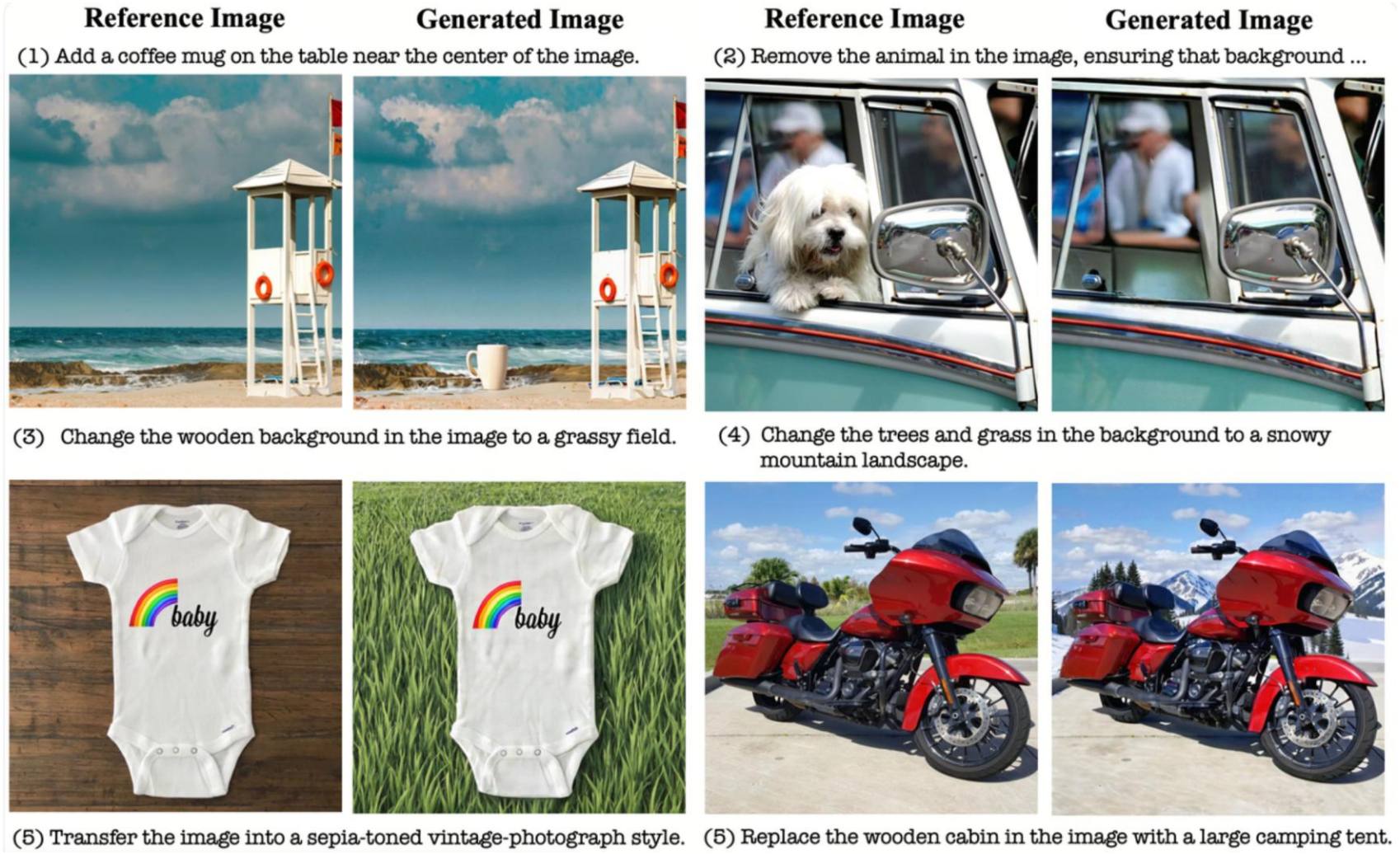
## 1. Encoder-Free Design Preserves Both Semantic and Pixel Representations




| Method | Downsampling Ratio | Resolution | PSNR↑ | SSIM↑ |
|---|---|---|---|---|
| SDXL VAE [103] | 8 | 256 | 25.76 | 0.76 |
| SD3 VAE [86] | 8 | 256 | 29.47 | 0.86 |
| FLUX.1-dev VAE [62] | 8 | 256 | 30.43 | 0.93 |
| RAE (DINOv2-B) [168] | 14 | 256 | 18.36 | 0.47 |
| UniFlow (DINOv2-L) [161] | 14 | 256 | 30.66 | 0.94 |
| UAE (DINOv2-L) [35] | 14 | 256 | 32.74 | 0.94 |
| FLUX.1-dev VAE [62] | 8 | 512 | 31.56 | 0.93 |
| Neo-unify (2B) | 32 | 512 | 31.56 | 0.85 |

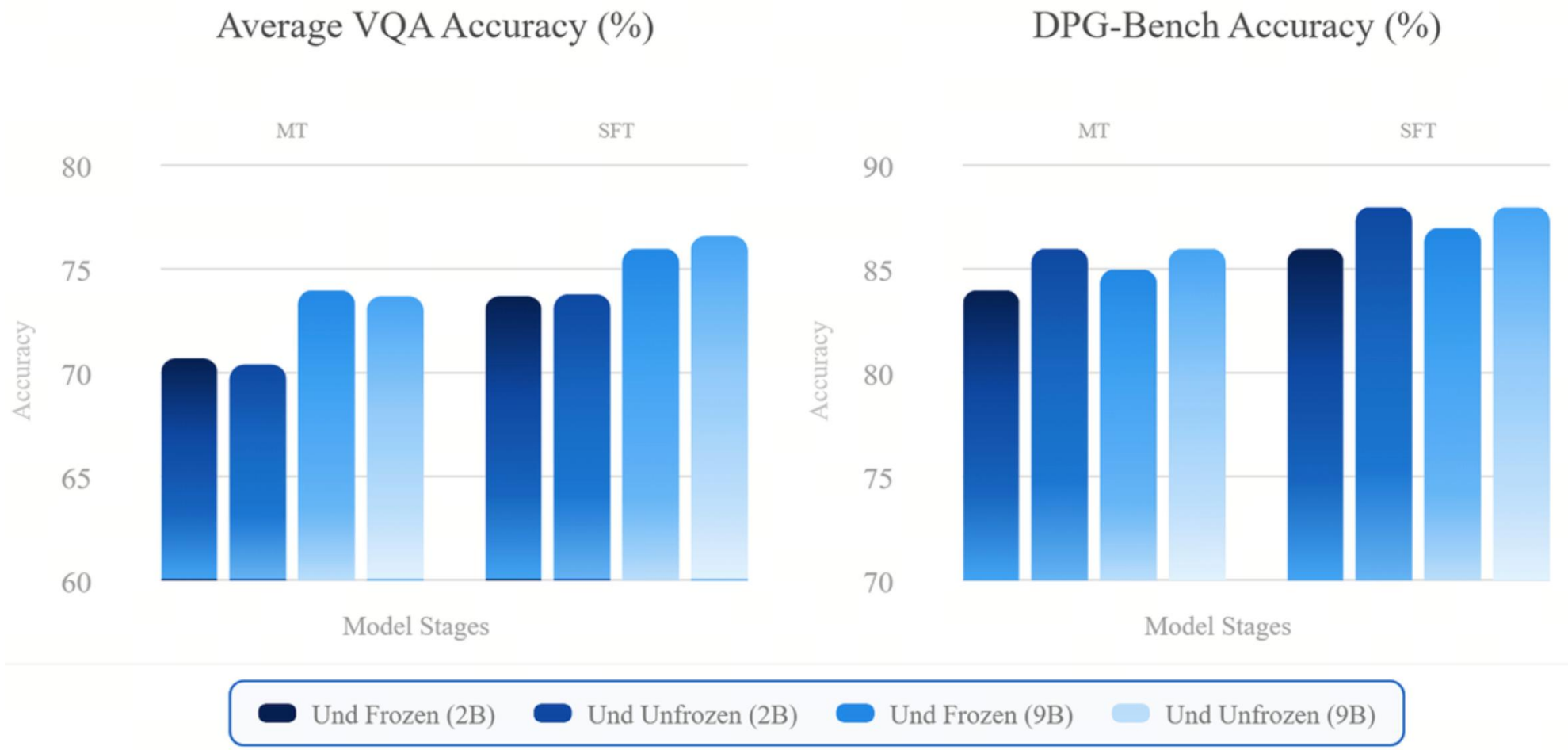
# Key Insights

## 1. Encoder-Free Design Preserves Both Semantic and Pixel Representations



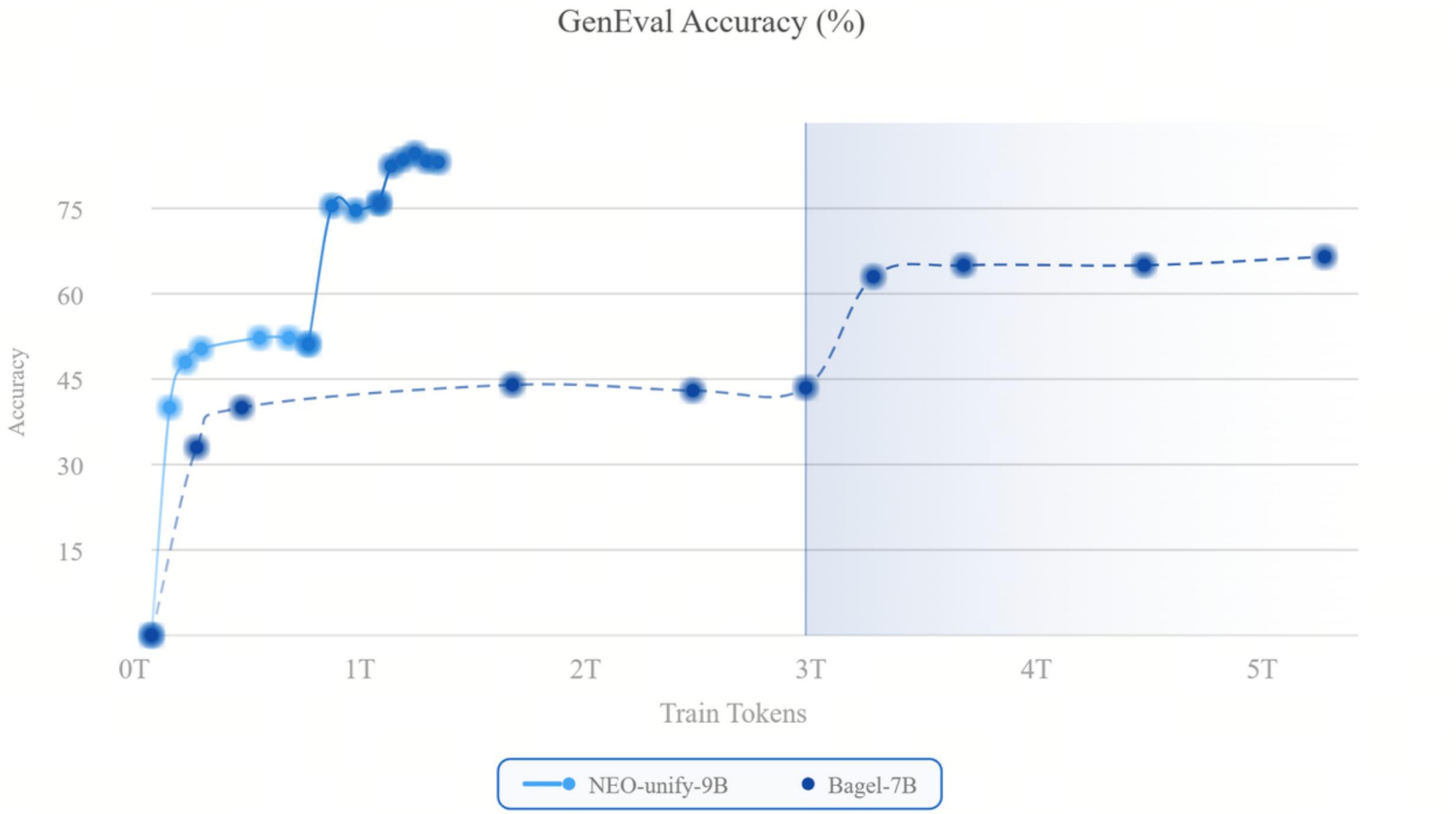
# Key Insights

## 2. Encoder-Free Design Synergizes with MoT Backbone with Minimal Intrinsic Conflict

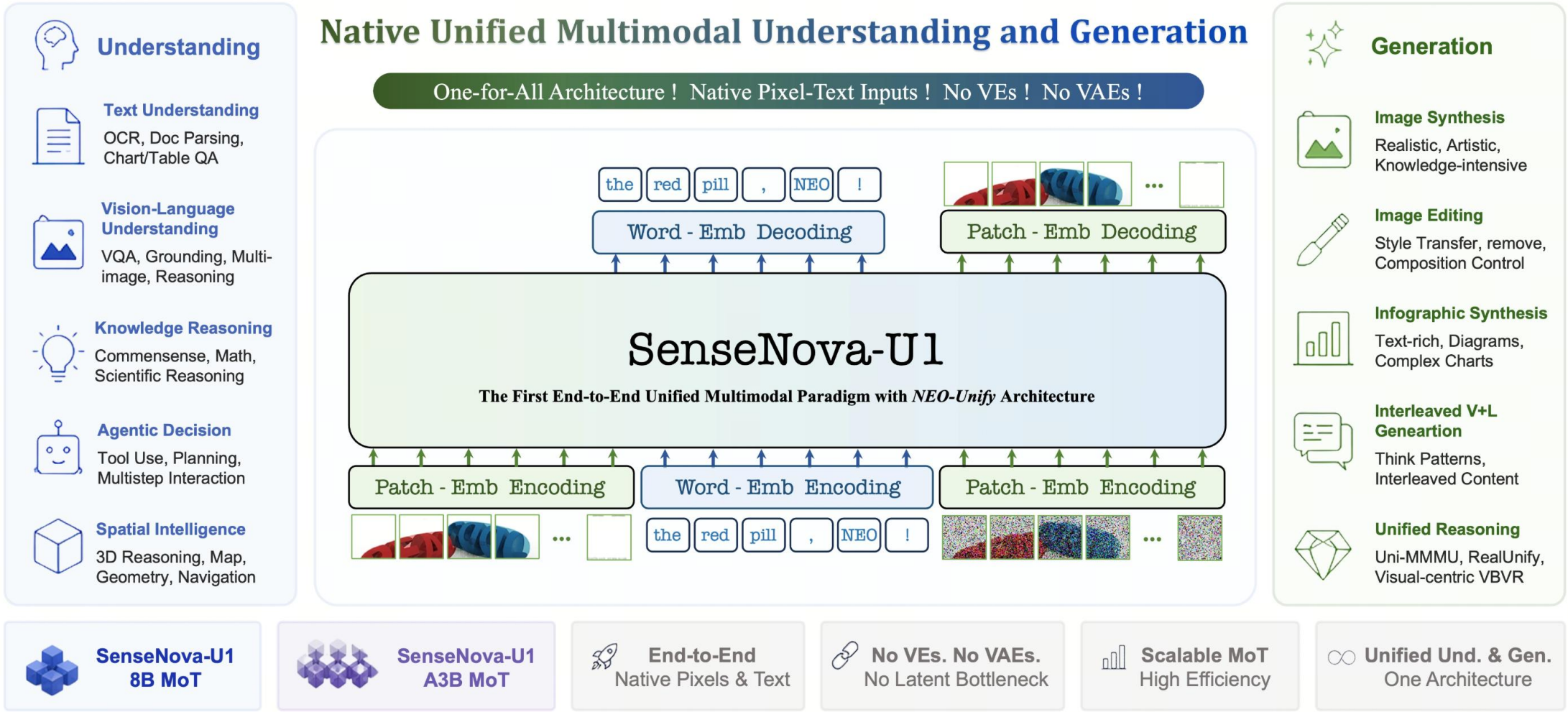


# Key Insights

### 3. Encoder-Free Design Shows High Data-scaling Efficiency

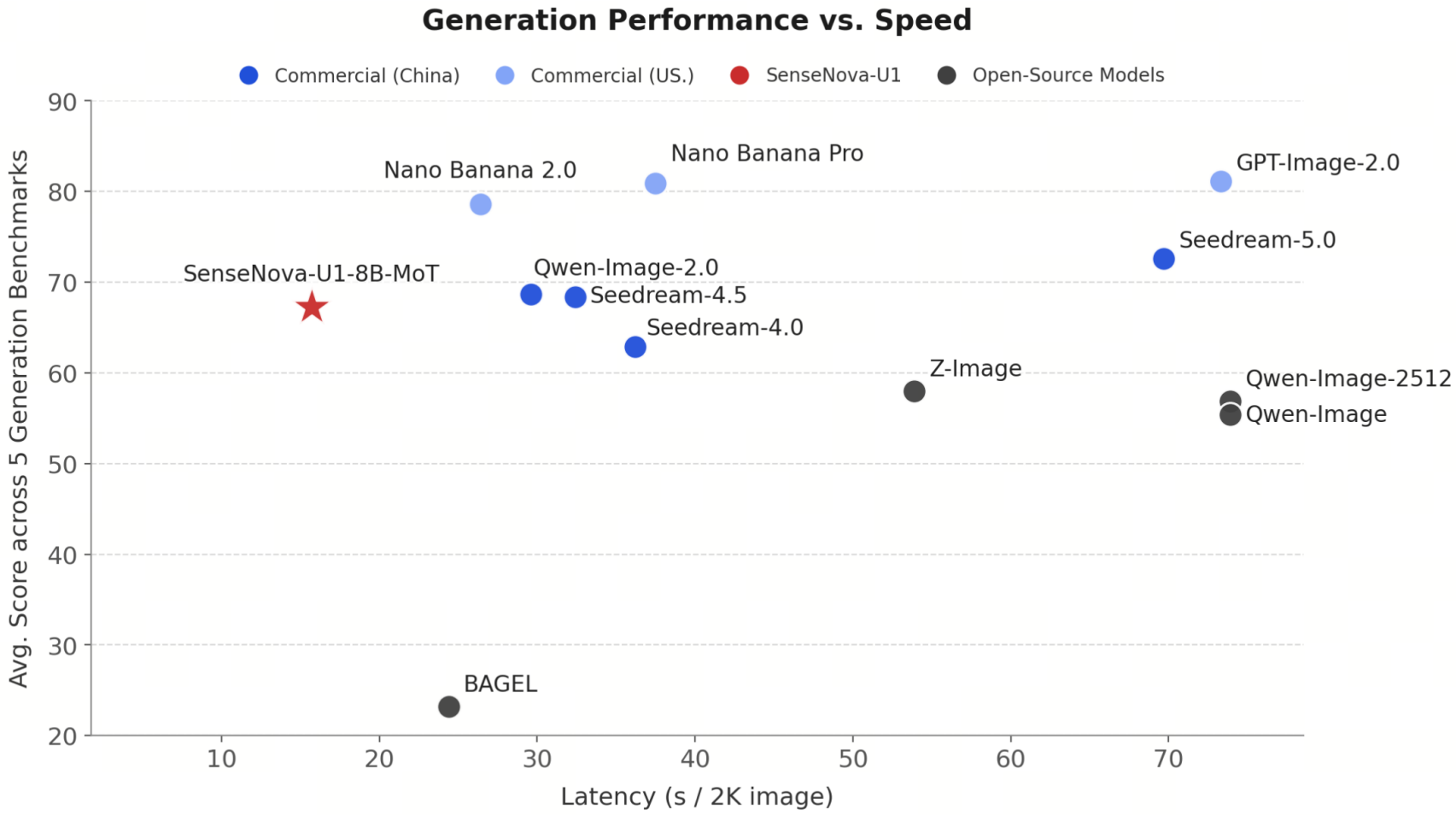


# SenseNova-U1: Further Scaling Data, Capacity, Applications




Arxiv: https://arxiv.org/abs/2605.12500 Github: https://github.com/OpenSenseNova/SenseNova-U1

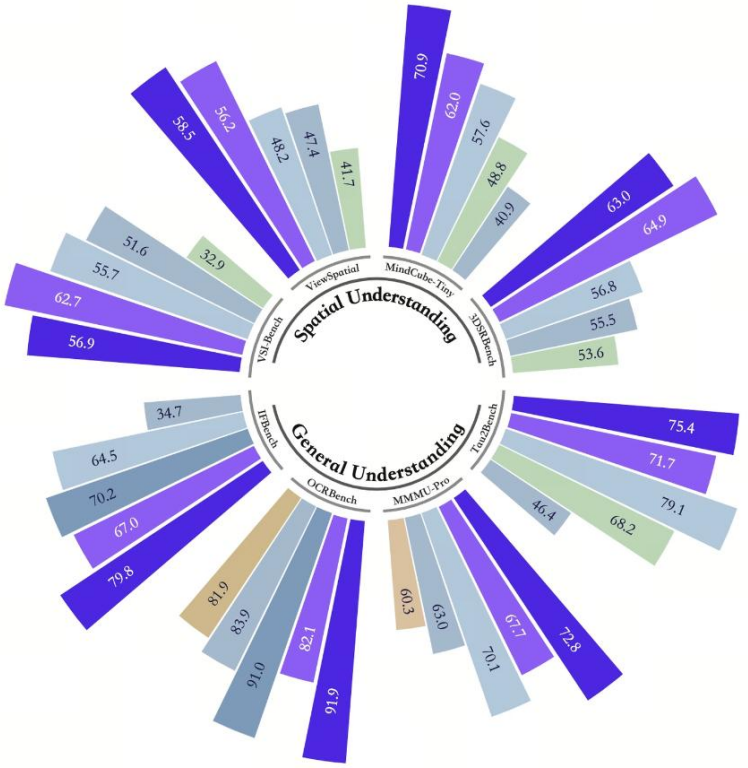
# Efficient & Powerful

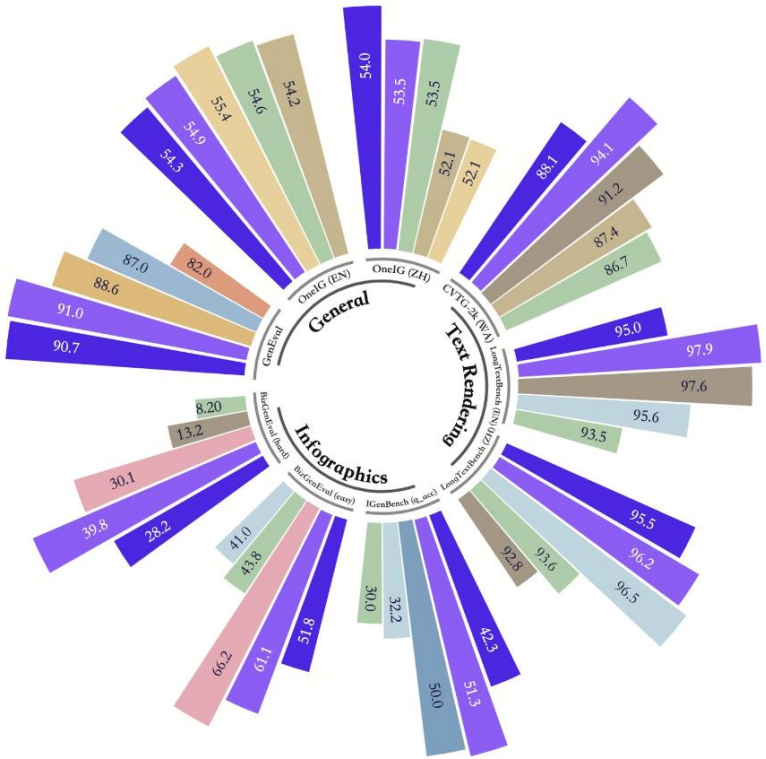


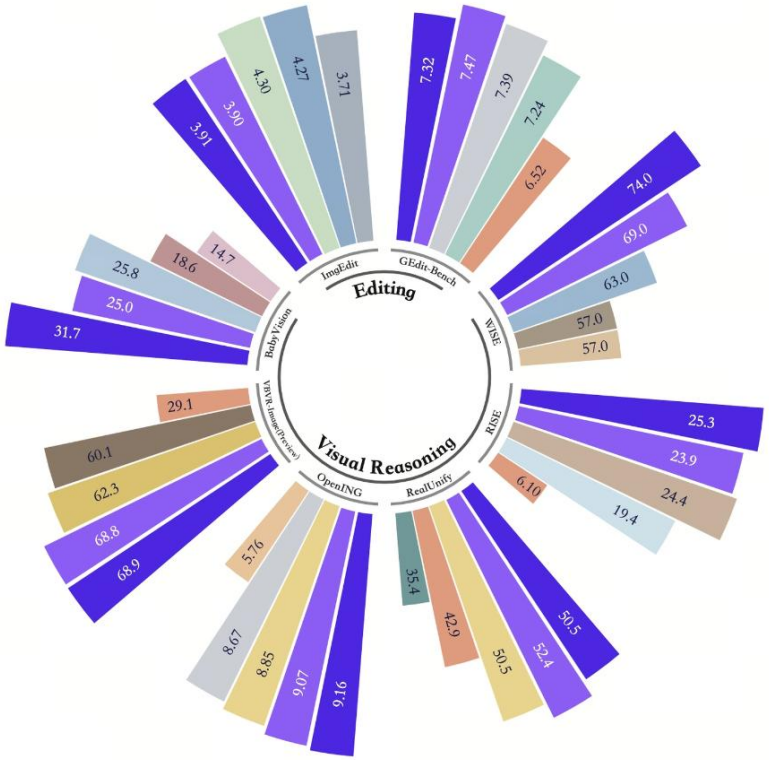

**Speed Analysis:**

1. 32x downsampling ratio
2. Only one visual context
3. Refine QK attention dim

# Main Results on 43 Understanding & Generation Tasks



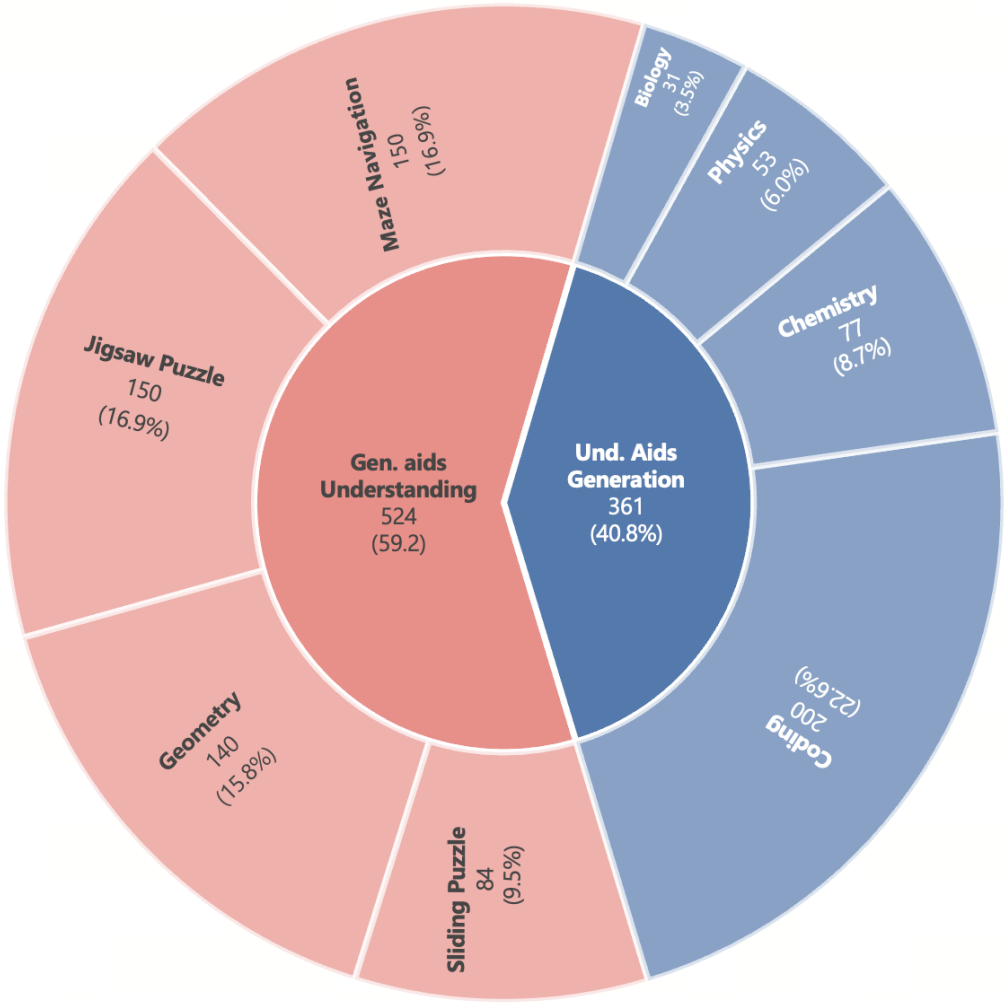




SenseNova-U1 shows **SOTA open-source results** across Visual Understanding / Reasoning, Spatial Intelligence,

Visual Generation, Visual Editing, Interleaved Generation, Reasoning Generation, Text Rendering & Infographic

# Main Results on Unification Benchmarks: Uni-MMMU

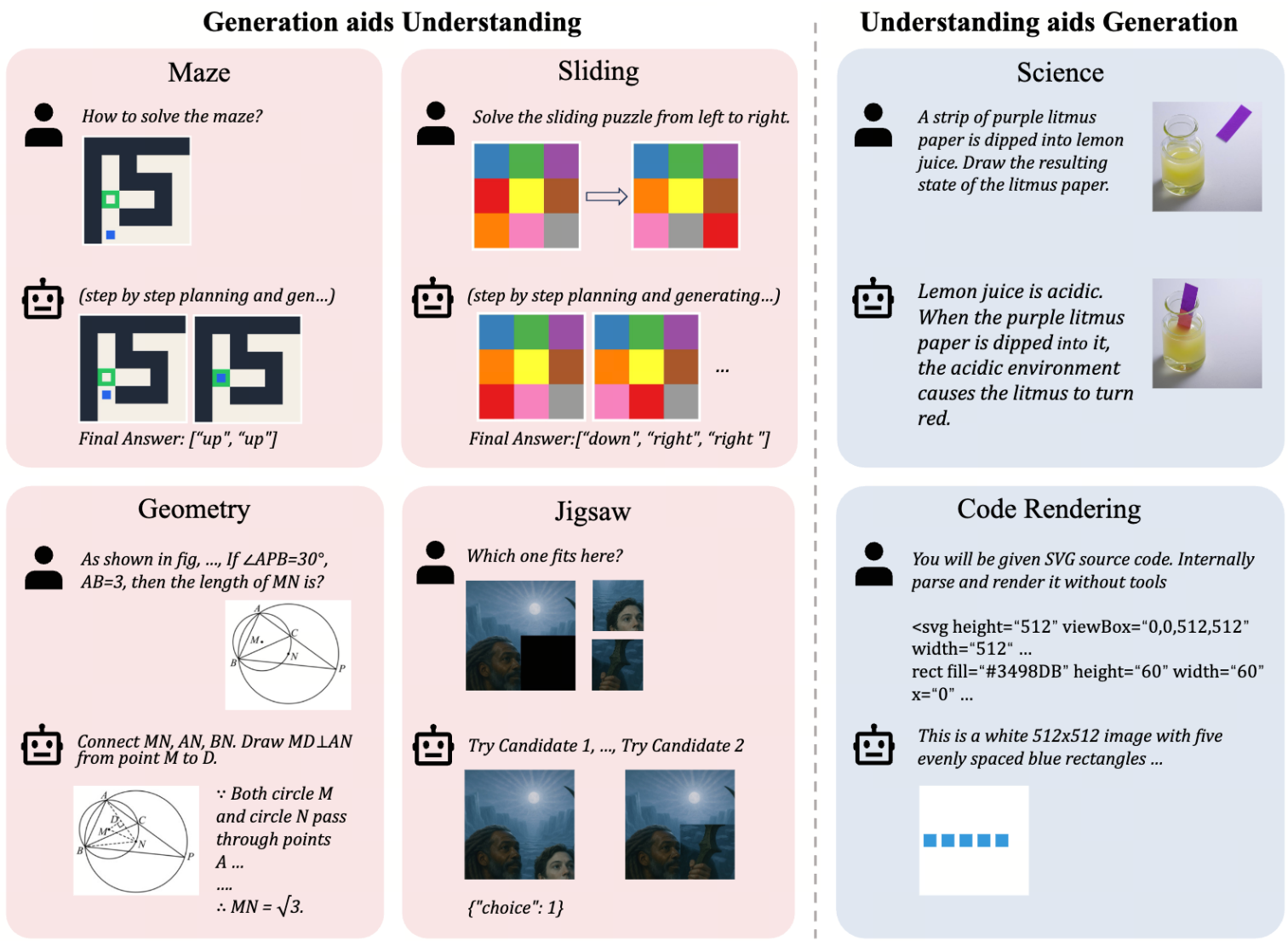



| | MMU | Gen&Edit | Multi-Turn | Dual Eval |
|---|---|---|---|---|
| MMMU | ✓ | ✗ | ✗ | ✗ |
| WISE | ✗ | ✗ | ✗ | ✗ |
| RISEBench | ✗ | ✓ | ✗ | ✗ |
| OpenING | ✓ | ✓ | ✓ | ✗ |
| MME-Unify | ✓ | ✓ | ✗ | ✗ |
| UniEval | ✓ | ✗ | ✗ | ✗ |
| **Uni-MMMU** | ✓ | ✓ | ✓ | ✓ |

**Uni-MMMU: A Massive Multi-discipline Multimodal Unified Benchmark (ACL2026 Oral)**

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, Ziwei Liu

# Main Results on Unification Benchmarks: Uni-MMMU



# Main Results on Unification Benchmarks: Uni-MMMU

| Model | # Params | Jigsaw-T | Maze-T | Sliding-T | Geometry-T | Avg↑ |
|---|---|---|---|---|---|---|
| *Closed-source Models* | | | | | | |
| Nano-Banana [26] | - | 57.0 | 4.7 | 0.0 | 47.8 | 27.4 |
| *Open-source Models* | | | | | | |
| **SenseNova-U1-SFT** | **8B** | 87.3 | 28.6 | 0.0 | 24.2 | **35.0** |
| **SenseNova-U1-SFT** | **8BA3B** | 88.0 | 34.0 | 1.2 | 7.1 | **32.6** |
| BAGEL [28] | 7B | 48.0 | 0.0 | 1.2 | 32.8 | 20.5 |
| Ovis-U1 [130] | 1.2B | 53.0 | 0.0 | 0.0 | 3.5 | 14.1 |
| OmniGen2 [141] | 4B | 48.0 | 0.0 | 0.0 | 5.7 | 13.4 |
| Qwen-Image-Edit [139] | 20B | 43.3 | 0.7 | 0.0 | 8.5 | 13.1 |

SenseNova-U1 shows strong capability for **Generation and Understanding Synergy.**

# Visualized Results : Interleaved Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

INPUT 香氛蜡烛的产品宣传图，多图

为您设计了六张不同维度的香氛蜡烛产品宣传图，旨在通过视觉语言传达产品的高级感与治愈氛围：

1 清晨仪式感

利用百叶窗的自然光影，营造纯净、高级的早晨氛围。


2 材质工艺细节

微距视角展示陶瓷杯身的磨砂质感与平整的蜡面，体现高品质。

3 点燃瞬间

捕捉跳动的火焰与融化的蜡池，传达温暖与生命力。


4 香调视觉化

通过森林系元素（如尤加利叶）的搭配，唤起嗅觉联想。

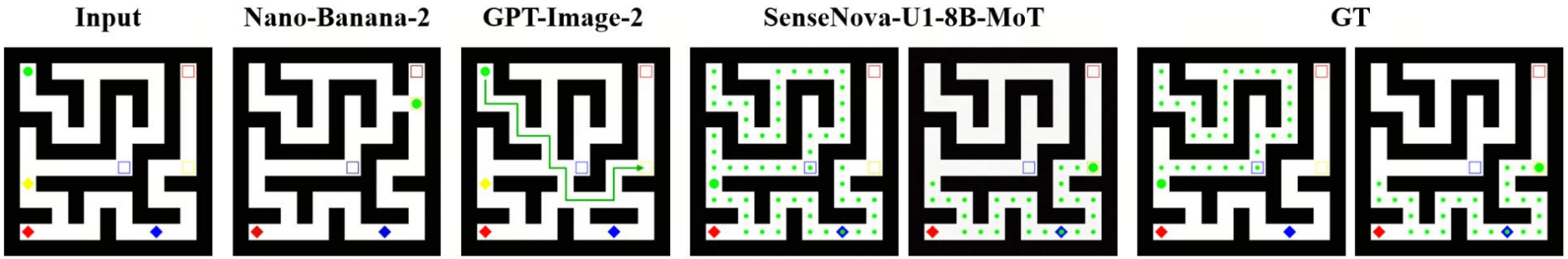
5 深夜静谧陪伴

在床头场景中，利用暖色调营造私密、安全的居家氛围。

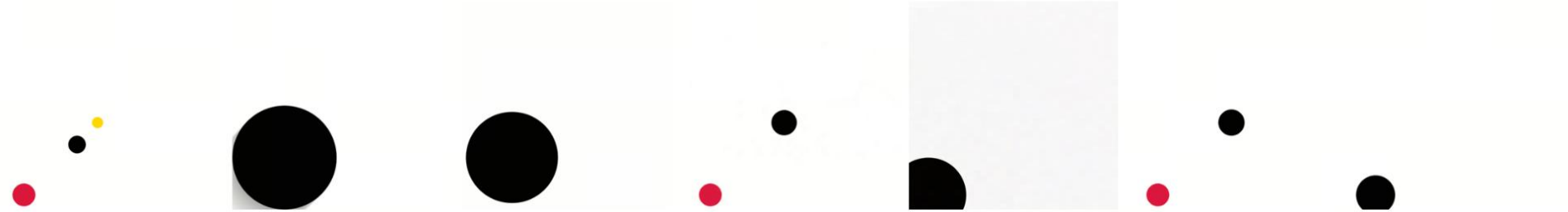

6 多场景适配

展示产品在现代简约浴室或玄关中的百搭美学。

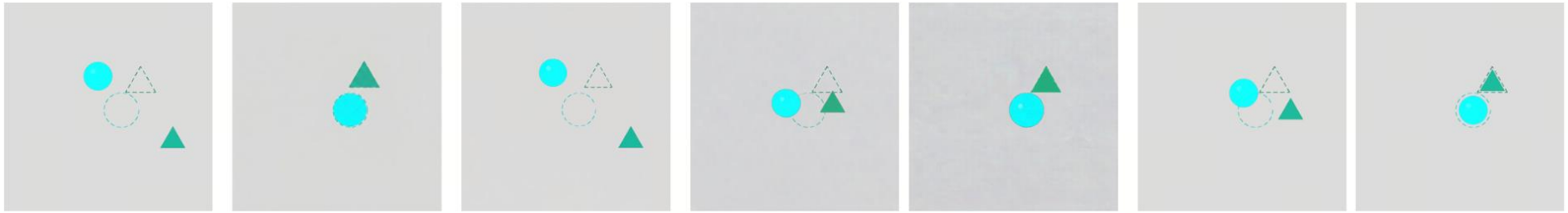
# Visualized Results : Visual Reasoning




**Prompt** The scene shows a maze with a green circular agent, colored diamond-shaped keys, and colored hollow rectangular doors. Find the Yellow key and then navigate to the matching Yellow door, showing the complete movement process step by step.

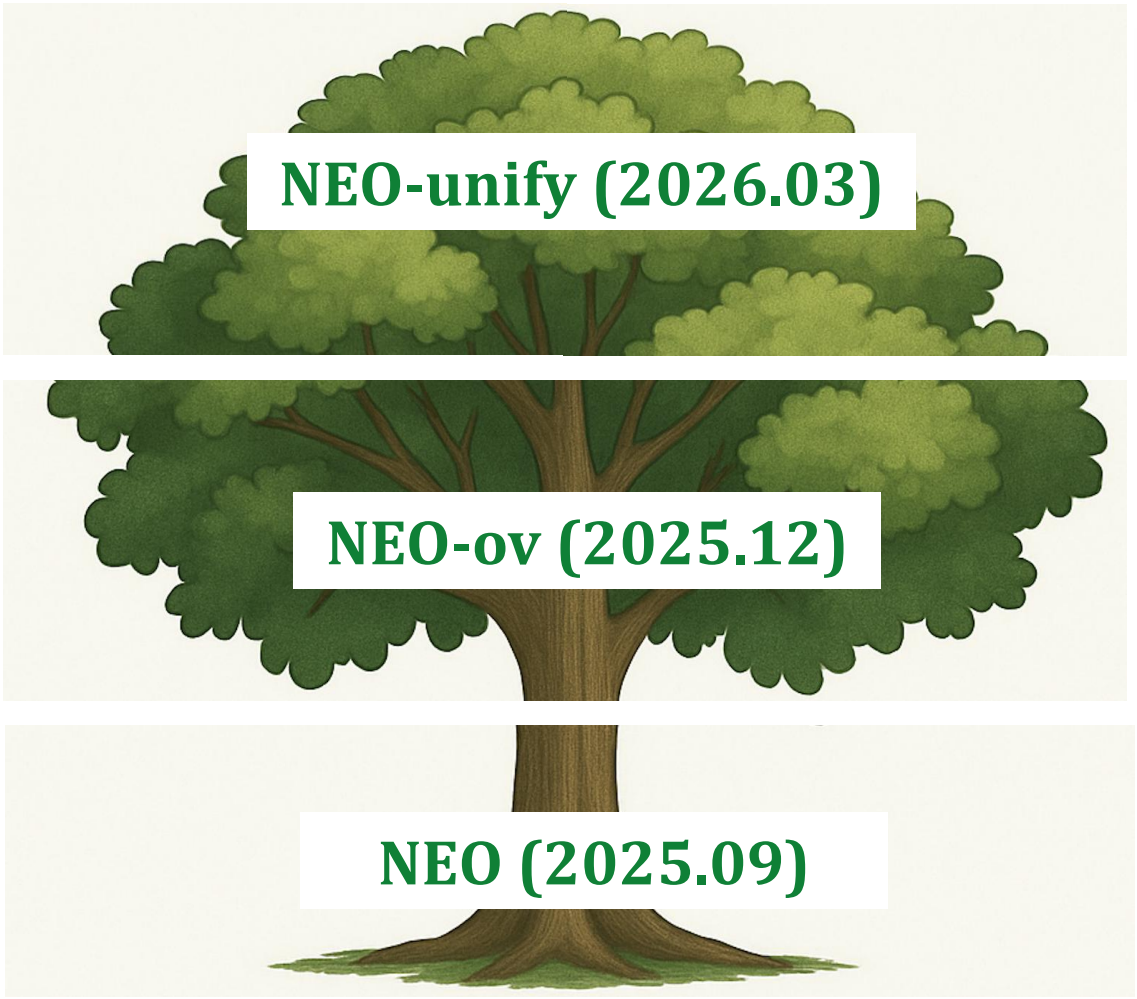


**Prompt** In the scene, there is a black ball and several colored balls of different sizes. The black ball can eat balls that are smaller than itself. After eating a ball, the black ball grows larger. Find the correct sequence to eat all colored balls step by step.



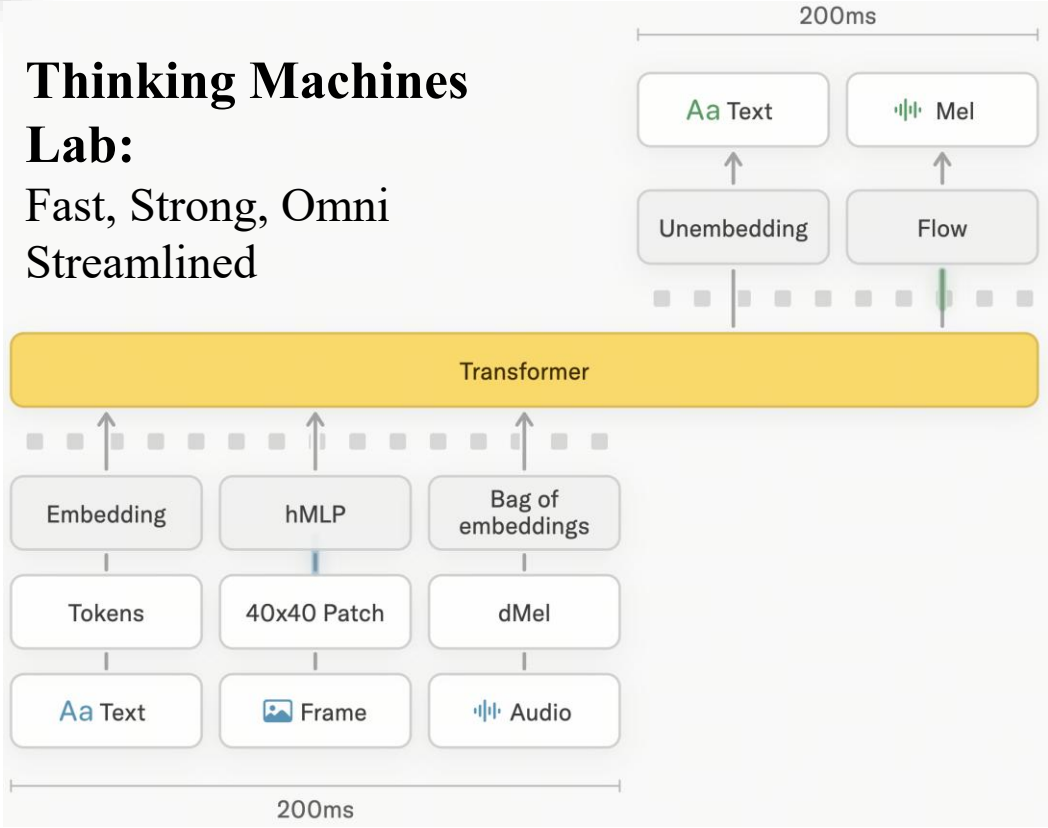
**Prompt** In the scene there are two objects and their corresponding target outlines; each outline matches its object in color and shape. Move each object to its matching outline via shortest path. Show the movement step by step.

# Outlook： Native Omni-Model



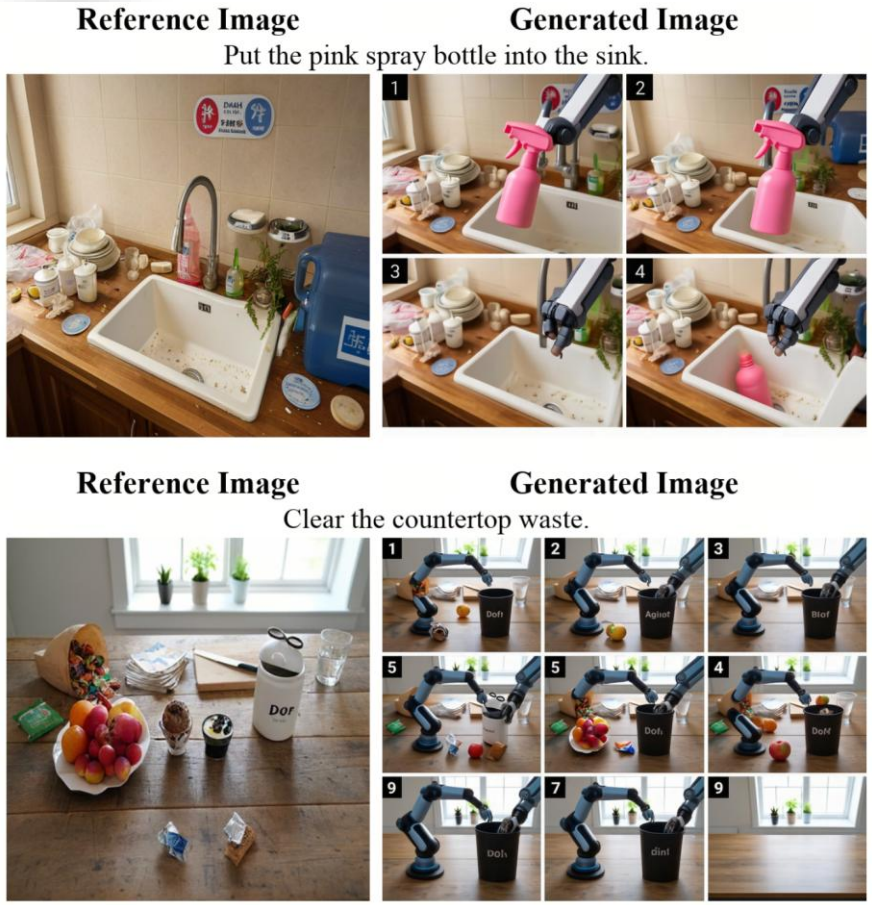

## Native Interaction models

**Thinking Machines Lab:**
Fast, Strong, Omni
Streamlined



# Outlook： Native World Model or Action Model

## Native World Model




## Native Action Models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

# Thank You

Ziwei Liu

Nanyang Technological University

https://liuziwei7.github.io