FROM PIXELS TO WORDS – TOWARDS NATIVE VISION-LANGUAGE PRIMITIVES AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

The edifice of native Vision-Language Models (VLMs) has emerged as a rising contender to typical modular VLMs, shaped by evolving model architectures and training paradigms. Yet, two lingering clouds cast shadows over its widespread exploration and promotion: (-) What fundamental constraints set native VLMs apart from modular ones, and to what extent can these barriers be overcome? (-) How to make research in native VLMs more accessible and democratized, thereby accelerating progress in the field. In this paper, we clarify these challenges and outline guiding principles for constructing native VLMs. Specifically, one native VLM primitive should: (i) effectively align pixel and word representations within a shared semantic space; (ii) seamlessly integrate the strengths of formerly separate vision and language modules; (iii) inherently embody various cross-modal properties that support unified vision-language encoding, aligning, and reasoning. Hence, we launch **NEO**, a novel family of native VLMs built from first principles, greatly narrowing the gap with top-tier modular counterparts across diverse real-world scenarios. With only 390M image-text examples, **NEO** efficiently develops visual perception from scratch while mitigating vision-language conflicts inside a dense and monolithic model crafted from our elaborate primitives. We position **NEO** as a cornerstone for scalable and powerful native VLM development, paired with a rich set of reusable components that foster a cost-effective and extensible ecosystem. Code and weights will be publicly available to promote further research.

1 INTRODUCTION

Recently, large Vision-Language Models (VLMs) (Bai et al., 2025; Zhu et al., 2025; Wang et al., 2025b; xAI, 2025; Anthropic, 2025; DeepMind, 2025; Hurst et al., 2024; OpenAI, 2025) have emerged as a major breakthrough, extending the strong text-processing capabilities of Large Language Models (LLMs) to multimodal understanding. Contemporary VLMs typically follow a modular design that integrates a pre-trained Visual Encoder (VE) (Radford et al., 2021; Chen et al., 2024f; Fang et al., 2023; Tschannen et al., 2025), a Projector (Alayrac et al., 2022; Liu et al., 2024a; Dai et al., 2024), and an LLM (Touvron et al., 2023; Yang et al., 2025; DeepSeek-AI et al., 2025), using next-token prediction as the primary objective. Through complex multi-stage post-training at scale, they incrementally overcome limitations in image resolution, aspect ratio, and visual encoding flexibility. Yet, modular designs still contend with strong inductive biases in pre-trained visual semantics, as well as complex infrastructure and scaling laws needed to harmonize their components.

Against this backdrop, native VLMs have arisen as a new avenue of exploration, with Fuyu (Bavishi et al., 2023) and EVE (Diao et al., 2024) pioneering a promising and practical route towards encoder-free VLMs with a monolithic framework. Subsequent efforts seek to learn vision perception from scratch and mitigate vision-language conflicts via visual encoder distillation (Diao et al., 2024; Li et al., 2025b; Wang et al., 2025a; Li et al., 2025a), mixed training data (Lei et al., 2025; Li et al., 2025a), and modality-specific decomposition (Diao et al., 2025; Luo et al., 2024; 2025; Li et al., 2025a). Nonetheless, constructing visual representations via mapping functions inside pre-trained LLMs often hinders efficiency (Chen et al., 2024d; Luo et al., 2024), destabilizes optimization (Team, 2024; Wang et al., 2024b), and disrupts original linguistic knowledge (Diao et al., 2024; Chen et al., 2024d), even under decoupled designs or large computational budgets (Beyer et al., 2024). Besides, HoVLE (Tao et al., 2025) and HaploVL (Yan et al., 2025) address this by projecting vision-language inputs into a shared embedding space in advance. However, their modality-sharing modules, whether

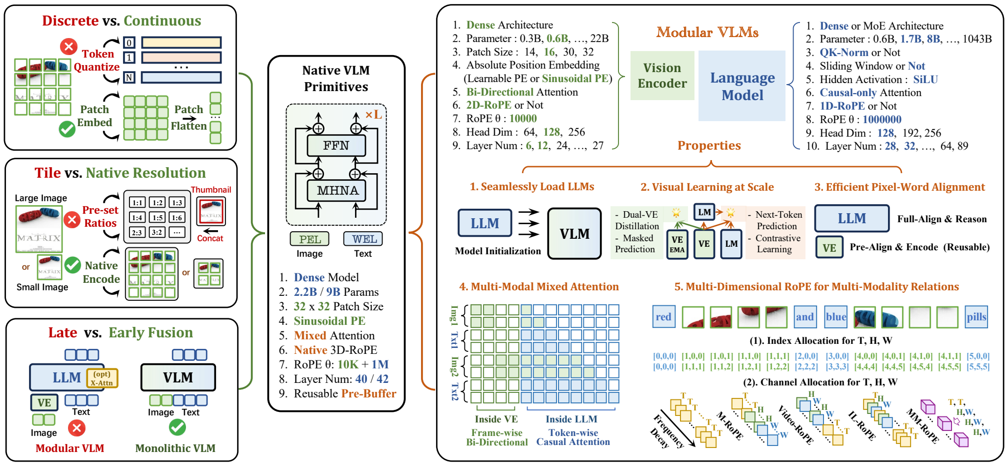

Figure 1: Overview of our native vision-language frameworks, which project arbitrary-resolution images into a continuous latent space, integrating the virtues of modular VLM architectures and enabling efficient vision-language encoding, alignment, and interaction in an early-fusion manner.

derived from the LLM or VE architectures, neglect intrinsic discrepancies in encoding and interaction across modalities, ultimately compromising VLM's capacity to unify visual-linguistic properties.

Figure 1 outlines a central question: *What properties must native VLMs possess to compete with modular ones?* Modular VLMs decouple vision encoders from language models, allowing each to exploit modality-specific characteristics, *e.g.*, bi-directional versus causal attention, distinct positional embeddings, and varied network configurations. This separation accelerates the development of visual and linguistic competencies and permits flexible combinations of individual components. However, it fragments the training procedure, increases alignment costs, and leaves the intermodal balance unresolved. Motivated by these analyses, we formulate the following strategies accordingly:

(1) Native VLM Primitive. Native VLMs should embody a unified vision–language primitive that simultaneously integrates encoding, alignment, and reasoning across modalities in one single module. Its design should encompass three principles: (i) a Flexible Position Encoding scheme that generalizes effectively to dynamic spatial structures; (ii) a Multi-Head Native Attention (MHNA) that jointly processes visual–textual connectivity; (iii) Native Rotary Position Embedding (Native-RoPE) that preserves compatibility with pretrained LLM while absorbing VE's interaction patterns. Guided by these tenets, we evolve the LLM blocks into native VLM primitives with brand-new RoPE designs and modality-aware interaction patterns, thereby capturing multi-dimensional relationships for fine-grained and comprehensive correspondence from an intrinsically multimodal perspective.

(2) Pre-Buffer and Post-LLM. The next crucial issue is to efficiently scale visual training while securing consistent pixel-word alignment. Here, we partition the monolithic backbone into pre-Buffer and post-LLM layers during pre-training, each rooted in identical native primitive architectures. This transient stage enables pretrained LLMs to steer visual learning and establish coherent relevance with later stages. As mid-training and supervised fine-tuning advance, the partition dissolves, yielding a unified architecture that autonomously allocates the VLM's capacities to their respective functions. This end-to-end training reduces semantic biases of separate pretraining and large overheads of post-stage alignment, effectively bridging native and modular VLMs. Crucially, pre-Buffer persists as a reusable pretrained asset, facilitating sustainable resources for native VLM development.

We launch **NEO**, an innovative native VLM that reimagines multi-modal integration from first principles. Unlike typical modular designs, **NEO** rests on unified primitives that natively encode, align, and reason across modalities, forming coherent pixel–word correspondences from the outset. Through streamlined end-to-end training on 390M image–text samples, **NEO** acquires strong visual perception and approaches leading modular VLMs of comparable scale across diverse benchmarks. Beyond these results, **NEO** offers reusable components that simplify subsequent development and reduce barriers to promoting native exploration. This reveals that next-generation multimodal systems could also originate from architectures that are native, unified, and intrinsically multimodal.

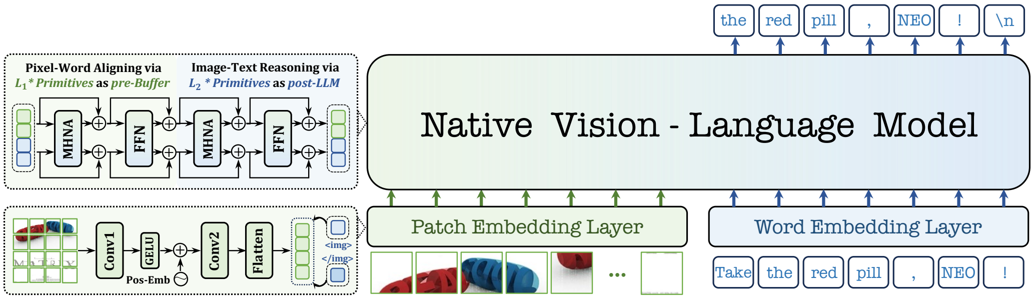

Figure 2: Overview of our proposed NEO architecture. We begin with lightweight patch and word embedding layers that encode images and text into token sequences, which are subsequently processed by a monolithic decoder-only architecture. The pre-Buffer and post-LLM components, each stacked with multiple native primitives, facilitate efficient and precise pixel–word alignment and reasoning.

2 Related Works

2.1 Modular Vision-Language Models

Current Vision–Language Models (VLMs) have converged on a modular paradigm, where a pretrained Vision Encoder (VE) is paired with a Large Language Model (LLM) via lightweight adapters, *e.g.*, projection layers (Li et al., 2024a;b) or cross-attention mechanisms (Alayrac et al., 2022; Dai et al., 2024). This encoder-based architecture underlies various popular and leading vision-language systems, including InternVL (Zhu et al., 2025; Wang et al., 2025b), Qwen-VL (Wang et al., 2024a; Bai et al., 2025), Seed-VL (Guo et al., 2025), GLM-V (Hong et al., 2025), and Grok (xAI, 2024; 2025). By harnessing the complementary strengths of vision and language components, modular architectures, adhering to the "ViT-MLP-LLM" pipeline, achieve unprecedented performance across diverse multimodal benchmarks and have emerged as the dominant design principle in the field.

Despite empirical successes, modular VLMs remain constrained by multi-stage training and heterogeneous structures. Extensive post-training interventions are often required to mitigate rigid inductive biases in pretrained VEs (Wang et al., 2024a), which limit resolution flexibility, erode fine-grained details, and blunt sensitivity to features across scales. Besides, pretraining semantic biases and capacity trade-offs between VEs and LLMs collectively impede design simplicity, deployment efficiency, and seamless integration of vision and language, underscoring the urgent need for a monolithic backbone.

2.2 Native Vision-Language Models

Native VLMs embrace early-fusion integration rather than grafting VEs onto LLMs. Early Fuyu (Bavishi et al., 2023), EVE (Diao et al., 2024), and SOLO (Chen et al., 2024d), embed image patches via linear projections, whereas Chameleon (Team, 2024), MoMA (Lin et al., 2024), and MoT (Liang et al., 2024) transform images into symbolic sequences via discrete tokenizers. Later studies (Luo et al., 2024; Diao et al., 2025; Li et al., 2025b; Luo et al., 2025; Li et al., 2025a) leverage Mixture-of-Experts (MoE) or Divide-and-Conquer (DaC) strategies to suppress vision-language interference, while others (Diao et al., 2024; Li et al., 2025b; Wang et al., 2025a; Li et al., 2025a) upgrade visual encoder supervision to accelerate the acquisition of visual concepts. Empirical evidence (Beyer et al., 2024; Luo et al., 2024; Lei et al., 2025) reveals that, with sufficient data and progressive training, native VLMs rapidly approach modular counterparts, corroborating recent scaling-law insights (Shukor et al., 2025b;a). Besides, recent methods (Tao et al., 2025; Yan et al., 2025; Xiao et al., 2025) indicate that multi-modality encoding modules with the LLM or VE style slightly resolve vision-language misalignment, yet fail to fully integrate the distinct properties of each modality.

Notably, NEO redefines native VLMs as a unibody system built from first-principle primitives. Every network component, from native rotary position embeddings to multi-modality interaction patterns, ensures full compatibility with the intrinsic modeling patterns of VEs and LLMs. Meanwhile, NEO differs from existing modular VLMs via modality-agnostic pre-Buffer and end-to-end training, dramatically enhancing pixel-word alignment and pushing the frontier of native VLM research.

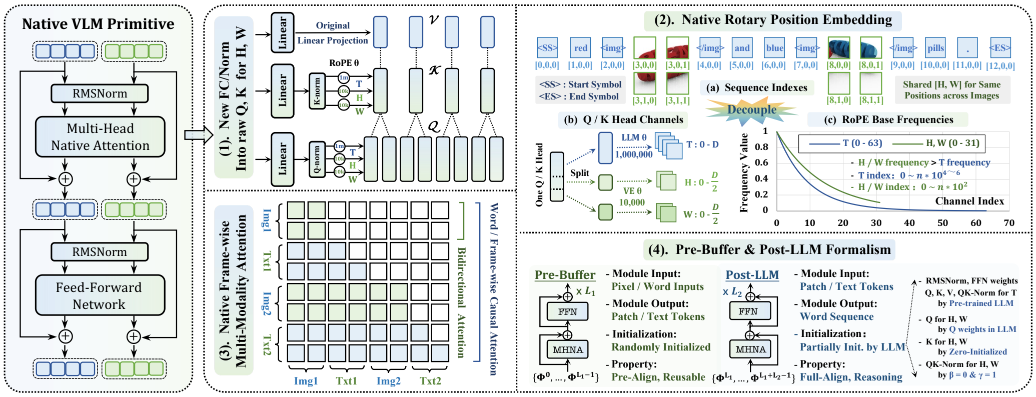

Figure 3: Overview of our native primitive, which integrates native attention with bi-directional dependencies within images and word / frame-wise causal interactions, together with native rotary position embeddings parameterized by modality-specific frequency, channel, and index allocation. It is inherently unified and intrinsically multimodal, substantially enhancing pixel–word correspondence.

3 Methodology

3.1 Model Architecture

Figure 2 illustrates the proposed NEO framework, which comprises lightweight patch and word embedding layers, a pre-Buffer, and a post-LLM, built upon stacked native VLM primitives.

Patch and Word Embeddings. Given an image $\mathbf{I}$, we convert it into token sequences via a lightweight Patch Embedding Layer (PEL) with two Convolutional layers (Conv1–2) (Krizhevsky et al., 2012) and a Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel, 2016). For input text $\mathbf{T}$, we encode it into word tokens using the original LLM Tokenizer as Word Embedding Layer (WEL):

$$\boldsymbol{x_v} = \text{Conv2}(\text{GELU}(\text{Conv1}(\mathbf{I})) + \mathbf{PE}), \quad \boldsymbol{x_t} = \text{Tokenizer}(\mathbf{T}), \tag{1}$$

where $\boldsymbol{x_v} \in \mathbb{R}^{(h \times w) \times d}$ / $\boldsymbol{x_t} \in \mathbb{R}^{n \times d}$ denote visual / textual tokens, and $\mathbf{PE}$ is 2D Sinusoidal Positional Encoding (Dosovitskiy et al., 2021). The stride of Conv1 / Conv2 is 16 / 2, *i.e.*, each visual token corresponds to a 32 × 32 image patch. Notably, Conv2 performs token folding like pixel unshuffle (Chen et al., 2024e), with the special `<img>` and `</img>` tokens inserted at the boundaries of visual tokens, while mapping position and patch embeddings into a unified space. Afterward, visual and textual tokens are merged and propagated through the unified backbone.

Native VLM Primitive. It adopts RMSNorm (Zhang & Sennrich, 2019) and SwiGLU (Dauphin et al., 2017) consistent with LLM layers. Unlike prior methods that collapse visual tokens into 1D representations (Zhu et al., 2025; Wang et al., 2025b) or merely reallocate pre-trained LLM head dimensions across temporal (T), height (H), and width (W) (Wang et al., 2024a; Bai et al., 2025), we enlarge Query (Q) and Key (K) head dimensions and decouple H, W, and T relations in Figure 3(1), adding ∼10% more parameters over the raw Transformer block. The T dimension is retained, and new H and W dimensions are added with their respective QK normalization (Yang et al., 2025).

This philosophy aligns with our **Native Rotary Position Embedding (Native-RoPE)** in Figure 3(2). *(a) Index Allocation*: For text, T index is retained while H / W indexes are zeroed. For images, each visual token has a constant T index, with unique H / W indexes encoding spatial location. Videos, treated as sequences of frames, increment T index per frame, while H / W indexes follow the same spatial scheme as images. In multimodal inputs, each modality's T index starts from the maximum ID of the preceding modality, ensuring continuous and unambiguous positional encoding across modalities. This serves two purposes: (-) For image pairs, H / W indexes start independently from (0,0), and tokens at corresponding positions share identical dependency, strongly reinforcing correlations and interactions across matching regions (Liao et al., 2025; Wu et al., 2025); (-) For image-text pairs, H / W indexes are decoupled from T index and bounded within (0,0) and (H,W), preventing large T index growth from disproportionately affecting H / W indexes (Wang et al., 2024a; Bai et al., 2025) and thereby keeping spatial dependencies between long-range text and images.

Another key aspect is *(b) Channel and (c) Frequency Allocation*. Unlike recent 3D-RoPE methods (Bai et al., 2025; Wei et al., 2025; Yuan et al., 2025; Liao et al., 2025), we fully decompose the channel and frequency allocation of H, W, and T, equipped with additional Q/K head dimensions for H and W. This resolves two issues: (-) Zeroing H / W indexes for pure text would disrupt the modeling patterns and linguistic capacity of the LLM if restricted to its original channels. Repairing this disruption requires substantial resources; (-) Even with interleaved or segmented reallocation, H and W are theoretically equivalent but are assigned different frequencies. Meanwhile, the RoPE frequency in LLMs is far lower than that of visual encoders in Figure 3(2c). This mismatch limits the modeling of relative distances and local semantics. The problem is exacerbated by the disparity in scales, with temporal ranges spanning up to one million and spatial ranges only a few hundred.

Specifically, Native-RoPE assigns distinct base frequencies to T, H, and W within their own dimensions, *i.e.*, original LLM head dimension for T and new head dimension for H / W as follows:

$$\Theta_T = \left\{ \beta_T^{-\frac{2k}{d}} \mid k \in [0, \frac{d}{2}) \right\}, \ \Theta_H = \left\{ \beta_H^{-\frac{4i}{d}} \mid i \in [0, \frac{d}{4}) \right\}, \ \Theta_W = \left\{ \beta_W^{-\frac{4j}{d}} \mid j \in [0, \frac{d}{4}) \right\} \qquad (2)$$

where β and Θ indicate the base and rotation frequency across H, W, and T. Notably, temporal T dimension captures both local and long-range relations, whereas spatial H / W dimensions emphasize local dependencies. This also opens avenues for broader applications, *e.g.*, video understanding (Wei et al., 2025), multimodal generation (Deng et al., 2025b), and editing (Deng et al., 2025a).

Inspired by prior works (Lei et al., 2025; Deng et al., 2025b; Li et al., 2025a; Beyer et al., 2024), we also treat one single image as a unified meta-unit for autoregressive modeling, denoted as **Native Multi-Modal Attention** with mixed masking in Figure 3(3). Text tokens adhere to standard causal attention, attending only to preceding tokens to maintain autoregressive generation. In contrast, image tokens employ full bidirectional attention, enabling exhaustive interactions among all visual tokens, akin to a visual encoder. This design captures rich spatial and contextual dependencies within images and facilitates vision-language correspondences, thereby supporting complex multimodal reasoning. We use FlexAttention (Dong et al., 2024) to minimize memory overhead and increase throughput, as variable-length block-wise attention is fully optimized through CUDA kernel modifications.

Pre-Buffer and Post-LLM. In Figure 3(4), we develop a modality-shared pre-Buffer to translate pixel–word inputs into a unified representation with minimal disturbance to the post-LLM, which inherits the linguistic proficiency and reasoning capabilities of pre-trained LLM. The layer depths L_1 and L_2 primarily refer to parameter counts and scaling properties (Tian et al., 2025) of existing VEs and LLMs to balance accuracy and efficiency. Here, we formulate one primitive Φ^l as follows:

$$\boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l}'} = \boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l}} + \text{MHNA}(\text{RMSNorm}(\boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l}})), \quad \boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l+1}} = \boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l}'} + \text{FFN}(\text{RMSNorm}(\boldsymbol{x}_{\boldsymbol{m}}^{\boldsymbol{l}'})), \qquad (3)$$

where $\boldsymbol{m} \in \{\boldsymbol{v}, \boldsymbol{t}\}$ indicates input modality. Besides, $\{\Phi^0, ..., \Phi^{L_1-1}\}$ and $\{\Phi^{L_1}, ..., \Phi^{L_1+L_2-1}\}$ denotes pre-Buffer and post-LLM, respectively. Notably, we randomly initialize the entire pre-buffer, while the post-LLM inherits RMSNorm, Feed-Forward Network (FFN), and Q/K/QK-Norm parameters along the temporal dimension from a pretrained LLM. The temporal Q is reused to initialize Q for the H and W dimensions, their K weights are zero-initialized, and the corresponding QK-Norm is initialized with $\beta = 0$ and $\gamma = 1$. We further match the attention scaling to that of the pretrained LLM, thereby preserving its pre-training paradigm from the outset and enabling a progressive emergence of multimodal spatial reasoning within the post-LLM. Crucially, this separation exists only during pre-training. After that, these components are merged into a monolithic backbone that autonomously allocates capacity for encoding, alignment, and reasoning.

3.2 Training Procedure

Figure 4 illustrates the whole training pipeline, where the entire model is optimized end-to-end.

Pre-Training Stage. In this phase, NEO acquires fundamental visual concepts and contextual dependencies from scratch, guided by pre-trained patterns from LLMs. Training leverages 345M web-scale and synthetic image-caption pairs, including 100M English and 20M Chinese pairs from LAION-400M (Schuhmann et al., 2021), 150M English pairs from COYO-700M (Byeon et al., 2022), 20M long-caption examples from BLIP3o (Chen et al., 2025), and 5M short-caption pairs from OpenImages (Kuznetsova et al., 2018), recaptioned with a pre-trained InternVL2-8B model. The dataset is further enriched with 30M samples from LAION-COCO (Schuhmann et al., 2022)

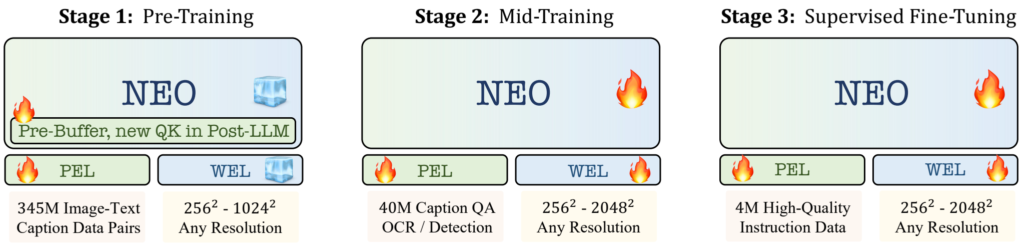
Stage 1: Pre-Training
NEO
Pre-Buffer, new QK in Post-LLM
PEL
WEL
345M Image-Text Caption Data Pairs
256^2 - 1024^2 Any Resolution
Stage 2: Mid-Training
NEO
PEL
WEL
40M Caption QA OCR / Detection
256^2 - 2048^2 Any Resolution
Stage 3: Supervised Fine-Tuning
NEO
PEL
WEL
4M High-Quality Instruction Data
256^2 - 2048^2 Any Resolution

Figure 4: Overview of the entire training recipe. During pre-training, NEO learns visual perception from massive web-scale and synthetic image-caption pairs with frozen LLM weights to preserve linguistic knowledge. In mid-training and supervised fine-tuning, the full model is progressively optimized end-to-end using caption, conversation, OCR, detection, and high-quality instruction data.

and 20M examples from Wukong (Gu et al., 2022) with rich Optical Character Recognition (OCR) annotations. A 3:7 ratio of language to multi-modal data is incorporated to reconstruct text projections in the pre-Buffer. Only the patch embedding layer, the pre-Buffer, and extra QK linear weights and normalization in post-LLM, along with H and W, are optimized with a simple next-token prediction objective. Notably, the new QK heads not only counteract the LLM's strong language bias that limits visual specialization but also safeguard its capabilities against the effects of low-quality data.

Mid-Training Stage. The objective at this stage is to strengthen the alignment between visual and linguistic capabilities while progressively enhancing recognition of high-resolution images, complex scenes, object scales, spatial grounding, and compact OCR content. The training data is drawn from the pre-training corpus of InternVL-1.5 (Chen et al., 2024f), comprising 40M samples across image captioning, conversation, detection, and OCR data, which account for approximately 66%, 11%, 8%, and 15% of the total, respectively. A 3:7 ratio of language to multi-modal data is again applied. The entire architecture is updated with the same loss functions to consolidate vision-language alignment, thereby equipping NEO with the foundational abilities required for various visual scenarios.

Supervised Fine-Tuning Stage. During the SFT stage, NEO's ability to follow complex linguistic instructions and varied dialogue patterns is further enhanced, a critical step towards real-world deployment. The full network is optimized across diverse high-quality, multi-source instruction datasets. Following Mono-InternVL (Luo et al., 2024), we employ about 4M bilingual instructions for supervised learning, covering tasks such as visual question answering, multimodal dialogue, mathematics, and knowledge reasoning. Details of the instruction data are provided in the Appendix.

4 Experiments

4.1 Training Settings

Our NEO models are built on Qwen3-1.7B and Qwen3-8B (Yang et al., 2025) as the LLMs. The pre-Buffer employs $L_1 = 12$ primitive layers for NEO-2.2B and $L_1 = 6$ for NEO-9B. We extend only the QK head dimension in raw transformer layers, introducing roughly 10% extra parameters over the original design. The base RoPE frequencies β_T, β_H, and β_W are set to 1×10^6, 1×10^4, and 1×10^4, respectively. NEO is trained on sixteen 8-GPU (80G) nodes using the AdamW optimizer (Loshchilov & Hutter, 2019). The maximum learning rates for pre-training, mid-training, and SFT are 8×10^{-4}, 4×10^{-5}, and 5×10^{-5}, with a warm-up ratio of 0.01 and a cosine decay scheduler across all stages.

4.2 Main Results

We conduct standard evaluations with VLMEvalKit (Duan et al., 2024) on diverse benchmarks, covering chart, diagram, and document understanding tasks, *e.g.*, AI2D (Kembhavi et al., 2016), DocVQA (Clark & Gardner, 2018), ChartQA (Masry et al., 2022), InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), and OCRBench (Liu et al., 2023e); visual perception and challenging reasoning tasks, *e.g.*, MMMU (Yue et al., 2024), MMBench-EN (MMB) (Liu et al., 2024b), MMVet (Yu et al., 2024), MMStar (Chen et al., 2024c), SEEDBench-IMG (SEED-I) (Li et al., 2023a); hallucination tasks, *e.g.*, POPE (Li et al., 2023b) and HallusionBench (HallB) (Guan et al., 2024).

Table 1: **Comparison with modular and native VLMs on general vision-language benchmarks.** "# Data" denotes the dataset scale during pre-training, mid-training, and supervised fine-tuning. † indicates models that employ reinforcement learning (RL). **Bold** highlights the highest performance.

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
Qwen2.5-VL†	Qwen2.5-1.5B	– / – / –	**51.2**	79.1	61.8	55.9	–	–	**46.3**
InternVL3†	Qwen2.5-1.5B	>6B / 100M / 22M	48.6	**81.1**	**62.2**	**60.7**	–	89.6	42.5
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**
▼ *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

Following InternVL3 (Zhu et al., 2025), we construct the *Encoder-Based* by combining Qwen3 (Yang et al., 2025) and InternViT-300M (Zhu et al., 2025). In the mid-training stage, we first train the projector on 10M samples, and further unfreeze the vision encoder utilizing another 30M samples.

Comparison with Modular VLMs. In Table 1 and Table 2, NEO achieves highly competitive performance against Encoder-Based counterparts at the 2B and 8B scales. Impressively, NEO largely narrows the performance gap with top-tier modular VLMs, *e.g.*, Qwen2-VL (Wang et al., 2024a), InternVL2.5 (Chen et al., 2024e), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025) across multiple benchmarks, despite using relatively limited training data and without reinforcement learning. These results highlight the effectiveness of an end-to-end training strategy and a unified model design with Native-RoPE and multi-modality interaction patterns. Moreover, the performance gap between Encoder-Based variants and state-of-the-art methods on MMMU, MMVet, TextVQA, and *etc*, indicates that NEO still suffers from the limitation in training data scale and quality.

Comparison with Native VLMs. From Table 1 and Table 2, NEO delivers substantial gains on visual-centric benchmarks over the best competitors, *e.g.*, Mono-InterVL (Luo et al., 2024; 2025), HoVLE (Tao et al., 2025), OneCAT (Li et al., 2025a), EVE (Diao et al., 2024; 2025), Emu3 (Wang et al., 2024b), BREEN (Li et al., 2025b), VoRA (Wang et al., 2025a), and SAIL (Lei et al., 2025). By seamlessly integrating post-LLM components with the pre-Buffer for large-scale visual learning, NEO aligns visual inputs with textual features from scratch and supports complex visual reasoning, even without visual encoder supervision (Diao et al., 2024; Tao et al., 2025; Li et al., 2025a; Wang et al., 2025a; Li et al., 2025b), highlighting the strengths of its native primitive designs and training strategies. These design choices allow NEO to surpass many native VLMs using fewer training resources, demonstrating the advantages of our primitives with efficient data-scaling capability.

Table 2: **Comparison with modular and native VLMs on visual question answering benchmarks.** Any Res., Tile-wise, Any Rat., and Fix Res. refer to any resolution, image tile splitting, any aspect ratio, and fixed resolution. MoE and DaC are Mixture-of-Experts and Divide-and-Conquer models.

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
Qwen2.5-VL†	Any Res.	M-RoPE	Dense	**81.6**	**93.9**	**84.0**	**77.1**	79.3	79.7
InternVL3†	Tile-wise	1D-RoPE	Dense	78.7	88.3	80.2	66.1	77.0	**83.5**
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
▼ *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

Despite strong results, NEO lags on knowledge-/OCR-heavy tasks, *e.g.*, MMMU, InfoVQA, and TextVQA. *Interestingly, NEO-9B does not surpass NEO-2B on DocVQA and InfoVQA*, indicating limitations in our current training corpus. Even so, NEO performs well under constraints, highlighting the native VLM as a scalable paradigm. Larger datasets and resources can unlock its full potential.

4.3 Ablation Studies

Unless otherwise specified, we report the average evaluation results, denoted as **Avg.**, across ten vision-language benchmark datasets in Table 3. The pre-Buffer and new head dimensions in the post-LLM are trained on 20M pre-training samples, followed by full-backbone fine-tuning on 2M SFT instruction data. These constitute the standard training settings for our ablation studies.

Hyperparameters of the Pre-Buffer Layer. Figure 5 illustrates the relationship between the number of pre-Buffer layers and the model's average accuracy, using Qwen3-1.7B as the post-LLM. Performance improves consistently as the layer count increases, but gains begin to saturate beyond eight layers. Inspired by the parameter counts of popular VEs (Chen et al., 2024f; Radford et al., 2021; Zhai et al., 2023) and the scaling properties (Tian et al., 2025) between existing VEs and LLMs, we select 12 and 6 layers for NEO-2.2B and NEO-9B to balance the trade-off between performance and efficiency.

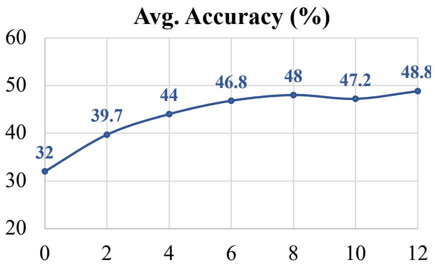

Figure 5: Configurations of pre-Buffer.

Table 3: Configurations of attention and RoPE. MMS, CQA, IVQA, and OCRB denote MMStar, ChartQA, InfoVQA, and OCRBench. ⋆ indicates that the base RoPE frequencies for height and width are set to 1M. To ensure fairness, we add new head dimensions of equal size across all models.

Model	Attention	RoPE	MMMU	MMB	MMS	SEED-I	AI2D	CQA	IVQA	TVQA	OCRB	POPE	Avg.
A	Causal	1D-RoPE	40.2	48.6	36.1	55.3	63.6	16.1	22.5	16.2	13.9	78.6	39.1
B	Mixed	1D-RoPE	**40.8**	48.8	36.4	57.3	63.7	16.0	21.9	17.4	16.0	79.2	39.8
C	Mixed	IL-RoPE	40.0	47.3	36.3	57.6	62.0	18.8	23.4	17.9	13.2	78.8	39.5
D	Mixed	M-RoPE	40.3	49.6	37.2	57.8	64.2	23.7	25.2	20.4	18.8	79.3	41.7
E	Mixed	MM-RoPE	40.5	50.8	37.6	58.2	**65.8**	25.7	**26.3**	22.1	18.2	78.8	42.4
F	Mixed	Video-RoPE	40.6	**51.3**	**37.8**	**58.8**	64.3	**27.4**	26.1	**23.7**	**21.3**	**81.0**	**43.2**
G	Causal	Native-RoPE	40.2	49.2	36.3	57.1	63.7	19.2	23.5	19.5	16.7	77.8	40.3
H	Mixed	Native-RoPE	**40.7**	**51.9**	**38.2**	**58.9**	**65.8**	**30.6**	**26.9**	**24.1**	**23.2**	**80.0**	**44.0**
I	Mixed	Native-RoPE⋆	40.4	50.4	36.9	57.0	64.1	25.6	25.2	21.7	20.1	78.7	42.0

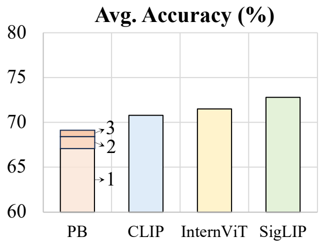

Figure 6: Pre-Buffer vs. VEs.

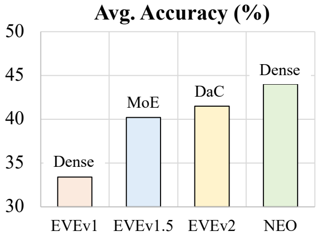

Figure 7: Neo vs. EVE series.

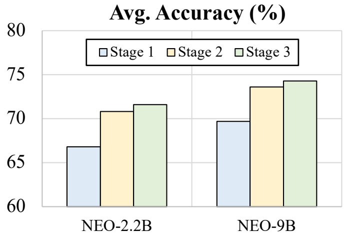

Figure 8: Three training processes.

Configurations of Native Primitives. Table 3 compares various attention and RoPE designs. The pre-Buffer depth is 4, and the post-LLM is initialized with Qwen3-1.7B. All models share the same new QK head dimensions and normalization. *(1) Attention mode*. Comparing models A/B and G/H reveals consistent gains of mixed attention over causal one, reflecting its stronger capacity to model comprehensive dependencies and cross-modal alignment. *(2) RoPE mode*. Native-RoPE outperforms 1D-RoPE (Zhu et al., 2025), IL-RoPE (Liao et al., 2025), M-RoPE (Bai et al., 2025), MM-RoPE (Yuan et al., 2025), and Video-RoPE (Wei et al., 2025), with at least a 0.8% gain. This validates the importance of disentangling height, width, and temporal components in RoPE to enhance spatial–temporal representations and fine-grained interactions. By contrast, setting the base RoPE frequency to 1M for height and width severely impairs the ability to perceive local semantics.

Comparison between Pre-Buffer and Vision Encoders. In Figure 6, PB 1–3 denotes the Pre-Buffer obtained from stage 1–3. For fairness, all post-LLMs are initialized by Qwen3-1.7B (Yang et al., 2025) combined by pre-trained pre-Buffer, CLIP-vit-large-patch14 (Radford et al., 2021), InternViT-300M (Chen et al., 2024e), and SigLIP-so400m-patch14-384 (Zhai et al., 2023). During pre-training, we train the projector for encoder-based methods and the newly added QK parameters for all models. During SFT, we unfreeze the entire backbone. After two-stage re-training, PB3 delivers strong performance, trailing CLIP / InternViT / SigLIP by only 1.7 / 2.4 / 3.7% on average across diverse benchmarks. Notably, despite using only 22M training samples, PB3 is just 2.5% below the full NEO model, substantially reducing the training cost of developing native VLMs for future research.

Comparison between NEO and Native VLM variants. Existing native VLMs in Table 1 and 2 differ dramatically in training pipelines, data volume/quality, and base LLM choice. To isolate model architectural effects, we compare our NEO with representative EVEv1.0 (Dense), EVEv1.5 (Mixture-of-Experts), and EVEv2.0 (Divide-and-Conquer), all built on Qwen3-1.7B. From Figure 7, EVEv1.0, EVEv1.5, and EVEv2.0 achieves 33.4%, 40.2%, and 41.5%, respectively, compared with 44.0% for our NEO. This confirms that the gains stem from our pre-buffer, native RoPE design, and interaction patterns, rather than from a newer backbone or larger and higher-quality data alone.

Performance Gains across Stages. Figure 8 presents the result evolution across training stages. In Stages 1 and 2, the model is fine-tuned on 2M SFT examples. Performance improves consistently as training data scales increase across 2.2B and 9B model sizes. Following progressive training, NEO shows strong multimodal capabilities, enabling robust performance across diverse real-world tasks.

5 Conclusion

We introduce NEO, a native VLM that seamlessly integrates vision and language into a single unified framework, eliminating the need for separate visual encoders or ad-hoc alignment modules. By leveraging hybrid attention and modality-aware rotary position embeddings, NEO captures rich, fine-grained interactions between pixels and words from the outset. Its pre-Buffer and post-LLM training paradigm ensures efficient convergence and robust alignment while maintaining end-to-end learning. Experiments show that this unified design not only advances multimodal understanding and reasoning but also lays the foundation for reusable, scalable components. Our native primitives highlight a promising path toward intrinsically multimodal, unified, and adaptable architectures.

Ethics Statement

All resources are drawn from open-access datasets with explicitly defined usage policies. Our work seeks to advance multimodal learning capabilities without introducing ethical or safety concerns beyond those already associated with existing models. Nevertheless, risks such as dataset biases and potential misuse cannot be entirely ruled out. We emphasize the importance of careful data curation, responsible deployment, and transparent reporting as essential practices to mitigate these challenges.

Reproducibility Statement

We place strong emphasis on reproducibility, providing detailed descriptions to facilitate replication and validation. Information about dataset selection, training strategies, and evaluation settings is provided in Sec. 3.2 and Sec. 4.1. We commit to releasing the code, model weights, and detailed documentation to allow the community to reproduce our findings in future research.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances of Neural Information Processing Systems*, New Orleans, LA, USA, 2022.

Anthropic. Claude 3.7 sonnet: A hybrid reasoning ai model, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

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

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Junting Zhou, Ziqiang Liu, Feiteng Fang, Mingshan Chang, Tianyu Zheng, Xincheng Zhang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *CoRR*, abs/2403.18058, 2024.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bosnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan

Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: a versatile 3b vlm for transfer. *CoRR*, abs/2407.07726, 2024.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *IEEE International Conference on Computer Vision*, pp. 4290–4300, Seoul, Korea (South), 2019.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022. URL https://github.com/kakaobrain/coyo-dataset.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *International Conference on Computational Linguistics*, pp. 1511–1520, 2022.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: harnessing gpt4v-synthesized data for a lite vision-language model. *CoRR*, abs/2402.11684, 2024a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *CoRR*, abs/2505.09568, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: improving large multi-modal models with better captions. In *European Conference on Computer Vision*, volume 15075, pp. 370–387, Milan, Italy, 2024b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances of Neural Information Processing Systems*, Vancouver, BC, Canada, 2024c.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *CoRR*, abs/2407.06438, 2024d.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271, 2024e.

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

Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *International Conference on Document Analysis and Recognition*, pp. 1571–1576, 2019.

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*, pp. 845–855, Melbourne, Australia, 2018.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: open frontier-class multimodal llms. *CoRR*, abs/2409.11402, 2024.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1080–1089, Honolulu, HI, USA, 2017.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, volume 70, pp. 933–941, Sydney, NSW, Australia, 2017.

Google DeepMind. Gemini 2.5 pro: Google's most advanced reasoning model, 2025. URL https://deepmind.google/models/gemini/pro/.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *CoRR*, abs/2505.14683, 2025a.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, Singapore, 2025b.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *CoRR*, abs/2406.11832, 2024.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. *CoRR*, abs/2502.06788, 2025.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *CoRR*, abs/2412.05496, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*, Austria, 2021.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM International Conference on Multimedia*, pp. 11198–11201, Melbourne, VIC, Australia, 2024.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: exploring the limits of masked visual representation learning at scale. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, Vancouver, BC, Canada, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6334, Honolulu, HI, USA, 2017.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *Advances of Neural Information Processing Systems*, New Orleans, LA, USA,, 2022.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, Seattle, WA, USA, 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *CoRR*, abs/2505.07062, 2025.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *CoRR*, abs/2308.10755, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *CoRR*, abs/1606.08415, 2016.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *CoRR*, abs/2507.01006, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, Long Beach, CA, USA, 2019.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. *CoRR*, abs/2410.21276, 2024.

Jimmycarter. Textocr gpt-4v dataset, 2023. URL https://huggingface.co/datasets/jimmycarter/textocr-gpt4v.

Kushal Kafle, Brian L. Price, Scott Cohen, and Christopher Kanan. DVQA: understanding data visualizations via question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, Salt Lake City, UT, USA, 2018.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision*, volume 9908, pp. 235–251, Amsterdam, The Netherlands, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4999–5007, 2017.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, volume 13688, pp. 498–517, Tel Aviv, Israel, 2022.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances of Neural Information Processing Systems*, pp. 1106–1114, Lake Tahoe, Nevada, US, 2012.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018.

LAION. Gpt-4v dataset, 2023. URL https://huggingface.co/datasets/laion/gpt4v-dataset.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *CoRR*, abs/2504.10462, 2025.

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

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: stronger llms supercharge multimodal capabilities in the wild, 2024a. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: easy visual task transfer. *CoRR*, abs/2408.03326, 2024b.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125, 2023a.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *CoRR*, abs/2509.03498, 2025a.

Tianle Li, Yongming Rao, Winston Hu, and Yu Cheng. BREEN: bridge data-efficient encoder-free multimodal learning with learnable queries. *CoRR*, abs/2503.12446, 2025b.

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

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14963–14973, 2023c.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *CoRR*, abs/2411.04996, 2024.

Chao Liao, Liyang Liu, Xun Wang, Zhengxiong Luo, Xinyu Zhang, Wenliang Zhao, Jie Wu, Liang Li, Zhi Tian, and Weilin Huang. Mogao: An omni foundation model for interleaved multi-modal generation. *CoRR*, abs/2505.05472, 2025.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: efficient early-fusion pre-training with mixture of modality-aware experts. *CoRR*, abs/2407.21770, 2024.

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

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *CoRR*, abs/2311.10774, 2023c.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances of Neural Information Processing Systems*, New Orleans, LA, USA, 2023d.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, Seattle, WA, USA, 2024a.

Xi Liu, Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. *CoRR*, abs/1912.09641, 2019.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: is your multi-modal model an all-around player? In *European Conference on Computer Vision*, volume 15064, pp. 216–233, Milan, Italy, 2024b.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *CoRR*, abs/2305.07895, 2023e.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, LA, USA, 2019.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *CoRR*, abs/2105.04165, 2021.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: multimodal reasoning via thought chains for science question answering. In *Advances of Neural Information Processing Systems*, volume 35, pp. 2507–2521, New Orleans, LA, USA, 2022a.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *CoRR*, abs/2209.14610, 2022b.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *CoRR*, abs/2410.08202, 2024.

Gen Luo, Wenhan Dou, Wenhao Li, Zhaokai Wang, Xue Yang, Changyao Tian, Hao Li, Weiyun Wang, Wenhai Wang, Xizhou Zhu, Yu Qiao, and Jifeng Dai. Mono-internvl-1.5: Towards cheaper and faster monolithic multimodal large language models. *CoRR*, abs/2507.12566, 2025.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11–20, Las Vegas, NV, USA, 2016.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: a visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, Vienna, Austria, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: a benchmark for question answering about charts with visual and logical reasoning. In *Annual Meeting of the Association for Computational Linguistics*, pp. 2263–2279, Dublin, Ireland, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 2582–2591, Waikoloa, HI, USA, 2022.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: visual question answering by reading text in images. In *International Conference on Document Analysis and Recognition*, pp. 947–952, Sydney, Australia, 2019.

OpenAI. Gpt-5: A unified multimodal model, 2025. URL https://openai.com/gpt-5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763, virtual, 2021.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021.

Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en, 2022. URL https://laion.ai/blog/laion-coco/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: a benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, volume 13668, pp. 146–162, Tel Aviv, Israel, 2022.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 8876–8884, 2019.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *International Conference on Document Analysis and Recognition*, volume 1, pp. 1429–1434. IEEE, 2017.

Mustafa Shukor, Louis Béthune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *CoRR*, abs/2507.09404, 2025a.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua M. Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. *CoRR*, abs/2504.07951, 2025b.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, volume 12347, pp. 742–758, Glasgow, UK, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, Long Beach, CA, USA, 2019.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *International Conference on Document Analysis and Recognition*, pp. 1557–1562, 2019.

Chenxin Tao, Shiqian Su, Xizhou Zhu, Chenyu Zhang, Zhe Chen, Jiawen Liu, Wenhai Wang, Lewei Lu, Gao Huang, Yu Qiao, and Jifeng Dai. Hovle: Unleashing the power of monolithic vision-language models with holistic vision-language embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14559–14569, Nashville, TN, USA, 2025.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Chameleon Team. Chameleon: mixed-modal early-fusion foundation models. *CoRR*, abs/2405.09818, 2024.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5.

Changyao Tian, Hao Li, Gen Luo, Xizhou Zhu, Weijie Su, Hanming Deng, Jinguo Zhu, Jie Shao, Ziran Zhu, Yunpeng Liu, et al. Navil: Rethinking scaling properties of native multimodal large language models under data constraints. *CoRR*, abs/2510.08565, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *CoRR*, abs/1601.07140, 2016.

Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. Vision as lora. *CoRR*, abs/2503.20680, 2025a.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: prompting GPT-4V for better visual instruction tuning. *CoRR*, abs/2311.07574, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: next-token prediction is all you need. *CoRR*, abs/2409.18869, 2024b.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Videorope: What makes for good video rotary position embedding? *CoRR*, abs/2502.05173, 2025.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation. *CoRR*, abs/2506.18871, 2025.

xAI. Grok-1.5 vision preview, 2024. URL `https://x.ai/blog/grok-1.5v`.

xAI. Grok 3: xAI's flagship ai model, 2025. URL `https://x.ai/news/grok-3`.

Yicheng Xiao, Lin Song, Rui Yang, Cheng Cheng, Zunnan Xu, Zhaoyang Zhang, Yixiao Ge, Xiu Li, and Ying Shan. Haploomni: Unified single transformer for multimodal video understanding and generation. *CoRR*, abs/2506.02975, 2025.

Rui Yan, Lin Song, Yicheng Xiao, Runhui Huang, Yixiao Ge, Ying Shan, and Hengshuang Zhao. Haplovl: A single-transformer baseline for multi-modal understanding. *CoRR*, abs/2503.14694, 2025.

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, Jian Yang, Jiaxi Yang, 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. *CoRR*, abs/2505.09388, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, volume 9906, pp. 69–85, Amsterdam, The Netherlands, 2016.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *CoRR*, abs/2309.12284, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, Vienna, Austria, 2024.

Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, Hao Luo, Jiasheng Tang, Fan Wang, and Yi Yang. Lumos-1: On autoregressive video generation from a unified model perspective. *CoRR*, abs/2507.08801, 2025.

Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3):509–521, 2019.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, Seattle, WA, USA, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE International Conference on Computer Vision*, pp. 11941–11952, Paris, France, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances of Neural Information Processing Systems*, pp. 12360–12371, Vancouver, BC, Canada, 2019.

Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: scaling up visual instruction tuning. *CoRR*, abs/2307.04087, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances of Neural Information Processing Systems*, 36:46595–46623, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025.

A Appendix

Usage of Large Language Models

During manuscript preparation, large language models were used solely as writing assistants. They helped to check grammar, refine sentence structure, and provide style alternatives. All content related to methodology, experiments, and conclusions was developed entirely by the authors. LLM outputs were reviewed critically, and only human-verified edits were incorporated into the final text.

A.1 Supervised Fine-tuning Datasets

Table 4: Dataset summary in supervised fine-tuning stage.

Task	Dataset
Captioning	TextCaps (en) (Sidorov et al., 2020), ShareGPT4V (en&zh) (Chen et al., 2024b)
General QA	VQAv2 (en) (Goyal et al., 2017), GQA (en) (Hudson & Manning, 2019), OKVQA (en) (Marino et al., 2019), VSR (en) (Liu et al., 2023a), VisualDialog (en) (Das et al., 2017)
Science	AI2D (en) (Kembhavi et al., 2016), ScienceQA (en) (Lu et al., 2022a), TQA (en) (Kembhavi et al., 2017)
Chart	ChartQA (en) (Masry et al., 2022), MMC-Inst (en) (Liu et al., 2023c), DVQA (en) (Kafle et al., 2018), PlotQA (en) (Methani et al., 2020), LRV-Instruction (en) (Liu et al., 2023b)
Mathematics	GeoQA+ (en) (Cao & Xiao, 2022), TabMWP (en) (Lu et al., 2022b), MathQA (en) (Yu et al., 2023), CLEVR-Math/Super (en) (Lindström & Abraham, 2022; Li et al., 2023c), Geometry3K (en) (Lu et al., 2021)
Knowledge	KVQA (en) (Shah et al., 2019), A-OKVQA (en) (Schwenk et al., 2022), ViQuAE (en) (Lerner et al., 2022), Wikipedia (en&zh) (He et al., 2023)
OCR	OCRVQA (en) (Mishra et al., 2019), InfoVQA (en) (Mathew et al., 2022), TextVQA (en) (Singh et al., 2019), ArT (en&zh) (Chng et al., 2019), COCO-Text (en) (Veit et al., 2016), CTW (zh) (Yuan et al., 2019), LSVT (zh) (Sun et al., 2019), RCTW-17 (zh) (Shi et al., 2017), ReCTs (zh) (Liu et al., 2019), SynthDoG (en&zh) (Kim et al., 2022), ST-VQA (en) (Biten et al., 2019)
Document	DocVQA (en) (Clark & Gardner, 2018), Common Crawl PDF (en&zh)
Grounding	RefCOCO/+/g (en) (Yu et al., 2016; Mao et al., 2016), Visual Genome (en) (Krishna et al., 2017)
Conversation	LLaVA-150K (en&zh) (Liu et al., 2023d), LVIS-Instruct4V (en) (Wang et al., 2023), ALLaVA (en&zh) (Chen et al., 2024a), Laion-GPT4V (en) (LAION, 2023), TextOCR-GPT4V (en) (Jimmycarter, 2023), SVIT (en&zh) (Zhao et al., 2023)
Text-only	OpenHermes2.5 (en) (Teknium, 2023), Alpaca-GPT4 (en) (Taori et al., 2023), COIG-CQIA (zh) (Bai et al., 2024), ShareGPT (en&zh) (Zheng et al., 2023)

A.2 Implementation Details

Table 5: Implementation details in the pre-training, mid-training and supervise fine-tuning.

Configuration	Pre-Training	Mid-Training	Supervised Fine-Tuning
Resolution	$256^2 - 1,024^2$	$256^2 - 2,048^2$	$256^2 - 2,048^2$
Optimizer	AdamW		
Optimizer hyperparameters	$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad eps = 1e^{-8}$		
Learning rate schedule	cosine with min lr	cosine with min lr	cosine decay
Peak learning rate	$8e^{-4}$	$4e^{-5}$	$5e^{-5}$
Min learning rate ratio	0.05	0.1	–
Weight decay	0.01		
Training steps	190k	50k	6k
Warm-up steps	2k	200	200
Max sample length	$8,192$	$8,192$	$8,192$
Global batch size	$2,560$	$1,200$	650
Text-only ratio	0.3	0.3	–
Numerical precision	`bfloat16`		

A.3 Limitation and Discussion

In this study, we innovate network architectures and training strategies for efficiently building native vision-language models. The full promise of NEO has remained largely untapped, hindered by scarce training data and limited computational resources, especially in knowledge-intensive and OCR-focused domains. Yet, strikingly, our NEO rivals state-of-the-art VLMs despite these severe constraints. We envision subsequent directions of NEO for the native VLM community as follows:

Contextual relevance to recent advancements. Recent models such as Qwen3VL highlight concepts that resonate with our design choices, including dense linking of visual-language features, relative positional encodings, and architectural details like patch embedding and bias. In particular, the DeepStack approach underscores the importance of establishing strong pixel-word associations from the earliest stages, reinforcing the significance of densely integrated visual-language representations.

Maximizing the potential via large investment. It is in great demand for continuously investing substantial resources, especially during the pre-training stage, to fully unlock NEO's performance and approach the upper bound of the native model. At the same time, selectively open-sourcing key components during intermediate development can reduce follow-up training costs for future researchers and attract more research to native visual-language models. Moreover, the fundamental models from this work provide a valuable baseline for advancing reinforcement learning research.

Explorations of full-spectrum model capacities. Expanding the full model sizes remains a critical factor in advancing various real-world applications. Even with limited resources, NEO-2.2B closely matches those of modular visual-language models with equivalent capacity, suggesting that the design philosophy of models in the 0.6 to 8 billion parameter range has matured. Such architectures not only achieve high performance but also facilitate the deployment of lightweight models at the edge, which is crucial for scenarios with limited computational resources or strict real-time requirements.

Upgrading architectures and applications. To date, our work has focused on dense models for image-text understanding, while a sparse divide-and-conquer architecture is simultaneously under active development. Notably, we regard NEO not merely as an autoregressive VLM but as a new paradigm for visual-language intelligence. Its principle is to leverage end-to-end training within a unified architecture, eliminating manually imposed biases and scaling-up complexities by allowing data and models to dictate the learning process. Besides, our efforts are designed not merely to improve performance but to establish a definitive baseline for visual-language generation, long video understanding, and embodied AI. Crucially, NEO's architecture systematically integrates the demands of video generation and related tasks, including attention mechanisms and rotary positional encodings, from the ground up. Although currently focused on text and images, NEO is poised to push the boundaries of what is possible across a wide spectrum of application scenarios and input modalities.