

VIVAS: Vitalizing Visual Perception in VLM

Pre-training via Vision-language Unified Autoregressive Supervision

Zhehan Kan^{1,3,†}, Yubo Zhu^{2,3,†}, Xinghua Jiang^{3,†}, Zhixiang Wei^{3,†},
Shifeng Liu^{3,†}, Wei Tong^{2,✉}, Sheng Zhong², Qingmin Liao¹,
Wenming Yang^{1,✉}, Xin Li^{3,✉}, Yinsong Liu³, Deqiang Jiang³, and Xing Sun³

¹Tsinghua University ²Nanjing University ³Tencent Youtu Lab

Abstract. While Vision–Language Models (VLMs) demonstrate strong capabilities, they continue to suffer from a critical limitation: insufficient fine-grained visual perception, which fundamentally limits their multimodal understanding. We attribute this bottleneck to text-dominant optimization biases during pre-training, which encourage the model to overlook fine-grained visual details, thereby limiting the capability of multimodal understanding. We investigate that overcoming this bottleneck requires two key elements: (1) a unified token space paradigm that ensures stable training dynamics, and (2) a modality-aligned dense visual supervision signal enriched with both structural granularity and semantic information to capture critical visual representations. Based on these insights, we propose VIVAS, a framework built upon the unified token space paradigm, which introduces a dense-structural–semantic vision tokenizer, which expands the textual vocabulary into a unified vision–language vocabulary by incorporating a visual vocabulary. During pretraining, VIVAS performs vision–language unified autoregressive supervision over both visual details and linguistic content, thereby enhancing visual perception to improve multimodal understanding. Trained end-to-end on 12.4T tokens, VIVAS achieves state-of-the-art performance across 7 tasks and 39 multimodal benchmarks.

Keywords: VLM · Multimodal Understanding · Autoregressive Model

1 Introduction

Vision-Language Models (VLMs) have emerged as a promising paradigm for advancing toward artificial general intelligence. Conventionally, these architectures rely on a vision encoder and an alignment module to project visual representations into the linguistic space, leveraging a Large Language Model (LLM) for subsequent decoding [2, 4, 12–14, 22]. Despite their widespread success, current VLMs exhibit a critical limitation: a severe deficiency in fine-grained visual perception that intrinsically bounds their multimodal understanding to a coarse

[†] Equal contribution. Work done during Zhehan Kan and Yubo Zhu’s internship at Tencent Youtu Lab.

^{1,2,3} All institutions are located in China.

✉ Corresponding author

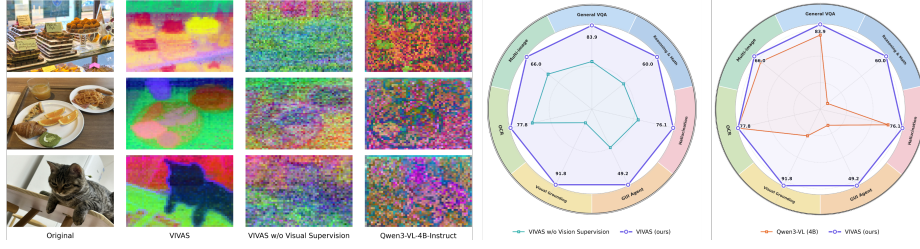


Fig. 1: Left: Principal Component Analysis (PCA) visualizations of hidden states extracted from the output-layer vision tokens, VIVAS demonstrates significantly more fine-grained visual perception. **Right:** Comprehensive multimodal understanding evaluation of VIVAS against the SOTA Qwen3-VL (4B) and an ablated baseline without visual supervision. VIVAS yields substantial performance improvements.

level. We argue that this fundamental flaw originates from a pervasive text-dominant optimization bias inherent in standard training paradigms, where visual signals are treated merely as passive conditioning inputs. Consequently, models are implicitly incentivized to discard visual details deemed redundant for high-level text generation. This inevitably creates a visual information bottleneck, severely impeding their dense perception capabilities.

To address this, recent efforts have sought to enhance the visual perception of VLMs by injecting visual supervision during the post-training stage [18, 19, 21, 40, 42, 51]. However, at this late phase, the model’s visual representations have already been deeply entrenched by text-dominant, large-scale pretraining. Moreover, since visual supervision acts merely as an auxiliary alignment constraint during post-training, its impact is inherently limited, yielding marginal improvements. Therefore, we posit that the critical key to unlocking fine-grained visual perception lies in introducing visual supervision into the pre-training stage, thereby intrinsically optimizing visual representation learning fundamentally.

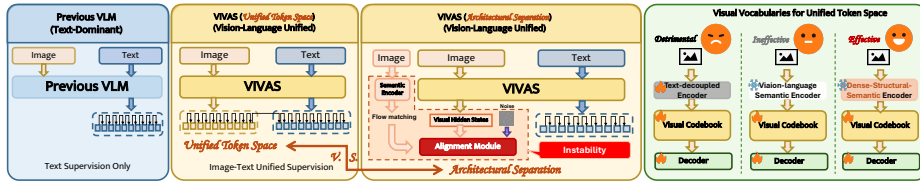


Fig. 2: Left: Previous text-dominant VLMs lack direct visual supervision. **Middle:** To introduce visual supervision, VIVAS formulates a unified token space, fundamentally avoiding the instability inherent in architectural separation approaches. **Right:** Within this unified space, we introduce the dense-structural-semantic visual supervision signal, starkly outperforming detrimental pixel-level or ineffective semantic-level signals.

In this paper, as illustrated in Figure 2, we first investigate the architectural paradigm requirements for effective visual supervision. We observe that the ar-

chitectural separation paradigm, which introduces alignment modules for visual reconstruction during pretraining, leads to training instability and convergence difficulties. In contrast, the unified token space paradigm, built upon a standard architecture, expands the textual vocabulary into a unified vision–language vocabulary by incorporating a visual vocabulary. This expansion aligns visual tokens within a cohesive vision–language predictive stream, enabling stable and rapid convergence.

Furthermore, we study what characteristics of a visual vocabulary can effectively enhance multimodal understanding. We find that text-decoupled visual vocabularies exhibit a substantial modality disconnect from text, which hinders unified supervision; introducing such representations into visual supervision can even degrade multimodal understanding. In contrast, vision–language semantic visual vocabularies derived from image–text contrastive learning encode sparse semantic features that largely overlap with those produced by the VLM’s vision encoder, providing limited complementary information for enhancing fine-grained visual perception and capturing dense structural granularity in visual representations.

To obtain a modality-aligned dense visual vocabulary enriched with both structural granularity and semantic information, we leverage the complementary properties of two representation learning paradigms. Specifically, DINOv3 [35], trained via self-supervised distillation, produces dense and structured visual features, while SigLIP2 [39], trained with image–text contrastive learning, provides strong semantic representations. Based on this observation, we propose a dense-structural-semantic vision tokenizer. This tokenizer fuses the dense structural features from DINOv3 with the semantic features from SigLIP2 through a cross-attention mechanism, and derives the visual vocabulary via a visual reconstruction training objective.

Built upon this visual vocabulary and the unified token space paradigm, we introduce **VIVAS**, which incorporates vision–language unified autoregressive supervision over both visual details and linguistic content during pretraining, thereby enhancing visual perception to improve multimodal understanding. We conduct end-to-end training on 12.4T tokens of data, including pre-training, supervised fine-tuning (SFT), and reinforcement learning (RL). Extensive experiments across seven tasks and 39 benchmarks demonstrate that our approach significantly improves multimodal understanding by strengthening visual perception, while incurring only 3.8% additional inference latency.

In summary, our contributions are as follows:

- We identify a key limitation of existing VLMs: a **text-dominant optimization bias** that leads to the loss of fine-grained visual details and creates a visual information bottleneck.
- We study architectural paradigms for visual supervision in pretraining and show that a **unified token space paradigm** enables stable training by integrating visual tokens into a shared vision–language predictive stream.

- We propose a **dense-structural-semantic vision tokenizer** that fuses DINOv3 structural features and SigLIP2 semantic features via cross-attention to construct a modality-aligned visual vocabulary.
- We introduce **VIVAS**, a vision–language pre-training framework with unified autoregressive supervision, achieving significant improvements on 7 tasks and 39 benchmarks with only 3.8% additional latency.

2 Related Work

2.1 Multimodal Understanding

Multimodal understanding refers to the capability of computational models to process, integrate, and interpret semantic information across diverse data modalities. To facilitate this paradigm, VLMs, including Seed1.5-VL, GLM-4.5, Qwen3-VL, and InternVL3.5 [1, 11, 23, 37, 44], predominantly rely on pre-trained vision encoders (e.g., CLIP and SigLIP [32, 53]) to extract continuous visual features. Optimized via contrastive learning on extensive image-text pairs, these encoders establish foundational cross-modal alignments. However, while projecting these features into the LLM’s latent space enables conditional text generation, this inherently text-centric optimization severely bottlenecks the retention of dense visual details.

2.2 Visual Supervision incorporated in VLMs

Recent research explores the incorporation of visual supervision into VLMs to enhance visual understanding during the post-training phase. One trajectory of inquiry introduces this supervision via vision-text alignment [21], whereas an alternative paradigm focuses on incorporating visual supervision by predicting or reconstructing image features produced by a vision tokenizer [18, 19, 40–42, 51]. However, the efficacy of these methodologies is inherently constrained. Given that the latent representations are predominantly crystallized by text-centric large-scale pretraining prior to this stage, visual signals function merely as auxiliary alignment, yielding marginal improvements. Conversely, the initial pre-training phase constitutes the critical window for fundamentally optimizing the representational capacity of VLMs. To address these limitations, we propose VIVAS, a framework that leverages dense semantic visual signals to formulate a vision-language unified autoregressive supervision, thereby significantly enhancing fine-grained visual perception.

3 Methodology

3.1 Preliminary

Given a standard VLM parameterized by θ , visual inputs v and text queries x are mapped into a shared embedding space via a vision encoder and an alignment

module. The model autoregressively generates a response.

$$y_t \sim p_\theta(y_t \mid v, x, y_{<t}) \propto \exp [\text{logit}_\theta(y_t \mid v, x, y_{<t})]. \quad (1)$$

In this paradigm, optimizing solely via text-generation objectives treats visual signals as passive conditions. Consequently, models tend to discard fine-grained visual details unnecessary for coarse-grained text generation, creating an information bottleneck that impedes dense visual perception capabilities. To overcome this limitation, Section 3.2 establishes the requisite visual supervision for VLM pre-training. Subsequently, Section 3.3 introduces how to obtain such visual supervision, and Section 3.4 details our proposed unified vision-language autoregressive training pipeline. Crucially, our pre-training entails the end-to-end optimization of the vision encoder, projector, and LLM decoder.

3.2 What Visual Supervision Are Needed for VLM Pre-training?

To jointly ensure stable optimization while capturing essential visual representations, this section investigates the mechanisms required for visual supervision during pretraining from two primary perspectives: *the visual supervision architectural paradigm* and *the characteristics of the visual supervision signal*.

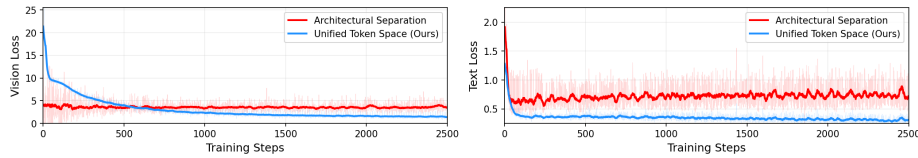


Fig. 3: Comparison of training loss between architectural separation and unified token space. Vision and text losses correspond to the supervised training objectives for the visual and textual modalities, respectively.

Visual Supervision Architectural Paradigm. In visual supervision for VLM pre-training, two different architectural paradigm formulations can be adopted: architectural separation and unified token space.

Architectural Separation: Under the architectural separation paradigm, visual supervision is introduced through an auxiliary alignment module attached to the backbone VLM. Specifically, the output-layer hidden states produced by the model are fed into this module, which maps them to predefined visual targets using a feature-space regression loss. The alignment module is optimized jointly with the backbone during pre-training.

Unified Token Space: In the unified token space paradigm, the textual vocabulary is expanded into a unified vision-language vocabulary by incorporating a visual vocabulary. The VLM is then trained autoregressively over this unified vocabulary, where visual supervision is formulated as next-token prediction under a standard cross-entropy objective.

As illustrated in the Fig. 3, evaluating the training loss trajectories highlights a fundamental divergence between architectural separation and unified token space paradigms. In the architectural separation paradigms, an external Diffusion Transformer (DiT) [31] is appended to the hidden states of the VLM’s output-layer. These hidden states function as conditions to guide the denoising process of corrupted images. Empirical observations indicate that the architectural separation paradigm suffers from severe training instability and convergence difficulties. This primarily stems from the early stage of pretraining, when the model’s visual capability has not yet been established; introducing additional training modules at this stage to provide visual supervision disrupts optimization and exacerbates training instability. In contrast, optimization within a unified token space yields stable and rapid convergence. Consequently, these findings establish the unified token space paradigm as a crucial prerequisite for achieving stable training convergence in effective visual supervision.

Table 1: Performance comparison of various visual supervision strategies. All models are pre-trained on a 0.5T-token corpus and evaluated under a 5-shot setting across diverse multimodal benchmarks.

Methods	VisuLogic	<i>MMBENCH</i>	RealWorldQA	MME	<i>CRPEEXIST</i>	<i>CRPERELATION</i>	<i>CVBENCHD</i>	HallusionBench	TextVQA	DocVQA	RefCOCO _{val}	RefCOCO _{test}
w/o Visual Supervision	22.2	53.0	43.0	48.2	65.7	57.0	51.0	55.3	52.7	60.2	62.1	59.4
Text-decoupled vocabularies	21.6	51.2	41.8	46.9	63.5	55.1	49.3	53.9	51.4	58.6	60.7	57.8
Vision-language semantic vocabularies	23.6	53.6	43.2	47.7	66.2	57.6	51.4	54.8	53.1	59.3	63.0	60.7
Ours	24.4	54.9	46.5	49.8	72.2	59.0	56.3	56.5	54.2	63.4	65.2	63.1

Characteristics of the Visual Supervision Signal. Under the unified token space paradigm, we categorize visual supervision signals into two groups: *text-decoupled vocabularies* and *vision-language semantic vocabularies*.

Text-decoupled vocabularies. Text-decoupled signals are learned strictly through visual supervision, without explicit text alignment. These representations typically stem from pixel-level tokenizers (e.g., VQ-VAE) or self-supervised encoders (e.g., DINOv3). Despite varied training paradigms, both approaches capture purely visual features. In this study, we extract these representations using a standard VQ-VAE vision tokenizer optimized for image reconstruction.

Vision-language semantic vocabularies. In contrast, semantic signals derived from image-text contrastive learning encode high-level abstractions aligned with language semantics. To obtain such representations, we adapt the VQ-VAE framework by leveraging a frozen, pretrained semantic vision encoder (SigLIP2) to produce semantic features while retaining the reconstruction objective.

Standard VLM benchmarks predominantly employ zero-shot evaluation settings. However, pre-training checkpoints typically lack instruction alignment, resulting in formatting mismatches that obscure the intrinsic capabilities of the

model. To decouple representation quality from instruction-following proficiency, we adopt a few-shot evaluation protocol aligned with recent assessment methodologies for foundational language models [6]. Specifically, we leverage in-context learning to standardize output formats, thereby facilitating a rigorous assessment of the base model independent of explicit instruction tuning.

Utilizing this protocol, we investigate the impact of different visual supervision signals during VLM pre-training. As summarized in Table 1, incorporating either text-decoupled or vision-language semantic supervision results in suboptimal performance compared to a purely text-supervised baseline. Specifically, models trained with text-decoupled supervision exhibit a noticeable degradation in multimodal understanding. In contrast, semantic supervision derived from image-text contrastive models provides only marginal improvements and occasionally leads to slight performance drops across standard benchmarks. These observations reveal two fundamental representational limitations. First, text-decoupled supervision, learned purely from visual signals without alignment to the language modality, introduces a substantial semantic gap that hinders alignment with high-level textual concepts. This misalignment leads to a pronounced modality disconnect. Second, vision-language semantic supervision derived from contrastive vision-language encoders generates representations that are highly correlated with those already captured by the VLM’s vision encoder. Consequently, such supervision provides limited complementary information and fails to introduce the structural granularity required for improving fine-grained visual perception.

3.3 Dense-Structural-Semantic Visual Supervision

Motivated by the limitations discussed in Section 3.2, we aim to design a visual supervision mechanism that preserves training stability while providing a dense supervision signal enriched with both structural granularity and semantic information during pretraining. To achieve this, we propose a dense-structural-semantic vision tokenizer. This vision tokenizer initially synthesizes dense, modality-aligned representations enriched with explicit spatial anchoring, subsequently discretizing them via a compact codebook to derive the requisite visual supervisory signals.

DINOv3, optimized via self-distillation, excels in delineating local geometry, high-frequency structures, and precise object-part boundaries; nevertheless, it inherently lacks modality-aligned semantics. Conversely, SigLIP2 yields robust category-level semantics through rigorous vision-language alignment, yet its spatial representations remain intrinsically coarse. This spatial deficiency frequently precipitates blurred boundaries and semantic bleeding across adjacent regions. To synthesize visual representations that are concurrently structurally dense and semantically aligned, we leverage DINOv3 features as structural queries ($\mathbf{Q}$) to establish rigorous spatial templates, while treating SigLIP2 features as the semantic key-value repository ($\mathbf{K}, \mathbf{V}$) for the retrieval of language-grounded attributes.

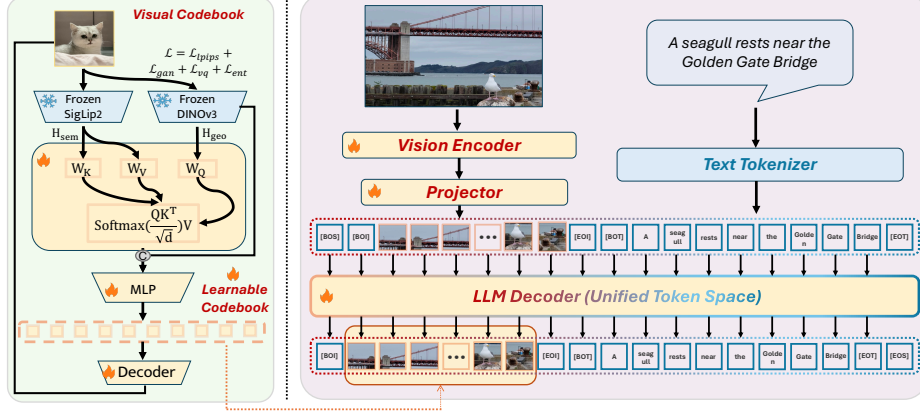


Fig. 4: Overview of VIVAS. **Left:** The dense structural-semantic vision tokenizer constructs a visual codebook by fusing semantic and geometric features through cross-attention. **Right:** During end-to-end pretraining, VIVAS performs unified autoregressive supervision over both the textual vocabulary and the vision vocabulary.

Given an input image, it is processed through dual frozen expert encoders to extract patch-level feature embeddings. Specifically, let $\mathbf{H}_{geo} \in \mathbb{R}^{N \times d_d}$ and $\mathbf{H}_{sem} \in \mathbb{R}^{N \times d_s}$ denote the sequences of DINOv3 tokens and SigLIP-2 tokens, capturing geometry-aware and language-aligned semantic features, respectively. Here, N represents the number of spatial patches, while d_d and d_s correspond to their feature dimensions. To integrate these representations, a cross-attention fusion mechanism is employed to probe semantic features under structural constraints. By projecting the feature maps into a shared latent space, the Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) matrices are defined as:

$$\mathbf{Q} = \mathbf{H}_{geo} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{H}_{sem} \mathbf{W}_K, \quad \mathbf{V} = \mathbf{H}_{sem} \mathbf{W}_V, \quad (2)$$

whereafter the synergistic representation $\mathbf{Z}_{syn}$ is computed via cross-attention:

$$\mathbf{Z}_{syn} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}. \quad (3)$$

Prior to quantization, $\mathbf{Z}_{syn}$ is concatenated with the structural features $\mathbf{H}_{geo}$ along the channel dimension. The composite representation is projected via a multi-layer perceptron (MLP) and discretized using Index Backpropagation Quantization (IBQ) [34]. This operation maps continuous vectors to the nearest prototype within a learnable codebook $\mathcal{C} = \{c_k\}_{k=1}^K$, characterized by a vocabulary size of $K = 150,000$ and an embedding dimension of $D = 768$.

The tokenizer is optimized end-to-end to reconstruct the input images from the discrete codes. A compound objective function $\mathcal{L}_{tok}$ is formulated to balance perceptual fidelity with codebook utilization:

$$\mathcal{L}_{tok} = \underbrace{\lambda_p \mathcal{L}_{lpips} + \lambda_g \mathcal{L}_{gan}}_{\text{Perceptual \& Adversarial Fidelity}} + \underbrace{\mathcal{L}_{vq} + \lambda_e \mathcal{L}_{ent}}_{\text{Codebook Optimization}} \quad (4)$$

where $\mathcal{L}_{\text{lip}}$ enforces textural realism through perceptual similarity, and $\mathcal{L}_{\text{gan}}$ represents the adversarial discriminator loss. To preclude codebook collapse, a vector quantization loss $\mathcal{L}_{\text{vq}}$ is integrated alongside an entropy regularization term $\mathcal{L}_{\text{ent}}$, with loss weights empirically set to $\lambda_p = 1$, $\lambda_g = 1$, and $\lambda_e = 0.1$. A critical design choice in this optimization strategy is the deliberate exclusion of the standard pixel-wise ℓ_1 reconstruction loss. Minimization of the ℓ_1 norm inherently induces a “texture bias”, promoting the memorization of high-frequency noise while bypassing high-level semantic abstraction. By relying exclusively on perceptual and adversarial constraints, the codebook is compelled to encode structural semantics rather than mere pixel statistics. Empirically, this configuration yields a codebook utilization rate of 97.7% on the composite dataset. Consequently, the learned indices form a robust visual vocabulary $\mathcal{V}_{\text{img}}$ to provide visual supervision.

3.4 Vision-Language Unified Autoregressive Pre-training

Unified Training Objective. We formulate multimodal learning by interleaving discrete visual tokens $V = \{v_1, \dots, v_M\}$ and textual tokens $T = \{t_1, \dots, t_N\}$ into a unified sequence $\mathcal{S}$. Under an autoregressive framework, the model is optimized to maximize the likelihood of each token s_i conditioned on its preceding context $s_{<i}$, regardless of modality. The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \lambda \mathcal{L}_{\text{vision}}, \quad (5)$$

where λ balances the relative contribution of visual token prediction, which we empirically set to 0.5. Both components are optimized using token-level cross-entropy:

$$\mathcal{L}_{\text{text}} = - \sum_{i \in \mathcal{I}_t} \log P(t_i \mid s_{<i}; \theta), \quad \mathcal{L}_{\text{vision}} = - \sum_{j \in \mathcal{I}_v} \log P(v_j \mid s_{<j}; \theta), \quad (6)$$

where $\mathcal{I}_t$ and $\mathcal{I}_v$ denote the positional indices of textual and visual tokens in $\mathcal{S}$, respectively.

Progressive Training Curriculum. To systematically endow the model with both robust reasoning and fine-grained visual grounding, the training process is organized into three sequential stages.

Stage I: Language Backbone Pre-training. The language backbone is first trained on large-scale pure text corpora to establish a rigorous linguistic foundation. This unimodal phase focuses on acquiring general reasoning abilities and domain-specific knowledge. During this stage, the model is optimized solely using the textual objective:

$$\mathcal{L}_{\text{Stage}_1} = \mathcal{L}_{\text{text}}. \quad (7)$$

Stage II, III: Multimodal Pre-training. In this stage, the entire architecture, including the vision encoder, projector, and LLM, is trained end-to-end to align

visual and linguistic modalities. Crucially, we introduce visual supervision during this phase to guide the representation learning. By activating the joint objective:

$$\mathcal{L}_{\text{Stage}_{2,3}} = \mathcal{L}_{\text{text}} + \lambda \mathcal{L}_{\text{vision}}. \quad (8)$$

By explicitly incorporating $\mathcal{L}_{\text{vision}}$, the model is compelled to simultaneously predict visual details and linguistic content. This strategy effectively mitigates the text-centric optimization bias inherent in traditional MLLM pre-training and preserves dense visual perception capabilities.

Stage IV: Multimodal Post-training. The final stage focuses on aligning the model with human preferences and specializing it for diverse multimodal tasks through Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). This phase refines the model’s instruction-following and conversational capabilities. To ensure robust instruction alignment and superior dialogue quality, the visual supervision is deactivated during this stage:

$$\mathcal{L}_{\text{Stage}_4} = \mathcal{L}_{\text{text}}. \quad (9)$$

4 Experiment

4.1 Implementation Details

Training Recipe. We adopt a four-stage training pipeline with progressively increasing data diversity and task complexity.

Stage I: Language Backbone Pre-training. The language backbone is trained through two consecutive phases: (1) commonsense-oriented pre-training and (2) STEM- and coding-centric pre-training. These two phases collectively comprise approximately 10T tokens of pure textual data, covering general knowledge, mathematical reasoning, scientific domains and coding tasks.

Stage II, III: Multimodal Pre-training. In these stages, the entire architecture undergoes end-to-end training to develop robust visual perception and versatile task capabilities, processing approximately 2.4T tokens in total. The process begins with foundational pre-training (Stage II, 1.8T tokens) on a mixture of image-caption pairs, vision-centric tasks, and high-quality textual data to balance cross-modal comprehension with linguistic proficiency. Subsequently, the model undergoes task adaptation (Stage III, 0.6T tokens) on diverse multimodal instructions covering general VQA, OCR, STEM, GUI navigation, and fine-grained perception (e.g., detection, grounding). We also incorporate synthetic short Chain-of-Thought (CoT) data to bolster long-context understanding and logical reasoning.

Stage IV: Multimodal Post-training. SFT. We construct a high-quality multimodal instruction dataset through two strategies, comprising approximately 4B tokens: (1) mining instruction-following samples from the pre-training corpus using VLM-based quality filtering and taxonomy-guided balancing; (2) refining open-source datasets by regenerating detailed, reasoning-enhanced responses via a rewrite-and-expand pipeline. *RL.* We further optimize the model with reinforcement learning in three aspects, comprising approximately 0.2B tokens: (i)

perception RL for fine-grained visual tasks such as grounding and detection; (ii) reasoning RL on curated STEM data with verifiable rewards; (iii) general RL on diverse VQA tasks including OCR and chart understanding to improve generalization.

Model Components. The architectural framework of VIVAS integrates a LLM backbone predicated on Youtu-LLM [25] with a visual perception module driven by the SigLIP2-so400m-patch16-naflex encoder.

Evaluation. To rigorously validate the efficacy of VIVAS, an extensive evaluation is conducted across a comprehensive suite of 39 general multimodal benchmarks. This diverse evaluation protocol encompasses a broad spectrum of tasks, including visual grounding, object detection, instance counting, visual query answering, optical character recognition (OCR), and graphical user interface (GUI) interactions. For the ablation studies concerning the codebook configurations, specifically its vocabulary size, embedding dimension, and the applied loss function, all variant models are pre-trained on a corpus of 0.5T tokens and subsequently evaluated under a 5-shot setting.

Table 2: Ablation study of applying visual supervision at different training stages.

Pre-train VS	Post-train VS	VisuLogic	MMBench _{EN}	CRPE _{relation}	HallusionBench
×	×	23.1	83.2	69.5	57.3
✓	×	25.7	83.9	72.2	59.1
×	✓	23.3	82.9	69.9	57.7
✓	✓	23.6	83.3	70.1	58.1

4.2 Quantitative Results

As delineated in Table 3, a comprehensive evaluation across 39 diverse multimodal benchmarks reveals that VIVAS establishes SOTA performance on the most multimodal understanding tasks. Notably, the proposed method demonstrates pronounced improvements on tasks necessitating fine-grained visual perception, such as visual grounding. Furthermore, ablation comparisons against the baseline variant devoid of pre-training visual supervision confirm substantial performance gains attributable to the proposed mechanism. Crucially, these representational enhancements are achieved with strict computational efficiency; the requisite vocabulary expansion incurs a nominal inference time overhead of merely 3.8%.

4.3 Out-of-Distribution and Fine-Grained Generalization

To further examine whether the improvements of VIVAS generalize beyond the curated multimodal pre-training distribution, we evaluate the model on an out-

Table 3: Performance comparison between VIVAS and SOTA VLMs across multiple benchmarks. All scores adhere to official evaluation protocols. The results reproduced via our internal evaluation system are marked with ‘*’.

Benchmarks	LLaVA-OV 1.5	Qwen3-VL instruct	LLaVA-OV 1.5	InternVL-3.5	Qwen3-VL instruct	VIVAS w/o Vision Sup.	VIVAS (Ours)
Params	8B	8B	4B	4B	4B	4B	4B
General VQA							
MMBench _{CN} [24]	81.0	84.7	76.9	–	83.5	82.7	83.6
MMBench _{EN} [24]	84.1	84.5	84.2	80.3	83.9	83.2	83.9
MMStar [3]	67.7	70.9	64.9	65.0	69.8	70.4	71.1
MME (/2800) [7]	–	–	–	2272	2309*	2360	2384
CVBench _{2d} [38]	–	–	–	–	79.1*	74.2	80.4
CVBench _{3d} [38]	–	–	–	–	92.4*	87.7	93.0
ScienceQA _{val} [27]	–	–	–	–	94.7*	95.8	97.0
SEEDBench _{IMG} [17]	77.3	–	76.6	–	77.0*	75.8	76.9
Multimodal Reasoning & Math							
VisuLogic [50]	–	22.5	–	–	19.0	23.1	25.7
MathVista _{mini} [26]	69.6	77.2	67.9	77.1	73.7	75.1	76.5
MathVerse _{mini} [54]	–	62.1	–	45.8	46.8	53.1	56.5
LogicVista [48]	–	55.3	–	41.8	53.2	50.9	52.4
VLMsAreBlind [33]	–	74.0	–	–	71.9	85.4	88.9
Hallucination							
HallusionBench [10]	–	61.1	–	44.8	57.6	57.3	59.1
CRPE _{exist} [45]	–	–	–	–	95.6*	95.2	96.9
CRPE _{relation} [45]	–	–	–	75.0	71.0*	69.5	72.2
GUI Agent							
ScreenSpot Pro [20]	–	54.6	–	–	59.5	57.1	59.6
OSWorld [49]	–	33.9	–	–	26.2	36.4	38.8
Visual Grounding							
RefCOCO _{val} [52]	–	–	–	92.5	90.7	88.1	93.6
RefCOCO _A [52]	–	–	–	94.3	92.2	91.4	95.2
RefCOCO _B [52]	–	–	–	88.2	86.7	85.1	90.8
RefCOCO ₊ _{val} [52]	–	–	–	87.6	82.9	87.6	90.1
RefCOCO ₊ _A [52]	–	–	–	92.3	89.4	90.1	93.9
RefCOCO ₊ _B [52]	–	–	–	81.6	75.6	81.3	85.4
RefCOCO _g _{val} [52]	–	–	–	89.6	87.3	89.5	92.2
RefCOCO _g _{test} [52]	–	–	–	89.3	87.7	89.0	92.9
OCR-related Understanding							
AI2D _{test} [15]	84.2	85.7	83.6	82.6	84.1	84.2	85.6
InfoVQA _{val} [29]	78.4	83.1	76.1	78.0	80.3	77.8	79.1
TextVQA _{val} [36]	–	–	–	77.9	80.8*	79.4	79.6
DocVQA _{val} [30]	95.0	96.1	94.4	92.4	95.3	94.1	94.4
ChartQA _{test} [28]	86.5	89.6	87.1	86.0	84.6	84.7	85.3
OCRBench [8]	829	896	800	822	881	804	813
SEEDBench2 _{plus} [16]	69.2	–	68.9	69.4	71.5*	69.8	71.3
CharXiv _{DQ} [46]	70.9	83.0	63.8	71.1	76.2	76.5	79.4
CharXiv _{RQ} [46]	–	46.4	–	39.6	39.7	41.6	43.8
Multi-image & Real-world							
BLINK [9]	–	69.1	–	58.1	65.8	62.7	64.3
RealWorldQA	68.1	71.5	67.8	66.3	70.9	72.9	74.6
MMERealWorld _{EN} [55]	61.7	–	61.6	–	63.0*	60.7	61.5
MMERealWorld _{CN} [55]	56.1	–	49.6	59.8	61.3*	61.2	63.5

Table 4: Ablation study on the codebook size of the visual tokenizer.

Codebook Size	VisuLogic	MMBench _{EN}	CRPE _{relation}	HallusionBench	Utilization (%)
50k	23.6	53.5	57.3	55.6	98.3
100k	24.0	54.1	58.1	56.1	98.6
150k	24.4	54.9	59.0	56.5	97.7
200k	24.2	54.5	58.7	56.3	96.6

Table 5: Ablation study on the codebook embedding dimension of the visual tokenizer.

Codebook Dim	VisuLogic	MMBench _{EN}	CRPE _{relation}	HallusionBench	Utilization (%)
64	22.7	53.1	57.4	55.5	95.1
256	23.6	53.7	58.6	55.9	97.8
768	24.4	54.9	59.0	56.5	97.7
1024	23.9	54.6	58.6	56.4	94.3

of-distribution domain and additional fine-grained perception benchmarks. As shown in Table 6, VIVAS consistently outperforms the baseline without visual supervision on GEOBench-VLM, VLMsAreBiased, and V*. In particular, the improvement on GEOBench-VLM, which contains satellite imagery that is underrepresented in standard multimodal training corpora, suggests that the proposed visual supervision does not merely fit the original data distribution. The gains on VLMsAreBiased and V* further support that the visual supervision strengthens fine-grained visual perception capabilities.

4.4 Ablation Studies

As delineated in Table 2, an analysis is conducted to investigate the impact of integrating visual supervision across distinct training stages. Empirical evidence indicates that confining this supervision exclusively to the post-training phase yields only marginal improvements, whereas its integration across both pre-training and post-training phases provides merely modest gains. This is primarily because introducing visual supervision during post-training interferes with the model’s instruction-following capabilities. Conversely, isolating the visual supervision entirely within the pre-training stage results in substantial performance enhancements. This observation robustly substantiates that the pre-training phase serves as the pivotal window for optimizing the foundational representational capacity of VLMs.

To further examine the effect of modifying input-side visual features, we evaluate an encoder-side baseline that combines SigLIP2 with DINOv3 features through a learnable projector, while using the same 0.5T-token corpus setting and without visual-token supervision. As shown in Table 8, this baseline fails to outperform VIVAS w/o visual supervision, remains clearly inferior to full VIVAS, and introduces additional inference cost. We attribute this to two factors: 1) under text-dominant supervision, simply enriching the input visual features is insufficient to learn fine-grained visual signals; and 2) incorporating text-decoupled

Table 6: Out-of-distribution and fine-grained perception evaluation.

Benchmark	Qwen3-VL-4B	VIVAS w/o VS	VIVAS	Gain
GEOBench-VLM [5]	40.21	44.64	46.37	+1.73
VLMsAreBiased [43]	21.31	18.33	22.16	+3.83
V* [47]	80.63	78.53	81.15	+2.62

Table 7: Ablation study on adding L1 loss during codebook training.

L1 Loss	VisuLogic	MMBench _{EN}	CRPE _{relation}	HallusionBench
×	24.4	54.9	59.0	56.5
✓	23.7	54.2	57.9	56.1

DINOv3 features may dilute the text-aligned SigLIP2 representation, making vision-language alignment more difficult. In contrast, VIVAS keeps the standard SigLIP2-based continuous visual input unchanged and instead introduces explicit visual supervision during pre-training, thereby improving fine-grained perception without disrupting input-side alignment.

Furthermore, Tables 4 and 5 detail a hyperparameter analysis regarding the capacity, specifically the vocabulary size and embedding dimension, of the visual codebook derived from the synergistic vision tokenizer. The findings reveal that peak performance is achieved with a vocabulary size of 150,000 and an embedding dimension of 768, a configuration that concurrently sustains an exceptionally high codebook utilization rate of 97.7%.

In addition, while L1 loss serves as a foundational component in traditional codebook training, Table 7 investigates the efficacy of its inclusion when optimizing specifically for visual understanding. The experimental results demonstrate that omitting the L1 loss yields superior performance. This is because the minimization of the L1 loss inherently induces a “texture bias,” compelling the codebook to allocate its representational capacity toward high-frequency noise at the expense of high-level semantic abstraction.

4.5 Analysis of Data Scaling

We demonstrate the scalability and stability of our 2.4T-token pre-training regime through empirical scaling laws. Figure 5a shows a monotonic increase in multimodal performance (averaged over 27 benchmarks) from 0.43 to >0.74, spanning both Stage 2 (Foundation) and Stage 3 (Task Adaptation) without early saturation.

By mapping evaluation error against compute (GFLOPs), Figure 5 confirms a predictable power-law decay, $L(C) \propto C^{-\alpha}$. The model exhibits a steep scaling exponent of $\alpha \approx 0.102$ during Stage 2, indicating rapid information absorption from mixed corpora. In Stage 3, the exponent naturally moderates to $\alpha \approx 0.079$.

Table 8: Comparison with a combined visual encoder baseline. The combined visual encoder baseline uses SigLIP2 and DINOv3 features with a learnable projector, but does not use visual-token supervision.

Method	VisuLogic	CVBench2D	MME	RefCOCO _{val}	DocVQA	HallusionBench
VIVAS w/o VS	22.2	51.0	48.2	62.1	60.2	55.3
w/ combined visual encoder	22.6	50.7	48.5	60.9	60.5	54.6
VIVAS	24.4	56.3	49.8	65.2	63.4	56.5

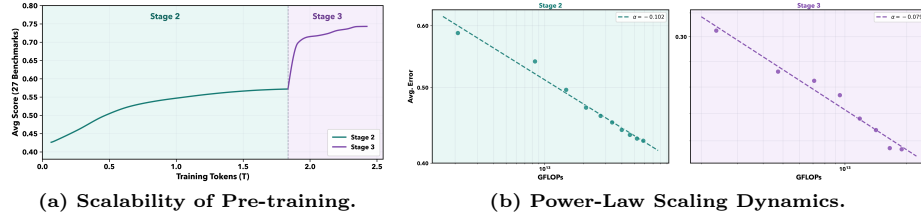


Fig. 5: Scaling analysis of VIVAS. (a) The evolution of average performance across 27 multimodal benchmarks as a function of training tokens. The results reveal steady and consistent capability improvements during training across different data regimes, indicating that the model effectively exploits large-scale data without showing signs of empirical saturation. (b) Examination of the mean evaluation error ($1 - \text{Score}$). The optimization dynamics in both Stage 2 and Stage 3 follow neural scaling laws, with scaling exponents estimated at $\alpha \approx 0.102$ and $\alpha \approx 0.079$, respectively.

This sustained log-linear scaling underscores the sample efficiency of our framework, effectively converting massive computational investment into predictable capability gains.

5 Conclusion

In this paper, we study the limitation of existing Vision–Language Models (VLMs) in fine-grained visual perception and attribute it to the text-dominant optimization bias in standard training paradigms. To address this issue, we introduce visual supervision into the pretraining stage and show that the unified token space paradigm enables stable and effective training for vision–language autoregressive models. Based on this paradigm, we propose a dense-structural-semantic vision tokenizer that integrates dense structural features from DINOv3 and semantic representations from SigLIP2 to construct a modality-aligned visual vocabulary. Building on this tokenizer, we develop VIVAS, a vision–language pretraining framework with unified supervision over visual and textual tokens. Extensive experiments demonstrate that VIVAS consistently improves multimodal understanding while introducing minimal inference overhead.

Acknowledgments

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(No.KJZD20231023094700001) and the Shenzhen-Tsinghua Special Project for Fundamental & Frontier Research in Artificial Intelligence(No.AI2026018).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), <https://arxiv.org/abs/2511.21631>
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024), <https://arxiv.org/abs/2403.20330>
4. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
5. Danish, M.S., Munir, M.A., Shah, S.R.A., Kuckreja, K., Khan, F.S., Fraccaro, P., Lacoste, A., Khan, S.: Geobench-vlm: Benchmarking vision-language models for geospatial tasks (2025), <https://arxiv.org/abs/2411.19325>
6. DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J.L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R.J., Jin, R.L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S.S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W.L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X.Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhu, Y.X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong,

- Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z.F., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., Pan, Z.: Deepseek-v3 technical report (2025), <https://arxiv.org/abs/2412.19437>
7. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), <https://arxiv.org/abs/2306.13394>
 8. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., Bai, X.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning (2025), <https://arxiv.org/abs/2501.00321>
 9. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive (2024), <https://arxiv.org/abs/2404.12390>
 10. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu.e.g., X.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models (2024), <https://arxiv.org/abs/2310.14566>
 11. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., Chen, J., Huang, J., Lei, K., Yuan, L., Luo, L., Liu, P., Ye, Q., Qian, R., Yan, S., Zhao, S., Peng, S., Li, S., Yuan, S., Wu, S., Cheng, T., Liu, W., Wang, W., Zeng, X., Liu, X., Qin, X., Ding, X., Xiao, X., Zhang, X., Zhang, X., Xiong, X., Peng, Y., Chen, Y., Li, Y., Hu, Y., Lin, Y., Hu, Y., Zhang, Y., Wu, Y., Li, Y., Liu, Y., Ling, Y., Qin, Y., Wang, Z., He, Z., Zhang, A., Yi, B., Liao, B., Huang, C., Zhang, C., Deng, C., Deng, C., Lin, C., Yuan, C., Li, C., Gou, C., Lou, C., Wei, C., Liu, C., Li, C., Zhu, D., Zhong, D., Li, F., Zhang, F., Wu, G., Li, G., Xiao, G., Lin, H., Yang, H., Wang, H., Ji, H., Hao, H., Shen, H., Li, H., Li, J., Wu, J., Zhu, J., Jiao, J., Feng, J., Chen, J., Duan, J., Liu, J., Zeng, J., Tang, J., Sun, J., Chen, J., Long, J., Feng, J., Zhan, J., Fang, J., Lu, J., Hua, K., Liu, K., Shen, K., Zhang, K., Shen, K., Wang, K., Pan, K., Zhang, K., Li, K., Li, L., Li, L., Shi, L., Han, L., Xiang, L., Chen, L., Chen, L., Li, L., Yan, L., Chi, L., Liu, L., Du, M., Wang, M., Pan, N., Chen, P., Chen, P., Wu, P., Yuan, Q., Shuai, Q., Tao, Q., Zheng, R., Zhang, R., Zhang, R., Wang, R., Yang, R., Zhao, R., Xu, S., Liang, S., Yan, S., Zhong, S., Cao, S., Wu, S., Liu, S., Chang, S., Cai, S., Ao, T., Yang, T., Zhang, T., Zhong, W., Jia, W., Weng, W., Yu, W., Huang, W., Zhu, W., Yang, W., Wang, W., Long, X., Yin, X., Li, X., Zhu, X., Jia, X., Zhang, X., Liu, X., Zhang, X., Yang, X., Luo, X., Chen, X., Zhong, X., Xiao, X., Li, X., Wu, Y., Wen, Y., Du, Y., Zhang, Y., Ye, Y., Wu, Y., Liu, Y., Yue, Y., Zhou, Y., Yuan, Y., Xu, Y., Yang, Y., Zhang, Y., Fang, Y., Li, Y., Ren, Y., Xiong, Y., Hong, Z., Wang, Z., Sun, Z., Wang, Z., Cai, Z., Zha, Z., An, Z., Zhao, Z., Xu, Z., Chen, Z., Wu, Z., Zheng, Z., Wang, Z., Huang, Z., Zhu, Z., Song, Z.: Seed1.5-v1 technical report (2025), <https://arxiv.org/abs/2505.07062>
 12. Kan, Z., Li, X., Liu, Y., Yang, X., Jiang, X., Liu, Y., Jiang, D., Sun, X., Liao, Q., Yang, W.: RAR: Reversing visual attention re-sinking for unlocking potential in multimodal large language models. In: The Fourteenth International Conference on Learning Representations (2026), <https://openreview.net/forum?id=xTbFeDhdRG>

13. Kan, Z., Liu, Y., Yin, K., Jiang, X., Li, X., Cao, H., Liu, Y., Jiang, D., Sun, X., Liao, Q., Yang, W.: Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms (2025), <https://arxiv.org/abs/2505.20777>
14. Kan, Z., Zhang, C., Liao, Z., Tian, Y., Yang, W., Xiao, J., Li, X., Jiang, D., Wang, Y., Liao, Q.: Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms (2024), <https://arxiv.org/abs/2411.12713>
15. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images (2016), <https://arxiv.org/abs/1603.07396>
16. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench-2: Benchmarking multimodal large language models (2023), <https://arxiv.org/abs/2311.17092>
17. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension (2023), <https://arxiv.org/abs/2307.16125>
18. Li, F., Song, W., Zhao, H., Wang, J., Ding, P., Wang, D., Zeng, L., Li, H.: Spatial forcing: Implicit spatial representation alignment for vision-language-action model. arXiv preprint arXiv:2510.12276 (2025)
19. Li, H., Zhang, X., Peng, Q., Luo, B., Hu, H., Jiang, D., Ye, H.J., Zhang, T., Jin, H.: Unleashing the intrinsic visual representation capability of multimodal large language models. arXiv preprint arXiv:2512.06281 (2025)
20. Li, K., Meng, Z., Lin, H., Luo, Z., Tian, Y., Ma, J., Huang, Z., Chua, T.S.: Screenspot-pro: Gui grounding for professional high-resolution computer use (2025), <https://arxiv.org/abs/2504.07981>
21. Li, M., Su, N., Qu, F., Zhong, Z., Chen, Z., Li, Y., Tu, Z., Li, X.: Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. arXiv preprint arXiv:2505.10917 (2025)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485>
24. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), <https://arxiv.org/abs/2307.06281>
25. Lu, J., Qin, J., Qiao, L., Li, Y., Dai, X., Ke, B., He, J., Qiao, R., Yin, D., Sun, X., Wu, Y., Liu, Y., Liu, S., Tang, M., Lin, H., Kuang, J., Meng, F., Tang, X., Xi, Y., Huang, J., Yang, H., Shen, Z., Li, Y., Zhang, Q., Yu, Y., An, S., Dong, J., Wang, Q., Wang, J., Chen, K., Wen, W., Guo, T., Shen, Z., Yu, D., Li, J., Li, K., Li, Z., Tan, X.: Youtu-llm: Unlocking the native agentic potential for lightweight large language models (2026), <https://arxiv.org/abs/2512.24618>
26. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts (2024), <https://arxiv.org/abs/2310.02255>
27. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering (2022), <https://arxiv.org/abs/2209.09513>
28. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning (2022), <https://arxiv.org/abs/2203.10244>
29. Mathew, M., Bagal, V., Tito, R.P., Karatzas, D., Valveny, E., Jawahar, C.V.: Infographicvqa (2021), <https://arxiv.org/abs/2104.12756>

30. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for vqa on document images (2021), <https://arxiv.org/abs/2007.00398>
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023), <https://arxiv.org/abs/2212.09748>
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), <https://arxiv.org/abs/2103.00020>
33. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind: Failing to translate detailed visual features into words (2025), <https://arxiv.org/abs/2407.06581>
34. Shi, F., Luo, Z., Ge, Y., Yang, Y., Shan, Y., Wang, L.: Scalable image tokenization with index backpropagation quantization (2025), <https://arxiv.org/abs/2412.02692>
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025), <https://arxiv.org/abs/2508.10104>
36. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read (2019), <https://arxiv.org/abs/1904.08920>
37. Team, ., Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., Wang, K., Zhong, L., Liu, M., Lu, R., Cao, S., Zhang, X., Huang, X., Wei, Y., Cheng, Y., An, Y., Niu, Y., Wen, Y., Bai, Y., Du, Z., Wang, Z., Zhu, Z., Zhang, B., Wen, B., Wu, B., Xu, B., Huang, C., Zhao, C., Cai, C., Yu, C., Li, C., Ge, C., Huang, C., Zhang, C., Xu, C., Zhu, C., Li, C., Yin, C., Lin, D., Yang, D., Jiang, D., Ai, D., Zhu, E., Wang, F., Pan, G., Wang, G., Sun, H., Li, H., Li, H., Hu, H., Zhang, H., Peng, H., Tai, H., Zhang, H., Wang, H., Yang, H., Liu, H., Zhao, H., Liu, H., Yan, H., Liu, H., Chen, H., Li, J., Zhao, J., Ren, J., Jiao, J., Zhao, J., Yan, J., Wang, J., Gui, J., Zhao, J., Liu, J., Li, J., Li, J., Lu, J., Wang, J., Yuan, J., Li, J., Du, J., Du, J., Liu, J., Zhi, J., Gao, J., Wang, K., Yang, L., Xu, L., Fan, L., Wu, L., Ding, L., Wang, L., Zhang, M., Li, M., Xu, M., Zhao, M., Zhai, M., Du, P., Dong, Q., Lei, S., Tu, S., Yang, S., Lu, S., Li, S., Li, S., Shuang-Li, Yang, S., Yi, S., Yu, T., Tian, W., Wang, W., Yu, W., Tam, W.L., Liang, W., Liu, W., Wang, X., Jia, X., Gu, X., Ling, X., Wang, X., Fan, X., Pan, X., Zhang, X., Zhang, X., Fu, X., Zhang, X., Xu, Y., Wu, Y., Lu, Y., Wang, Y., Zhou, Y., Pan, Y., Zhang, Y., Wang, Y., Li, Y., Su, Y., Geng, Y., Zhu, Y., Yang, Y., Li, Y., Wu, Y., Li, Y., Liu, Y., Wang, Y., Li, Y., Zhang, Y., Liu, Z., Yang, Z., Zhou, Z., Qiao, Z., Feng, Z., Liu, Z., Zhang, Z., Wang, Z., Yao, Z., Wang, Z., Liu, Z., Chai, Z., Li, Z., Zhao, Z., Chen, W., Zhai, J., Xu, B., Huang, M., Wang, H., Li, J., Dong, Y., Tang, J.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models (2025), <https://arxiv.org/abs/2508.06471>
38. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, Z., Fergus.e.g, R.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms (2024), <https://arxiv.org/abs/2406.16860>
39. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmesen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders

- with improved semantic understanding, localization, and dense features (2025), <https://arxiv.org/abs/2502.14786>
40. Wang, D., Song, W., Wang, Y., Wang, S., Yu, K., Wei, Z., Wang, J.: Autoregressive semantic visual reconstruction helps vlms understand better. arXiv preprint arXiv:2506.09040 (2025)
 41. Wang, D., Song, W., Wang, Y., Wang, S., Yu, K., Wei, Z., Wang, J.: Autoregressive semantic visual reconstruction helps vlms understand better (2026), <https://arxiv.org/abs/2506.09040>
 42. Wang, H., Zheng, A., Zhao, Y., Wang, T., Ge, Z., Zhang, X., Zhang, Z.: Reconstructive visual instruction tuning. arXiv preprint arXiv:2410.09575 (2024)
 43. Wang, S., Cao, X., Zhang, J., Yuan, Z., Shan, S., Chen, X., Gao, W.: Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model (2026), <https://arxiv.org/abs/2406.14194>
 44. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), <https://arxiv.org/abs/2508.18265>
 45. Wang, W., Ren, Y., Luo, H., Li, T., Yan, C., Chen, Z., Wang, W., Li, Q., Lu, L., Zhu, X., et al.: The all-seeing project v2: Towards general relation comprehension of the open world. arXiv preprint arXiv:2402.19474 (2024)
 46. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., Chevalier, A., Arora, S., Chen, D.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms (2024), <https://arxiv.org/abs/2406.18521>
 47. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms (2023), <https://arxiv.org/abs/2312.14135>
 48. Xiao, Y., Sun, E., Liu, T., Wang, W.: Logicvista: Multimodal llm logical reasoning benchmark in visual contexts (2024), <https://arxiv.org/abs/2407.04973>
 49. Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T.J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., Yu, T.: Oworld: Benchmarking multimodal agents for open-ended tasks in real computer environments (2024), <https://arxiv.org/abs/2404.07972>
 50. Xu, W., Wang, J., Wang, W., Chen, Z., Zhou, W., Yang, A., Lu, L., Li, H., Wang, X., Zhu, X., Wang, W., Dai, J., Zhu, J.: Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models (2025), <https://arxiv.org/abs/2504.15279>
 51. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
 52. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions (2016), <https://arxiv.org/abs/1608.00272>
 53. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), <https://arxiv.org/abs/2303.15343>

- 54. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., Li, H.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? (2024), <https://arxiv.org/abs/2403.14624>
- 55. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R., Tan, T.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? (2025), <https://arxiv.org/abs/2408.13257>