arXiv:2508.12466v1 [cs.CV] 17 Aug 2025

Inverse-LLaVA: Eliminating Alignment Pre-training Through Text-to-Vision Mapping

Xuhui Zhan
Data Science Institute
Vanderbilt University
xuhui.zhan@vanderbilt.edu

Tyler Derr
Computer Science Department
Vanderbilt University
tyler.derr@vanderbilt.edu

Abstract

Traditional multimodal learning approaches require expensive alignment pre-training to bridge vision and language modalities, typically projecting visual features into discrete text token spaces. We challenge both fundamental assumptions underlying this paradigm by proposing **Inverse-LLaVA**, a novel approach that eliminates alignment pre-training entirely while inverting the conventional mapping direction. Rather than projecting visual features to text space, our method maps text embeddings into continuous visual representation space and performs fusion within transformer intermediate layers. Through selective additive components in attention mechanisms, we enable dynamic integration of visual and textual representations without requiring massive image-text alignment datasets. Comprehensive experiments across nine multimodal benchmarks demonstrate nuanced performance trade-offs: Inverse-LLaVA achieves notable improvements on reasoning-intensive and cognitive tasks (MM-VET: +0.2%, VizWiz: +1.8%, ScienceQA: +0.2%, cognitive reasoning: +27.2%), while showing expected decreases in perception tasks requiring memorized visual-text associations (celebrity recognition: -49.5%, OCR: -21.3%). These results provide the first empirical evidence that alignment pre-training is not necessary for effective multimodal learning, particularly for complex reasoning tasks. Our work establishes the feasibility of a new paradigm that reduces computational requirements by 45%, challenges conventional wisdom about modality fusion, and opens new research directions for efficient multimodal architectures that preserve modality-specific characteristics. Our project website with code and additional resources is available at https://inverse-llava.github.io.

1 Introduction

The rapid evolution of large language models has fundamentally reshaped our understanding of artificial intelligence. Through the seemingly simple paradigm of next-token prediction, models like GPT-3 [4], PaLM [8], and GPT-4 [30] have demonstrated that scaling neural networks can yield remarkable capabilities in reasoning, comprehension, and generation. This success has naturally prompted researchers to extend these powerful language backbones into the visual domain, leading to the emergence of *large vision-language models* (LVLMs) that promise to bridge the gap between textual understanding and visual perception.

The path toward multimodal intelligence began with foundational works that established the core architectural principles still dominant today. Flamingo [2] pioneered the integration of vision encoders with frozen language models through sophisticated cross-attention mechanisms, while BLIP-2 [22] introduced the influential Q-Former architecture for bridging visual and textual modalities. Building upon these early innovations, more recent models such as LLaVA [25, 23], Qwen2-VL [39], and SEED-1.5-VL [15] have demonstrated that sophisticated multimodal reasoning can emerge from relatively straightforward architectural designs. Across all these approaches, a consistent paradigm

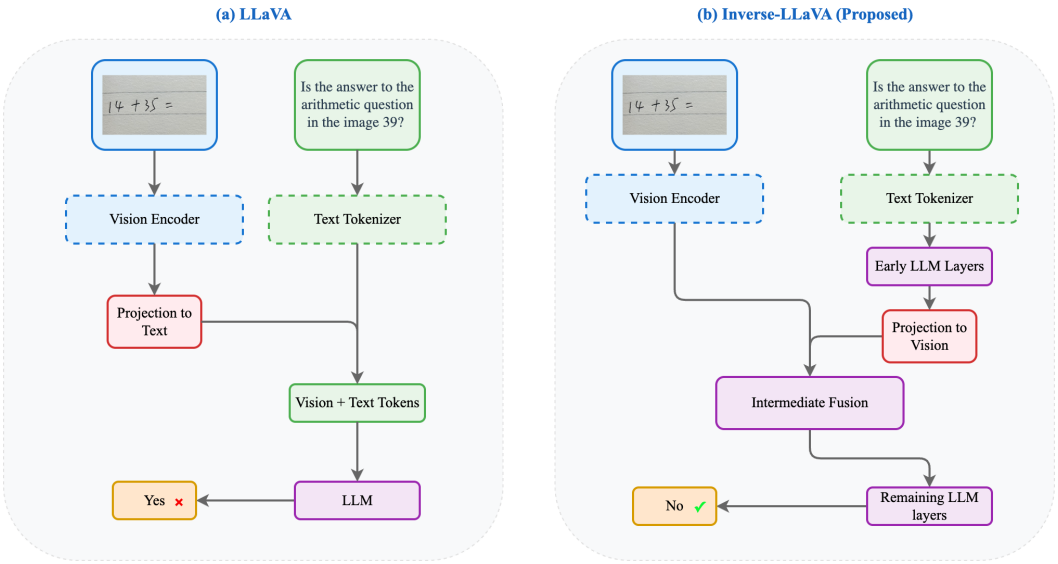


Figure 1: **High-level overview of the proposed Inverse-LLaVA framework on an MME benchmark example [13].** (a) LLaVA projects visual features into discrete text space via an explicit projection, requiring alignment pre-training, and produces the wrong answer ("Yes"). (b) Inverse-LLaVA maps text embeddings into continuous vision space for fusion, eliminating alignment pre-training, and yields the correct answer ("No"). Blue indicates vision flow, green indicates text flow, red denotes explicit projection, purple represents LLM components and orange indicates LLM output.

has emerged: visual information is processed by specialized encoders and then mapped into the discrete token space that language models can process.

However, this conventional wisdom rests on a fundamental assumption that may be unnecessarily constraining. The standard approach requires a computationally expensive two-stage training process where continuous visual features must first undergo *alignment training* to conform to the discrete distributional structure of text embeddings. During this alignment phase, the rich, continuous nature of visual representations is forced to match the inherently discrete characteristics of linguistic tokens, which originate from finite vocabulary lookups. This distributional conformity potentially limits the preservation of fine-grained spatial and photometric details that could be crucial for visual understanding. Moreover, this alignment stage typically demands hundreds of millions of paired image-text examples, dominating the computational budget and creating significant barriers for resource-constrained research environments.

What if this entire paradigm could be inverted? Rather than constraining continuous visual features to conform to discrete text distributions, we propose mapping text embeddings into the richer *continuous visual space*. This architectural inversion preserves the continuous nature of visual information while maintaining compatibility with existing language model architectures. Our approach, which we call **Inverse-LLaVA**, as shown in Figure 1, operates within the attention mechanism itself through a LoRA-inspired projection strategy. Text embeddings are mapped to match the dimensionality of visual features, concatenated along the feature dimension, and projected back to the original hidden state dimensionality. This design eliminates the need for expensive alignment pre-training entirely while preserving the expressive power of continuous visual representations.

In summary, our contributions are threefold:

- We propose an inverse mapping approach that projects text embeddings into continuous visual space rather than constraining visual features to discrete token distributions, preserving visual information continuity within standard transformer architectures.
- We demonstrate that this inverse mapping eliminates the alignment pre-training stage while reducing training computational requirements by 45%, using only instruction tuning data without the millions of image-text pairs required for alignment.

- We provide empirical evidence across nine benchmarks showing that text-to-visual projection achieves strong performance on reasoning-intensive tasks (MM-VET: +0.2%, ScienceQA: +0.2%, cognitive tasks: +27.2% improvement) while acknowledging trade-offs in perception tasks requiring memorized visual-text associations (overall perception: -12.4%), validating our theoretical framework about when inverse mapping excels.

2 Preliminaries

2.1 Multimodal Training Paradigms and Alignment Bottlenecks

Current large vision-language models universally follow a two-stage training paradigm that has become the de facto standard. Building on foundational work in contrastive vision-language learning [31], modern LVLMs adopt generative approaches that integrate visual understanding with language generation capabilities. Flamingo [2] pioneered cross-attention mechanisms for frozen language models, while BLIP-2 [22] introduced the Q-Former architecture for efficient visual feature extraction. The LLaVA family [25, 23] streamlined this with simple linear projections, achieving remarkable performance with architectural simplicity.

The first stage involves alignment pre-training, where visual features are mapped into a semantic space compatible with language models through massive image-text datasets [25, 23, 39, 15]. This process typically requires hundreds of millions of examples and dominates computational budgets. The second stage applies instruction tuning to teach multimodal task performance. This alignment stage has proven crucial for the performance of current architectures, as demonstrated by significant improvements achieved through proper vision-language alignment techniques [1].

However, while alignment is essential for existing paradigms, it creates both computational and representational bottlenecks that limit scalability and accessibility. Beyond the enormous resource requirements, the alignment process fundamentally forces continuous visual features to conform to the discrete distributional structure of text embeddings, which originate from finite vocabulary lookups. Research has shown that significant gaps can emerge between image and text embeddings during training, particularly when using standard next-token prediction objectives [1], illustrating the fundamental challenges in current alignment approaches. This distributional constraint potentially limits the preservation of fine-grained visual information that exists naturally in continuous visual manifolds.

Various efficiency improvements have been explored, including parameter-efficient training methods and architectural innovations. Vision-as-LoRA (VoRA) [38] represents a notable departure by eliminating external vision modules entirely, incorporating vision capabilities directly into language models through vision-specific low-rank adaptation layers. However, these approaches typically still require expensive alignment stages or introduce other training complexities, leaving the fundamental computational bottleneck unaddressed.

2.2 Continuous Representations in Multimodal Models

The treatment of continuous versus discrete representations represents a fundamental design choice in multimodal architectures. While language models operate on discrete tokens, their internal representations are inherently continuous. Theoretical work has demonstrated that Transformer-based language models implicitly learn to represent sentences as continuous-time functions defined over continuous input spaces [28], suggesting that these models naturally operate in continuous embedding spaces despite discrete input tokenization.

This perspective challenges the standard practice of constraining visual features to conform to discrete text distributions. Visual features naturally exist in continuous manifolds that can preserve nuanced spatial and semantic information when left unconstrained by distributional alignment requirements. Alternative approaches have explored deeper integration strategies, such as incorporating visual expert modules within transformer layers [40] or enabling modality collaboration through shared functional modules [41], suggesting that multimodal processing may be more natural when performed within the continuous representation spaces of the model's internal layers rather than only at input boundaries.

Our approach builds on these insights by inverting the traditional mapping direction within the continuous representation spaces of intermediate layers. Rather than constraining visual representations to match discrete text distributions at the input level, we project text embeddings into continuous visual spaces and perform fusion within the model's hidden states, where both modalities naturally exist as continuous representations. This design preserves the inherent continuity of visual information while leveraging the continuous nature of transformer internal processing, eliminating both the computational expense of alignment training and the representational constraints imposed by distributional alignment requirements.

2.3 Problem Formulation

Consider a multimodal input consisting of an image $\mathcal{I}$ and text sequence $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$. Traditional approaches first extract visual features $\mathbf{V} = f_{\text{vis}}(\mathcal{I}) \in \mathbb{R}^{d_v \times p}$ where p is the number of visual patches, then project them into text embedding space:

$$\mathbf{V}_{\text{proj}} = \mathbf{W}_{\text{v2t}}\mathbf{V} + \mathbf{b}_{\text{v2t}} \tag{1}$$

where $\mathbf{W}_{\text{v2t}} \in \mathbb{R}^{d_h \times d_v}$ maps visual features to the language model's hidden dimension d_h.

This conventional approach requires expensive alignment pre-training to learn $\mathbf{W}_{\text{v2t}}$ on massive image-text pairs, forcing continuous visual representations to conform to discrete text distributions.

3 Methodology

We propose **Inverse-LLaVA**, a novel approach that eliminates alignment pre-training by inverting the traditional modality mapping direction. Rather than projecting visual features into discrete text token space, our method maps text embeddings into continuous visual representation space and performs fusion within the transformer's intermediate layers (See Figure 2 for the detailed architecture).

The motivation for this inversion stems from a fundamental asymmetry between modalities. Visual features extracted from continuous pixel values naturally form dense, high-dimensional representations where spatial relationships and fine-grained details are encoded through continuous variations. In contrast, text embeddings originate from discrete vocabulary lookups, representing a finite set of semantic concepts. When we force continuous visual signals into discrete text space—as done in conventional approaches—we inevitably lose information through quantization. This loss is particularly problematic for tasks requiring precise spatial reasoning, numerical understanding, or fine-grained visual details. By instead expanding discrete text into the richer continuous visual space, we preserve the full expressiveness of visual information while text embeddings can naturally expand to fill the available representational capacity, potentially enabling more nuanced cross-modal interactions.

3.1 Inverse Mapping Strategy

Our key insight is to preserve visual continuity by inverting this mapping direction. We maintain the visual features $\mathbf{V}$ in their native continuous space and instead project text embeddings to match the visual dimensionality:

$$\mathbf{T}_{\text{proj}} = \mathbf{W}_{\text{t2v}}\mathbf{T} + \mathbf{b}_{\text{t2v}} \tag{2}$$

where $\mathbf{T} \in \mathbb{R}^{d_h \times n}$ are the text embeddings and $\mathbf{W}_{\text{t2v}} \in \mathbb{R}^{d_v \times d_h}$ projects text into visual space.

3.2 Vision-Text Fusion Mechanism

We implement this inverse mapping through a novel vision-text fusion mechanism that, while inspired by LoRA's parameter-efficient philosophy, differs significantly in its architectural design and fusion strategy. Unlike standard LoRA that applies low-rank decomposition to existing weight matrices, our approach introduces selective additive fusion components that enable dynamic integration of visual and textual representations.

For selected attention layers $l \in \mathcal{S} \subseteq \{1, 2, \ldots, L\}$, we augment the standard query, key, and value projections with vision-fused components:

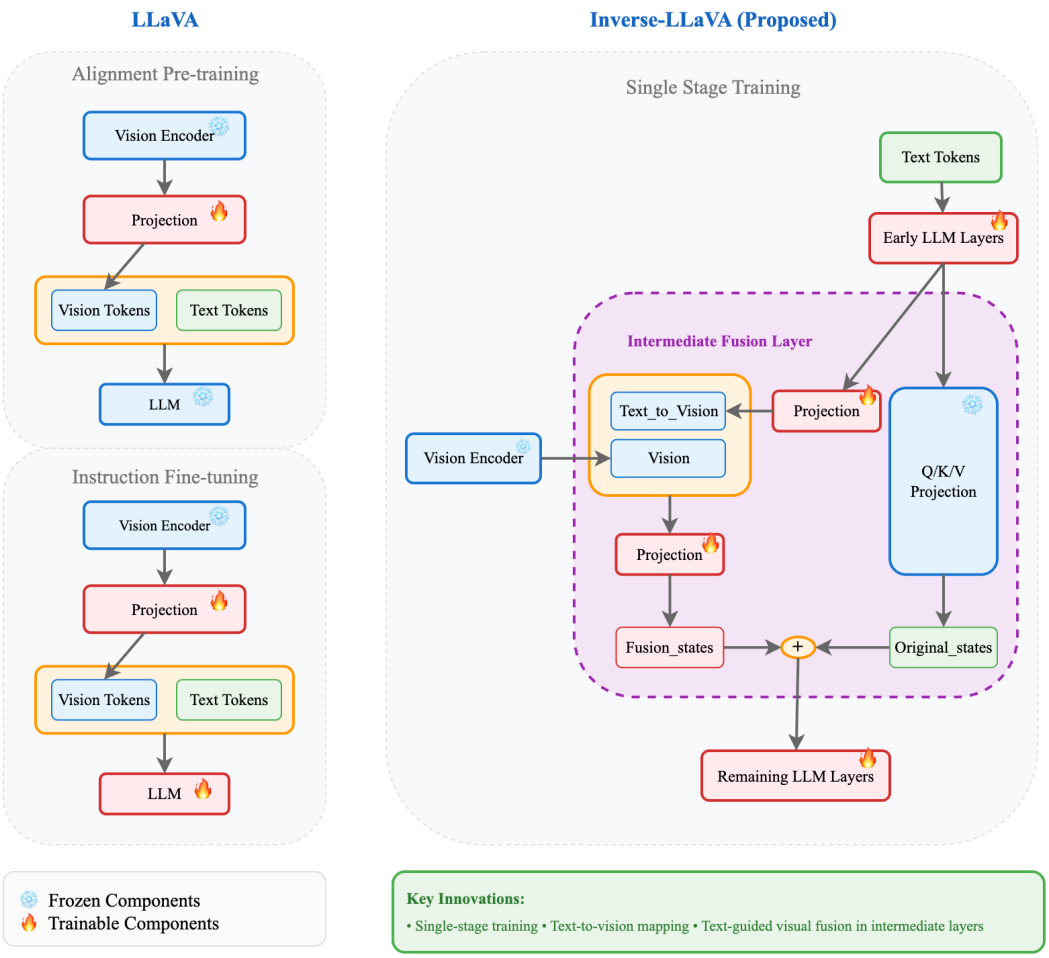


Figure 2: **Architecture comparison between LLaVA and Inverse-LLaVA.** LLaVA employs a two-stage training approach with alignment pretraining followed by instruction fine-tuning, where vision and text tokens are concatenated before being fed to the LLM. In contrast, Inverse-LLaVA uses single-stage training with text-guided visual fusion in intermediate layers, where vision information is integrated through learnable text-to-vision projections and combined with original hidden states via residual connections.

$$\mathbf{Q}_l = \mathbf{W}_Q^{(l)}\mathbf{H}_l + \alpha_Q^{(l)}\mathbf{W}_{\text{concat}}^{Q,(l)}\text{concat}(\mathbf{W}_{\text{t2v}}^{(l)}\mathbf{H}_l, \mathbf{V}_{\text{emb}}) \tag{3}$$

$$\mathbf{K}_l = \mathbf{W}_K^{(l)}\mathbf{H}_l + \alpha_K^{(l)}\mathbf{W}_{\text{concat}}^{K,(l)}\text{concat}(\mathbf{W}_{\text{t2v}}^{(l)}\mathbf{H}_l, \mathbf{V}_{\text{emb}}) \tag{4}$$

$$\mathbf{V}_l = \mathbf{W}_V^{(l)}\mathbf{H}_l + \alpha_V^{(l)}\mathbf{W}_{\text{concat}}^{V,(l)}\text{concat}(\mathbf{W}_{\text{t2v}}^{(l)}\mathbf{H}_l, \mathbf{V}_{\text{emb}}) \tag{5}$$

where:

- $\mathbf{W}_Q^{(l)}, \mathbf{W}_K^{(l)}, \mathbf{W}_V^{(l)}$ are frozen pre-trained projection matrices
- $\alpha_Q^{(l)}, \alpha_K^{(l)}, \alpha_V^{(l)} \in \mathbb{R}$ are learnable scaling parameters
- $\mathbf{W}_{\text{t2v}}^{(l)} \in \mathbb{R}^{d_v \times d_h}$ projects hidden states to visual dimensionality
- $\mathbf{V}_{\text{emb}} \in \mathbb{R}^{d_v \times p}$ are the visual embeddings from the vision encoder
- $\mathbf{W}_{\text{concat}}^{\{Q,K,V\},(l)} \in \mathbb{R}^{d_h \times 2d_v}$ are learnable concatenation projection matrices
- $\mathcal{S}$ is a configurable subset of layers where fusion is applied (can be one or more layers)

Algorithm 1 $\mathbf{Z} \leftarrow$ INVERSE-LLAVA-FORWARD$(\mathcal{I}, \mathcal{T} \mid \mathcal{W}, \mathcal{S})$

Inputs: images $\mathcal{I}$; text sequence $\mathcal{T}$, length n.
Output: logits $\mathbf{Z} \in \mathbb{R}^{|\mathcal{V}| \times n}$ per-text-position vocabulary logits (probabilities via $\mathrm{softmax}(\mathbf{Z})$).
Hyperparams: L = number of LLM layers; $\mathcal{S} \subseteq \{1, \ldots, L\}$ = fusion layer indices.
Components (frozen unless noted):
f_{vis} — vision encoder extracting visual token sequence.
Embedding layer + LLM (self-attention + FFNs).
$\mathbf{W}_{t2v}^{(l)}, \mathbf{W}_{\text{concat}}^{Q,K,V,(l)}, \alpha_{Q,K,V}^{(l)}$ — fusion params at layer $l \in \mathcal{S}$ (trainable).
$\mathbf{W}_{\text{lm}}, \mathbf{b}_{\text{lm}}$ — LM head mapping hidden states to vocabulary logits.
Complementary slotting: choose disjoint $\mathcal{P}_{\text{text}}, \mathcal{P}_{\text{vis}} \subset \{1, \ldots, n\}$, covering all positions.

$\mathbf{V}_{\text{vis}} \leftarrow f_{\text{vis}}(\mathcal{I})$ ▷ vision token sequence
$\mathbf{H}_0 \leftarrow \mathrm{Embed}(\mathcal{T})$ ▷ text token embeddings
Build complementary padded streams:
$\mathbf{T}_{\text{pad}}$: keep text embeddings at $\mathcal{P}_{\text{text}}$, zeros elsewhere.
$\mathbf{V}_{\text{pad}}$: place vision tokens at $\mathcal{P}_{\text{vis}}$, zeros at text positions.
for $l = 1$ **to** L **do**
  if $l \in \mathcal{S}$ **then** ▷ fusion layer
    $\mathbf{T}_{\text{proj}} \leftarrow \mathbf{W}_{t2v}^{(l)} \mathbf{H}_{l-1}$ ▷ text → vision projection
    $\mathbf{F}_{\text{concat}} \leftarrow \mathrm{concat}\big(\mathbf{T}_{\text{proj}}^{\text{pad}}, \mathbf{V}_{\text{pad}}\big)$
    Compute $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l$ from Eqs. (3)–(5)
    $\mathbf{H}_l \leftarrow \mathrm{Attention}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l)$
  else
    $\mathbf{H}_l \leftarrow \mathrm{Attention}_l(\mathbf{H}_{l-1})$
  end if
  $\mathbf{H}_l \leftarrow \mathrm{FFN}_l(\mathbf{H}_l) + \mathbf{H}_{l-1}$ ▷ residual
end for
$\mathbf{Z} \leftarrow \mathbf{W}_{\text{lm}} \mathbf{H}_L + \mathbf{b}_{\text{lm}}$ ▷ per-position logits
return $\mathbf{Z}$

The key distinction from LoRA lies in our fusion strategy: instead of decomposing existing weights, we introduce an additive pathway that processes concatenated visual-textual features. The text-to-vision projection $\mathbf{W}_{\text{t2v}}^{(l)}$ maps from the language model's hidden dimension d_h to the visual dimension d_v, where typically $d_v = \frac{d_h}{2}$ to $\frac{d_h}{4}$, creating a more compact yet expressive visual-aligned representation space.

3.3 Training Objective

Our approach eliminates the alignment stage entirely. We train end-to-end using standard autoregressive language modeling loss:

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{T}_{\text{target}}|} \log P(t_i | \mathcal{I}, t_{<i}; \theta) \tag{6}$$

where $\mathcal{T}_{\text{target}}$ is the target text sequence and θ represents all trainable parameters including the LoRA adaptation matrices.

3.4 Advantages of the Inverse Approach

This design offers several key advantages: (1) **Preserves visual continuity**: Visual features remain in their native continuous manifold without discretization constraints. (2) **Eliminates alignment cost**: No separate pre-training stage is required, reducing computational overhead by orders of magnitude. (3) **Parameter efficiency**: Only trains lightweight projection matrices and scaling parameters rather than full model weights. (4) **Maintains compatibility**: The approach works with any pre-trained vision encoder and language model combination without architectural modifications. (5) **Adaptive fusion**: The learnable scaling parameters α allow the model to dynamically balance between original language model representations and vision-fused features.

4 Experiments

We conduct controlled experiments to evaluate the effectiveness of our inverse mapping paradigm, comparing Inverse-LLaVA with baseline LLaVA-1.5[23] across nine multimodal benchmarks. Our experimental design isolates the architectural effect by maintaining identical training data, backbone models, and optimization settings. We analyze performance patterns across diverse vision-language tasks to understand how inverse mapping affects reasoning capabilities, visual understanding, and the fundamental need for alignment pre-training. Through detailed performance analysis, particularly on the MME[13] benchmark, we reveal significant performance gains in cognitive tasks while identifying specific trade-offs in visual perception domains.

4.1 Experimental Setup

We implement Inverse-LLaVA using Vicuna-7B-v1.5[43] as the language model backbone and CLIP-ViT-Large-Patch14-336[31] as the vision encoder, maintaining identical foundation components to LLaVA-1.5 for controlled comparison. All experiments are conducted on 8 NVIDIA A100 GPUs, matching the computational setup used for LLaVA-1.5 training.

Training Configuration. To ensure rigorous comparison, we adopt identical training protocols to LLaVA-1.5: the LLaVA-v1.5-mix665k instruction tuning dataset containing 665,000 image-text pairs, single-epoch training with cosine learning rate scheduling (learning rate 2×10^{-4}, weight decay 0.0, warmup ratio 0.03), and batch size of 32 samples per device. Both baseline and proposed methods employ LoRA fine-tuning [17] with rank $r = 128$ and scaling factor $\alpha = 256$. Critically, our approach applies LoRA adaptation only to transformer layers without visual information injection—specifically, all layers except the first layer where vision-text fusion occurs.

Architectural Distinctions. The fundamental difference between our approach and LLaVA-1.5 lies in the modality fusion strategy. LLaVA-1.5 employs a learnable MLP projector that maps visual features $\mathbf{V} \in \mathbb{R}^{d_v \times p}$ to language model dimensions d_h, requiring expensive alignment pre-training on massive image-text pairs. In contrast, Inverse-LLaVA eliminates vision-to-text projection entirely and instead projects text embeddings to visual space via learnable matrices $\mathbf{W}_{\text{t2v}}^{(l)} \in \mathbb{R}^{d_v \times d_h}$, performing fusion within the first transformer layer through our proposed concatenation mechanism.

Vision Feature Variants. We investigate two visual representation strategies: Inverse-LLaVA utilizes final CLIP hidden states ($d_v = 1024$), while Inverse-LLaVA-HD concatenates final and penultimate CLIP hidden states ($d_v = 2048$) for richer visual representation. Note that LLaVA-1.5 conventionally uses penultimate CLIP features; our investigation explores optimal feature selection for inverse mapping compatibility.

Evaluation Protocol. We assess performance across nine established multimodal benchmarks: MM-VET [42], MMBENCH and MMBENCH-CN [26], MME [13], VizWiz [16], ScienceQA [27], VQAV2 [14], TextVQA [35], and GQA [18], following standard evaluation protocols for each dataset.

4.2 Paradigm Validation: Inverse Mapping Without Alignment

We investigate two fundamental questions about multimodal learning:

1. Does the conventional vision-to-text mapping constrain the rich continuous nature of visual representations, limiting multimodal understanding?
2. Can inverse text-to-vision mapping preserve visual information richness while eliminating the need for expensive alignment pre-training?

Visual Information Preservation Through Inverse Mapping. Table 1 demonstrates that our inverse mapping approach achieves competitive performance across nine vision-language benchmarks while fundamentally rethinking how modalities interact. Unlike conventional methods that compress visual features to match text distributions, we project text embeddings into the richer visual space, preserving spatial relationships and fine-grained visual details.

The results reveal selective advantages: Inverse-LLaVA outperforms LLaVA-1.5 on MM-VET (31.2 vs 31.1), VizWiz (50.95 vs 50.0), and notably on ScienceQA-IMG (67.84 vs 66.80). However, we observe consistent degradation on tasks requiring precise visual-text alignment, including TextVQA (52.02 vs 58.2) and GQA (58.46 vs 62.0). This performance pattern suggests our approach excels at reasoning tasks while struggling with direct visual-linguistic correspondence.

Table 1: **Performance comparison of vision-language models using Vicuna-7B[43] as the backbone LLM across multiple vision-language benchmarks.** We evaluate models on MM-VET [42]; VizWiz [16]; SQA^{I}: ScienceQA-IMG[27]; MMB: MMBench[26]; MMB^{CN}: MMBench-CN [26]; MME^{P}: MME-Perception[13]; VQA^{V2}: VQA-v2 [14]; VQA^{T}: TextVQA [35]; GQA [18]. **Bold** indicates the best performance and underlined indicates the second-best performance in each benchmark.

Model	MM-VET	VizWiz	SQA^{I}	MMB	MMB^{CN}	MME^{P}	VQA^{V2}	VQA^{T}	GQA
LLaVA-1.5[23]	31.1	50.0	66.80	**64.3**	**58.3**	1510.7	78.5	**58.2**	62.0
InstructBLIP[9]	26.2	34.5	60.5	36.0	23.7	-	-	50.1	49.2
InternVL-Chat[6]	-	**52.5**	-	-	-	**1521.1**	**79.3**	57.0	**62.9**
EVE-7B[10]	25.6	41.8	63.0	49.5	-	1217.3	75.4	51.9	60.8
Inverse-LLaVA	**31.2**	50.95	**67.84**	54.55	41.84	1293.15	74.76	52.02	58.46
Inverse-LLaVA-HD	-	-	-	-	-	1335.67	-	-	59.33

"-" denotes unreported results or experiments not conducted due to computational constraints. LLaVA-1.5 results from [23]; InstructBLIP, InternVL-Chat, and EVE-7B results from [10]. Inverse-LLaVA-HD evaluated on limited benchmarks due to computational constraints.

Table 2: **Training data comparison.** #Samples indicates the number of samples used in alignment pre-training and instruction fine-tuning stages respectively.

Model	Alignment #Samples	Finetune #Samples
LLaVA-1.5[23]	558K	665K
InstructBLIP[9]	129M	1.2M
InternVL-Chat[6]	4.98B	665K
EVE-7B[10]	33M	665K
Inverse-LLaVA	-	665K

Alignment-Free Learning Validation. Table 2 starkly illustrates our paradigm shift: while LLaVA-1.5 requires 558K alignment samples, InstructBLIP uses 129M, and InternVL-Chat employs a massive 4.98B samples, Inverse-LLaVA achieves competitive performance with *zero* pre-training samples. This 100% reduction in alignment requirements validates our hypothesis that architectural design can substitute for data-intensive alignment procedures.

The performance patterns across benchmarks support this validation. Despite the absence of alignment training, our approach shows advantages in reasoning-heavy tasks (ScienceQA-IMG: 67.84 vs 66.80) while exhibiting systematic degradation on recognition and grounding tasks (MMB: 54.55 vs 64.3, TextVQA: 52.02 vs 58.2, GQA: 58.46 vs 62.0). This dichotomy validates our hypothesis that inverse mapping favors abstract reasoning over direct visual-text correspondence.

4.3 Performance Pattern Analysis: Visual Richness vs. Alignment Constraints

The MME benchmark breakdown (Figure 3) provides granular insights into how our architectural choices manifest across different task types. Comparing Inverse-LLaVA models against LLaVA-1.5-7B-LoRA (all using LoRA for fair comparison), we observe distinct performance signatures.

Emergent Advantages in Cognitive Tasks. Most strikingly, Inverse-LLaVA achieves a **27.2% higher cognition score** (328.21 vs 257.86), with dramatic improvements in numerical calculation (+69.2%) and text translation (+125%). These gains were unexpected—our design prioritized visual representation preservation, not cognitive enhancement. This emergence suggests that maintaining continuous visual features enables richer cross-modal reasoning than anticipated.

The cognitive advantages persist in code reasoning (55.0 vs 52.5) and remain competitive in commonsense reasoning (105.71 vs 122.86). The HD model shows further refinement, improving commonsense reasoning to 122.14 while maintaining strong text translation performance (97.5).

Recognition Task Trade-offs. As predicted by our theoretical framework, perception tasks show clear trade-offs (1293.15 vs 1477.05). The gaps are most pronounced in tasks requiring learned visual-text associations: celebrity recognition (-49.5%), OCR (-21.3%), and artwork identification (-20.0%). These tasks benefit from the discrete pattern matching that alignment training optimizes for, where specific visual patterns map to memorized text tokens.

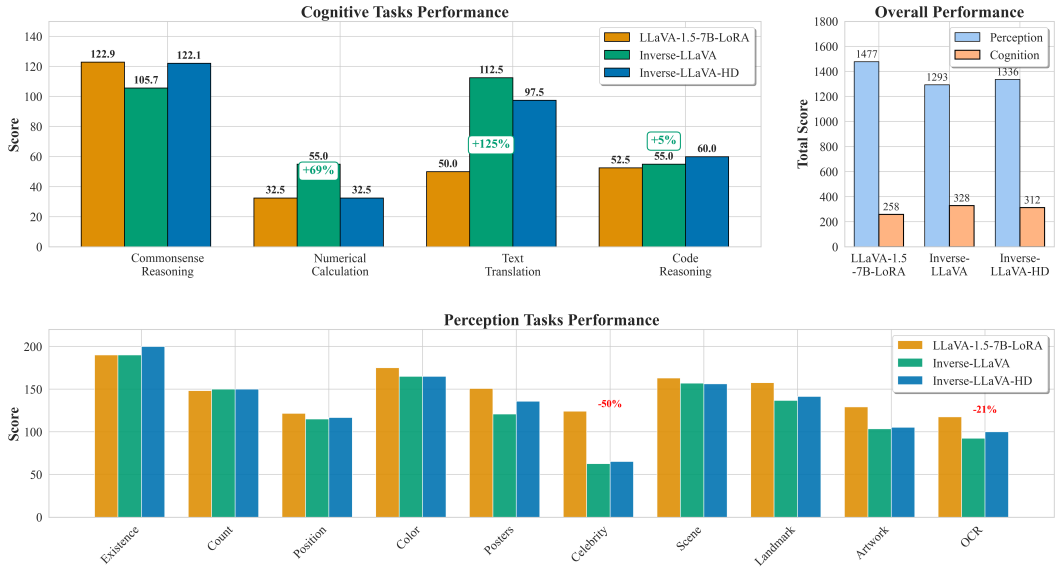


Figure 3: MME Benchmark Analysis comparing LLaVA-1.5-7B-LoRA, Inverse-LLaVA, and Inverse-LLaVA-HD across cognitive and perception tasks in the MME benchmark[13]. **Left (top):** Cognitive tasks performance showing Inverse-LLaVA achieving superior performance in numerical calculation (+69%) and text translation (+125%) compared to the baseline LLaVA-1.5-7B-LoRA model. **Right (top):** Overall performance comparison. **Bottom:** Perception tasks evaluation shows that Inverse-LLaVA variants excel in Existence and Count tasks, with Inverse-LLaVA-HD achieving perfect performance on Existence tasks. However, significant performance drops in Celebrity recognition (-50%) and OCR tasks (-21%) primarily account for the overall perception score gap. The results indicate that inverse training maintains strong cognitive capabilities while showing task-specific effects on perception.

However, the progression from Inverse-LLaVA to Inverse-LLaVA-HD (1293.15 $\rightarrow$ 1335.67) demonstrates that enhanced visual encoders can partially mitigate these gaps. HD achieves perfect existence detection (200.0) and improves across multiple perception metrics, suggesting that richer visual features can compensate for the lack of explicit alignment when properly utilized.

4.4 Architecture Analysis: Continuous Representations and Training Dynamics

Representation Space Asymmetry. Our inverse mapping exploits a fundamental asymmetry between modalities: while text exists as discrete tokens from a finite vocabulary, visual features form a continuous manifold in high-dimensional space. By projecting discrete text embeddings into the continuous visual space rather than the reverse, we avoid the quantization error inherent in discretizing continuous visual information.

Formally, let $\mathcal{V} \subset \mathbb{R}^{d_v}$ denote the visual feature space and $\mathcal{T} \subset \mathbb{R}^{d_t}$ the text embedding space. Traditional approaches learn $f : \mathcal{V} \rightarrow \mathcal{T}$, requiring continuous features to approximate discrete text distributions. Our inverse mapping $g : \mathcal{T} \rightarrow \mathcal{V}$ preserves the full expressiveness of $\mathcal{V}$ while maintaining injectivity for text features—ensuring no information loss from the text modality.

This design choice has measurable consequences: the preserved visual feature dimensionality enables the model to maintain fine-grained spatial information (evidenced by strong ScienceQA performance) and supports unexpected cognitive capabilities (demonstrated in Figure 3).

Training Dynamics and Convergence Properties. The elimination of alignment pre-training fundamentally alters the optimization landscape. Traditional approaches require two-stage training: first learning cross-modal correspondences through contrastive or generative objectives, then fine-tuning for downstream tasks. This creates a sequential dependency where suboptimal alignment limits downstream performance.

Our single-stage training directly optimizes for task performance without the alignment bottleneck. The training efficiency—achieving competitive results with 45% fewer total samples than LLaVA-1.5—suggests that the inverse mapping creates a more favorable optimization landscape. The consistent convergence across diverse benchmarks (Table 1) indicates that the expanded text representations naturally align with visual features during task-specific training.

Preserved Language Model Priors. A critical design choice in our approach is preserving pre-trained language model representations. By adapting text through learned projections rather than modifying the core LLM weights during alignment, we maintain the linguistic priors acquired during large-scale language pre-training. However, this preservation shows mixed effects: while certain language-reasoning tasks benefit (Text Translation: 112.5), others requiring tight visual-text coupling suffer (TextVQA: 52.02 vs 58.2).

This divergence reveals a nuanced trade-off: our projection mechanism excels when language understanding can leverage expanded representations for reasoning (translation, calculation), but struggles when tasks demand precise visual-to-text grounding (OCR, TextVQA). The expansion of text embeddings into visual space, while preserving linguistic structure, may dilute the sharp visual-textual associations needed for reading text in images.

4.5 Implications for Multimodal Learning

Our results challenge the conventional wisdom that alignment pre-training is essential for vision-language models. The evidence suggests that:

1. **Architectural innovation can substitute for data scale**: Proper design choices eliminate the need for millions of alignment samples while maintaining competitive performance.
2. **Visual richness preservation enables emergent capabilities**: Maintaining continuous visual representations unexpectedly enhances cognitive reasoning tasks, suggesting deeper connections between representation quality and reasoning ability.
3. **Trade-offs reflect fundamental design choices**: The performance patterns validate our theoretical framework—tasks requiring memorized associations favor alignment-based approaches, while complex reasoning benefits from preserved visual richness.

These findings have broader implications for multimodal foundation model design, suggesting that future architectures should prioritize preserving each modality's natural characteristics rather than forcing convergence through alignment.

5 Limitations and Future Directions

Our experiments reveal both fundamental constraints and untapped potential of the inverse mapping paradigm, offering insights for multimodal learning beyond vision-language tasks. Central to understanding these findings is what we term the **representational bias hypothesis**, which explains the systematic performance patterns observed across different task categories.

The Representational Bias Hypothesis. We propose that the performance dichotomy between inverse mapping and alignment-based approaches reflects a fundamental difference in multimodal representation learning [31, 22]. Current VLMs, through their alignment and pre-training stages, effectively quantize continuous visual features into discrete tokens that are by design processed as textual puzzles—alignment forces visual understanding to be encoded in the same linguistic dimension as text [22, 24]. This process, while computationally efficient, transforms genuine visual understanding into sophisticated puzzle-solving mechanisms where the backbone LLM learns to decode visual information that has been artificially constrained to textual representations, rather than processing the underlying continuous visual dynamics in their native dimensionality [2, 8].

This hypothesis explains several key observations in our MME benchmark results [13]. Current VLMs demonstrate diminishing returns or even performance degradation when exposed to increased training data—not because they lack capacity, but because they are fundamentally solving textual puzzles embedded with visual patterns rather than understanding dynamic visual signals in their native form [43, 42]. The token-based representation, while capturing task-relevant patterns, inherently loses the rich continuous information necessary for genuine multimodal reasoning by forcing visual

understanding through the constraints of textual encoding [4, 30]. Consequently, alignment-based models excel at correspondence tasks like Celebrity recognition and OCR (where pattern matching within textual representations suffices) but struggle with complex reasoning requiring integration of fine-grained visual details processed in their continuous form, as evidenced by our substantial gains in Numerical Calculation (+69.2%) and Text Translation (+125%).

Performance Dichotomy Explained. Our results show a clear split: correspondence tasks needing precise visual-text matching (Celebrity: -49.8%, OCR: -21.3%) favor alignment-based, while reasoning tasks benefit from preserved visual richness (Numerical Calculation: +69.2%, Text Translation: +125%). This dichotomy stems from the fundamental difference between pattern recognition within constrained textual representations and genuine understanding across separate modality dimensions.

Critically, the nature of bias in multimodal representations manifests differently across task domains [26, 16]. Like color-blindness affecting specific visual processing while leaving other capabilities intact, our model exhibits selective limitations—notably in color-related tasks and OCR where subtle color distinctions may be crucial—while excelling in domains requiring rich continuous information. Our approach demonstrates a form of "color-blindness" in the MME benchmark, showing reduced performance in Color tasks and OCR (where close color similarities may confound text recognition), yet this selective limitation does not impair reasoning capabilities that benefit from preserved visual richness.

This selective bias contrasts fundamentally with alignment-based approaches, where the bias toward discrete textual representations is systemic and by design [22, 9, 2, 31]. Current VLMs like LLaVA are architecturally constrained to process visual information through textual tokens, creating overfitting to pattern-matching tasks that can be solved within these representational limits [25, 23]. This design-level bias cannot be easily overcome through additional training, as it is embedded in the fundamental processing paradigm. Conversely, our model's current limitations (such as color sensitivity) represent training-addressable challenges rather than architectural constraints, more targeted training could resolve these specific deficits while preserving the advantages of continuous processing.

Alignment-based methods create efficient shortcuts for matching discrete visual patterns that have been encoded as textual tokens, making them superior for tasks where such correspondences suffice. However, when tasks [27, 13] demand reasoning across complex visual relationships or extracting nuanced information that cannot be adequately captured through textual encoding, the preserved continuous signals processed in separate dimensions in our approach provide crucial advantages.

The high-dimensional (HD) version's mixed performance in cognitive tasks further validates this hypothesis. While some cognitive tasks show slight degradation, this reflects the expected learning curve of processing richer, more dynamic visual signals without pre-filtering through task-specific abstractions. Significantly, the HD version achieves perfect scores in Existence tasks, demonstrating that when given access to higher-dimensional continuous information, the model can achieve superior understanding for tasks requiring precise visual analysis. This suggests that as models scale and training strategies optimize, continuous approaches will increasingly outperform discrete methods across broader task categories [8, 30].

Architectural Sensitivity and Precision Requirements. The inverse mapping approach shows remarkable sensitivity to design choices, where simple modifications—replacing concatenation with addition, or introducing gating mechanisms—cause substantial performance degradation. This brittleness stems from operating directly in high-dimensional visual space, where preserving information flow requires precise architectural configurations. Additionally, our approach requires maintaining exact numerical precision from visual encoders throughout the pipeline, contrasting with conventional methods' tolerance for precision variations [19]. While this validates our information preservation claims, it complicates deployment in resource-constrained environments.

Technical Constraints. Two implementation challenges emerged. Multi-layer injection attempts (simultaneously at layers 1 and 3) resulted in significant performance degradation, producing unstable outputs with gradient instabilities. Without alignment pre-training to establish cross-modal coordination, the model cannot effectively reconcile multimodal signals across different processing depths—though hierarchical integration remains theoretically appealing, mirroring human sensory processing [12, 33]. This limitation reflects the model's need to develop new mechanisms for processing continuous multimodal signals rather than relying on pre-established discrete correspondences.

Scaling Potential and Future Promise. Despite current limitations, our approach demonstrates clear improvement trajectories, indicating substantial headroom remains. The continuous nature of our signals creates what we term a more "information-preserving representational framework"—one that maintains access to the full spectrum of visual information rather than pre-filtering it through task-specific statistical abstractions [37, 34]. This paradigm resembles providing visual signals to "wisdom that was once blind": the model may initially appear less efficient than specialized puzzle-solvers, requiring more computational resources to process uncompressed information, but possesses fundamentally greater potential for genuine multimodal understanding [3, 21].

Three factors support scaling optimism. First, stronger LLM backbones should yield proportional gains given our preservation of linguistic capabilities [8, 30]. Second, modern techniques like AnyRes [24] naturally complement our approach since we maintain rather than compress visual information and process modalities in their native dimensions [11, 39, 1, 38]. Third, our ablation revealing modest degradation when simulating alignment-free LLaVA suggests the performance gap may narrow with optimized training strategies, challenging assumptions about alignment necessity.

Toward Continuous Multimodal Integration. The temporary performance trade-offs observed in correspondence tasks represent not fundamental limitations but rather the expected learning curve of developing genuine understanding versus pattern matching [7, 29]. As foundation models continue scaling [32, 36, 22, 5], we hypothesize that continuous approaches will increasingly outperform discrete methods, particularly as tasks demand more sophisticated reasoning beyond simple visual-textual correspondence. This positions our approach as complementary rather than competitive to alignment methods in the near term, with hybrid architectures potentially applying each strategy selectively—alignment for correspondence, preservation for reasoning—while transitioning toward fully continuous integration as computational resources and training strategies advance.

Our findings delineate when inverse mapping excels: rich continuous modalities, reasoning-dominant tasks, and sufficient computational resources. Beyond vision-language tasks, this principle extends to other inherently continuous modalities—molecular structures, audio spectrograms, scientific measurements—that may gain more from preservation than forced discretization. Vision-Language-Action models [20]for robotics may particularly benefit from preserving continuous sensorimotor dynamics for physical reasoning.

As foundation models expand beyond vision and text, the principle of preserving modality-specific characteristics becomes increasingly vital. The inverse mapping paradigm thus represents not just an alternative technique, but a fundamental rethinking of how to respect and leverage each modality's intrinsic properties in the pursuit of truly multimodal intelligence—moving from sophisticated pattern matching to authentic multimodal understanding.

6 Conclusion

We presented Inverse-LLaVA, which inverts the conventional multimodal paradigm by projecting text embeddings into continuous visual space rather than constraining visual features to discrete textual representations. This architectural innovation processes text and vision in separate dimensions through feature-wise concatenation, eliminating the need for alignment pre-training entirely while achieving competitive performance across nine benchmarks with zero alignment samples.

Our empirical results reveal that preserving continuous visual signals in separate processing dimensions can break current bottlenecks in multimodal learning. The substantial improvements in reasoning tasks—Numerical Calculation (+69.2%) and Text Translation (+125%)—demonstrate that continuous processing enables enhanced understanding capabilities, while correspondence tasks show selective limitations that appear training-addressable rather than architectural constraints. The performance dichotomy suggests different approaches may be optimal for different task categories.

This work establishes a new research direction for multimodal architecture design, moving from pattern matching to authentic multimodal intelligence. As foundation models expand to incorporate diverse continuous signals, our results suggest that preserving each modality's intrinsic characteristics in separate processing dimensions offers a more principled path toward Artificial General Intelligence (AGI) than forcing convergence through textual bottlenecks.

References

[1] K. E. Ak, J. Mohta, D. Dimitriadis, S. Manchanda, Y. Xu, and M. Shen. Aligning vision language models with contrastive learning. In *Computer Vision – ECCV 2024 Workshops: Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVIII*, page 32–45, Berlin, Heidelberg, 2025. Springer-Verlag.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[5] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023.

[6] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[7] F. Chollet. On the measure of intelligence, 2019.

[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, et al. Palm: Scaling language modeling with pathways, 2022.

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[10] H. Diao, Y. Cui, X. Li, Y. Wang, H. Lu, and X. Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[12] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, Jan 1991. Funding by NIH.

[13] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[14] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] D. Guo, F. Wu, F. Zhu, F. Leng, G. Shi, H. Chen, H. Fan, J. Wang, et al. Seed1.5-vl technical report, 2025.

[16] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018.

[17] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[18] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model, 2024.

[21] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[22] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[23] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023.

[24] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[25] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[26] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[27] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[28] S. Marro, D. Evangelista, X. A. Huang, E. L. Malfa, M. Lombardi, and M. J. Wooldridge. Language models are implicitly continuous. In *The Thirteenth International Conference on Learning Representations*, 2025.

[29] M. Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, June 2021.

[30] OpenAI et al. Gpt-4 technical report, 2024.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[32] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A generalist agent, 2022.

[33] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature neuroscience*, 3 Suppl:1199–204, 12 2000.

[34] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *ArXiv*, abs/1703.00810, 2017.

[35] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[36] G. Team et al. Gemini: A family of highly capable multimodal models, 2025.

[37] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.

[38] H. Wang, Y. Ye, B. Li, Y. Nie, J. Lu, J. Tang, Y. Wang, and C. Huang. Vision as lora, 2025.

[39] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

[40] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[41] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

[42] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024.

[43] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.