LaViT: Aligning Latent Visual Thoughts for Multi-modal Reasoning

Linquan Wu[1], Tianxiang Jiang[2], Yifei Dong[3], Haoyu Yang[4], Fengji Zhang[1], Shichang Meng[1], Ai Xuan[1], Linqi Song[1], and Jacky Keung[1(✉)]

[1] City University of Hong Kong, Hong Kong SAR, China
linquan.wu@my.cityu.edu.hk
[2] University of Science and Technology of China, Hefei, China
[3] Utrecht University, Utrecht, The Netherlands
[4] University of Electronic Science and Technology of China, Chengdu, China

Abstract. Current multimodal latent reasoning often relies on external supervision (e.g., auxiliary images), ignoring intrinsic visual attention dynamics. In this work, we identify a critical **Perception Gap** in distillation: student models frequently mimic a teacher's textual output while attending to fundamentally divergent visual regions, effectively relying on language priors rather than grounded perception. To bridge this, we propose **LaViT**, a framework that aligns *latent visual thoughts* rather than static embeddings. LaViT compels the student to autoregressively reconstruct the teacher's visual semantics and attention trajectories prior to text generation, employing a *curriculum sensory gating* mechanism to prevent shortcut learning. Extensive experiments show that LaViT significantly enhances visual grounding, achieving up to **+16.9%** gains on complex reasoning tasks and enabling a compact 3B model to rival larger models on selected evaluated benchmarks.

Keywords: Latent visual reasoning · Knowledge distillation · MLLMs

1 Introduction

Multimodal Large Language Models (MLLMs) have advanced rapidly in recent years [1,8,37], early multimodal reasoning models primarily *thinking about images*, captioning visual inputs into text and reasoning mainly in the language space via explicit chains of thought [19,31,52]. Recent work instead emphasizes *thinking with images*, more tightly integrating visual evidence into the reasoning process [18,35,57], leading to improved performance on complex visual reasoning tasks [13,44,51,54]

Linquan Wu and Tianxiang Jiang have contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-37574-2_19.

P. Favaro et al. (Eds.): ECCV 2026, LNCS 17021, pp. 345–363, 2026.
https://doi.org/10.1007/978-3-032-37574-2_19

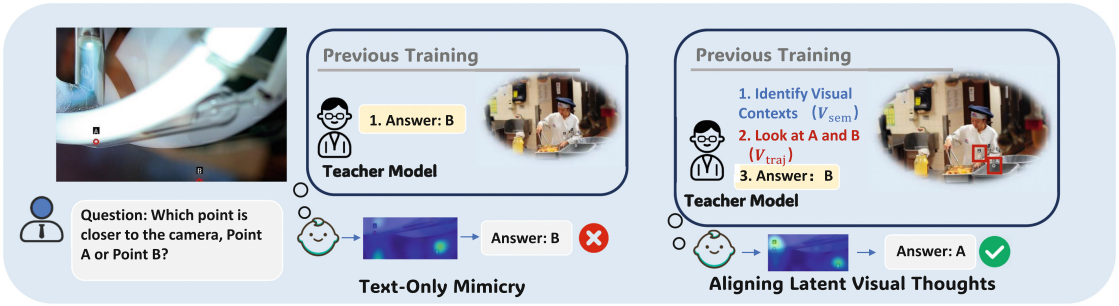

Fig. 1. Conceptual illustration of our proposed method LaViT.

Subsequently, latent reasoning [16] has emerged as a complementary direction, compressing intermediate reasoning into continuous hidden states rather than explicit CoT. This paradigm has been extended to MLLMs to model abstract visual thoughts within latent tokens [25,53]. While effective, existing methods largely rely on *manually designed* visual supervision, such as auxiliary images or annotated regions, leaving intrinsic visual attention dynamics during reasoning unexplored.

These limitations motivate the use of knowledge distillation [17] as a lens to analyze and transfer visual reasoning behaviors in MLLMs. Distilling a high-capacity teacher into a compact student enables us to probe not only *what* knowledge is transferred, but also *how* visual reasoning is internally realized. However, existing multimodal distillation methods mainly align final textual outputs or distributions [3,33], implicitly assuming that reproducing answers or static representations suffices to inherit multimodal reasoning ability.

To examine this assumption, we conduct empirical analyses, revealing a pronounced mismatch between textual alignment and visual reasoning: **(I)** Correct multimodal reasoning is strongly associated with focused visual attention: when models fail to attend to relevant regions, hallucinated or unreliable responses emerge. **(II)** Even when student models closely match teacher outputs under standard distillation, their visual attention trajectories can diverge substantially, particularly for tasks requiring fine-grained visual grounding. Together, these findings expose a fundamental **Perception Gap** in multimodal distillation: students often learn *what to say* without learning *where to look*, instead relying on language priors rather than grounded visual evidence.

Motivated by this insight, we propose **LaViT**, a distillation framework that aligns latent visual thoughts rather than static visual embeddings. LaViT trains the student to autoregressively generate continuous latent tokens that reconstruct the teacher's internal visual semantics and attention trajectories prior to textual response generation, explicitly transferring both what visual concepts to encode and where to attend during reasoning. Figure 1 gives an overview of the proposed latent visual-thought alignment pipeline.

To prevent shortcut learning through direct access to visual features, we introduce **Curriculum Sensory Gating**, which progressively restricts and then relaxes visual input during training. This strategy enforces a latent bottleneck

early on, compelling reliance on latent visual reasoning while avoiding training–inference mismatch.

Extensive experiments demonstrate that LaViT substantially improves both visual grounding and multimodal reasoning. LaViT-3B achieves up to **+5.0%** gains on fine-grained perception benchmarks MMVP [41] and substantial improvement on BLINK [13], outperforming strong baselines and rivaling larger models on the evaluated benchmarks. The code, models, and data are available at https://github.com/Svardfox/LaViT.

2 Related Work

2.1 Visual Chain-of-Thought

Originating from text-only LLMs [48], Chain-of-Thought (CoT) has expanded to multimodal contexts [30]. Following DeepSeek-R1 [15], recent works enhance multi-step visual reasoning via RL-style optimization [11,19,21,31,52]; however, these methods primarily rely on indirect textual proxies rather than intrinsic visual understanding. Conversely, a parallel stream shifts to *thinking with images* by orchestrating tools [34,43,50,55,56], utilizing executable programs [18] or iterative region grounding [57]. For streaming video understanding, recent work also studies semantic-aware visual memory management to retain critical visual transitions under limited token budgets [42]. Furthermore, unified architectures now support interleaved generation, exemplified by Chameleon's unified tokens [38], MVoT's multimodal trajectories [26], and other general-purpose frameworks [4,5,9,14,40]

2.2 Latent Reasoning

Recent advances have shifted the reasoning paradigm from discrete token sequences to continuous hidden states, effectively enhancing both computational efficiency and flexibility [16,32,49]. Extending this concept to multimodal learning, current MLLMs align specialized latent tokens with visual embeddings derived from auxiliary supervision signals, such as helper images [53] or annotated bounding boxes [7,25]. To further improve grounding, CoVT [29] integrates fine-grained perceptual priors from models like DINO [28] and SAM [24], while other approaches explore interleaved patterns to mimic internal visual imagination [39,45]. However, these existing methods primarily constrain latent tokens using static encoder features, critically overlooking the dynamic guidance offered by attention maps.

2.3 Knowledge Distillation

Knowledge distillation [17], which transfers capabilities from a high-capacity teacher to a compact student, has been widely adopted in LLMs via logit matching [22,36]. Extending this paradigm to multimodal models, DistillVLM [10] performs transformer distillation by using an MSE loss to match the teacher and

student's hidden attention distributions and feature maps. In contrast, MAD [46] emphasizes aligning visual and textual token features between teacher and student, leveraging token selection to guide the matching. More recent research explores distillation tailored to MLLMs for specific downstream tasks, including visual grounding [3,12] and compositional learning [23].

3 Empirical Analysis of Perception Gap

We conduct a pilot study to quantify the misalignment between textual generation and visual attention, centered on two research questions: **(RQ1)** Is correct visual reasoning strongly tied to focused visual attention? (i.e., *Does looking at the right place precondition the right answer?*) **(RQ2)** Does a significant "alignment gap" exist in the visual trajectories between teacher and student models, even when their textual outputs are similar?

3.1 Visual Attention Mediates Reasoning Bounds

Definition 3.1 (*Visual Focusing Score*) Let I be the image and B_{gt} the target bounding box. Given the model's aggregated attention trajectory $\mathcal{A}_{traj} \in \mathbb{R}^{H \times W}$, which accumulates attention weights across all layers and heads, the visual focusing score S_{focus} is defined as:

$$S_{focus} = \frac{\sum_{(u,v) \in B_{gt}} \mathcal{A}_{traj}(u,v)}{\sum_{(u,v) \in I} \mathcal{A}_{traj}(u,v)} \tag{1}$$

where $\mathcal{A}_{traj}(u,v)$ denotes the attention intensity at spatial coordinate (u,v). The denominator represents the total attention mass distributed across the entire image I. A larger S_{focus} indicates a stronger dependency of the reasoning process on the verified visual evidence, implying that the model is actively "looking" at the semantically correct region rather than relying on language priors or blind guessing.

Building upon the above metric, we analyze the attention trajectories of Qwen2.5-VL-32B [2] on 1,000 randomly sampled instances from the Visual-CoT [30]. As illustrated in Fig. 2, reasoning outcomes are strongly tied to the intensity of visual attention:

- **Monotonic Performance Gain**: We observe that reasoning accuracy improves monotonically as the S_{focus} threshold increases. Statistical analysis confirms that correct samples maintain a significantly higher average S_{focus} (**18.78%**) compared to incorrect ones (**9.18%**), indicating a substantial relative gap of ∼104%. This validates that higher visual energy is a strong predictor of reasoning success.
- **Visual Absence and Hallucination**: Conversely, in samples with negligible S_{focus} ($< 1\%$), we predominantly observe responses that are completely irrelevant to the visual content or contain severe hallucinations. This pattern

suggests that without active visual grounding, the model relies on language priors to blindly guess, which proves to be a highly unreliable strategy for complex visual tasks.

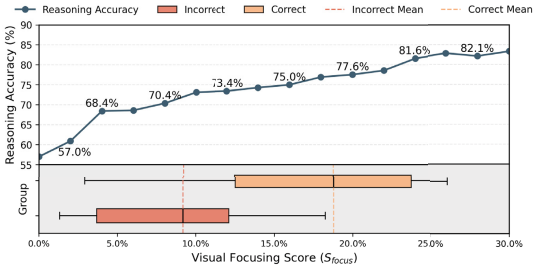

Fig. 2. Impact of visual attention on reasoning accuracy. The monotonic increase in accuracy with higher Visual Focusing Score (S_{focus}) thresholds suggests that effective visual grounding is a critical foundation for correct reasoning.

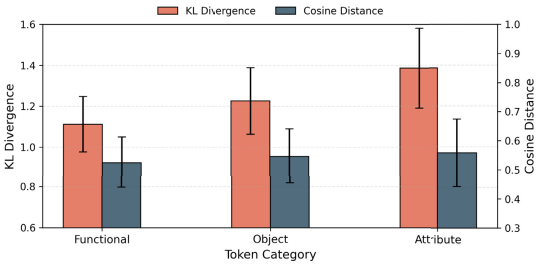

Fig. 3. The perception-reasoning gap. While the student aligns closely with the teacher in textual representations (stable Cosine Distance), their visual attention trajectories diverge significantly on attribute-heavy tokens (rising KL Divergence).

Observation 1: ***Focused visual attention is a necessary mediator, not merely an interpretive artifact.*** The positive correlation indicates that focused visual grounding (S_{focus}) is a strong prerequisite for reasoning success. Models struggle to reason reliably without "looking" at the right evidence.

3.2 Perception Gap Between Teachers and Students

To investigate whether standard Supervised Fine-Tuning (SFT) enables student models to inherit the teacher's intrinsic visual thinking process, we conduct a controlled experiment using a Naive-SFT baseline trained on the textual components of our proposed LaViT-15k dataset (detailed in Sect. 4). Specifically, the student model is fine-tuned to predict the teacher's ground-truth responses using standard cross-entropy loss, without any explicit visual alignment constraints. We then analyze the resulting attention dynamics on a hold-out set of 1,000 samples following the experimental setup in Sect. 3.1.

We fed identical reasoning prompts to both the Teacher and the Student. We categorized the generated tokens into three groups based on their semantic reliance on visual evidence: **Functional** (e.g., stop words), **Object** (nouns), and **Attribute** (adjectives, spatial relations). We then computed the Kullback-Leibler (KL) divergence between their normalized attention maps, alongside the Cosine Distance of their hidden states.

As shown in Fig. 3, we observe a decoupling between textual alignment and visual attention:

- **Attention Drift on Visual Concepts**: The attention divergence exhibits a distinct monotonic increase as the semantic reliance on vision deepens. While Functional tokens maintain a lower average divergence ($\mu = 1.11$), Attribute tokens, which require precise visual grounding, show the highest misalignment ($\mu \approx 1.39$). This indicates that when describing fine-grained details (e.g., color, texture), the student struggles to focus on the same regions as the teacher.
- **The "Blind Guessing" Phenomenon**: Crucially, while the attention divergence surges, the *Cosine Distance* of the hidden states remains relatively stable across categories (ranging from 0.52 to 0.55). This implies that the student can mimic the teacher's textual representations (learning *what* to say) without correctly aligning its visual attention (learning *where* to look), effectively relying on language priors rather than active observation.

Observation 2: ***Textual mimicry does not guarantee visual understanding***. SFT trains the student to reproduce the teacher's *words* but fails to transfer the underlying *visual trajectory*. This "Perception Gap" suggests that the student model is often "guessing" based on language context rather than actively "observing" the image.

4 Method

4.1 LaViT-SFT-15K

We construct **LaViT-SFT-15K**, comprising 15K tuples $\langle I, Q, A, \mathcal{A}_{traj}, \mathcal{V}_{sem} \rangle$, to distill the **Internal Cognitive States** of the teacher (Qwen2.5-VL-32B [2]). LaViT assumes white-box access to a transformer-based teacher, which is available for many open-weight VLMs. Unlike tool-based approaches, we directly extract intrinsic reasoning traces. Data quality is enforced via a Three-Stage Filtering pipeline: (1) **Correctness**: retaining samples matching ground truth; (2) **Difficulty**: removing instances solvable by a text-only model; and (3) **Alignment**: rejecting samples with $< 20\%$ aggregated attention mass falling within the target regions delineated by the ground-truth bounding box annotations in Visual-CoT [30], thereby excluding non-visually grounded hallucinations. These boxes are used only for offline data filtering, not as training supervision and not during inference.

We extract two white-box signals. First, **Dynamic Visual Gaze** ($\mathcal{A}_{traj}$) represents the attention trajectory. Let $Y = \{y_1, y_2, \ldots, y_T\}$ denote the response sequence generated by the teacher model given the input image I and instruction Q. To capture the intrinsic visual reasoning process, we aggregate the cross-attention weights $W^{(l,h)} \in \mathbb{R}^{T \times P}$ between the generated text tokens in Y and the visual patches P. Specifically, for each visual patch j, we compute the raw saliency score S_j by averaging the attention weights across all layers L, heads H, and sequence length T:

$$S_j = \frac{1}{L \cdot H \cdot T} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t=1}^{T} W_{t,j}^{(l,h)} \tag{2}$$

where $W_{t,j}^{(l,h)}$ represents the attention weight of the t-th token in the response sequence Y attending to the j-th visual patch in layer l and head h.

We further apply Min-Max normalization to obtain the final gaze probability:

$$\mathcal{A}_{traj}(j) = \frac{S_j - \min(S)}{\max(S) - \min(S) + \epsilon} \tag{3}$$

Crucially, unlike static visual features extracted from a frozen vision encoder, our target $\mathcal{V}_{sem}$ is derived from the Teacher's last transformer layer. Due to the self-attention mechanism, these image token representations have effectively interacted with the textual instructions Q. Therefore, $\mathcal{V}_{sem}$ represents contextualized visual thoughts—reflecting not just what is in the image, but how the Teacher interprets the visual content specifically in response to the given query. To distill the most salient visual cues, we subject $\mathcal{A}_{traj}$ to Top-N ($N = 8$) sparsification based only on attention magnitude, thereby ensuring sparse and noise-free supervision.

4.2 White-Box Trajectory Distillation

We establish the foundational reasoning capability of our model through a Supervised Fine-tuning (SFT) stage defined as **Latent Teacher Forcing**. Unlike standard SFT that maps inputs directly to text, we train the model to **autoregressively generate** a sequence of K continuous latent tokens, $\mathbf{V} = \{< v - trace_1 >, \ldots, < v - trace_k >\}$, prior to producing the textual response.

These latent tokens serve as **Visual Information Containers**. By leveraging a white-box distillation approach, we force $\mathbf{V}$ to explicitly capture and compress the teacher's high-dimensional visual semantics and gaze patterns. Formally, given an input $[I, Q]$, the model generates the latent and textual sequences to form the complete trajectory $X = [I, Q, \mathbf{V}, A]$, where $\mathbf{V}$ acts as the indispensable cognitive bridge supplying visual evidence for the subsequent response A.

Curriculum Sensory Gating A naive attention mechanism allows response tokens A to attend directly to image patches I, enabling the model to bypass $\mathbf{V}$ (shortcut learning). Conversely, a permanent hard mask creates a training-inference distribution shift. To resolve this, we propose Curriculum Sensory Gating, which modulates the direct visual perception path via a time-dependent scalar $\gamma(t) \in [\epsilon, 1]$.

We implement gating within the attention bias of the Transformer. Let Q_{txt} denote queries from response tokens and K_{img} denote keys from image patches. The attention scores are computed as:

$$\begin{aligned} \text{Attn}(Q_{txt}, K_{img}) &= \text{Softmax}\left(\frac{Q_{txt}K_{img}^{\top}}{\sqrt{d}} + \mathbf{B}_{gate}(t)\right), \\ \text{s.t. } \mathbf{B}_{gate}(t) &= \ln(\gamma(t)). \end{aligned} \tag{4}$$

To ensure a structured internalization process mentioned above, we introduce a warm-up period T_w. The gating scalar $\gamma(t)$ is defined as:

$$\gamma(t) = \begin{cases} \epsilon + \frac{1-\epsilon}{2}\left[1 - \cos\left(\frac{\pi t}{T_w}\right)\right], & t < T_w \\ 1, & t \geq T_w \end{cases} \tag{5}$$

where T_w denotes the warm-up steps. This schedule defines two distinct operational phases governed by the training progress:

- **Phase 1: Sensory Warm-up ($t < T_w$).** The direct visual path opens gradually, following the cosine curve. Initially, $\gamma \approx \epsilon$ (set to 1e-6.). Setting ϵ to this value not only yields an extremely large negative bias ($B_{gate} \ll 0$) to block the direct visual path—effectively forming an implicit bottleneck—but also maintains the validity of the logarithmic operation, ensuring absolute convergence stability in the early training phase. This creates a strict **Latent Bottleneck**, mathematically compelling the model to compress necessary visual information into $\mathbf{V}$. The subsequent smooth relaxation prevents optimization shock while establishing a strong dependency on latent reasoning.
- **Phase 2: Fully Observable ($t \geq T_w$).** The gate becomes fully open ($\gamma = 1$), reducing the bias $\mathbf{B}_{gate}$ to zero. The direct visual path functions as a **Residual Perception** connection, allowing the model to attend to fine-grained pixel details that complement the high-level reasoning encoded in $\mathbf{V}$. This configuration matches the standard inference topology, ensuring zero distribution shift.

Optimization Objectives To ensure the student strictly internalizes the teacher's cognition, we employ a dual-stream distillation scheme with explicit gradient flow controls.

1. **Semantic Reconstruction ($\mathcal{L}_{concept}$).**
   We align the student's latent hidden states h_z with the teacher's holistic visual concepts V_{sem}. Since the teacher's representations encapsulate high-quality visual semantics, we treat them as *fixed semantic anchors*. We employ a projection head ϕ_{mlp} to map the student's latent manifold into the teacher's semantic space:

$$\mathcal{L}_{concept} = 1 - \frac{1}{B}\sum_{i=1}^{B} \text{CosSim}\left(\phi_{\text{mlp}}(h_z^{(i)}), V_{\text{sem}}^{(i)}\right) \tag{6}$$

   This objective compels the student's latent tokens $\mathbf{V}$ to act as informative containers, actively capturing and compressing the visual information necessary to reconstruct the teacher's visual semantics.
2. **Trajectory Alignment ($\mathcal{L}_{traj}$).**
   We define the reasoning trajectory as the distribution of attention weights

over visual patches. We treat the teacher's attention map A_{traj} as the target, following prior observations that distilling teacher attention maps can effectively guide student visual alignment [10]. We constrain the student's attention A_{student} (originating from $\mathbf{V}$) to match this target via KL Divergence:

$$\mathcal{L}_{traj} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{K}\mathcal{D}_{\text{KL}}\left(A_{\text{traj}}^{(i,j)} \| A_{\text{student}}^{(i,j)}\right). \tag{7}$$

This ensures the latent tokens learn where to look, inheriting the teacher's visual search strategy. Crucially, we apply this supervision only to the final latent token V_{trace_k}. This constraint forces the preceding $k-1$ tokens to act as a compression bottleneck, aggregating and refining visual information into a single, cohesive representation before alignment.

3. **Response Generation with Dynamic Gradient Transition ($\mathcal{L}_{ntp}$).** The Next-Token Prediction loss drives the response generation. Crucially, the gradient flow from this loss is intrinsically modulated by our gating mechanism. By chain rule, the sensitivity of the loss with respect to direct visual features is proportional to the attention weight:

$$\left\|\frac{\partial \mathcal{L}_{ntp}}{\partial I}\right\| \propto \text{Attn}(Q_{txt}, K_{img}) \approx \gamma(t). \tag{8}$$

During Phase 1, gradients are fully channeled through $\mathbf{V}$, establishing the bottleneck. As $\gamma \to 1$, the gradients transition to a synergistic flow, optimizing both the latent abstraction and the residual perception paths jointly.

Joint Training Dynamics Unlike complex multi-stage pipelines [25,45] that require careful hyperparameter scheduling to avoid collapse, our Curriculum Sensory Gating provides a robust structural constraint. This physically enforced bottleneck naturally regulates the learning difficulty, eliminating the need for dynamic loss weighting.

We employ a streamlined **Joint Training** paradigm with a fixed distillation weight. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ntp} + \lambda \cdot (\mathcal{L}_{concept} + \mathcal{L}_{traj}), \tag{9}$$

By maintaining a constant but moderate alignment pressure, we ensure the latent tokens $\mathbf{V}$ remain semantically consistent with the teacher in the curriculum, while allowing the NTP loss to primarily drive the generation quality as the sensory gate opens.

5 Experiment

5.1 Experiment Setup

SFT Settings. In Phase 1, we train the model for 400 steps with the sensory gating scalar γ initialized at 1e-6. In Phase 2, we continue training for an additional 600 steps. We set $\lambda = 0.3$ across all experiments. The model is initialized

Table 1. Performance comparison on multimodal benchmarks across three categories: Fine-grained Perception, Visual Reasoning, and Multimodal Robustness

Method	Params	Fine-grained Perception		Visual Reasoning				Robustness
		MMVP	V* Attrib.	Rel. Depth	IQ-Test	Rel. Ref.	Spatial	MMStar
Proprietary Model								
GPT-4o [20]	-	58.33	72.5	64.52	30.0	38.8	76.9	63.9
Open Source Model								
Qwen2.5-VL [2]	7B	66.7	77.39	71.77	26.0	38.8	87.4	58.9
Qwen2.5-VL [2]	32B	**75.33**	82.61	75.81	30.0	**55.97**	85.31	**69.8**
Open Source Reasoning MLLMs								
Naive SFT	3B	65.33	80.87	70.16	28.0	34.33	81.82	55.53
PAPO [47]	3B	50.0*	22.61*	59.68	31.33*	33.6	76.92	52.7
DMLR [27]	3B	61.33	46.96	54.84	26.0	23.88	72.03	51.2
LVR_RL [25]	3B	55.3*	69.6*	64.52	30.7*	42.54	77.62	53.73
R1-OneVision [19]	7B	67.0*	-	-	-	-	-	52.1*
LVR [25]	7B	72.0	**84.4**	76.61	28.7	42.5	**89.5**	59.4
Baseline & Our Model								
Qwen2.5-VL (Baseline)	3B	62.33	81.74	61.29	24.0	29.85	81.12	50.2*
LaViT	3B	67.33 ↑5.0	82.61 ↑0.87	**78.23** ↑16.9	**32.0** ↑8.0	45.52 ↑15	86.82 ↑0.7	54.07 ↑3.87

The best and second-best results are marked in **bold** and underlined, respectively. Green values indicate the absolute gains of **LaViT** over the Qwen2.5-VL-3B baseline. Asterisks (*) denote results reported by papers.[13,25,27]

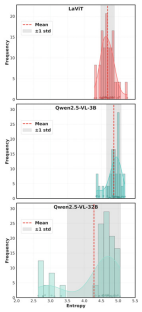

Fig. 4. Attention entropy distribution.

from Qwen2.5-VL-3B and finetuned on the LaViT-SFT-15K. We set the number of latent visual tokens V to 4. More details are shown in the appendix.

Evaluated Benchmarks. We evaluate LaViT on diverse benchmarks. For subtle visual details, we use MMVP [41] and the Attribute Recognition subset of V* [51], which target CLIP-blind patterns and object attributes. For higher-level cognition, we adopt BLINK [13] tasks on Relative Depth, IQ-Test, Relative Reflectance, and Spatial Relation, which require mental manipulation and geometric reasoning. We also include MMStar [6] to test robustness against language priors and ensure genuine visual understanding.

5.2 Main Results

Table 1 presents the comparative performance of LaViT against state-of-the-art MLLMs. As shown, our method achieves consistent improvements across all benchmarks, demonstrating the efficacy of latent visual thinking.

Cross-Scale Competitiveness. LaViT significantly enhances its backbone, achieving substantial gains of +15.67% on Relative Reflectance and +16.94% on Relative Depth. Despite its compact 3B scale, it demonstrates strong cross-scale competitiveness, outperforming the larger Qwen2.5-VL-7B on five out of seven evaluated benchmarks. Moreover, LaViT surpasses SOTA 7B models on selected evaluated tasks, beating LVR-7B on Relative Depth (78.23% vs. 76.61%) and R1-OneVision-7B on MMStar. These results indicate that optimizing latent thinking can be more parameter-efficient than simply scaling up model size for these benchmarks.

Advantage in Complex Visual Reasoning. LaViT excels in perception-intensive BLINK benchmarks by leveraging continuous latent reasoning to preserve spatial structures, effectively addressing the limitations of standard models in abstract geometric manipulation. Consequently, LaViT-3B achieves 78.23% on Relative Depth and 32.0% on IQ-Test, outperforming GPT-4o on these selected tasks (64.52% and 30.0%) and reasoning-enhanced baselines. Notably, it even

surpasses the SOTA latent reasoning model LVR-7B on Relative Depth (+1.62%) and Relative Reflectance (+3.0%), validating its superior capability in capturing structural visual semantics.

Fine-Grained Perception and Robustness. LaViT effectively mitigates "CLIP-blindness," achieving 67.33% on MMVP and substantially outperforming DMLR (61.33%) and PAPO (50.0%). By refining visual features to correct encoding errors rather than trading perception for reasoning, LaViT ensures robust visual grounding. This is further evidenced by its 54.07% score on MMStar (vs. 50.2% baseline), confirming that performance gains stem from genuine visual understanding rather than language hallucinations.

5.3 In-Depth Analysis of Attention Dynamics

To investigate the underlying mechanisms of LaViT, we conduct a deep dive into the attention distribution on the BLINK *Relative Depth*. We combine quantitative metrics with qualitative visualizations to analyze **Concentration** and **Stability**.

Quantifying Attention Concentration (Entropy). We employ Information Entropy (H) to measure the "sharpness" of the model's visual focus. Formally, for a given image, the attention entropy is defined as:

$$H = -\sum_{i=1}^{N} p_i \log(p_i) \tag{10}$$

where p_i is the normalized attention weight of the i-th patch. To quantify these observations, we analyze the attention statistics on the BLINK Relative Depth dataset (Table 2). We define Salient Regions as patches with attention weights surpassing 1.5× the global mean, serving as a proxy for the area of active visual processing. Additionally, we report the Coefficient of Variation (CV) (σ/μ) to measure the stability of the attention mechanism. These entropy/CV statistics capture distribution-level concentration, while the region-aware S_{focus} in Sect. 3.1 measures whether attention falls on annotated target regions. As visualized in Fig. 4 and detailed in Table 2, the Base 3B model exhibits a heavy tail towards high entropy ($H = 4.870$), suggesting it lacks a clear visual target. LaViT significantly shifts this distribution leftward, reducing the mean entropy to **4.686**. This improvement not only approximates the Teacher's focused state ($H = 4.284$) but is also visually corroborated by Fig. 5, where LaViT exhibits pinpoint focus on critical depth markers compared to the scattered gaze of the Base model.

Takeaway 1: Sharpening Visual Focus

LaViT effectively transitions the student from a "diffuse" observation mode to a "focused" reasoning mode. By reducing attention entropy (Figure 4), the student learns to ignore irrelevant background noise and concentrate on task-relevant regions.

Table 2. Statistical analysis of attention distribution on BLINK Relative Depth.

Model	Salient Regions Count				Attention Entropy (H)		
	Mean	Std (σ)	Range	CV	Mean	Std (σ)	CV
Qwen2.5-32B	39.9	15.6	8-65	0.392	**4.284**	0.787	0.1837
Qwen2.5-3B	53.8	9.0	36-77	0.191	4.870	0.216	0.0444
LaViT	47.3	**5.6**	**43-69**	**0.102**	4.686	**0.204**	**0.0435**

Table 3. Ablation study on key components of LaViT.

Method	MMVP	Rel. Depth	IQ-test	Rel. Ref	Spatial
LaViT-3B	**67.33**	**78.23**	**32**	45.52	**81.82**
w/o Traj. Align (L_{traj})	64.33	75	30.67	44.03	78.32
w/o Sem. Recon ($L_{concept}$)	65.33	75.81	30.67	42.54	76.92
w/o Curr. Gate	59.33	71.77	27.33	**48.51**	79.72
w/o Latent Tokens	64.33	77.42	25.33	40.30	81.12
Single Stage Training	65.67	71.77	28	38.81	79.02

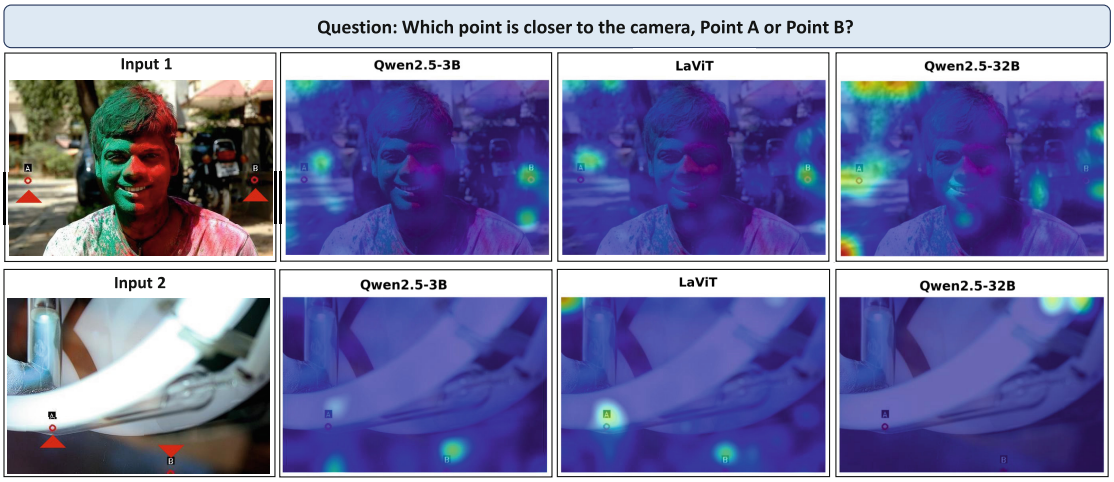

Fig. 5. Visualization of attention distributions across Qwen2.5-VL-3B, 32B (Teacher), and LaViT on two representative samples from the BLINK. ▲ indicates the task-relevant critical regions required for correct reasoning.

Achieving Superior Stability via Sparsification. While the Teacher model is highly focused, it exhibits significant variance in its viewing strategy across samples (Salient Regions CV = 0.392). We observe that LaViT not only inherits the Teacher's focus but actually achieves a much more stable attention pattern (CV = **0.102**), surpassing the teacher in consistency.

We attribute this distilled stability directly to our **White-box Trajectory Distillation** design:

- **Top-N Sparsification**: By explicitly supervising the student with only the Top-N ($N = 8$) strongest attention points from the teacher, we enforce a hard constraint that filters out the teacher's low-confidence "attentional noise" or hesitation.
- **Data Filtering**: Our preprocessing pipeline excludes samples where attention does not align with ground-truth regions, ensuring only high-quality traces are learned.

Consequently, LaViT does not merely mimic the Teacher's raw output; it distills the **most robust visual cues**, resulting in a student model that is consistently focused on critical regions without the variance observed in the large teacher model.

Takeaway 2: The Denoising Effect

LaViT acts as a **semantic filter**. Instead of inheriting the teacher's instability, LaViT leverages Top-N sparsification to retain only the core attention patterns, reducing the Coefficient of Variation (CV) from 0.392 to 0.102. This results in a student who is surprisingly more decisive and stable than its teacher.

5.4 Ablation Study

We evaluate LaViT on MMVP and BLINK subsets, comparing it against variants including single-stage training and the removal of curriculum gating.

Impact of Alignment Components. To dissect the contribution of each module, we report the ablation results in Table 3. We investigate the impact of removing key components:

- **Trajectory Alignment (w/o Traj.)**: Removes the attention alignment loss (L_{traj}).
- **Semantic Reconstruction (w/o Sem.)**: Excludes the concept reconstruction objective ($L_{concept}$).
- **Curriculum Sensory Gating (w/o Gate)**: Replaces our progressive schedule with a hard switch ($\gamma = 0$ in Phase 1, $\gamma = 1$ in Phase 2).
- **Single Stage Training**: Conducts training in a single phase where visual tokens are always fully visible ($\gamma(t) = 1$).
- **Latent Tokens (w/o Latent)**: Masks the generated latent visual tokens during inference to test dependency.

As shown in Table 3, removing either Trajectory Alignment or Semantic Reconstruction leads to significant performance drops across all tasks. Furthermore, masking latent tokens during inference precipitates a marked decline, verifying the model's genuine reliance on the generated visual thoughts. This confirms explicit supervision on "where to look" and "what to see" is essential for the student to inherit the teacher's visual cognitive patterns effectively.

Effectiveness of Curriculum Sensory Gating. The performance degradation observed without this module (e.g., MMVP accuracy drops to 59.33%) indicates that a simple hard switch of visual visibility is insufficient. Our progressive gating strategy is crucial for preventing shortcut learning and forcing the model to rely on deep latent reasoning.

Impact of Training Strategy. The *Single Stage Training* baseline ($\gamma(t) = 1$) consistently underperforms the full model, particularly on Relative Reflectance (38.81% vs. 45.52%). This demonstrates that progressively exposing visual information is key to avoiding sub-optimal convergence and building robust reasoning capabilities.

6 Conclusion

In this work, we identify the "Perception Gap" in multimodal distillation, where student models often mimic textual outputs without inheriting the teacher's visual attention patterns. To bridge this, we propose **LaViT**, a framework that aligns latent visual thoughts via White-box Trajectory Distillation and Curriculum Sensory Gating. By utilizing latent tokens as cognitive containers, LaViT compels the student to reconstruct the teacher's visual semantics and gaze before response generation. Extensive experiments demonstrate that LaViT-3B significantly outperforms SFT baselines and rivals 7B-scale models on the evaluated reasoning-intensive benchmarks.

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., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2. 5-vl technical report (2025). arXiv preprint arXiv:2502.13923
3. Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., Wang, C., Xue, Z., Liu, Y., Bai, X.: Llava-kd: A framework of distilling multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 239–249 (2025)
4. Cao, H., Chu, H., Li, Y., Zhang, Y., Li, C., Lyu, J., Wang, Y., Zhao, Y., Wu, F.: ClueAegis: Heuristic-to-reasoning cognitive-skill learning for unified evidence-based synthetic image detection (2026). https://arxiv.org/abs/2605.25009
5. Cao, H., Mei, Q., Li, Z., Li, Y., Meng, Z., Zhang, Y., Li, C., Zhang, Z., Ding, X., Wang, Y., Lyu, J., Wu, F.: Reveal: reasoning-enhanced forensic evidence analysis for explainable AI-generated image detection (2026). https://arxiv.org/abs/2511.23158
6. 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? Adv. Neural. Inf. Process. Syst. **37**, 27056–27087 (2024)
7. Chi, H., Qiu, D., Su, H., Liu, H., Li, Z., Zhang, H., Lv, C.: Driver-WM: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout (2026). https://arxiv.org/abs/2605.05092
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025). arXiv preprint arXiv:2507.06261

9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining (2025). arXiv preprint arXiv:2505.14683
10. Fang, Z., Wang, J., Hu, X., Wang, L., Yang, Y., Liu, Z.: Compressing visual-linguistic model via knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1428–1438 (2021)
11. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-R1: reinforcing video reasoning in MLLMS (2025). arXiv preprint arXiv:2503.21776
12. Feng, Q., Li, W., Lin, T., Chen, X.: Align-KD: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 4178–4188 (2025)
13. 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. In: European Conference on Computer Vision, pp. 148–166. Springer (2024)
14. Gu, J., Hao, Y., Wang, H.W., Li, L., Shieh, M.Q., Choi, Y., Krishna, R., Cheng, Y.: ThinkMorph: emergent properties in multimodal interleaved chain-of-thought reasoning (2025). arXiv preprint arXiv:2510.27492
15. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature **645**(8081), 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z
16. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Haotraining large language models to reason in a continuous latent space (2024). arXiv preprint arXiv:2412.06769
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015). arXiv preprint arXiv:1503.02531

18. Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: sketching as a visual chain of thought for multimodal language models. Adv. Neural. Inf. Process. Syst. **37**, 139348–139379 (2024)
19. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models (2025). arXiv preprint arXiv:2503.06749
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D.P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F.P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H.P., Ren, H., Chang, H., Chung, H.W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J.G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J.Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J.W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M.J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M.O.T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige,

N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R.G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., Malkov, Y.: GPT-4o system card (2024). arXiv preprint arXiv:2410.21276

21. Jiang, T., Xia, S., Xu, Y., Wu, L., Zeng, X., Wang, L., Qiao, Y., Wang, Y.: VKnowU: Evaluating visual knowledge understanding in multimodal LLMS (2025). arXiv preprint arXiv:2511.20272
22. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: distilling bert for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4163–4174 (2020)
23. Kim, J., Kim, K., Seo, S., Park, C.: CompoDistill: attention distillation for compositional reasoning in multimodal LLMs (2025). arXiv preprint arXiv:2510.12184
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3992–4003. (2023). https://doi.org/10.1109/ICCV51070.2023.00371. Segment anything
25. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning (2025). arXiv preprint arXiv:2509.24251
26. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: multimodal visualization-of-thought (2025). arXiv preprint arXiv:2501.07542
27. Liu, C., Yang, Y., Fan, Y., Wei, Q., Liu, S., Wang, X.E.: Reasoning within the mind: dynamic multimodal interleaving in latent space (2025). arXiv preprint arXiv:2512.12623
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: learning robust visual features without supervision (2023). arXiv preprint arXiv:2304.07193
29. Qin, Y., Wei, B., Ge, J., Kallidromitis, K., Fu, S., Darrell, T., Wang, X.: Chain-of-visual-thought: teaching VLMs to see and think better with continuous visual tokens (2025). arXiv preprint arXiv:2511.19418

30. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual CoT: unleashing chain-of-thought reasoning in multi-modal language models. CoRR (2024)
31. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: VLM-R1: a stable and generalizable R1-style large vision-language model (2025). arXiv preprint arXiv:2504.07615
32. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: CODI: Compressing chain-of-thought into continuous space via self-distillation (2025). arXiv preprint arXiv:2502.21074
33. Shu, F., Liao, Y., Zhuo, L., Xu, C., Zhang, L., Zhang, G., Shi, H., Chen, L., Zhong, T., He, W., Fu, S., Li, H., Li, B., Yu, Z., Liu, S., Li, H., Jiang, H.: LLaVA-MoD: making LLaVA tiny via moe knowledge distillation (2024). arXiv preprint arXiv:2408.15881
34. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: OpenThinkIMG: learning to think with images via visual tool reinforcement learning (2025). arXiv preprint arXiv:2505.08617
35. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., Li, L., Cheng, Y., Ji, H., He, J., Fung, Y.R.: Thinking with images for multimodal reasoning: foundations, methods, and future frontiers (2025). arXiv preprint arXiv:2506.23918
36. Sun, S., Cheng, Y., Gan, Z., Liu, J.: Patient knowledge distillation for BERT model compression, (2019). arXiv preprint arXiv:1908.09355
37. Team, B.S.: Seed1.5-VL technical report (2025). arXiv preprint arXiv:2505.07062
38. Team, C.: Chameleon: mixed-modal early-fusion foundation models (2024). arXiv preprint arXiv:2405.09818
39. Tong, J., Gu, J., Lou, Y., Fan, L., Zou, Y., Wu, Y., Ye, J., Li, R.: Sketch-in-latents: eliciting unified reasoning in MLLMs (2025). arXiv preprint arXiv:2512.16584
40. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17001–17012 (2025)
41. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? Exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9568–9578 (2024)
42. Wang, C., Tan, X., Cao, J., Li, K., Chen, T.: CurveStream: boosting streaming video understanding in MLLMs via curvature-aware hierarchical visual memory management (2026). arXiv preprint arXiv:2603.19571
43. Wang, H., Su, A., Ren, W., Lin, F., Chen, W.: Pixel reasoner: incentivizing pixel-space reasoning with curiosity-driven reinforcement learning (2025). arXiv preprint arXiv:2505.15966
44. Wang, Q., He, J., Pan, Y., Yeo, S.Y., Yang, X., Li, S.: MonoSR: open-vocabulary spatial reasoning from monocular images (2025). https://arxiv.org/abs/2511.19119
45. Wang, Q., Shi, Y., Wang, Y., Zhang, Y., Wan, P., Gai, K., Ying, X., Wang, Y.: Monet: reasoning in latent visual space beyond images and language (2025). arXiv preprint arXiv:2511.21395
46. Wang, Z., Codella, N., Chen, Y.C., Zhou, L., Dai, X., Xiao, B., Yang, J., You, H., Chang, K.W., Fu Chang, S., Yuan, L.: Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks (2022). arXiv preprint arXiv:2204.10496

47. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., Ji, H.: Perception-aware policy optimization for multimodal reasoning (2025). arXiv preprint arXiv:2507.06448
48. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
49. Wei, X., Liu, X., Zang, Y., Dong, X., Cao, Y., Wang, J., Qiu, X., Lin, D.: SIM-CoT: Supervised implicit chain-of-thought, (2025). arXiv:2509.20317 arXiv preprint
50. Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., Nahrstedt, K.: VTool-R1: VLMs learn to think with images via reinforcement learning on multimodal tool use (2025). arXiv preprint arXiv:2505.19255
51. Wu, P., Xie, S.: V*: guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13084–13094. (2024)
52. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., Zhang, B., Chen, W.: R1-onevision: advancing generalized multimodal reasoning through cross-modal formalization (2025). arXiv preprint arXiv:2503.10615
53. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: empower multimodal reasoning with latent visual tokens (2025). arXiv preprint arXiv:2506.17218
54. Zhang, F., Wu, L., Bai, H., Lin, G., Li, X., Yu, X., Wang, Y., Chen, B., Keung, J.: Humaneval-v: Evaluating visual understanding and reasoning abilities of large multimodal models through coding tasks (2024). arXiv preprint arXiv:2410.12381
55. Zhang, X., Gao, Z., Zhang, B., Li, P., Zhang, X., Liu, Y., Yuan, T., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Chain-of-focus: adaptive visual search and zooming for multimodal reasoning via RL (2025). arXiv preprint arXiv:2505.15436
56. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., Fan, H., Chen, K., Chen, J., Ding, H., Tang, K., Zhang, Z., Wang, L., Yang, F., Gao, T., Zhou, G.: Thyme: Think beyond images (2025). arXiv preprint arXiv:2508.11630
57. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning (2025). arXiv preprint arXiv:2505.14362