ProLaViT: Learning Progressive Latent Visual Thoughts in Structured Latent Space

Peiming Li[1*], Yifan Wang[1*], Xiaotian Zhang[2], Zhiyuan Hu[3], Shiyu Li[1], Zheng Wei[1†], and Yang Tang[1‡†]

[1] Basic Algorithm Center, PCG, Tencent
{peimingli, wyattyfwang, shyuli, hemingwei, ethanntang}@tencent.com
[2] Zhejiang University
xiaotian.24@intl.zju.edu.cn
[3] Peking University
zhiyuanhu@stu.pku.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) have achieved remarkable progress but still struggle with complex visual reasoning tasks requiring multi-step perception and logical deduction. While explicit visual generation incurs prohibitive computational costs, existing latent approaches often rely on external experts or lack rigorous cognitive logic. In this paper, we introduce ProLaViT (**Pro**gressive **La**tent **Vi**sual **T**hought), a framework empowering MLLMs to perform structured visual derivation in the continuous latent space. Unlike works dependent on heterogeneous external models, ProLaViT leverages an endogenous self-distillation mechanism, utilizing the model's own visual encoder to supervise latent thoughts. To facilitate this, we construct a scalable programmatic synthesis pipeline enabling the model to internalize algorithmic precision without inference-time tools. We design two reasoning paradigms: (1) Coarse-to-Fine Causal Chain for spatial tasks, guiding attention from global context to local targets. (2) Dialectical Reasoning Chain for logical tasks, incorporating counterfactual thinking for verification. Furthermore, we propose a Distance-Weighted Diversity Loss to impose topology-aware constraints, preventing feature degeneration by enforcing semantic distinctiveness. Extensive experiments demonstrate that ProLaViT outperforms baselines on vision-centric benchmarks, achieving superior accuracy and interpretability with high efficiency.

Keywords: Multimodal Large Language Models · Latent Visual Reasoning · Progressive Visual Derivation

1 Introduction

The advent of Multimodal Large Language Models (MLLMs) [1, 17, 20] has revolutionized the intersection of vision and language, enabling systems to describe

[*] Equal contribution. [†] Corresponding author. [‡] Project Lead.

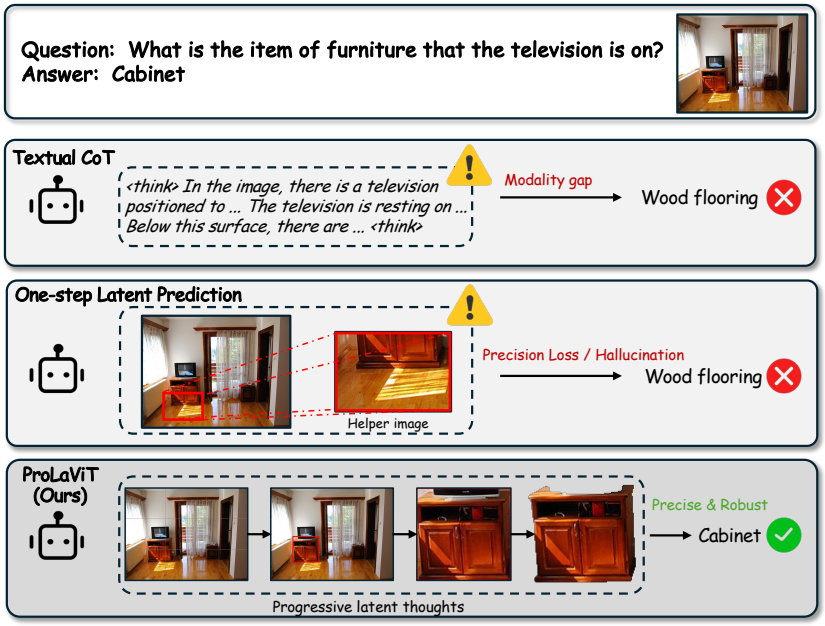


Fig. 1: Comparison of multimodal reasoning paradigms. Top: Textual CoT suffers from a *modality gap*, failing to accurately ground spatial relationships (e.g., mistaking the wood flooring for the cabinet). **Middle:** One-step Latent Prediction attempts to identify the target region in a single leap but suffers from *precision loss*, often attending to irrelevant background areas (hallucination). **Bottom: ProLaViT (Ours)** decomposes the task into a *progressive latent chain* (Locate → Focus → Isolate), enabling precise target isolation and robust reasoning.

images and answer open-ended questions with unprecedented fluency. Nevertheless, despite these achievements, current MLLMs frequently exhibit limitations when addressing visual reasoning tasks that necessitate multi-step spatial perception, geometric analysis, or logical deduction [7,19,28,29]. The core challenge lies in the reasoning mechanism, specifically in how to effectively bridge the gap between high-level linguistic semantics and low-level visual features.

Prevalent approaches rely on Textual Chain-of-Thought (CoT) [31,38] to decompose complex tasks. However, as illustrated in Fig. 1, reasoning solely in text suffers from a fundamental modality gap. Text is inherently abstract and often fails to capture fine-grained spatial nuances. For instance, when asked to identify the furniture beneath a television, the model fails to ground the spatial relationship, hallucinating "wood flooring" instead of the correct "cabinet". To address this, recent works have explored incorporating visual information into the reasoning chain. Explicit visual generation methods such as Anole [3], Bagel [4], Zebra-CoT [15] and ThinkMorph [9] generate intermediate pixel-level images to aid reasoning. While effective, the computational cost of iterative image generation is prohibitive for real-time applications. Conversely, implicit latent reasoning

approaches [6,10,16,23,30,36] attempt to model reasoning within the continuous feature space. However, as illustrated in Fig. 1, these approaches typically adhere to a One-step Latent Prediction paradigm. Attempting to directly localize a specific visual region such as a helper image crop within a single step often exceeds the representational capacity of the model. This limitation frequently results in significant precision loss where the model focuses on irrelevant background elements like the floor rather than the intended target, ultimately leading to fragile representations and erroneous predictions.

In this work, we argue that effective visual reasoning requires neither expensive pixel generation nor unstructured latent guessing, but rather **Progressive Visual Derivation** within a structured latent space. We introduce **ProLaViT**, a novel framework that decomposes complex visual tasks into a structured chain of latent thoughts. Instead of jumping to a conclusion, it mimics human cognitive processes, following a *Locate* $\rightarrow$ *Focus* $\rightarrow$ *Isolate* causal chain to progressively refine visual attention, ensuring precise object isolation and robust reasoning.

A key challenge in training such multi-step latent models is the absence of ground truth for intermediate images. Unlike prior works that rely on external vision experts such as [13, 21, 35], ProLaViT employs an **Endogenous Self-Distillation** mechanism. This approach effectively avoids the domain gaps and pipeline complexity often introduced by external dependencies. We leverage the MLLM's own frozen vision encoder to extract features from a sequence of auxiliary images. Crucially, these auxiliary sequences are generated via a scalable programmatic synthesis pipeline. By training on these rigorous, code-generated visual trajectories, ProLaViT effectively internalizes algorithmic precision, learning to perform latent operations in its latent space without needing external tools during inference.

Furthermore, a common pathology in sequential latent reasoning is *latent collapse*, where the representations of subsequent steps become indistinguishable, degrading the reasoning chain into redundancy. To mitigate this, we propose a novel **Distance-Weighted Diversity Loss**. This objective function explicitly constrains the topology of the latent space, penalizing cosine similarity between reasoning steps. Crucially, the penalty is weighted by the causal distance in the chain, forcing the latent state of a segmentation step to be significantly distinct from the initial global view step, thereby ensuring that each step contributes unique and progressive information to the inference process. Our main contributions are summarized as follows:

- We propose ProLaViT, a framework that enables MLLMs to perform Progressive Visual Derivation in the latent space, overcoming the efficiency bottlenecks of explicit generation and the precision limitations of unstructured latent methods.
- We introduce an Endogenous Self-Distillation strategy supported by a programmatic data synthesis pipeline, allowing the model to internalize rigorous algorithmic logic without external vision experts.

- We design a Distance-Weighted Diversity Loss to prevent latent collapse, enforcing semantic distinctiveness between reasoning steps proportional to their causal distance.
- Extensive experiments demonstrate that ProLaViT achieves SOTA performance, offering a superior balance of accuracy, interpretability, and efficiency.

2 Related Work

2.1 Multimodal Chain-of-Thought Reasoning

Chain-of-Thought (CoT) prompting [31] enhances LLM reasoning by decomposing problems into logical steps. Multimodal CoT [11, 12, 38] extends this to vision-language tasks, and works like Qwen-VL [1] and LLaVA-CoT [34] demonstrate that reasoning data improves performance. However, these methods rely on projecting visual information into discrete text. Recent studies [7,29] indicate this creates a modality gap, as abstract text often fails to capture fine-grained spatial nuances. This leads to hallucinations where reasoning is linguistically coherent yet visually inaccurate. In contrast, ProLaViT performs reasoning directly in the continuous latent space, preserving visual fidelity.

2.2 Visual-Augmented Reasoning with External Tools

To bridge the modality gap, researchers have augmented MLLMs with external tools [18, 25, 27, 32, 37] or explicit visual generation [3, 5, 8, 9, 14]. Specifically, methods like CoVT [23, 24] employ experts such as SAM [13] or DepthAnything [35] for supervision. However, these approaches face two critical limitations: (1) *Prohibitive Cost* due to pixel-level generation or external inference latency; and (2) *Pipeline Complexity* arising from heterogeneous dependencies. In contrast, ProLaViT adopts an *endogenous self-distillation* mechanism, leveraging the MLLM's own frozen encoder to supervise latent thoughts, eliminating external overhead.

2.3 Latent Visual Reasoning

Recent works explore reasoning within continuous latent spaces to balance efficiency and expressiveness. While language-centric methods [10,26] introduce continuous tokens, they lack spatial grounding. Visual adaptations like Mirage [36] and LVR [16] map visual semantics into LLMs but typically adopt an *unstructured* or *single-step* paradigm. As our analysis reveals, this often leads to *latent collapse*, where intermediate representations degenerate into redundancy. ProLaViT advances this direction by introducing Structured Progressive Derivation. Instead of black-box transitions, we enforce a rigorous causal chain (*Locate* $\rightarrow$ *Focus* $\rightarrow$ *Isolate*) constrained by a Distance-Weighted Diversity Loss, ensuring that latent thoughts remain topologically distinct and logically progressive.

3 Method

3.1 Overview

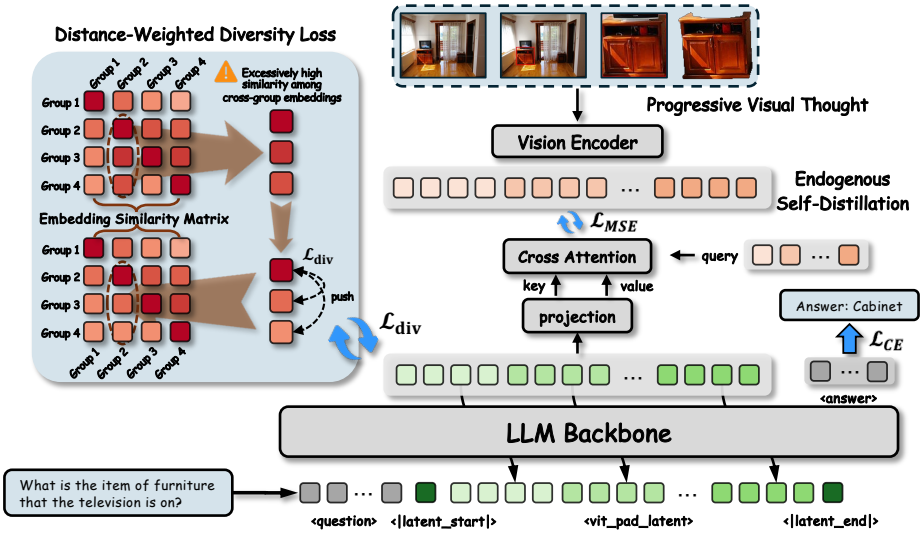


Fig. 2: Overview of ProLaViT. The framework generates progressive latent thoughts (green tokens) supervised by an Endogenous Self-Distillation mechanism (*Right*), where latent states are aligned with visual features from synthesized auxiliary images via $\mathcal{L}_{MSE}$. To prevent representation degeneration, a Distance-Weighted Diversity Loss ($\mathcal{L}_{div}$, *Left*) penalizes excessive similarity between reasoning steps, enforcing topological distinctiveness in the latent space.

Fig. 2 illustrates the overall architecture of ProLaViT. Given an input image I and a textual query Q, ProLaViT performs visual reasoning through a progressive latent thought chain rather than a single-step prediction. The framework consists of three core components: (1) A **Progressive Latent Visual Thought** module that decomposes complex visual tasks into a structured sequence of K intermediate reasoning steps. (2) An **Endogenous Self-Distillation** mechanism that leverages the model's own vision encoder to provide supervision signals. (3) A **Distance-Weighted Diversity Loss** that prevents latent collapse by enforcing topology-aware constraints on the reasoning chain. Formally, let $\mathcal{M} = (\mathcal{V}, \mathcal{L})$ denote a Multimodal Large Language Model with vision encoder $\mathcal{V}$ and language model $\mathcal{L}$. Given an input image I and query Q, ProLaViT generates a sequence of latent thoughts $\mathcal{Z} = \{z_1, z_2, \ldots, z_K\}$ that progressively refine the visual understanding before producing the final answer A.

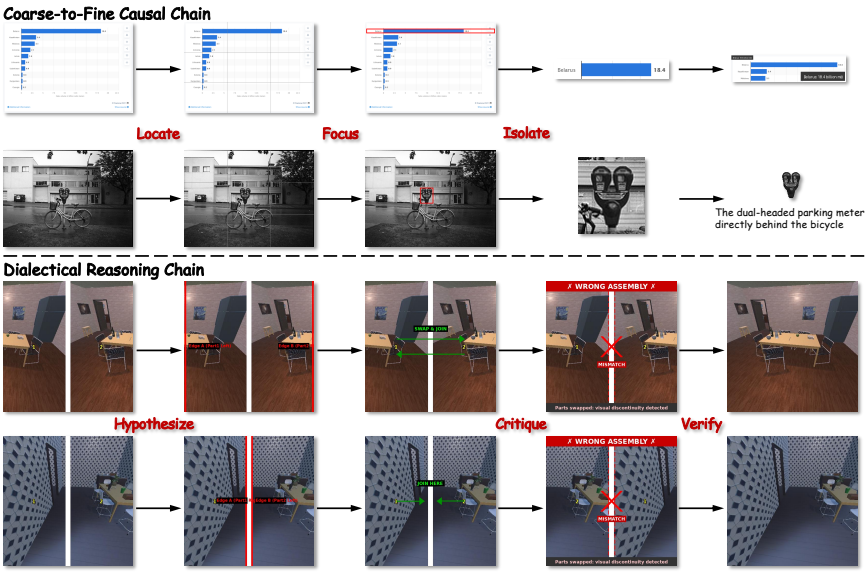


Fig. 3: Illustration of the two reasoning paradigms in ProLaViT. Top: The *Coarse-to-Fine Causal Chain* for spatial tasks progressively narrows attention from global context to specific targets (*Locate* → *Focus* → *Isolate*). **Bottom:** The *Dialectical Reasoning Chain* for logical tasks (e.g., jigsaw puzzles) employs a trial-and-error approach (*Hypothesize* → *Critique* → *Verify*) to validate visual consistency.

3.2 Progressive Latent Visual Thought

Unlike prior methods that attempt to predict complex visual cues in a single step, ProLaViT decomposes visual reasoning into a structured chain of latent thoughts, as illustrated in Fig. 3. We design two distinct reasoning paradigms tailored for different task types.

Coarse-to-Fine Causal Chain. For spatial tasks such as visual search and object localization, we adopt a *Locate* → *Focus* → *Isolate* paradigm (Fig. 3, Top). Each reasoning step progressively narrows the attention scope:

$$z_k = \mathcal{T}_k(z_{k-1}, I, c_k), \quad k \in \{1, 2, \ldots, K\}, \tag{1}$$

where $\mathcal{T}_k$ denotes the k-th reasoning transformation, and c_k represents the task-specific context. Specifically, we implement four canonical operations that instantiate this paradigm:

- **Grid View (z_1):** Establishes global context with spatial grid overlay (Preparation).
- **Bounding Box (z_2):** Performs the **Locate** step via region localization.
- **Crop (z_3):** Executes the **Focus** step for detailed inspection of the target region.

– **Segmentation (z_4):** Completes the **Isolate** step with fine-grained object boundary delineation.

Dialectical Reasoning Chain. For logical tasks such as jigsaw puzzle assembly, we employ a *Hypothesize* → *Critique* → *Verify* loop that incorporates counterfactual thinking (Fig. 3, Bottom). This paradigm enables the model to evaluate multiple hypotheses before committing to a solution. Specifically, we instantiate this loop through four distinct visual operations:

– **Edge Enhancement (z_1):** Highlights high-frequency boundary features to identify potential mating surfaces, forming the initial **Hypothesis**.
– **Directional Guidance (z_2):** Visualizes the proposed movement vector using arrow overlays to suggest the assembly orientation.
– **Counterfactual Simulation (z_3):** Generates a plausible but incorrect assembly state (e.g., misalignment) to trigger the **Critique** mechanism, teaching the model to distinguish subtle errors.
– **Solution Verification (z_4):** Presents the correctly assembled state to **Verify** the logical consistency and visual coherence of the final solution.

Token Representation. Each reasoning step is represented by a set of learnable tokens in the LLM's embedding space. Specifically, we allocate N_{tokens} special tokens (denoted as `<|vit_pad_latent|>`) per reasoning step. For $K = 4$ steps with $N_{\text{tokens}} = 4$ tokens each, the total latent thought sequence contains $K \times N_{\text{tokens}} = 16$ tokens. These tokens are embedded in the input sequence and trained to capture the intermediate visual states:

$$\mathbf{H}_{\text{latent}} = \text{LLM}(\mathbf{E}_{\text{input}} \oplus \mathbf{E}_{\text{thought}}), \tag{2}$$

where $\mathbf{E}_{\text{thought}} \in \mathbb{R}^{(K \cdot N_{\text{tokens}}) \times d}$ represents the latent thought embeddings, and $\oplus$ denotes concatenation.

3.3 Endogenous Self-Distillation Mechanism

A key challenge in training latent reasoning models is the lack of ground truth for intermediate mental images. Prior works rely on external vision experts (e.g., SAM, DINO), which introduces domain gaps and pipeline complexity. Instead, ProLaViT employs an Endogenous Self-Distillation strategy that leverages the MLLM's own frozen vision encoder as the teacher.

Programmatic Synthesis Pipeline. To generate supervision signals, we construct a scalable programmatic pipeline that produces rigorous visual trajectories. For each training sample, we programmatically generate K auxiliary images $\{I_1^{\text{aux}}, I_2^{\text{aux}}, \ldots, I_K^{\text{aux}}\}$ corresponding to each reasoning step (e.g., applying `crop(I, bbox)` or `segment(I, target)`). These auxiliary images are generated automatically using geometric transformations and are guaranteed to be algorithmically precise. Specific implementation details are provided in the Supplementary Materials.

Teacher Feature Extraction. The frozen vision encoder $\mathcal{V}$ extracts features from each auxiliary image:

$$\mathbf{F}_k^{\text{teacher}} = \mathcal{V}(I_k^{\text{aux}}), \quad k \in \{1, \ldots, K\}, \tag{3}$$

where $\mathbf{F}_k^{\text{teacher}} \in \mathbb{R}^{L \times d}$ with L denoting the number of visual tokens per image. Crucially, gradients are not backpropagated through $\mathcal{V}$, ensuring the vision encoder remains frozen during training.

Cross-Attention Knowledge Distillation. To align the LLM-generated latent thoughts with the teacher features, we employ a single cross-attention over the full sequence. Let $\mathbf{H} \in \mathbb{R}^{N \times d}$ denote the hidden states of all reasoning-step tokens (e.g., $N = K \times n_{\text{tok}}$ with n_{tok} tokens per step). We project and normalize:

$$\mathbf{H}^{\text{proj}} = \text{Normalize}\big(\text{Linear}(\mathbf{H})\big). \tag{4}$$

A learnable query matrix $\mathbf{Q} \in \mathbb{R}^{L \times d}$ (one per image length L) is expanded to length $K \cdot L$ (e.g., via linear interpolation) so that the number of query positions matches the concatenated teacher sequence. The aligned representation is computed in one cross-attention:

$$\hat{\mathbf{F}} = \text{CrossAttn}\big(\mathbf{Q}_{\text{expand}}, \mathbf{H}^{\text{proj}}, \mathbf{H}^{\text{proj}}\big), \quad \hat{\mathbf{F}} \in \mathbb{R}^{K \cdot L \times d}, \tag{5}$$

where $\mathbf{Q}_{\text{expand}} \in \mathbb{R}^{K \cdot L \times d}$. Key and value are both $\mathbf{H}^{\text{proj}}$, thus every query position (across all K steps) attends to the same compressed latent tokens. We split $\hat{\mathbf{F}}$ into K segments of length L and denote the k-th segment by $\hat{\mathbf{F}}_k \in \mathbb{R}^{L \times d}$. The self-distillation loss is:

$$\mathcal{L}_{\text{distill}} = \frac{1}{K} \sum_{k=1}^{K} \big\| \hat{\mathbf{F}}_k - \mathbf{F}_k^{\text{teacher}} \big\|_2^2. \tag{6}$$

3.4 Distance-Weighted Diversity Loss

In our experiments, we observe a limitation in the sequential reasoning of ProLaViT, where representations of subsequent reasoning steps gradually become indistinguishable. This phenomenon is highlighted by the token similarity matrix in Fig. 4. Such convergence risks degrading the progressive reasoning chain into redundancy, a failure mode we term *latent collapse*. To mitigate the aforementioned degradation of the reasoning chain, we propose introducing a diversity loss during training. However, standard diversity losses often prove overly strict, as they indiscriminately penalize any positive similarity. Given that visual features naturally share common patterns (e.g., edges and textures), pursuing complete orthogonality is therefore unrealistic. Inspired by metric learning, we propose a margin-based diversity loss allowing reasonable similarity while penalizing representation collapse:

$$\mathcal{L}_{\text{div}}^{\text{margin}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max(0, \cos(\mathbf{g}_i, \mathbf{g}_j) - \tau), \tag{7}$$

where $\mathbf{g}_k$ denotes the centroid of the k-th reasoning step's tokens, $\mathcal{P}$ is the set of all step pairs, and $\tau \in [0, 1]$ is a learnable margin threshold. A key insight is that reasoning steps with larger causal distance should be more distinct. To capture this, we introduce a distance-weighted penalty:

$$w_{ij} = \epsilon + (1 - \epsilon) \cdot \left(\frac{d(i, j)}{d_{\max}} \right)^{\alpha}, \tag{8}$$

where $d(i, j) = |i-j|$ is the causal distance, $d_{\max} = K-1$, $\epsilon = 0.1$ is the minimum weight, and $\alpha \geq 1$ is a learnable parameter controlling steepness, parameterized as $\alpha = 1 + \text{softplus}(\theta_\alpha)$. The final loss is:

$$\mathcal{L}_{\text{div}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} w_{ij} \cdot \max(0, \cos(\mathbf{g}_i, \mathbf{g}_j) - \tau). \tag{9}$$

This formulation respects the hierarchical structure of the reasoning chain while effectively preventing representation collapse.

3.5 Training Objective and Strategy

The overall training objective combines language modeling with latent supervision:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{distill}} \mathcal{L}_{\text{distill}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}, \tag{10}$$

where $\mathcal{L}_{\text{LM}}$ is the standard cross-entropy loss. We set $\lambda_{\text{distill}} = 1.0$ and $\lambda_{\text{div}} = 0.2$.

To ensure the stable emergence of progressive thoughts and prevent representation collapse, we employ a multi-stage curriculum learning strategy:

- **Phase 1: Latent Anchor Initialization.** We freeze the vision encoder and LLM backbone, training only the embeddings of the special thought tokens. This phase acts as a "warm-up," initializing the latent tokens to a neutral state within the semantic space of the LLM, preventing early optimization instability.
- **Phase 2: Endogenous Knowledge Distillation.** We activate the **Self-Distillation** mechanism ($\mathcal{L}_{\text{distill}}$). Here, the model learns to align its generated latent thoughts with the target visual features extracted by its own frozen encoder. Crucially, we do not yet enforce the diversity constraint, allowing the model to focus solely on accurate feature reconstruction.
- **Phase 3: Structured Chain Evolution.** We introduce the **Distance-Weighted Diversity Loss** ($\mathcal{L}_{\text{div}}$) alongside the distillation loss. This is the critical phase where the reasoning chain transforms from a repetitive sequence into a structured, progressive derivation. The model is forced to differentiate adjacent reasoning steps (e.g., separating *Locate* from *Focus*), establishing the causal topology in the latent space.
- **Phase 4: Global Capability Integration.** Finally, we perform joint training with general VQA data. This prevents catastrophic forgetting of general instruction-following abilities while consolidating the learned visual reasoning patterns into the model's global behavior.

4 Experiment

4.1 Experimental Setup

We implement ProLaViT based on Qwen2.5-VL-7B-Instruct. The latent thought uses $K = 4$ reasoning steps with $N_{\text{tokens}} = 4$ tokens per step. Training is performed on 2 NVIDIA H20 GPUs. For Diversity Loss, $\lambda_{\text{div}} = 0.2$, margin τ initialized to 0.8.

Benchmarks. We evaluate ProLaViT on a comprehensive suite of perception-intensive visual reasoning benchmarks that span multiple dimensions of visual understanding: **MMVP** [29], **VisPuzzle** [9], **VStar** [33], **ChartQA** [19], **BLINK** [7], **CV-Bench** [28].

Table 1: Performance comparison on visual reasoning benchmarks. **ProLaViT (Full)** denotes our model trained with the proposed Distance-Weighted Diversity Loss. All values are reported in accuracy (%). Best results are in **bold**, second best are underlined.

Method	MMVP	VisPuzzle	VStar	ChartQA	BLINK	CV-Bench	BLINK-J	Avg.
Base Model								
Qwen2.5-VL-Instruct	77.33	34.75	76.44	78.45	54.49	73.61	59.33	66.00
Text-Space Reasoning								
SFT	77.00	67.50	78.01	76.74	55.60	77.36	49.33	69.86
CoT SFT	77.66	69.75	75.91	75.83	56.54	64.55	68.66	69.18
Latent-Space Reasoning								
CoVT	78.33	35.75	79.05	24.99	**57.49**	75.00	66.00	61.45
LVR	77.30	35.00	76.43	64.86	53.02	76.55	57.33	64.63
One-step Latent Pred.	**79.33**	72.00	78.01	63.02	56.28	76.10	66.00	70.85
Ours (Progressive Latent Visual Thought)								
ProLaViT (Base)	78.00	74.00	79.05	76.18	**57.49**	77.24	71.33	73.82
ProLaViT + $\mathcal{L}_{\text{div}}^{\text{margin}}$	78.33	**75.25**	79.05	75.66	56.33	**78.33**	67.33	73.58
ProLaViT + $\mathcal{L}_{\text{div}}$ (Full)	79.00	74.25	**80.11**	**78.99**	57.23	77.66	**76.00**	**75.11**

Baselines. We compare ProLaViT against the following methods:

- **Qwen2.5-VL-Instruct** [2]: The base MLLM without any fine-tuning, serving as the foundation model.
- **SFT**: Standard supervised fine-tuning on the same training data without latent reasoning.
- **CoT SFT**: Fine-tuning with explicit textual Chain-of-Thought reasoning annotations.
- **CoVT** [23]: Chain-of-Visual-Thought that distills knowledge from external lightweight vision experts.
- **LVR** [16]: Latent Visual Reasoning that enables reasoning in the visual embedding space.
- **One-step Latent Prediction**: An ablated variant that predicts all visual features in a single step without progressive decomposition.

4.2 Quantitative Results

Tab. 1 summarizes the performance comparison across diverse visual reasoning benchmarks. Our proposed framework, ProLaViT (equipped with the full Distance-Weighted Diversity Loss), consistently outperforms baselines, achieving the highest average accuracy of **75.11%**.

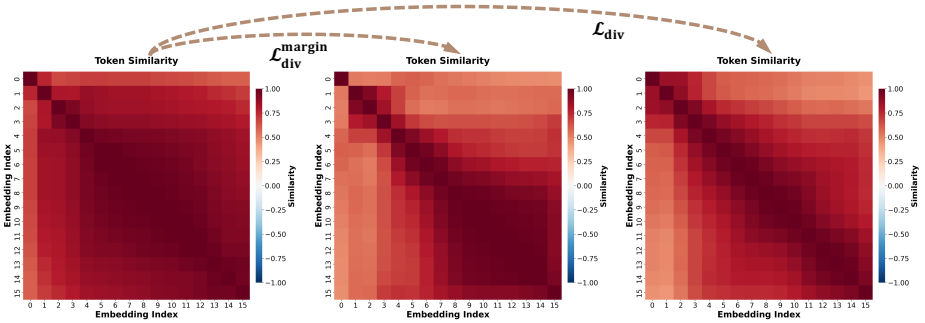
Analysis. We highlight four key observations from the results:

1. **Structured Decomposition vs. Monolithic Prediction.** Comparing ProLaViT with One-step Latent Prediction, we observe significant gains, particularly on complex tasks like VisPuzzle (+2.25%) and BLINK-Jigsaw (+10.00%). This confirms that monolithic latent prediction often hits a performance ceiling due to capacity overload. In contrast, our *Locate* $\rightarrow$ *Focus* $\rightarrow$ *Isolate* causal chain allows the model to incrementally refine its attention, leading to more precise visual grounding.
2. **Bridging the Modality Gap in Spatial Reasoning.** While CoT SFT improves performance on logical puzzles (VisPuzzle: 69.75%), it suffers a sharp decline on perception-intensive benchmarks (CV-Bench: 64.55%). This reflects the inherent limitation of text in capturing fine-grained spatial coordinates. ProLaViT overcomes this by reasoning in the continuous latent space, maintaining strong performance across both logical and perceptual tasks without the information loss associated with text projection.
3. **Robustness via Endogenous Alignment.** A critical failure mode of prior methods is revealed in the ChartQA benchmark. CoVT, which relies on external vision experts (e.g., SAM, DepthAnything), collapses to 24.99% accuracy, likely due to the domain gap between natural images (on which experts are trained) and abstract charts. ProLaViT avoids this pitfall by employing Endogenous Self-Distillation. By leveraging the MLLM's own vision encoder, our method ensures domain consistency, achieving a robust 78.99% on ChartQA.
4. **Efficacy of Dialectical Reasoning.** On the BLINK-Jigsaw benchmark, which demands rigorous logical verification, ProLaViT achieves 76.00%, a remarkable +16.67% improvement over the base model. This substantial margin validates the effectiveness of our *Hypothesize* $\rightarrow$ *Critique* $\rightarrow$ *Verify* paradigm, demonstrating that structured latent thoughts can successfully simulate trial-and-error processes for complex problem-solving.

4.3 Qualitative Analysis

To intuitively understand how ProLaViT structures its reasoning process, we visualize the cosine similarity matrices of the generated latent tokens in Fig. 4.

Latent Collapse without Constraints. As shown in Fig. 4(a), in the absence of diversity constraints, the model exhibits severe *latent collapse*. The representations for steps z_2 through z_4 share excessive similarity, effectively degenerating the progressive chain into redundancy. This explains why the baseline model fails to refine its attention, it is essentially stuck in the initial state.



(a) No Diversity Loss (b) Margin-Based Formulation (c) Distance-Weighted Loss

Fig. 4: Visualization of latent thought similarity. We plot the cosine similarity heatmaps between the token representations of different reasoning steps (z_1 to z_4). **(a) No Diversity Loss:** The chain suffers from *latent collapse*, where subsequent steps (z_2-z_4) become indistinguishable. **(b) Margin-Based Formulation:** Enforces distinctiveness equally across all pairs, potentially disrupting the semantic continuity of adjacent steps. **(c) Distance-Weighted Loss (Ours):** Respects the hierarchical structure, where adjacent steps retain moderate similarity for context, while distant steps are enforced to be distinct.

Margin-Based vs. Distance-Weighted Constraints. Applying the Margin-Based Formulation (Fig. 4(b)) successfully reduces global similarity, but it indiscriminately penalizes adjacent steps. This disrupts the natural pattern sharing required for causal reasoning. In contrast, our Distance-Weighted Diversity Loss (Fig. 4(c)) induces a structure consistent with our hypothesis: adjacent steps maintain moderate similarity to preserve context, while distant steps (e.g., z_1 vs. z_4) are forced to be distinct. This confirms that ProLaViT learns a non-redundant, progressive derivation path where steps with larger causal distance contribute unique perceptual information.

4.4 Ablation Study

To validate the effectiveness of our core design choices, we conduct comprehensive ablation studies focusing on two key aspects: the necessity of progressive decomposition and the impact of the distance-weighted diversity constraint.

Effect of Progressive Decomposition. We first validate the structural advantage of our multi-step derivation over monolithic prediction. As shown in Tab. 2, the One-step Latent Prediction baseline, which compresses all visual cues into a single latent state, suffers from a significant performance bottleneck. In contrast, ProLaViT's progressive approach yields substantial gains, particularly on tasks requiring logical depth such as ChartQA (+15.97%) and BLINK-Jigsaw (+10.00%). This empirical evidence confirms that complex visual reasoning overloads the capacity of a single latent state, whereas our structured chain (*Locate* $\rightarrow$ *Focus* $\rightarrow$ *Isolate*) effectively distributes the cognitive load, allowing for incremental and precise visual grounding.

Effect of Distance-Weighted Diversity Loss. Next, we isolate the impact of our diversity constraints. We first evaluate the Margin-Based Formulation ($\mathcal{L}_{\text{div}}^{\text{margin}}$) without distance weighting. While this formulation mitigates *latent collapse* by penalizing similarity above the threshold τ, it treats all reasoning step pairs equally. As observed in Tab. 3, this rigid constraint leads to a performance drop on BLINK-Jigsaw (-4.00%), likely because it disrupts the natural feature sharing (e.g., edges, textures) between adjacent steps. In contrast, our full Distance-Weighted Diversity Loss ($\mathcal{L}_{\text{div}}$) achieves the best overall performance. By incorporating the causal distance weight w_{ij}, it respects the hierarchical structure of the reasoning chain, it assigns lower penalty weights to adjacent steps to preserve context, while enforcing that steps with larger causal distance are more distinct. This design effectively prevents representation collapse while maintaining semantic continuity, as evidenced by the robust improvements on ChartQA (+2.81%) and BLINK-Jigsaw (+4.67%).

Table 2: Ablation on reasoning structure: **One-step Prediction** vs. **Progressive Latent Visual Thought** (Ours). The progressive decomposition yields consistent gains, verifying that complex visual reasoning requires multi-step derivation.

Structure	VisPuzzle	ChartQA	BLINK-J	CV-Bench	Avg.
One-step Latent Pred.	72.00	63.02	66.00	76.10	69.28
ProLaViT (Progressive)	**74.25**	**78.99**	**76.00**	**77.66**	**76.73**
Improvement	+2.25	+15.97	+10.00	+1.56	+7.45

Table 3: Ablation on diversity constraints. $\mathcal{L}_{\text{div}}^{\text{margin}}$: **Margin-Based Formulation** (without distance weights). $\mathcal{L}_{\text{div}}$: **Distance-Weighted Diversity Loss** (Ours).

Configuration	VisPuzzle	VStar	ChartQA	BLINK-J	Avg.
ProLaViT (No Diversity Loss)	74.00	79.05	76.18	71.33	75.14
+ $\mathcal{L}_{\text{div}}^{\text{margin}}$	75.25	79.05	75.66	67.33	74.32
+ $\mathcal{L}_{\text{div}}$ (Full)	**74.25**	**80.11**	**78.99**	**76.00**	**77.34**

Latent Reasoning vs. Text-Trace Supervision. To disentangle the benefit of latent-space reasoning from the programmatic training data, we convert the same synthesized trajectories into textual traces (`Crop [x1,y1,x2,y2]`→`Seg obj`) and fine-tune the base model. Text-Trajectory SFT gains only +0.9% over CoT SFT (73.41 vs. 72.54 avg.), while ProLaViT gains +4.8% (77.34 avg.). The net +3.9% arises purely from reasoning in continuous latent space, since text-tokenized coordinates discard sub-pixel appearance cues that ProLaViT's latent chain preserves.

Effect of Endogenous vs. Exogenous Distillation. Finally, we examine the necessity of our Endogenous Self-Distillation mechanism. A common alternative is to use external vision models to supervise the latent thoughts. To test this, we replace our native teacher (the Qwen2.5-VL vision encoder) with two representative external models: (1) DINOv2-Large [21] (ViT-L/14), a leading discriminative model; (2) SDXL-VAE [22], a generative autoencoder used in diffusion models. As shown in Table 4, our endogenous approach consistently outperforms both external variants. While DINOv2 performs well on basic perception tasks, it falls behind on fine-grained benchmarks like VStar. This suggests that aligning latent thoughts with a different feature space creates an *alignment gap*, whereas the native encoder offers naturally aligned supervision. Notably, SDXL-VAE shows a significant performance drop on VisPuzzle, despite good results on MMVP. This highlights a key limitation: VAE latent spaces are optimized for reconstructing pixels and textures, often sacrificing precise spatial logic. Consequently, structural details are lost during reasoning. These results confirm that Endogenous Self-Distillation is not only more efficient but also provides the most robust supervision for visual reasoning.

Table 4: Ablation on teacher selection: **Endogenous** (Native Encoder) vs. **Exogenous** (External Experts). While external experts perform adequately on basic perception (MMVP), they struggle with complex spatial reasoning (VisPuzzle), confirming the superiority of endogenous alignment.

Teacher Model	Type	MMVP	VisPuzzle	VStar	BLINK-J	Avg.
DINOv2 (ViT-L/14)	Discriminative	78.00	73.75	78.53	72.00	75.57
SDXL-VAE	Generative	78.66	45.50	75.91	72.00	68.02
ProLaViT (Native ViT)	**Endogenous**	**79.00**	**74.25**	**80.11**	**76.00**	**77.34**

Table 5: Ablation on reasoning step ordering. Random permutation validates the importance of the canonical coarse-to-fine progression.

Ordering	VisPuzzle	VStar	ChartQA	BLINK-J	CV-Bench	Avg.
Canonical Order	**74.25**	**80.11**	**78.99**	**76.00**	**77.66**	**77.40**
Random Order	68.00	74.50	73.50	70.66	69.14	71.16
Drop	-6.25	-5.61	-5.49	-5.34	-8.52	-6.24

Effect of Reasoning Step Ordering. To validate that the causal ordering of reasoning steps is critical, we randomly permute the order of the four steps while keeping all other components fixed. As shown in Tab. 5, the Canonical Order ($z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow z_4$) achieves superior performance (77.40% average), while Random Order suffers a substantial drop (-6.24% average), particularly

on spatial tasks like VisPuzzle (-6.25%) and CV-Bench (-8.52%). These findings provide strong evidence that the *progressive causal structure* is a fundamental requirement for effective visual reasoning.

Efficiency: Latent Reasoning vs. Explicit Tool-Use. A natural question is whether latent reasoning is preferable to explicitly invoking visual tools at each step. We fine-tune Qwen2.5-VL on identical traces with explicit tool tokens that invoke real operations at inference. As shown in Tab. 6, explicit tool-use re-invokes the ViT encoder every step, inducing 3.4× latency and +117% TFLOPs. In contrast, ProLaViT performs a single ViT pass and reasons entirely in latent space (1.21× latency, +12% TFLOPs) while achieving higher accuracy (VStar 80.11 vs. 79.60, BLINK-J 76.00 vs. 74.00), since latent reasoning avoids the feature loss from iterative image re-encoding. Training requires only ∼28 GPU-hours on 2×H20, compared to over 120h for generative-CoT methods [9].

Table 6: Explicit Tool-Use vs. Latent Distillation.

Method	VStar	BLINK-J	Latency	TFLOPs
Qwen2.5-VL (Base)	76.44	59.33	1.00×	1.00
Explicit Tool-Use SFT	79.60	74.00	3.40×	2.17
ProLaViT (Ours)	**80.11**	**76.00**	**1.21×**	**1.12**

5 Conclusion

In this paper, we presented ProLaViT, a novel framework empowering Multimodal Large Language Models to perform visual reasoning via progressive latent derivation. By transitioning reasoning from discrete text to a structured continuous latent space, ProLaViT effectively bridges the modality gap that limits traditional Chain-of-Thought approaches. Central to our contribution is the Endogenous Self-Distillation mechanism, which enables the model to internalize rigorous algorithmic logic from programmatically synthesized data, thereby eliminating reliance on computationally expensive external vision experts. Furthermore, our proposed Distance-Weighted Diversity Loss mitigates the critical issue of latent collapse, enforcing topological distinctiveness across the reasoning chain to ensure robust, non-redundant visual thoughts. Extensive experiments on perception-intensive benchmarks, including visual search and jigsaw assembly, demonstrate that ProLaViT achieves state-of-the-art performance with superior inference efficiency. While this work focuses on spatial and logical reasoning in static images, the principle of structured latent derivation holds significant potential for broader domains. Future work will extend ProLaViT to temporal video reasoning and complex geometric transformations, paving the way for more generalizable and interpretable multimodal intelligence.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S.: Qwen2.5-vl technical report (2025)
3. Chern, E., Su, J., Ma, Y., Liu, P.: Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135 (2024)
4. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
5. Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., Zhang, X., Ma, K., Yi, L.: DreamLLM: Synergistic multimodal comprehension and creation. In: The Twelfth International Conference on Learning Representations (2024)
6. Dong, S., Wang, S., Liu, X., Li, C., Hou, H., Wei, Z.: Interleaved latent visual reasoning with selective perceptual modeling. arXiv preprint arXiv:2512.05665 (2025)
7. 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. ArXiv **abs/2404.12390** (2024)
8. Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023)
9. 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)
10. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J.E., Tian, Y.: Training large language models to reason in a continuous latent space. ArXiv **abs/2412.06769** (2024)
11. Hao, Y., Gu, J., Wang, H.W., Li, L., Yang, Z., Wang, L., Cheng, Y.: Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. arXiv preprint arXiv:2501.05444 (2025)
12. Jiang, D., Zhang, R., Guo, Z., Li, Y., Qi, Y., Chen, X., Wang, L., Jin, J., Guo, C., Yan, S., et al.: Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621 (2025)
13. 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.B.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3992–4003 (2023)
14. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models (2023)
15. Li, A., Wang, C., Fu, D., Yue, K., Cai, Z., Zhu, W.B., Liu, O., Guo, P., Neiswanger, W., Huang, F., et al.: Zebra-cot: A dataset for interleaved vision language reasoning. arXiv preprint arXiv:2507.16746 (2025)
16. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. ArXiv **abs/2509.24251** (2025)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023)

18. Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.W., Wu, Y.N., Zhu, S.C., Gao, J.: Chameleon: Plug-and-play compositional reasoning with large language models. ArXiv **abs/2304.09842** (2023)
19. Masry, A., Long, D.X., Tan, J.Q., Joty, S.R., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. ArXiv **abs/2203.10244** (2022)
20. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L.: Gpt-4 technical report (2024)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023)
22. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023)
23. 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)
24. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems 37 (2024)
25. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face (2023)
26. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: Codi: Compressing chain-of-thought into continuous space via self-distillation. arXiv preprint arXiv:2502.21074 (2025)
27. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning (2023)
28. Team, C.: Cvbench: A benchmark for cross-video multimodal reasoning (2025)
29. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9568–9578 (2024)
30. 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. arXiv preprint arXiv:2511.21395 (2025)
31. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., Zhou, D.: Chain of thought prompting elicits reasoning in large language models. ArXiv **abs/2201.11903** (2022)
32. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models (2023)
33. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 13084–13094 (2023)
34. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2087–2098 (October 2025)

35. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10371–10381 (2024)
36. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens. ArXiv **abs/2506.17218** (2025)
37. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action (2023)
38. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models (2024)