
CoCoVa: CHAIN OF CONTINUOUS VISION-LANGUAGE THOUGHT FOR LATENT SPACE REASONING

A PREPRINT

Jizheng Ma^{1,2}, Xiaofei Zhou^{1,2,†}, Yanlong Song^{1,2}, Han Yan^{1,2}

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences

[†]Corresponding author: zhouxiaofei@iie.ac.cn

ABSTRACT

In human cognition, there exist numerous thought processes that are tacit and beyond verbal expression, enabling us to understand and interact with the world in multiple ways. However, contemporary Vision-Language Models (VLMs) remain constrained to reasoning within the discrete and rigid space of linguistic tokens, and neglect vision tokens that are necessary for reasoning. This discrete constraint restricts their ability to capture the continuous, high-dimensional subtleties inherent in visual perception and abstract thought. To bridge this gap, we propose CoCoVa (Chain of Continuous Vision-Language Thought), a novel framework for vision-language model that leverages continuous cross-modal reasoning for diverse vision-language tasks. The core of CoCoVa is an iterative reasoning cycle, orchestrated by our Latent Q-Former (LQ-Former) module, which functions as a dynamic reasoning engine rather than a static feature projector. This cycle iteratively refines a chain of latent thought vectors through gated cross-modal fusion, dynamically guided by a token selection mechanism that mimics attentional focus. To ensure semantic grounding, we employ a multi-task objective combining contrastive learning and diffusion-based reconstruction, enforcing alignment between latent representations and both visual and textual modalities. Empirically, CoCoVa demonstrates competitive performance across diverse vision-language understanding and reasoning benchmarks. CoCoVa notably improves reasoning accuracy while maintaining greater token efficiency compared to leading multimodal reasoning baselines. Based on a 1.5B-parameter LLM backbone, CoCoVa matches or surpasses the performance of 7B-scale models such as LLaVA-NeXT on four out of five benchmarks, while notably outperforming the 9B-parameter Qwen-VL-Chat on MMStar and MMVP. When scaled to different 7B LLM backbones, CoCoVa retains performance competitiveness even compared to state-of-the-art models. Further qualitative analysis validates that the learned latent space captures interpretable and structured reasoning patterns, highlighting the potential of CoCoVa to bridge the representational gap between discrete language processing and the continuous nature of visual understanding. Code and checkpoints will be released later.

We know more than we can tell.

—Michael Polanyi

1 Introduction

Vision-Language Models (VLMs) have achieved notable success across diverse multimodal domains, including visual-language understanding and reasoning[Xu et al., 2023, Li et al., 2025a]. A prevalent technique enhancing this success is the integration of Chain-of-Thought (CoT) reasoning[Wei et al., 2022, Zhou et al., 2024, Yao et al., 2023, OpenAI et al., 2024, Zhang et al., 2024a, Wang et al., 2022], which guides models to decompose problems and articulate intermediate reasoning steps in natural language. Despite its effectiveness, standard CoT reasoning confines the model’s

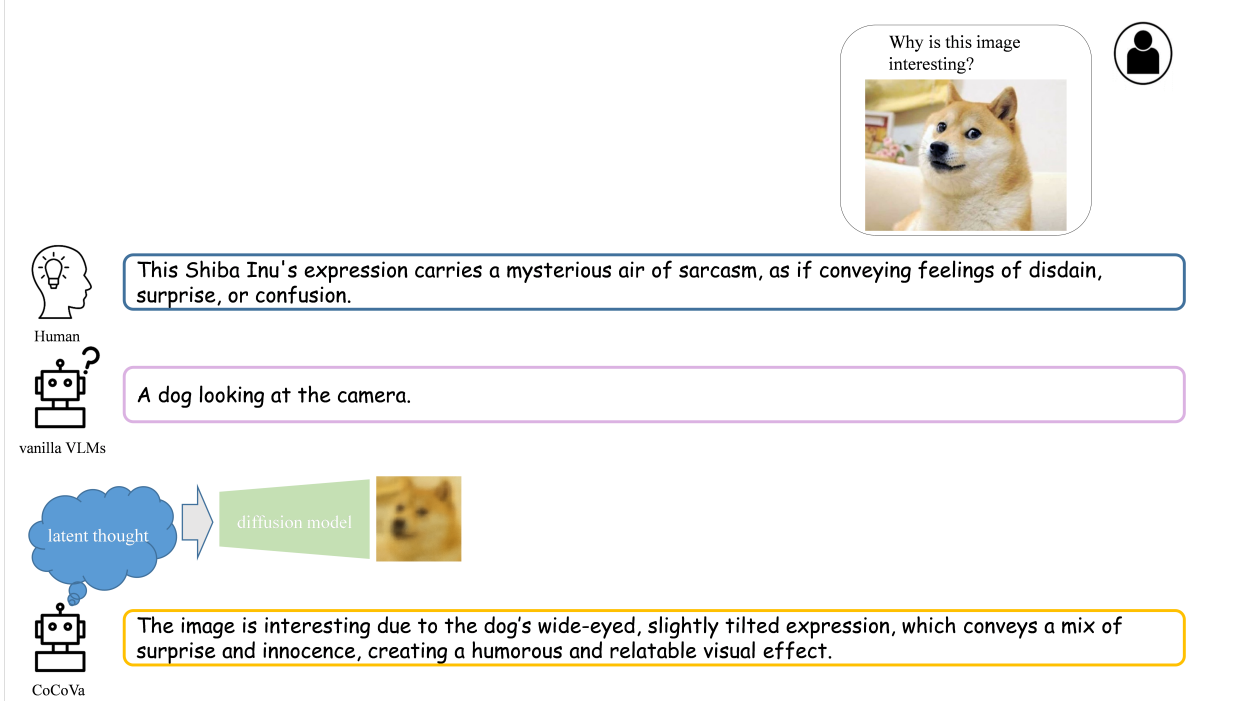


Figure 1: Illustrative comparison of interpretability across humans, vanilla vision–language models (VLMs), and our proposed CoCoVa framework. While humans perceive nuanced affective meanings and vanilla VLMs give shallow literal captions, CoCoVa leverages latent reasoning to produce richer and more context-aware interpretations and is able to yield both coarse visual reconstructions and richer textual interpretations with a diffusion-based reconstructor and LLM backbone.

internal processing to a sequence of discrete, predefined linguistic tokens, inherently bounded by the expressive limits of vocabulary[Hao et al., 2024, Shen et al., 2025a]. This paradigm also fails to adequately capture the continuous, high-dimensional nature of visual perception[Wei et al., 2021], forcing the rich information from image tokens through a restrictive linguistic bottleneck. Confining LLM reasoning to the discrete natural language tokens may fundamentally restrict their potential to represent and manipulate abstract multimodal concepts[Li et al., 2025b, Zhang et al., 2025a].

This limitation stands in stark contrast to human reasoning. Neuroscientific evidence suggests that the human brain often operates through a dynamic, context-dependent "stream of thought" within a high-dimensional conceptual space [Tanenhaus et al., 1995, Huth et al., 2012]. This implicit cognitive process—fluid and efficient—resists complete description using discrete symbols [Herd et al., 2021, Barack and Krakauer, 2021]. Therefore, to bridge this gap between artificial and human intelligence, a reasoning paradigm that operates within a continuous latent space is desirable.

The practical implications of this representational mismatch are readily apparent. Consider the popular “Doge” meme (Figure 1). Humans effortlessly perceive high-level attributes such as sarcasm, amusement, or internal monologue—nuances that exist in a continuous affective spectrum. In contrast, vanilla VLMs typically provide shallow, literal descriptions (e.g., “A dog looking at the camera”), failing to capture these latent cultural and affective dimensions. This example underscores a fundamental weakness: when reasoning is shackled to a discrete linguistic space, models struggle with the tacit, understood-but-not-verbalized knowledge that is crucial for deep visual understanding.

To bridge this gap, we propose CoCoVa (Chain of Continuous Vision-Language Thought), a framework that enables VLMs to reason within a continuous, high-dimensional latent space, imitating the fluid nature of human cognition. The core design principle of CoCoVa is the explicit separation of internal thought from external expression. Instead of being constrained to natural language tokens at every step, the model engages in a multi-step latent reasoning cycle. In this cycle, the model iteratively aggregates multimodal information to form latent thoughts, corresponding to high-dimensional vectors that are semantically aligned with both visual and textual modalities. A key innovation is that these thoughts are verifiable; they can be projected back into image reconstructions, providing a window into the model’s reasoning process. As illustrated in Figure 1, this approach allows CoCoVa to move beyond literal captioning and generate nuanced, human-aligned interpretations. By operating in a continuous space, CoCoVa circumvents the

discrete bottleneck of language, offering a more expressive substrate for the tacit knowledge crucial for advanced vision-language understanding.

From an architectural perspective, CoCoVa composes a set of practical, interoperable modules that realize the latent reasoning cycle. Like vanilla VLMs, CoCoVa builds upon pretrained vision encoders and Large Language Models (LLMs), connecting them through a Multilayer Perceptron (MLP). On top of this backbone we propose a novel multimodal latent fusion module called LQ-Former, which is based on a Q-Former[Li et al., 2023a] style cross-attention as its attention backbone. The LQ-Former takes as input the projected visual features and the last hidden states from the LLM’s every layer, and performs deep cross-modal fusion through its attention mechanism. This module orchestrates a multi-step latent reasoning cycle: at each step, it refines a continuous thought vector by attending to both visual information and the evolving contextual representation from the LLM. The output of this iterative process is a chain of continuous vision-language thoughts, a trajectory of latent states that embodies the model’s internal reasoning pathway. To ensure these latent thoughts remain semantically grounded, we employ a multi-objective training framework. To reconstruct latent thoughts into their corresponding representations, we employ a dual-pathway decoding approach: a diffusion[Sohl-Dickstein et al., 2015, Ho et al., 2020] model, for image reconstruction and the LLM backbone for textual decoding. Simultaneously, a symmetric InfoNCE[van den Oord et al., 2019] loss maximizes mutual information between the latent thoughts and the corresponding visual and textual representations, ensuring semantic alignment across modalities. This design allows the model to dynamically re-examine and integrate visual evidence over multiple steps before committing to a final, linguistically articulated output, thereby overcoming the single-pass limitation of standard VLMs.

Empirically, CoCoVa achieves competitive performance, outperforming strong baselines in both accuracy and token efficiency. Experimental results indicate that CoCoVa achieves consistent improvements over strong baselines, with the 1.5B backbone attaining an accuracy of 63.3% on MMBench while generating fewer tokens than competing baseline methods. Notably, our 1.5B model matches or exceeds the performance of 7B-scale models like LLaVA-NeXT on evaluated benchmarks. Furthermore, CoCoVa scales effectively to different 7B LLM backbones, maintaining competitive performance compared to state-of-the-art VLMs across diverse tasks. Ablation studies confirm the necessity of each component, and further analysis reveals semantically structured organization in the latent space. The main contributions of our work are as follows:

- We introduce CoCoVa, a novel framework that enables continuous latent space reasoning for vision-language tasks within a continuous, semantically-aligned space. By shifting from discrete token sequences to a continuous, high-dimensional reasoning space, CoCoVa effectively bridges the gap between the fluid nature of human thought and the rigid symbolic processing of conventional VLMs.
- We evolve the conventional Q-Former into our LQ-Former, shifting its role from a static connector for single-pass projection to that of a dynamic, iterative reasoning engine. This module enables deep, iterative cross-modal fusion of visual features and linguistic context, forming a verifiable chain of continuous vision-language thoughts.
- Experiments demonstrate that CoCoVa consistently outperforms strong baselines across diverse benchmarks. Furthermore, qualitative assessments reveal that the learned latent thoughts form semantically meaningful structures and enable verifiable reasoning trajectories, offering valuable insights for future exploration in multimodal latent space reasoning.

2 Related Work

Vision Language Models. In recent years, Vision-Language Models (VLMs) have made substantial progress in multimodal understanding and reasoning[Zhang et al., 2024b,c]. A common architectural paradigm has emerged for modern VLMs, which typically comprise three core components, including a Large Language Model (LLM), a visual module, and a connector that aligns the two modalities. In typical VLM architectures, a connector module is employed to map visual representations into the semantic space of the LLM. Standard choices for this role encompass Resamplers[Alayrac et al., 2022], Q-Formers[Li et al., 2023a, Bai et al., 2023], and MLPs[Liu et al., 2023a, 2024a, Li et al., 2024, Liu et al., 2024b]. While CoCoVa’s architecture may appear to synthesize elements from MLP and Q-Former based paradigms, such a view obscures its reconceptualization of the connector’s role. Standard VLMs utilize their connector for a single-purpose, one-off projection[Liu et al., 2023a, Li et al., 2023a] to transform a fixed set of visual features into a static set of visual tokens for the LLM to process. In contrast, CoCoVa repurposes and extends the Q-Former into an LQ-Former to serve a radically different function: that of a dynamic, iterative reasoning engine. This architectural innovation effectively decouples internal cross-modal reasoning from external linguistic expression, establishing a continuous computational substrate for the model’s thought process that is absent in single-pass projection-based VLMs.

Multimodal Chain-of-Thought Reasoning. To enhance the reasoning capabilities of Vision-Language Models (VLMs), multimodal Chain-of-Thought (CoT) methods have been developed, mirroring the technique used in LLMs by generating a sequence of intermediate reasoning steps[Wang et al., 2025a]. These approaches can be broadly categorized into prompt-based methods that elicit reasoning through in-context examples[Chen et al., 2023, Luo et al., 2024, Fei et al., 2024], plan-based methods that structure the reasoning process dynamically[Zheng et al., 2024, Yang et al., 2024], and learning-based methods that fine-tune models on datasets with annotated rationales[Lu et al., 2022a, Wang et al., 2024a, Tan et al., 2024]. A significant advancement is the incorporation of multimodal rationales[Wang et al., 2025a], where reasoning steps are not limited to text but may include visual tokens or sketches[Su et al., 2025a, Wei et al., 2025, Su et al., 2025b], in order to bridge logical gaps and support spatial understanding[Su et al., 2025c]. Despite the incorporation of multimodal rationales during reasoning, these methods remain predominantly reliant on discrete textual tokens, which constrains their potential for representing abstract multimodal concepts. Moreover, the effectiveness of these approaches is also fundamentally limited by their dependence on the quality and availability of external tools. In response to these limitations, CoCoVa introduces a continuous latent reasoning paradigm that iteratively fuses visual and linguistic context, eliminating the discrete bottleneck and dependency on external tools.

Latent Space Reasoning. Recent advances have shifted from explicit reasoning toward latent space reasoning, where intermediate steps are carried out internally in the latent representation space without being verbalized[Zhu et al., 2025]. Existing approaches to latent reasoning[Li et al., 2025b] encompass latent optimization across token-level[Tack et al., 2025, Sun et al., 2025, Gong et al., 2025] or trajectory-level[Hao et al., 2024, Shen et al., 2025b,a, Zhang et al., 2025a, Kong et al., 2025], signal-guided control[Herel and Mikolov, 2024, Zelikman et al., 2024, Li et al., 2025c] via thinking or pause tokens and layer-recurrent execution[Saunshi et al., 2025, Mohtashami et al., 2025] through iterative transformer forwarding. However, these pioneering methods remain predominantly confined to the unimodal, textual domain, operating solely within the linguistic manifold of the LLM itself. CoCoVa addresses this critical gap by extending latent space reasoning to the visual domain, introducing a continuous, semantically-aligned reasoning space where iterative cross-modal fusion occurs. Recent efforts have also explored latent space reasoning for visual modalities, such as Machine Mental Imagery (Mirage)[Yang et al., 2025] and Latent Visual Reasoning (LVR)[Li et al., 2025d], which attempt to ground intermediate reasoning in visual tokens or reconstructions. Yet these methods exhibit a lack of deep integration between visual and textual modalities, and typically rely on region-of-interest (ROI) supervision, limiting expressiveness and scalability. By contrast, CoCoVa establishes a deeply fused, continuous vision-language reasoning process through its LQ-Former module, which dynamically attends to and refines visual tokens within the latent space without requiring explicit ROI supervision.

3 Preliminaries

Vision-Language Model. VLMs typically consist of three core components: a visual encoder, a large language model (LLM), and a connector module that bridges the two modalities. Formally, given an input image and a text prompt, the visual encoder (e.g., CLIP [Radford et al., 2021]) first extracts visual features $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\} \in \mathbb{R}^{N \times d_v}$, where N is the number of visual tokens and d_v is the visual feature dimension. Simultaneously, the text prompt is tokenized and embedded into a sequence of token embeddings $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_M\} \in \mathbb{R}^{M \times d_t}$.

The connector module[Liu et al., 2023a, Li et al., 2023a]) then projects the visual features into the language model’s embedding space:

$$\mathbf{V}_T = \text{Connector}(\mathbf{V}) \in \mathbb{R}^{N' \times d_t} \quad (1)$$

where N' is the number of projected visual tokens and d_t is the LLM’s hidden dimension. These projected visual tokens $\mathbf{V}_T$ are then prepended to the text token embeddings and fed into the LLM to generate the response:

$$\mathbf{Y} = \text{LLM}([\mathbf{V}_T; \mathbf{T}]) \quad (2)$$

While effective for many tasks, this standard VLM architecture processes visual information in a single forward pass[Li et al., 2025d], often unable to capture the visual details most relevant to the text query.

Multimodal Chain-of-Thought Reasoning. Chain-of-Thought (CoT) reasoning has emerged as a powerful technique to enhance the reasoning capabilities of language models by decomposing complex problems into intermediate reasoning steps. In the context of VLMs, multimodal CoT extends this approach by generating a sequence of reasoning steps $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_K\} \in \mathbb{R}^{N \times d_t}$ before producing the final answer $\mathbf{a}$:

$$P(\mathbf{a} | \mathbf{V}_T, \mathbf{T}) = \prod_{i=1}^{|\mathbf{a}|} P(a_i | \mathbf{V}_T, \mathbf{T}, \mathbf{R}, a_{<i}) \quad (3)$$

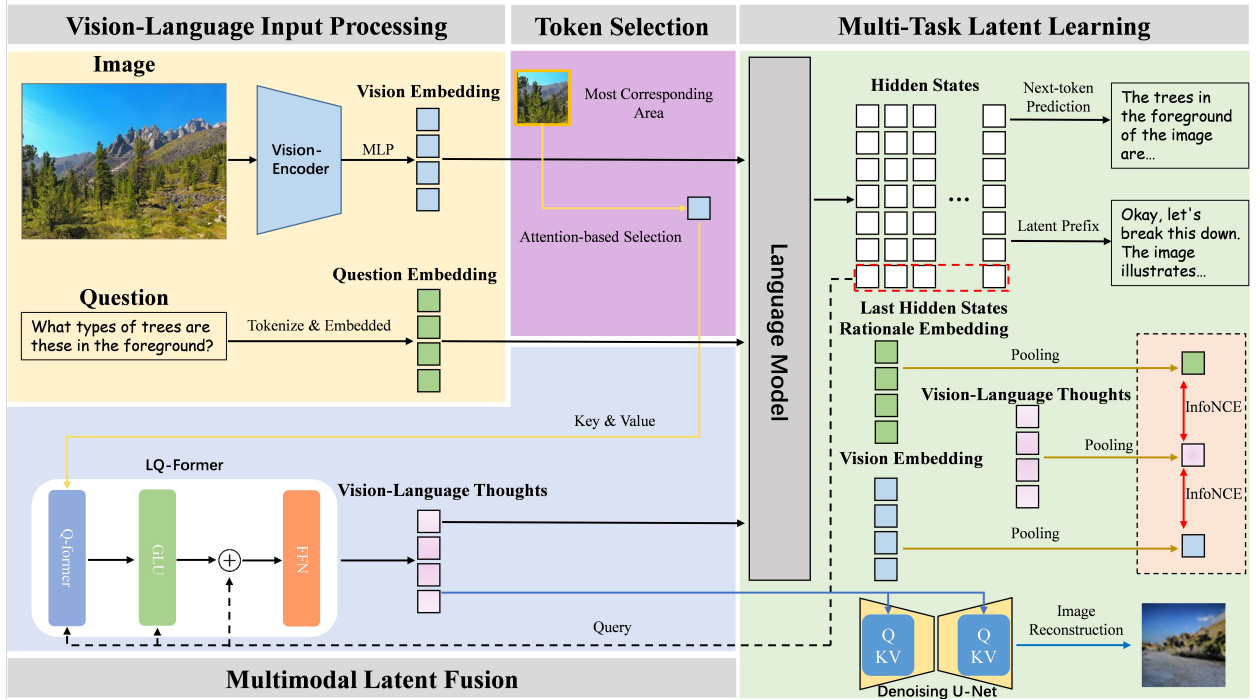


Figure 2: Overview of CoCoVa, which contains three core modules: (I) **Token Selection** — the model aggregates LLM attention and applies a $w \times w$ sliding window to select the most salient visual tokens; (II) **Multimodal Latent Fusion** — the LQ-Former takes the selected visual tokens and LLM’s last hidden states as inputs, iteratively generating latent thoughts that integrate visual and linguistic information over K reasoning steps; (III) **Multi-Task Representation Learning** — aims to learn a unified latent representation that consistently aligns visual and linguistic information across multiple tasks, encouraging vision-language models to form an interpretable and transferable multimodal reasoning capability.

The reasoning steps r_i are typically represented as discrete natural language tokens, making the reasoning process explicit and interpretable. However, this discrete representation inherently constrains the expressive capacity of the reasoning process, as it forces continuous, high-dimensional visual and conceptual information through the bottleneck of a finite vocabulary. This limitation becomes particularly apparent when dealing with abstract visual concepts, affective dimensions, and other forms of tacit knowledge that resist precise verbalization.

4 Methodology

In this section, we introduce CoCoVa, a novel framework that achieves modality-aligned latent reasoning by constructing a unified latent space where visual and linguistic representations interact and evolve through multi-step continuous reasoning dynamics. We first provide an overview of the CoCoVa framework, establishing its core workflow and design philosophy. Subsequently, in Section 4.1, we provide a detailed exposition of the model’s architectural components, delving into the mechanics of the token selection mechanism, the latent fusion process via the LQ-Former, and the multi-task learning objective. Finally, Section 4.2 elaborates on the multi-stage training pipeline, which progressively builds the model’s capability for continuous latent reasoning. The detailed process is shown in Algorithm 1 in Appendix.

Overview. Figure 2 gives detailed illustration of CoCoVa. Like standard VLMs, vision-language input that includes an image and a question is first processed into vision embedding $\mathbf{V}_T$ and question embedding $\mathbf{T}$, as illustrated in Section 3. However, diverging from the standard single-pass projection paradigm where $[\mathbf{V}_T; \mathbf{T}]$ is directly fed to the LLM, CoCoVa utilizes these representations through its core latent reasoning cycle. To enable this dynamic reasoning process, CoCoVa incorporates three specialized modules. First, the Token Selection module analyzes the LLM’s attention patterns to identify and retain the most salient visual tokens, pruning irrelevant information. These refined visual features then enter the Multimodal Latent Fusion module, where they undergo iterative, cross-modal fusion with linguistic representations to generate a sequence of continuous latent thoughts $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K\}$. Throughout this process, the Multi-Task Representation Learning objective ensures semantic alignment across modalities through a

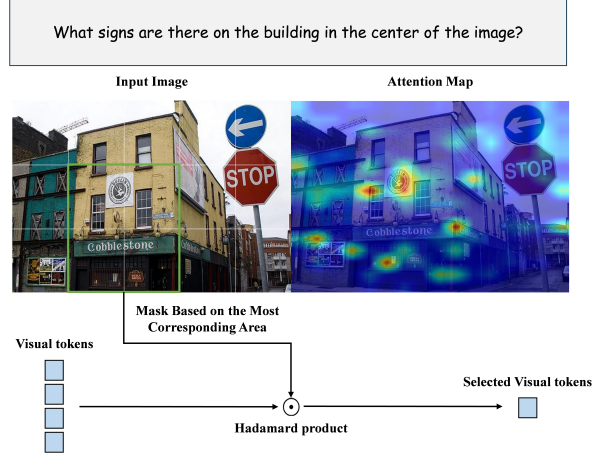


Figure 3: Dynamic visual token selection identifies the most corresponding image area based on attention map.

combination of contrastive and reconstruction losses. This coordinated pipeline enables the decoupling of internal reasoning from external expression, allowing the model to refine its understanding through latent simulation before generating the final answer $\mathbf{Y} = \text{LLM}([\mathbf{V}_T; \mathbf{T}; \mathbf{Z}])$.

4.1 CoCoVa: Reasoning in a Vision-Language Latent Space

Dynamic Visual Token Selection. The discrete linguistic bottleneck in standard Chain-of-Thought (CoT) reasoning not only limits expressive fidelity but also induces a static and biased visual perception in VLMs. This is evidenced by the fact that Standard VLMs exhibit a systematic bias in perceiving visual details, although their internal states already contain correct localization cues [Zhang et al., 2025b]. To emulate the human ability to dynamically redirect visual attention and to furnish our iterative latent-reasoning cycle with an ever-evolving visual context, we propose a dynamic token-selection mechanism. This mechanism ensures that at each reasoning step the model attends selectively to the visual information most relevant to its current internal thought state, rather than being overwhelmed by the full set of initial visual features.

Formally, let $\mathbf{A} \in \mathbb{R}^{L \times N \times 1 \times P}$ be the attention map of the current reasoning token to all image tokens in all layers of the backbone LLM, where L , N are the number of layers and heads-per-layer in the LLM, and P is the number of visual tokens provided to the LLM backbone. We first isolate the attention weights corresponding to the visual tokens and aggregate across heads and query positions to obtain a unified saliency map:

$$\mathbf{S} = \frac{1}{L \cdot N} \sum_{l=1}^L \sum_{n=1}^N \mathbf{A}[l, n, :, \text{img_start} : \text{img_end}] \quad (4)$$

where $\mathbf{S} \in \mathbb{R}^P$ represents the aggregate importance of each visual token, and $\text{img_start} : \text{img_end}$ denotes the slice indices corresponding to the visual tokens in the attention map. These 1D saliency scores are then reshaped into a 2D spatial grid $\mathbf{S}_{2D} \in \mathbb{R}^{H' \times W'}$, reconstructing the original spatial organization of the visual features from the vision encoder.

To identify coherent regions of high saliency rather than isolated points, we employ a sliding window approach. A kernel of size $w \times w$ is convolved over the saliency grid without padding:

$$\mathbf{W}_{\text{scores}} = \text{Conv2D}(\mathbf{S}_{2D}, \mathbf{K}_{w \times w}, \text{stride} = 1, \text{padding} = 0) \quad (5)$$

where $\mathbf{K}_{w \times w}$ is a kernel of ones. The window with the maximum cumulative score is selected:

$$(i^*, j^*) = \arg \max_{i, j} \mathbf{W}_{\text{scores}}[i, j] \quad (6)$$

Then a binary mask $\mathbf{M} \in \{0, 1\}^{B \times N}$ is generated by setting positions corresponding to the selected $w \times w$ region to 1 and all others to 0. This mask is applied to the original visual feature sequence via element-wise multiplication:

$$\mathbf{V}_{\text{selected}} = \mathbf{V} \odot \mathbf{M} \quad (7)$$

The resulting $\mathbf{V}_{\text{selected}}$ contains only the features from the most salient image region, which are then passed to the LQ-Former module. This selective mechanism not only reduces computational burden but, more importantly, enables the model to dynamically adjust its visual focus based on contextual cues.

Multimodal Latent Fusion with LQ-Former. The refined visual features $\mathbf{V}_{\text{selected}}$ are subsequently processed by the Latent Q-Former (LQ-Former). While architecturally rooted in the Q-Former design [Li et al., 2023a], the LQ-Former is repurposed from a static feature projector into a dynamic, iterative reasoning module. Its function evolves from modality alignment to orchestrating deep, multi-step fusion within a continuous latent space, producing a sequential refinement of the model’s internal state.

The LQ-Former operates over a series of K reasoning steps, orchestrating a closed-loop interaction between linguistic context and visual evidence. At each step k , it refines a latent thought vector $\mathbf{z}_k$ by jointly attending to contextual representations $\mathbf{h}_k$ from the last hidden states of the LLM’s every layer and the selected visual tokens $\mathbf{V}_{\text{selected}}$. Formally, the fusion process at step k is defined as:

$$\mathbf{h}_k = \mathbf{H}_{\text{LLM}}^{(k)}([\mathbf{V}_T; \mathbf{T}; \mathbf{Z}_{1:k-1}]) \quad (8)$$

$$\mathbf{z}_k = \text{LQ-Former}(\mathbf{h}_k, \mathbf{V}_{\text{selected}}) \quad (9)$$

Here, the current thought $\mathbf{h}_t$ serves as the query input to the LQ-Former’s cross-attention mechanism, while $\mathbf{V}_{\text{selected}}$ provides the keys and values. This enables the model to actively gather visual information most relevant to its evolving internal state. The module employs layer normalization to ensure training stability throughout multi-step reasoning. Moreover, we introduce a variant of the Gated Linear Unit (GLU) [Dauphin et al., 2017], a mechanism that dynamically modulates the fusion of cross-modal information based on the current reasoning state, followed by a feed-forward network to perform nonlinear transformation:

$$\mathbf{h}_{\text{fus}} = \text{CrossAttn}(\text{LayerNorm}(\mathbf{h}_k), \mathbf{V}_{\text{selected}}) \quad (10)$$

$$\mathbf{h}_{\text{glu}} = (\mathbf{h}_{\text{fus}} \mathbf{W}_1 + b) \otimes \sigma(\mathbf{h}_k \mathbf{W}_2 + c) \quad (11)$$

$$\mathbf{z}_k = \text{FFN}(\text{LayerNorm}(\mathbf{h}_k + \mathbf{h}_{\text{glu}})) \quad (12)$$

This iterative cycle begins with an initial thought state $\mathbf{h}_0$ derived from instruction embeddings and runs for K steps. The output is a **chain of continuous vision-language thoughts** $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K]$, representing the model’s internal reasoning trajectory. Finally, the latent rationale $\mathbf{Z}$ is prepended to the original input sequence. The LLM decodes the final answer conditioned on this enriched context:

$$\mathbf{Y} = \text{LLM}([\mathbf{V}_T; \mathbf{T}; \mathbf{Z}]) \quad (13)$$

In essence, the LQ-Former establishes a continuous reasoning pathway where the dynamic token selection provides spatially-focused visual evidence that is then iteratively integrated and refined through gated cross-modal fusion, mirroring the interplay [Luck and Ford, 1998] between visual attention and conceptual processing in human cognition.

Multi-Task Latent Learning. To ensure that the continuous latent thoughts $\mathbf{Z}$ remain semantically grounded in both visual and linguistic modalities, we employ a multi-task learning objective that jointly optimizes for semantic alignment and verifiable reasoning. This objective consists of three complementary loss functions that collectively enforce the coherence and faithfulness of the reasoning process.

First, we employ a **symmetric InfoNCE loss** [van den Oord et al., 2019] to maximize the mutual information between the latent thoughts and their corresponding visual representations, as well as between the latent thoughts and their corresponding textual representations. Let $\mathbf{f}_z = \text{Pooling}(\mathbf{Z})$, $\mathbf{f}_v = \text{Pooling}(\mathbf{V})$, and $\mathbf{f}_t = \text{Pooling}(\mathbf{R})$ denote the aggregated representations of the latent thoughts, original visual embeddings, and textual rationales extracted from the final layer of the LLM, respectively. The symmetric InfoNCE objective is formulated as:

$$\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{f}_z^i\}_{i=1}^B, \{\mathbf{f}_v^i\}_{i=1}^B) = -\frac{1}{B} \sum_{i=1}^B \log \left(\frac{\exp(\mathbf{f}_z^i \cdot \mathbf{f}_v^i / \tau)}{\sum_{j=1}^B \exp(\mathbf{f}_z^i \cdot \mathbf{f}_v^j / \tau)} \right) \quad (14)$$

$$\mathcal{L}_{\text{Symmetric-InfoNCE}} = \mathcal{L}_{\text{InfoNCE}}(\{\mathbf{f}_z^i\}_{i=1}^B, \{\mathbf{f}_v^i\}_{i=1}^B) + \mathcal{L}_{\text{InfoNCE}}(\{\mathbf{f}_z^i\}_{i=1}^B, \{\mathbf{f}_t^i\}_{i=1}^B) \quad (15)$$

where τ is a temperature hyperparameter, B is the batch size, and the superscripts i and j index samples within the batch. This bidirectional contrastive objective ensures that the semantic content of the latent thoughts aligns closely with their corresponding visual and textual representations.

Second, we introduce a **diffusion-based latent reconstruction loss** to enforce visual-grounding consistency within the latent reasoning process. Our approach employs a U-Net-based denoising network [Ronneberger et al., 2015] that learns to reconstruct the latent visual representations $\mathbf{I}_z$ of input images conditioned on the latent thoughts $\mathbf{Z}$, thereby compelling the preservation of semantically critical visual information while mitigating the spatial redundancy [He et al., 2022, Assran et al., 2023, Wang et al., 2024b] inherent in visual signal processing. The reconstruction objective is formulated as a denoising score matching loss:

$$\mathcal{L}_{\text{Recon}} = \mathbb{E}_{t, \mathbf{I}_z, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{I}_z + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, \mathbf{Z})\|_2^2] \quad (16)$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes standard Gaussian noise, ϵ_θ parameterizes the denoising network, and $\bar{\alpha}_t$ controls the noise schedule. This reconstruction task constrains the LQ-Former to maintain fine-grained visual details within the latent thoughts $\mathbf{Z}$, ensuring that the reasoning trajectory remains grounded in visual perceptual evidence rather than relying solely on linguistic associations.

Third, we incorporate the autoregressive latent-language modeling loss to maintain the linguistic coherence and task performance. Inspired by UPFT [Ji et al., 2025], we introduce **latent prefix fine-tuning**. We employ only latent thoughts $\mathbf{Z}$ as input to predict the prefix of the whole text sequence, thereby strengthening the guidance provided by the latent thoughts during the reasoning process. Additionally, we employ the conventional next-token prediction objective, which aims at predicting the complete token sequence based on both the vision-language inputs and the latent thoughts. With balancing coefficients λ_2 , the entire autoregressive latent-language modeling objective is defined as:

$$\mathcal{L}_{\text{AR}} = - \sum_{i=1}^{|\mathbf{Y}|} \log P(y_i | \mathbf{V}_T, \mathbf{T}, \mathbf{Z}, y_{<i}) - \lambda_1 \sum_{i=1}^{|\mathbf{R}|} \log P(r_i | \mathbf{Z}, r_{<i}) \quad (17)$$

This formulation employs two interconnected textual targets: $\mathbf{Y} = \{y_1, y_2, \dots, y_{|\mathbf{Y}|}\}$ represents the complete output sequence containing both reasoning steps and final answer, while $\mathbf{R} = \{r_1, r_2, \dots, r_{|\mathbf{R}|}\}$ specifically denotes the prefix of $\mathbf{Y}$ that constitutes the explicit textual rationale. Here, $\mathbf{R}$ corresponds precisely to the initial $|\mathbf{R}|$ tokens of $\mathbf{Y}$, forming the step-by-step reasoning process that precedes and leads to the final conclusion in the remaining sequence $\mathbf{Y}_{(|\mathbf{R}|+1):|\mathbf{Y}|}$.

The complete multi-task objective combines these losses with balancing coefficients λ_2 and λ_3 :

$$\mathcal{L}_{\text{Latent}} = \mathcal{L}_{\text{AR}} + \lambda_2 \mathcal{L}_{\text{Symmetric-InfoNCE}} + \lambda_3 \mathcal{L}_{\text{Recon}} \quad (18)$$

This multi-task framework, in conjunction with the dynamic token selection and iterative latent fusion, ensures that the latent thoughts $\mathbf{Z}$ are not only useful for language generation but also maintain semantic connections to both the visual input and linguistic context.

4.2 Training Pipeline

We propose a four-stage training pipeline for CoCoVa, as illustrated in Figure 4. Details are provided in the below.

Stage I: Vision–Language Alignment. The primary goal of this stage is to establish conceptual alignment between visual and linguistic elements within the embedding space, enabling the LLM to interpret the entities presented in images. During this foundational phase, we focus on training the vision projection layers while keeping both the visual encoder and LLM backbone frozen.

Stage II: Latent Alignment. This stage aims to establish a semantically-grounded latent reasoning space by training the LQ-Former module to generate continuous thought representations that preserve essential visual information. We jointly optimize the LQ-Former parameters alongside the denoising U-Net while maintaining the pre-trained vision encoder and LLM backbone in a frozen state to ensure training stability.

Stage III: Instruction Tuning. In this stage, we fine-tune the model using instruction-tuning datasets to enhance its capability for instruction-following and dialogue. All model parameters undergo optimization with the exception of the visual encoder, which remains frozen throughout this process.

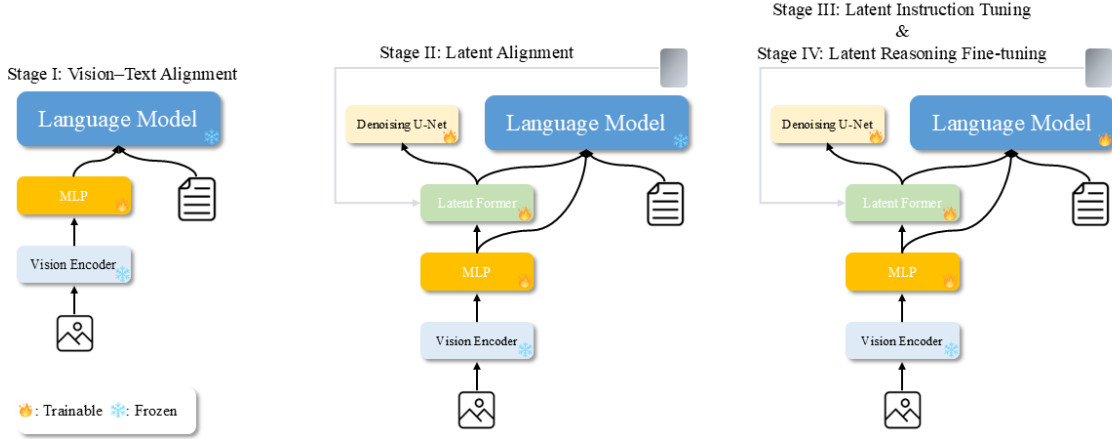


Figure 4: **CoCoVa four-stage training pipeline.** Flame and snowflake icons are used to indicate whether the corresponding module updates its parameters, respectively.

Stage IV: Latent Reasoning Fine-tuning. This final stage focuses on integrating latent reasoning capabilities with downstream task performance. As in Stage III, all parameters with the exception of the visual encoder undergo fine-tuning.

The training objectives are structured progressively across stages. In Stage I, optimization relies exclusively on the autoregressive language modeling loss. Subsequent stages (II–IV) integrate the latent learning loss $\mathcal{L}_{\text{Latent}}$, formulated in Section 4.1, to enforce coherent latent reasoning. Furthermore, across stages II through IV, we introduce the special tokens $\langle \text{bot} \rangle$ (beginning of thought) and $\langle \text{eot} \rangle$ (end of thought) to delimit the boundaries of the latent thought sequences, thereby providing explicit structural markers for the model’s internal reasoning process.

5 Experiments

5.1 Datasets

To assess our CoCoVa framework, we conduct comprehensive evaluations across four challenging benchmarks that collectively measure diverse capabilities in visual reasoning and understanding. **MMStar** [Chen et al., 2024a] evaluates advanced perceptual and reasoning abilities through carefully designed challenging questions, testing the model’s capacity for complex visual inference. **ScienceQA** [Lu et al., 2022a] test split provides a large-scale dataset for multimodal science question answering, requiring deep domain knowledge and structured reasoning. **MMBench** [Liu et al., 2024c] offers a comprehensive evaluation suite covering various perception and cognition dimensions, enabling holistic assessment of multimodal capabilities. **MMVP** [Tong et al., 2024] is a benchmark constructed from CLIP-blind pairs to expose systematic visual shortcomings in VLMs by testing them on straightforward questions about basic patterns like orientation and counting. Finally, **LLaVA-Bench-In-the-Wild** [Liu et al., 2024b] consists of diverse real-world images with open-ended instructions, testing the model’s generalization to practical scenarios.

5.2 Baselines

Multimodal CoT Methods. We compare CoCoVa against several representative Chain-of-Thought reasoning paradigms. The baseline approaches include: **Standard CoT**, which employs explicit step-by-step reasoning; **ICoT** [Gao et al., 2025], which generates sequential reasoning steps consisting of paired visual and textual rationales to infer the

final answer; and **MCOUT**[Pham and Ngo, 2025], which enables multimodal reasoning in a continuous latent space by generating reasoning steps consisting of hidden states.

Visual Reasoning Tuning Methods. We also compare CoCoVa against vision-language models that employ specialized tuning strategies for enhanced visual understanding and reasoning capabilities. The baseline approaches include: **Ross**[Wang et al., 2024b], which introduces reconstructive visual instruction tuning by denoising latent image representations to enhance fine-grained comprehension and reduce hallucinations; and **LVR**[Li et al., 2025d], which proposes latent visual reasoning to enable autoregressive reasoning directly in the visual embedding space by reconstructing key visual tokens, interleaved with text generation for improved perception-intensive tasks.

Vision-Language Models. We further compare CoCoVa against widely-adopted vision-language models that leverage large-scale pre-training and instruction tuning for multimodal understanding and reasoning. The baseline VLMs include: **BLIP-2**[Li et al., 2023a], which employs a Q-Former architecture to bridge frozen visual and linguistic encoders; **LLaVA-NeXT**[Liu et al., 2024a], representing an improved version of the LLaVA series with enhanced visual representation learning. Additionally, we include results from **Chameleon**[Team, 2024] and **Qwen-VL**[Bai et al., 2023] as reported in the literature.

5.3 Implementation Details

Model Architecture. Our CoCoVa framework is built upon the following core components. For the visual encoder, we employ CLIP-ViT-L/14[Radford et al., 2021] to extract visual features. We employ two different models as the LLM backbone, including Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct[Qwen et al., 2025], to evaluate the framework’s performance across different model scales. For the cross-attention module within the LQ-Former, we adopt the identical architectural configuration as the Q-Former in BLIP-2[Li et al., 2023a] and leverage its pre-trained parameters for initialization to expedite training convergence. As for the diffusion-based latent reconstruction, we adopt a lightweight U-Net architecture following the Latent Diffusion Model (LDM) paradigm [Rombach et al., 2022], which operates in the compressed latent space of a pre-trained VAE to ensure computational efficiency. Detailed settings of diffusion-based latent reconstruction can be found in Appendix A. By default, we configure the reasoning process to execute over 4 iterative steps and employ an 8×8 sliding window for dynamic visual token selection.

Training Configuration. We follow the four-stage training pipeline delineated in Section 4.2. Stages I and II utilize the LLaVA-Pretrain dataset [Liu et al., 2023a] for foundational alignment. Stage III employs the LLaVA-Instruct-150K dataset [Liu et al., 2023a] for instruction tuning, while Stage IV leverages the Vision-R1-cold dataset [Huang et al., 2025] for final-stage optimization. For most baseline methods, we employ the identical LLM backbone and visual encoder as used in CoCoVa and subject them to three-stage training on the same three datasets utilized by CoCoVa, thereby ensuring relatively fair comparison conditions. Regarding the LVR baseline, we maintain identical LLM backbones and visual encoders to CoCoVa for fair comparison. The model undergoes the same pretraining phase using the LLaVA-Pretrain dataset for feature alignment. During supervised fine-tuning and reinforcement learning stages, we employ the Visual CoT[Shao et al., 2024] and ViRL[Wang et al., 2025b] datasets respectively, following the original LVR implementation to ensure faithful reproduction of their method. All experiments were conducted on A100 GPUs, with detailed stage-specific hyperparameter configurations provided in Appendix A.

5.4 Main Results

We conducted a comprehensive evaluation of CoCoVa against multiple state-of-the-art baseline methods and leading open-source models across five diverse benchmarks.

Improved Accuracy. Our proposed CoCoVa consistently enhances accuracy across all evaluated benchmarks, demonstrating its broad effectiveness. As quantified in Table 1, CoCoVa achieves superior performance over all competing Chain-of-Thought and latent reasoning baselines at both model scales. On the MMBench benchmark, CoCoVa with the 1.5B backbone attains an accuracy of 63.3%, outperforming the strongest baseline, Ross, which scores 52.8%. This substantial margin of improvement is also evident in tasks requiring fine-grained visual understanding, such as MMVP, where CoCoVa-1.5B scores 42.8%, surpassing LVR by 2.9% points. Furthermore, CoCoVa’s gains are consistent across diverse reasoning types, from the scientific knowledge required by ScienceQA to the open-ended visual understanding in LLaVA-Bench. Notably, CoCoVa also competes favorably with significantly larger models. Across the five evaluated benchmarks, it consistently narrows or even surpasses the performance gap relative to 7B- and 9B-scale systems while operating on substantially smaller backbones. For instance, our 1.5B variant outperforms the 7B LLaVA-NeXT model on four benchmarks and matches its performance on MMBench, while our 0.5B model

Table 1: Results on five datasets: MMBench, MMStar, MMVP, ScienceQA and LLaVA-W. The best results of the same backbone are **bolded**. *We employ the Visual CoT [Shao et al., 2024] and ViRL [Wang et al., 2025b] datasets for LVR’s implementation, since it depends on ROI bounding box supervision for its training.

Backbone	Method	MMBench		MMStar		MMVP		ScienceQA		LLaVA-W	
		Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	ROUGE-L	# Tokens
Qwen2.5-0.5B-Instruct	Standard CoT	34.7	265.4	21.3	391.1	14.0	166.9	36.3	281.2	15.3	121.9
	ICoT	39.4	253.8	23.6	361.4	28.2	165.7	40.4	294.2	17.5	143.8
	MCOUt	34.9	265.3	19.9	405.4	26.2	167.6	32.4	283.3	16.4	136.6
	Ross	45.9	248.7	25.7	509.8	31.5	156.1	41.6	259.3	20.3	174.9
	LVR*	40.5	250.3	25.0	408.2	32.4	162.4	40.5	273.2	18.0	128.1
	CoCoVa (Ours)	51.2	225.1	32.2	366.1	36.2	153.0	42.3	256.9	23.4	101.7
Qwen2.5-1.5B-Instruct	Standard CoT	43.4	395.2	21.6	521.7	16.6	227.1	38.3	426.2	23.2	193.1
	ICoT	45.4	360.5	27.8	570.1	33.8	219.2	41.6	442.7	21.4	182.3
	MCOUt	41.3	362.3	24.8	565.2	34.4	197.5	36.5	397.2	23.1	159.0
	Ross	52.8	382.9	30.3	604.6	40.5	214.9	43.1	402.8	23.2	186.5
	LVR*	51.2	340.7	30.4	595.3	39.9	208.4	42.4	370.9	23.5	163.1
	CoCoVa (Ours)	63.3	256.7	36.0	465.1	42.8	171.9	45.6	330.1	25.7	122.2
BLIP-2-4B	—	31.6	—	30.0	—	17.1	—	38.8	—	12.5	—
LLaVA-NeXT-7B	—	64.2	—	33.3	—	36.7	—	42.2	—	17.3	—
<i>Literature reports</i>											
Chameleon-7B	—	16.7	—	—	—	—	—	45.0	—	13.1	—
Qwen-VL-Chat-9B	—	61.8	—	34.5	—	21.0	—	—	—	—	—

Table 2: Ablation study of CoCoVa on Qwen2.5-1.5B-Instruct backbone. The results demonstrate that each component contributes critically to the full model’s performance and efficiency.

Methods	Average Score	Average Tokens
CoCoVa	42.7	269.2
w/o LQ-Former	35.0 (-7.7)	311.5 (+42.3)
w/o Dynamic Selection	38.7 (-4.0)	280.3 (+11.1)
w/o Diffusion Recon. Loss	40.2 (-2.5)	274.0 (+4.8)
w/o Symmetric InfoNCE Loss	39.5 (-3.2)	276.9 (+7.7)
w/o Latent prefix fine-tuning	40.9 (-1.8)	263.6 (-5.6)

surpasses the 4B BLIP-2 by a wide margin across all tasks. These results demonstrate that the latent space reasoning paradigm implemented by CoCoVa enables stronger multimodal reasoning ability.

Token Efficiency. CoCoVa demonstrates superior token efficiency, a direct benefit of its continuous reasoning paradigm. As shown in Table 1, CoCoVa consistently generates fewer output tokens than competing CoT and latent reasoning baselines across almost all tasks and model scales. For instance, in the LLaVA-Bench-In-the-Wild (LLaVA-W), CoCoVa with the 0.5B backbone achieves a higher ROUGE-L score while using only 101.7 tokens, which is 16.6% fewer than Standard CoT and a remarkable 29.3% reduction compared to ICoT. This trend holds at the 1.5B scale, where CoCoVa uses 122.2 tokens, 33.0% fewer than ICoT’s 182.3 tokens. Notably, CoCoVa does not progressively replace linguistic reasoning steps with continuous latent representations during training [Hao et al., 2024], but is instead supervised directly on the complete reasoning chain. We hypothesize that the latent vision-language thoughts generated by the LQ-Former already contain elements of pre-reasoning prior to the generation of textual tokens. Such token efficiency is particularly advantageous for deployment in resource-constrained environments.

5.5 Ablation Study

Role of Each Module. As shown in Table 2, the most substantial degradation occurs when removing the LQ-Former, which effectively disables the iterative latent reasoning process. Performance drops significantly across all benchmarks—for instance, the average score decreases from 42.7% to 35.0%, while output length increases considerably. Removing the dynamic token selection mechanism results in moderate but consistent performance declines, particularly on perception-intensive tasks. This suggests that the ability to focus on salient image regions is especially valuable for tasks requiring fine-grained visual understanding. Removing the diffusion reconstruction loss leads to noticeable performance degradation across multiple benchmarks. For instance, the average score decreases from 42.7% to 40.2%, while output efficiency is also affected. Ablating the symmetric InfoNCE loss results in consistent performance drops, with average score reducing from 42.7% to 39.5%. Interestingly, removing latent prefix fine-tuning reduces performance

on text-generation quality while slightly improving token efficiency (average tokens drop from 269.2 to 263.6). This suggests that while this component is important for guiding the model to generate more reliable reasoning paths, the latent thoughts themselves may already encode substantial reasoning structure, highlighting the trade-off between output quality and efficiency in the continuous reasoning paradigm. Extended ablations on these modules can be found in Appendix B.1.

Impact of Reasoning Steps. In Figure 5, we systematically ablate the number of iterative reasoning steps K to understand its effect on CoCoVa’s performance. The results reveal a clear pattern of diminishing returns as the number of steps increases. Initially, performance improves substantially with additional reasoning steps, demonstrating the value of iterative refinement in the continuous latent space. Concurrently, output length decreases significantly, indicating more efficient reasoning. However, beyond a certain threshold, further increases in reasoning steps yield minimal performance gains while output length stabilizes. This saturation effect suggests that the model reaches a stable reasoning state after an optimal number of iterations.

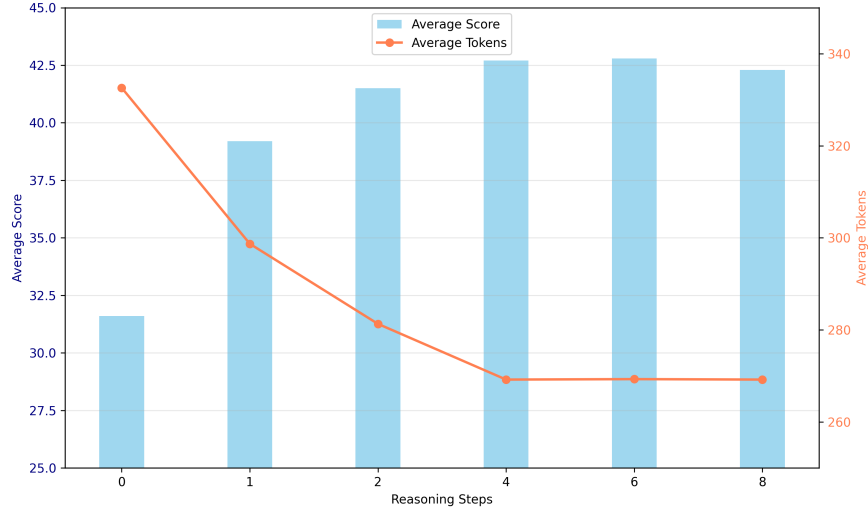


Figure 5: Impact of varying the number of reasoning steps on model performance and output efficiency. Performance improves with additional steps initially before saturating, while output length decreases and stabilizes, indicating an optimal balance between reasoning depth and computational efficiency.

5.6 In-Depth Analysis

Attention Analysis. We compute the attention scores from the first generated text token to all visual tokens on the MMStar dataset. As shown in Table 3, CoCoVa achieves higher attention scores than LLaVA. This indicates that CoCoVa effectively guides the model to focus more on the input image, assigning greater weight to visual evidence at the very beginning of the reasoning process, thereby enhancing its understanding of visual signals.

Table 3: Quantitative comparison on attention values. We employ T-test[Student, 1908] to compare the means and the Mann-Whitney U test[Mann and Whitney, 1947] to compare the medians of the two distributions. The results indicated that CoCoVa achieves higher values than LLaVA in both the mean and the median.

Statistic ($\times 10^{-4}$)	LLaVA	CoCoVa	P-value
Mean	4.47	10.84	1.08×10^{-7}
25th Percentile	4.03	9.96	–
Median	4.44	10.76	3.76×10^{-9}
75th Percentile	4.87	11.67	–
95th Percentile	5.54	13.03	–

Probing Analysis. To examine whether the continuous latent thoughts $\mathbf{Z}$ jointly encode linguistic reasoning and visual semantics, we conduct a probing analysis. We employ weak classifiers including a Linear SVM[Cortes and Vapnik, 1995] and a single-layer MLP[Rumelhart et al., 1986], in order to probe the intrinsic properties of the representation space. If weak classifiers can successfully differentiate whether a latent thought vector and a textual or visual representation originate from the same problem instance, it serves as strong evidence that the two representations

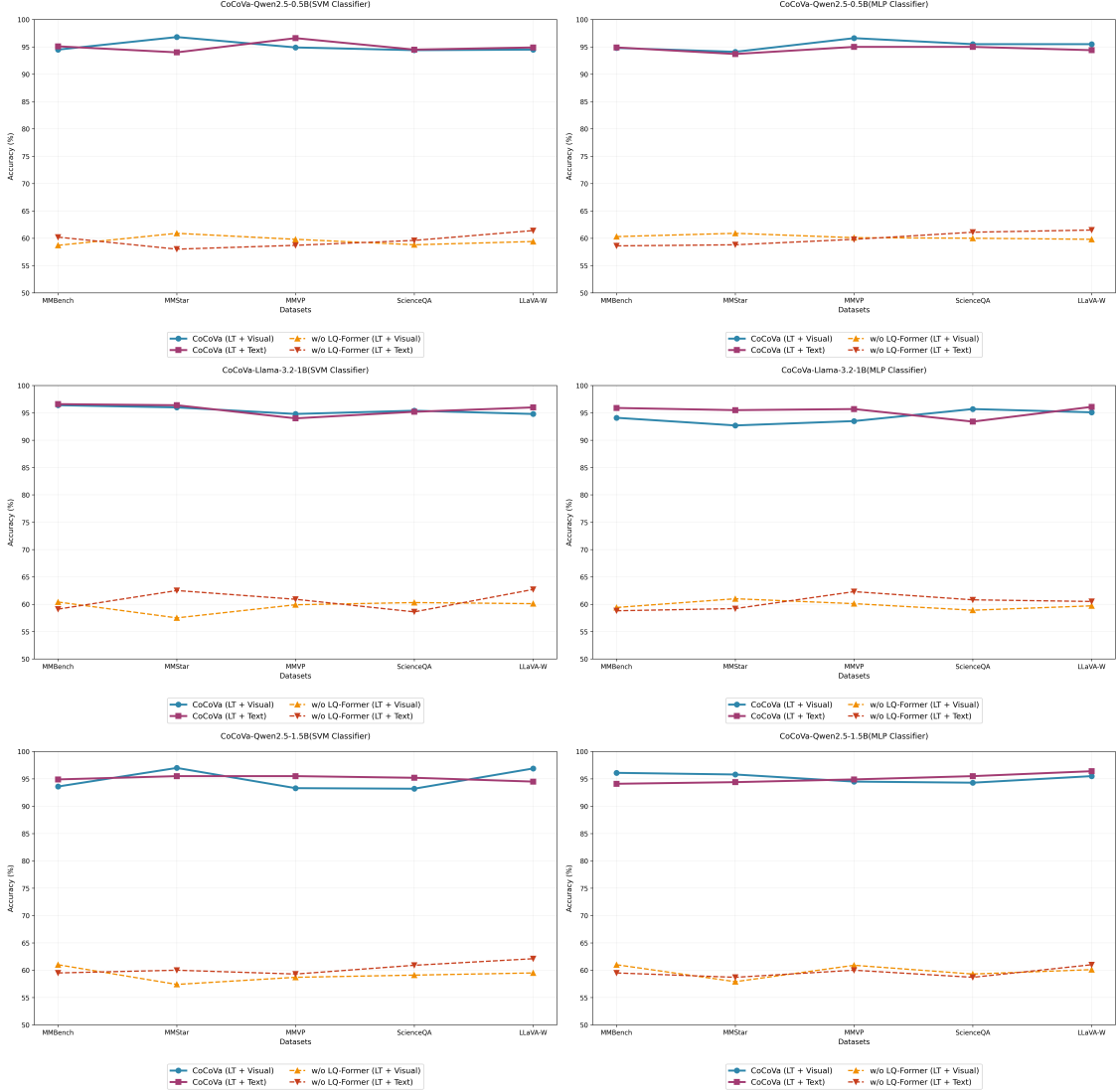


Figure 6: Accuracy of two weak classifiers.

share a high degree of semantic consistency regarding the core identity of the reasoning task. Detailed settings of this experiment can be found in Appendix A.

As shown in figure 6, our results confirm this hypothesis: the classification accuracy for pairing latent thoughts with their corresponding visual features, as well as with their corresponding textual rationales, both approached a remarkable 95%. This result indicates that the latent thoughts $\mathbf{Z}$ act as a compact yet highly informative semantic hub, effectively distilling and encapsulating the essential problem context from both visual and linguistic modalities. Crucially, we performed an ablation by replacing $\mathbf{Z}$ with the LLM’s last hidden states. The probe accuracy plummeted to approximately 60%, significantly lower than the 95% achieved by our model. This contrast underscores the role of the LQ-Former; it is not merely transmitting information but is structuring and refining cross-modal inputs into a unified, reasoning-centric representation within the continuous latent space.

Structure of Thoughts. To better understand the macroscopic organization of the latent reasoning space, we conducted a topological analysis of the learned thought representations. We extract the features of latent thoughts $\mathbf{Z}$ from samples of three distinct cognitive categories in the MMStar benchmark: *Fine-grained Perception*, *Logical Reasoning*, and *Mathematics*. The t-SNE[Maaten and Hinton, 2008] visualization of these high-dimensional vectors, as shown in Figure 7, reveals a compelling geometric structure. Instances from the same category tend to occupy coherent regions of the projection, with clear separation between perceptual, logical and mathematical clusters. Importantly, this

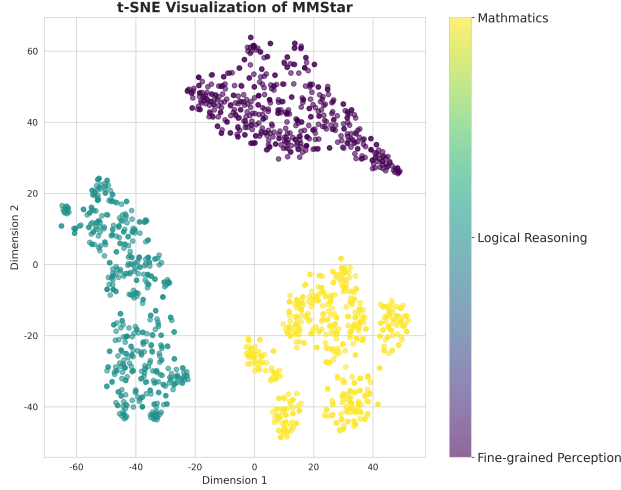


Figure 7: The t-SNE visualization of the final latent thoughts from the MMStar dataset, colored by problem category.

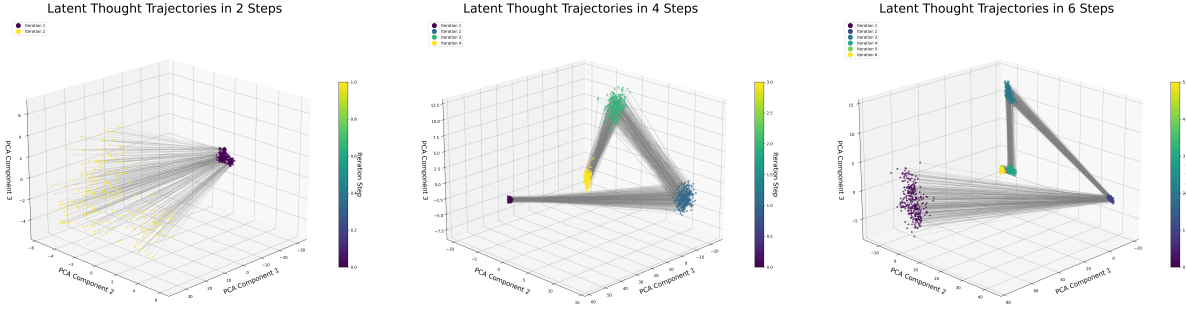


Figure 8: Visualization of latent thought trajectories in PCA space over 2, 4, and 6 reasoning steps (left to right). Points represent latent states at each iteration, and gray lines connect successive steps. As the reasoning unfolds, trajectories exhibit clearer clustering and reduced displacement, suggesting progressive convergence of latent thoughts.

organization emerges without explicit category supervision, indicating that the model’s iterative latent dynamics induce a form of self-organization in representation space. We could therefore view the latent manifold as more than a diffuse collection of activations: it appears to instantiate distinct, internally consistent modes of processing that reflect different classes of cognitive demand.

Reasoning Trajectory Analysis. The low-dimensional projections of CoCoVa’s reasoning trajectories reveal an emergent geometric structure that aligns with and elucidates the performance characteristics observed in our ablation studies. As illustrated in Figure 8, the initial reasoning steps exhibit exploratory "sweeping" motions, where latent thoughts diverge to cover a broad hypothesis space. This corresponds to the significant performance gains observed when increasing steps from 0 to 2 in our ablation experiments. As reasoning progresses, these trajectories organize into structured pathways with compact clusters, demonstrating a coarse-to-fine refinement process that mirrors the performance saturation beyond 4 steps. The convergence dynamics evidenced by progressively smaller displacements in later iterations, directly explain why additional steps yield diminishing returns in accuracy. Furthermore, the observed non-linear jumps in the trajectory space provide mechanistic insight into the model’s capacity for substantial hypothesis updates when integrating new cross-modal evidence. These geometric patterns collectively substantiate that CoCoVa’s continuous reasoning follows an organized trajectory from exploration to convergence, providing an intuitive visual explanation for the optimal step configuration identified empirically.

From Latent Thoughts to Images. We qualitatively examine the visual content of the latent thoughts via image reconstruction. The conditional denoising U-Net, lightly fine-tuned for merely 5 epochs on ImageNet-1K[Deng et al., 2009], successfully reconstructs coherent visual scenes from the latent thoughts $\mathbf{Z}$. As shown in Figure 9, the

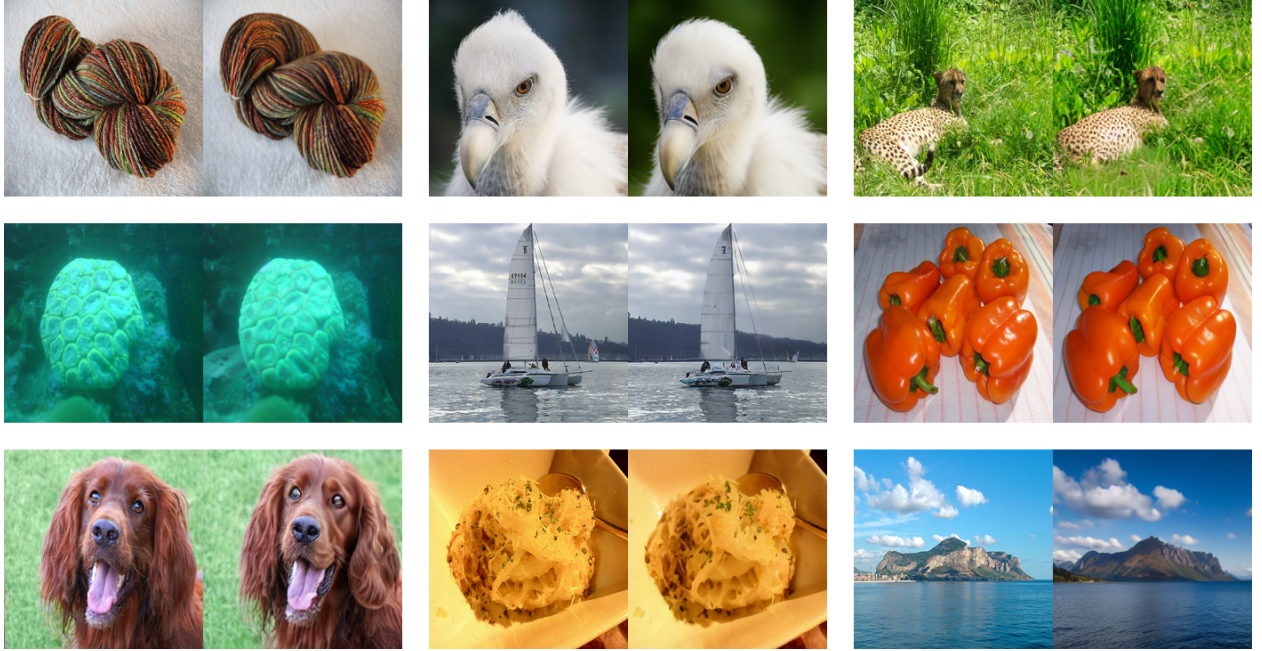


Figure 9: Visual reconstructions from the chain of latent thoughts. In each panel, the left image shows the original input (from the ImageNet-1K validation set), and the right image shows its reconstruction.

reconstruction results are reasonable, offering tangible proof that the chain of latent thoughts maintains a continuous and verifiable connection to the visual evidence throughout the reasoning process.

5.7 Explorations

To further investigate the properties and potential of the CoCoVa paradigm, we conduct several exploratory studies. More Extended experiments can be found in Appendix B.2.

Table 4: Performance of CoCoVa scaled to different 7B LLM backbones. The best results are **bolded** and the second best results are underlined. We evaluate the average accuracy of these models across the following benchmarks: AI2D[Hiippala et al., 2021], GQA[Hudson and Manning, 2019] test-dev-balanced split, MMBench[Liu et al., 2024c], MMStar[Chen et al., 2024a], MMMU[Yue et al., 2024] validation split, MMVP[Tong et al., 2024], POPE[Li et al., 2023b], and ScienceQA[Lu et al., 2022a] test split.

Models	AI2D	GQA	MMBench	MMStar	MMMU	MMVP	POPE	ScienceQA
<i>LLM Backbone: Qwen2.5-7B-Instruct</i>								
CoCoVa-7B_{qwen}	<u>73.4</u>	<u>65.0</u>	73.4	<u>51.6</u>	47.6	52.7	87.4	88.2
<i>LLM Backbone: Vicuna-7B-v1.5</i>								
CoCoVa-7B_{vicuna}	66.4	64.2	64.6	47.3	39.6	<u>50.9</u>	82.6	80.1
<i>Literature reports</i>								
DeepSeek-VL-7B[Lu et al., 2024]	64.9	–	73.2	37.1	36.6	–	85.8	–
InternVL-1.5-7B[Chen et al., 2024b]	69.8	61.2	72.5	50.3	48.9	–	86.2	<u>87.5</u>
Janus-pro-7B[Chen et al., 2025]	–	62.0	79.2	–	41.0	–	87.4	–
MiniCPM-V-2.6-8B[Yao et al., 2024]	70.1	63.8	70.3	51.4	48.2	–	86.7	86.9
LLaVA-NeXT-13B[Liu et al., 2024a]	72.4	65.4	70.0	43.9	36.2	37.3	86.2	–
GPT-4V[OpenAI, 2023]	78.2	36.8	<u>75.2</u>	57.1	53.8	50.0	75.4	–

Scalability to Larger LLM Backbones. We first investigate whether the benefits of CoCoVa’s latent reasoning scale with the capacity of the underlying Large Language Model. To this end, we instantiate our CoCoVa framework on top of the Qwen2.5-7B-Instruct[Qwen et al., 2025] and Vicuna-7B-v1.5[Chiang et al., 2023] backbone. All variants are trained following the same multi-stage curriculum described in Section 4.2 with the same datasets. Illustrated in Table 4, we compare our CoCoVa with state-of-the-art VLMs across eight multimodal benchmarks. The Qwen-based

Table 5: Performance comparison of adaptive reasoning mechanisms using different convergence criteria.

Methods	Average Steps	Average Score	Average Tokens
CoCoVa	4.0	42.7	269.2
+ Adaptive (Cosine)	5.9	42.6	265.1
+ Adaptive (L1 Distance)	2.8	42.0	279.8
+ Adaptive (L2 Distance)	2.9	42.5	270.5

variant achieves competitive results across multiple tasks, outperforming several established 7B-scale models such as DeepSeek-VL-7B and InternVL-1.5-7B on a majority of benchmarks, while also matching or surpassing the 13B-scale LLaVA-NeXT on several perception and reasoning tasks. Moreover, CoCoVa even outperforms GPT-4V on several benchmarks such as GQA, MMVP and POPE. These findings indicate that the continuous reasoning paradigm of CoCoVa scales effectively with model capacity and can leverage the enhanced representational power of larger LLMs.

Adaptive Reasoning Steps. To investigate whether CoCoVa can dynamically adjust its computational budget based on problem complexity, we implement an adaptive reasoning mechanism. We evaluate three distance measures including cosine similarity, L1 distance, and L2 distance between consecutive latent thoughts z_k and z_{k-1} to determine reasoning termination. As summarized in Table 5, the adaptive mechanisms successfully demonstrate that the latent reasoning process naturally converges, as evidenced by the reduction in inter-step distance. This observed challenge in adaptive reasoning directly corroborates and provides a deeper explanation for the findings from our ablation study on fixed reasoning steps. However, the results reveal a non-trivial trade-off between efficiency and performance. Adaptive reasoning steps do not bring significant gains in performance, and even lead to a decline in most metrics. These results suggest that while the latent space exhibits measurable convergence patterns, simply monitoring inter-step distance provides an incomplete signal for determining optimal stopping points. We expect future work to explore more sophisticated halting mechanisms that incorporate both local convergence signals and global reasoning state assessment to better balance efficiency and performance.

6 Conclusion

This paper identifies the discrete linguistic bottleneck as a fundamental limitation in modern VLMs’ reasoning capabilities. To address this, we propose CoCoVa, a novel framework that performs reasoning in a continuous latent space through iterative cross-modal interaction. By introducing the Latent Q-Former as a dynamic reasoning engine, CoCoVa enables multi-step evolution of latent thoughts that integrate both visual and linguistic information. Extensive experiments demonstrate that CoCoVa achieves superior accuracy while significantly improving token efficiency compared to strong baselines. Further analysis reveals interpretable reasoning trajectories and convergence patterns within the latent space, highlighting the potential of latent space reasoning as a scalable paradigm for future vision-language models.

References

- Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.
- Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025a. URL <https://arxiv.org/abs/2501.02189>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Víctor Gutiérrez-Basulto, Jeff Z Pan, and Hanjie Chen. Miceval: Unveiling multimodal chain of thought’s quality via image description and reasoning steps. *arXiv preprint arXiv:2410.14668*, 2024.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822, 2023.
- OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Ifimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz,

- Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiye Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. OpenAI o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.
- Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024a.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025a.
- Bing Wei, Yudi Zhao, Kuangrong Hao, and Lei Gao. Visual sensation and perception computational models for deep learning: State of the art, challenges and prospects. *arXiv preprint arXiv:2109.03391*, 2021.
- Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit reasoning in large language models: A comprehensive survey, 2025b.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025a.
- Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, June 1995. doi:10.1126/science.7777863. Erratum in: *Science*. 2005 Feb 11;307(5711):851.
- Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.
- Seth Herd, Kai Krueger, Ananta Nair, Jessica Mollick, and Randall O’Reilly. Neural mechanisms of human decision-making. *Cognitive, Affective, & Behavioral Neuroscience*, 21(1):35–57, 2021.

- David L Barack and John W Krakauer. Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22(6):359–371, 2021.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023a.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL <https://arxiv.org/abs/1807.03748>.
- Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024b.
- Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024c.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a. URL <https://arxiv.org/abs/2304.08485>.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025a.
- Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *arXiv preprint arXiv:2301.05226*, 2023.
- Xuwen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. In *International Conference on Neural Information Processing*, pages 62–76. Springer, 2024.
- Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*, 2024.
- Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428, 2024.
- Juncheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. Soft-prompting with graph-of-thought for multi-modal representation learning. *arXiv preprint arXiv:2404.04538*, 2024.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022a.
- Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170, 2024a.

- Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pages 305–322. Springer, 2024.
- Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.
- Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.
- Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinking: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.
- Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025c.
- Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A survey on latent reasoning, 2025. URL <https://arxiv.org/abs/2507.06203>.
- Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025.
- Yuchang Sun, Yanxi Chen, Yaliang Li, and Bolin Ding. Enhancing latent computation in transformers with latent tokens. *arXiv preprint arXiv:2505.12629*, 2025.
- Zhuocheng Gong, Jian Guan, Wei Wu, Huishuai Zhang, and Dongyan Zhao. Latent preference coding: Aligning large language models via discrete latent codes. *arXiv preprint arXiv:2505.04993*, 2025.
- Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025b.
- Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, et al. Scalable language models with posterior inference of latent thought vectors. *arXiv e-prints*, pages arXiv–2502, 2025.
- David Herel and Tomas Mikolov. Thinking tokens for language modeling. *arXiv preprint arXiv:2405.08644*, 2024.
- Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-star: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024.
- Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*, 2025c.
- Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. *arXiv preprint arXiv:2502.17416*, 2025.
- Amirkeivan Mohtashami, Matteo Pagliardini, and Martin Jaggi. Cotformer: A chain-of-thought driven architecture with budget-adaptive computation cost at inference. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025d.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021.
- Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *The Thirteenth International Conference on Learning Representations*, 2025b.
- Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

- Steven J Luck and Michelle A Ford. On the role of selective attention in visual perception. *Proceedings of the National Academy of Sciences*, 95(3):825–830, 1998.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.
- Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*, 2024b.
- Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, et al. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models. *arXiv preprint arXiv:2503.02875*, 2025.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024c.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.
- Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025.
- Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.
- Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-rl: Incentivizing reasoning capability in multimodal large language models, 2025. URL <https://arxiv.org/abs/2503.06749>.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *CoRR*, 2024.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025b.
- Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.
- Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

- Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55(3):661–688, 2021.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.
- Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024b.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- OpenAI. GPT-4V(ision) System Card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.
- Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.
- Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021a.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2025c.
- Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, November 2024.

- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021b.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.
- Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023c.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024c.
- Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2659, 2017.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520, 2022.
- Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022b.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.
- Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertr: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023b.

Appendix

Algorithm 1 Chain of Continuous Vision-Language Thought

Require: Input image I , text prompt T , number of reasoning steps K , special token $\langle |bot| \rangle$ & $\langle |eot| \rangle$

Ensure: Final answer Y

- 1: **Step 1: Input Processing**
 - 2: Extract visual features: $\mathbf{V} \leftarrow \text{VisionEncoder}(I)$
 - 3: Project visual features: $\mathbf{V}_T \leftarrow \text{MLP}_{\text{vision}}(\mathbf{V})$
 - 4: Encode text: $\mathbf{T} \leftarrow \text{Tokenizer}(T)$
 - 5: **Step 2: Dynamic Visual Token Selection**
 - 6: Get LLM attention maps: $\mathbf{A} \leftarrow \text{LLM}([\mathbf{V}_T; \mathbf{T}])$
 - 7: Aggregate attention: $\mathbf{S} \leftarrow \frac{1}{L \cdot N} \sum_{l=1}^L \sum_{n=1}^N \mathbf{A}[l, n, :, \text{img_start} : \text{img_end}]$
 - 8: Reshape to spatial grid: $\mathbf{S}_{2D} \leftarrow \text{reshape}(\mathbf{S}, [H', W'])$
 - 9: Compute window scores: $\mathbf{W}_{\text{scores}} \leftarrow \text{Conv2D}(\mathbf{S}_{2D}, \mathbf{K}_{w \times w})$
 - 10: Find optimal window: $(i^*, j^*) \leftarrow \arg \max_{i,j} \mathbf{W}_{\text{scores}}[i, j]$
 - 11: Generate binary mask $\mathbf{M}$ for window (i^*, j^*)
 - 12: Select visual tokens: $\mathbf{V}_{\text{selected}} \leftarrow \mathbf{V} \odot \mathbf{M}$
 - 13: **Step 3: Multimodal Latent Fusion**
 - 14: Initialize thought state: $\mathbf{h}_0 \leftarrow \text{InstructionEmbedding}$
 - 15: Initialize latent thoughts: $\mathbf{Z} \leftarrow [\langle |bot| \rangle]$
 - 16: **for** $k \leftarrow 1$ **to** K **do**
 - 17: Get contextual representation: $\mathbf{h}_k \leftarrow \mathbf{H}_{\text{LLM}}^{(k)}([\mathbf{V}_T; \mathbf{T}; \mathbf{Z}])$
 - 18: Compute cross-attention: $\mathbf{h}_{\text{fus}} \leftarrow \text{CrossAttn}(\text{LayerNorm}(\mathbf{h}_k), \mathbf{V}_{\text{selected}})$
 - 19: Apply GLU gating: $\mathbf{h}_{\text{glu}} \leftarrow (\mathbf{h}_{\text{fus}} \mathbf{W}_1 + b) \otimes \sigma(\mathbf{h}_k \mathbf{W}_2 + c)$
 - 20: Generate thought: $\mathbf{z}_k \leftarrow \text{FFN}(\text{LayerNorm}(\mathbf{h}_k + \mathbf{h}_{\text{glu}}))$
 - 21: Append to chain: $\mathbf{Z} \leftarrow \mathbf{Z} \cup \{\mathbf{z}_k\}$
 - 22: **end for**
 - 23: Add special token: $\mathbf{Z} \leftarrow \mathbf{Z} \cup [\langle |eot| \rangle]$
 - 24: **Step 4: Answer Generation**
 - 25: Generate final answer: $\mathbf{Y} \leftarrow \text{LLM}([\mathbf{V}_T; \mathbf{T}; \mathbf{Z}])$
 - 26: **return** $\mathbf{Y}$
-

A Implementation Details

Settings of Diffusion-based Latent Reconstruction. The diffusion process employs a linear noise schedule defined by $\beta_t \in [\beta_{\text{start}} = 1 \times 10^{-4}, \beta_{\text{end}} = 0.02]$ over $T = 1000$ iterations. At each training step, a random timestep $t \sim U(1, T)$ is sampled, and the input image x_0 is corrupted as $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$. The sampling process follows an iterative denoising procedure using the predicted noise and the reparameterization of the posterior distribution. The architecture of our denoising U-Net is illustrated in table 6.

Table 6: U-Net configuration.

Parameter	Value
sample_size	img_size (input spatial size)
in_channels	self.channel (input feature channels)
out_channels	self.channel (output feature channels)
layers_per_block	3
block_out_channels	(96, 192, 384, 512)
down_block_types	(DownBlock2D, CrossAttnDownBlock2D, CrossAttnDownBlock2D, DownBlock2D)
up_block_types	(UpBlock2D, CrossAttnUpBlock2D, CrossAttnUpBlock2D, UpBlock2D)
cross_attention_dim	768
attention_head_dim	64
norm_num_groups	32

Hyperparameters. The training process employs the AdamW optimizer [Loshchilov and Hutter, 2017] with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 0.05 across all stages. For Stage I, we utilize a learning rate of 1×10^{-3} for the vision

projection layers with the warmup ratio of 0.03, followed by cosine decay, training for one epoch with a global batch size of 256. Stage II employs a learning rate of 2×10^{-4} for the LQ-Former and reconstructor modules, training for one epoch with a reduced batch size of 128 to accommodate the additional computational requirements of the diffusion process. In Stages III and IV, all parameters except for the visual encoder are updated, while other hyperparameter settings remain identical to those in Stage II. As for the latent prefix fine-tuning, the prefix length is set to 32. The multi-task loss coefficients are set to $\lambda_1 = 1.0$ for the Latent prefix fine-tuning loss, $\lambda_2 = 0.9$ for the symmetric InfoNCE loss and $\lambda_3 = 0.9$ for the diffusion reconstruction loss, determined through grid search on the validation set. All models are trained using mixed precision (FP16) on 8 NVIDIA A100 80GB GPUs.

Image Resolution Configurations For image resolution configurations, all input images are resized to 224×224 pixels to match the standard input requirements of the CLIP-ViT-L/14 visual encoder. This resolution yields 256 visual tokens after processing through the vision encoder (a 16×16 grid of patch tokens). The diffusion-based reconstructor operates on latent representations with a feature map shape of $[4, 28, 28]$, corresponding to the compressed spatial dimensions while preserving essential visual information. During reconstruction evaluation, the output images are upsampled back to the original 224×224 resolution for both qualitative assessment and quantitative metric computation, ensuring consistent evaluation standards with the input domain.

Details of the Vision-R1-cold Dataset. Vision-R1-cold[Huang et al., 2025] is a high-quality multimodal Chain-of-Thought (CoT) dataset comprising 200K samples, constructed without human annotation. It leverages a modality bridging approach that converts visual information into textual form by first generating Pseudo-CoT descriptions using an existing MLLM, then refining them with DeepSeek-R1 to produce reasoning processes that incorporate human-like cognitive behaviors such as questioning and reflection. The dataset undergoes rule-based filtering to ensure logical consistency and semantic coherence, serving as effective cold-start initialization data for enhancing the reasoning capability of multimodal large language models. Details of this datasets are listed in Table 7.

Table 7: Data sources for the Vision-R1-cold dataset.

Category	Data Sources
Mathematical Data	GLLaVA [Gao et al., 2023], GEOS [Seo et al., 2015], UniGeo [Chen et al., 2022], GeoQA Plus [Chen et al., 2021], Geo3K [Lu et al., 2021a], MathVision [Wang et al., 2025c], GeoMverse [Kazemi et al., 2023], MathV360K [Shi et al., 2024], IconQA [Lu et al., 2021b], TabMWP [Lu et al., 2023], CLEVR [Johnson et al., 2017], CLEVR-Math [Lindström and Abraham, 2022], Super-CLEVR [Li et al., 2023c]
General QA Data	ShareGPT4V [Chen et al., 2024c], PISC [Li et al., 2017], VQA-AS [Antol et al., 2015], A-OKVQA [Schwenk et al., 2022], TextVQA [Singh et al., 2019], Vizwiz [Gurari et al., 2018], VQA2.0 [Goyal et al., 2017]
Science and Medical Data	From GeoQA+ [Cao and Xiao, 2022], CLEVR-Math [Lindström and Abraham, 2022], TQA [Kembhavi et al., 2017], AI2D [Kembhavi et al., 2016], ScienceQA [Lu et al., 2022b], VQA-RAD [Lau et al., 2018], PMC-VQA [Zhang et al., 2023]
Figure Understanding Data	From DVQA [Kafle et al., 2018], DocVQA [Mathew et al., 2021], FigureQA [Kahou et al., 2017], PlotQA [Methani et al., 2020], ChartQA [Masry et al., 2022], InfoVQA [Mathew et al., 2022], MultiHiertt [Zhao et al., 2022], LRV-Chart [Liu et al., 2023b]

Data Decontamination Procedure. To ensure the integrity of our evaluation and mitigate potential data contamination between the training set and the benchmark test sets, we implemented a rigorous two-stage decontamination process on the Vision-R1-cold dataset. First, we performed an exact match filtering based on textual question-answer pairs to explicitly remove all samples originating from the official training splits of the AI2D[Kembhavi et al., 2016] and ScienceQA[Lu et al., 2022a] benchmarks, which are known sources within Vision-R1-cold. Subsequently, to address more subtle, near-duplicate contamination at the image level, we employed a CLIP-ViT-L/14[Radford et al., 2021] model to compute image embeddings for all remaining Vision-R1-cold images and the images from the test splits of all our evaluation benchmarks. We then conducted a cross-dataset nearest-neighbor search. Any Vision-R1-cold sample whose image embedding had a cosine similarity of ≥ 0.92 with any test-set image embedding was identified as a semantic near-duplicate and removed. This conservative threshold was set to ensure the removal of visually and semantically highly analogous samples that could inflate performance metrics. In total, this procedure resulted in the removal of **4,479** samples from the original Vision-R1-cold dataset, thereby maximally ensuring that reported performance gains are attributable to our model’s improved reasoning capabilities rather than data leakage.

Table 8: Performance comparison with different sliding window sizes.

Methods	Avg. Score	Avg. Tokens	MMVP Accuracy
4×4	42.7	275.3	41.9
8×8	42.7	269.2	42.8
12×12	40.5	262.8	39.2

Table 9: Performance comparison of different β schedules in diffusion-based reconstruction.

Methods	Avg. Score	Avg. Tokens
Linear ($\beta_{\max} = 0.02$)	42.7	269.2
Linear ($\beta_{\max} = 0.05$)	41.8	265.2
Cosine ($\beta_{\max} = 0.02$)	42.7	271.2
Cosine ($\beta_{\max} = 0.05$)	42.1	262.8
Squared Cosine ($\beta_{\max} = 0.02$)	42.4	269.1
Squared Cosine ($\beta_{\max} = 0.05$)	42.6	272.3

Settings of Probing Analysis. To rigorously evaluate whether the continuous latent thoughts $\mathbf{Z}$ effectively encode joint linguistic reasoning and visual semantics, we conducted a comprehensive probing analysis following established methodologies in representation learning. The pooled latent vectors $\mathbf{z}_{\text{agg}} = \phi_{\text{pool}}(\mathbf{Z})$, textual rationale embeddings $\mathbf{r} = \phi_{\text{pool}}(\mathbf{R})$, and visual anchors $\mathbf{v} = \phi_{\text{pool}}(\mathbf{V})$ were computed using the same hybrid pooling module (mean-max pooling) employed during training. All representations were subsequently projected to a 256-dimensional space via a lightweight, trainable linear layer and L2-normalized prior to classification to ensure scale-invariant comparison. For binary classification, positive pairs were constructed from matching (image, question) instances, while negative pairs were sampled in a balanced manner from three distinct strategies to prevent trivial solutions: (i) different-question/same-image, (ii) different-image/same-question, and (iii) random in-batch negatives. The final probing dataset comprised 1,000 labeled pairs (500 positive/500 negative) with a stratified 70/30 train/test split based on negative sampling type. We employed two distinct probe architectures: (i) a Linear Support Vector Machine (SVM) with a linear kernel optimized via 5-fold cross-validation on the training set, and (ii) a single-layer Multi-Layer Perceptron (MLP) implemented in PyTorch consisting of one hidden layer, optimized with Adam ($\text{lr} = 1 \times 10^{-3}$) and early stopping based on validation loss.

B Some Supplement Experiments Results

B.1 More Ablations

Size of the Sliding Window. We further investigate the impact of the sliding window size in our dynamic token selection mechanism. As presented in Table 8, the window size significantly affects both model performance and efficiency. Extremely small windows (4×4) capture insufficient contextual information, leading to suboptimal performance across all benchmarks. As the window size increases to 8×8 , we observe substantial improvements in accuracy while maintaining efficient token usage. This suggests that an 8×8 region provides an optimal balance between focus and context. Further increasing the window size to 12×12 results in performance degradation, likely due to the inclusion of excessive irrelevant visual information that dilutes the salient features. These findings validate our choice of 8×8 as the default window size, demonstrating its effectiveness in selecting semantically meaningful visual regions for the reasoning process.

Schedule of β in Image Reconstruction. We investigate the sensitivity of CoCoVa to the noise schedule β used in the diffusion-based reconstruction objective. As shown in Table 9, we compare linear, cosine, and squared cosine schedules across different maximum β values. The cosine schedule with $\beta_{\max} = 0.02$ achieves the best balance between reconstruction quality and reasoning performance, yielding the highest average score of 42.7. While linear schedules demonstrate competitive performance, they exhibit slightly higher variance across tasks. The squared cosine schedule shows comparable results but requires more careful tuning of the maximum β value. These results indicate that CoCoVa is reasonably robust to different schedule choices, though smoother transitions in the noise schedule (as in cosine) appear to provide more stable training dynamics for the latent reasoning process.

Impact of Prefix Length. We investigate the sensitivity of CoCoVa to the length of the latent prefix used during training. As shown in Figure 10, model performance exhibits a clear pattern of improvement followed by saturation as prefix length increases from 2 to 128 tokens. Shorter prefixes (2-8 tokens) prove insufficient for capturing the complete

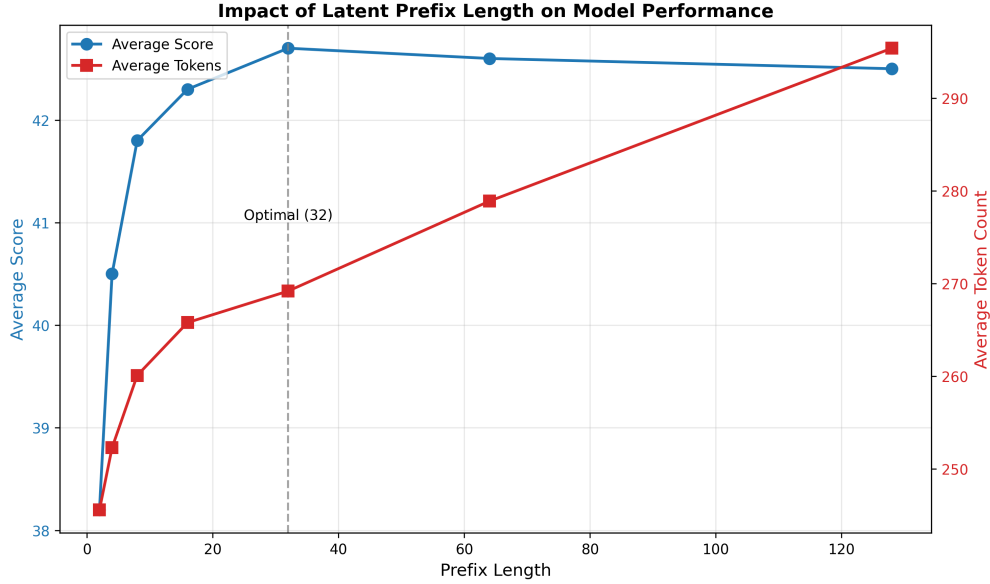


Figure 10: Effect of latent prefix length on model performance and efficiency. Performance improves with longer prefixes before saturating, while computational cost increases linearly. The optimal balance is achieved at 32 tokens.

Table 10: Performance of interleaved reasoning variants with different latent reasoning lengths.

Methods	MMBench		MMStar		MMVP		ScienceQA		LLaVA-W	
	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	Acc.	# Tokens	ROUGE-L	# Tokens
<i>Reasoning Steps: 2</i>										
CoCoVa	61.9	263.3	33.8	469.6	40.5	189.2	44.2	339.7	25.1	139.5
Interleaved CoCoVa	59.7	265.4	33.5	472.3	39.2	185.6	44.8	352.1	14.6	135.2
<i>Reasoning Steps: 4</i>										
CoCoVa	63.3	256.7	36.0	465.1	42.8	171.9	45.6	330.1	25.7	122.2
Interleaved CoCoVa	63.1	278.9	35.8	490.5	42.5	192.3	45.3	365.8	15.4	148.7
<i>Reasoning Steps: 6</i>										
CoCoVa	63.4	257.2	36.2	468.2	43.0	165.4	45.8	332.6	25.9	119.8
Interleaved CoCoVa	63.6	295.3	36.5	508.7	43.4	201.5	46.2	388.4	16.1	162.3
<i>Reasoning Steps: 8</i>										
CoCoVa	63.4	259.3	36.2	469.1	42.9	166.1	45.7	335.2	25.6	121.5
Interleaved CoCoVa	63.7	312.8	36.6	525.6	43.6	215.8	46.5	412.7	16.3	178.9

reasoning trajectory, resulting in suboptimal performance. The optimal balance is achieved at our default setting of 32 tokens, where the model captures sufficient contextual information without introducing unnecessary redundancy. Extending the prefix beyond this point yields diminishing returns, with performance plateauing and computational cost increasing linearly. These findings validate our architectural choice and demonstrate that a moderate-length latent prefix effectively balances representational capacity with training efficiency.

B.2 Other variants of CoCoVa

Interleaved Chain-of-Thought in Latent Space. We further explore an interleaved reasoning paradigm where textual rationales and latent thoughts alternate during the generation process. In this variant, a newline character "\n" serves as both a semantic and empirical delimiter[Gao et al., 2025] indicating the completion of the current reasoning segment and triggering the generation of subsequent latent thoughts. We investigate four configurations with latent reasoning lengths of 2, 4, 6, and 8 steps between textual segments. As shown in Table 10, the interleaved approach demonstrates an interesting trade-off: while it achieves competitive or even slightly improved accuracy on complex reasoning benchmarks like MMBench and MMStar with longer reasoning sequences, this comes at the cost

of significantly increased token counts across all configurations. Notably, the text generation quality as measured by ROUGE-L on LLaVA-Bench shows a substantial decrease in the interleaved setting, suggesting that the alternation between explicit textual reasoning and implicit latent reasoning may disrupt the coherence of the final output. The results suggest that the purely latent reasoning approach of standard CoCoVa provides a more efficient pathway to accurate conclusions, though the interleaved variant shows promise for tasks requiring deeper reasoning chains.

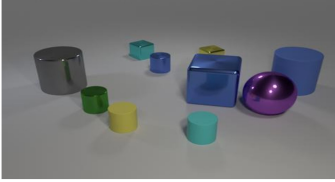
					
What will happen next? A: the car is gonna crash into the house B: the car is gonna fly C: the car is gonna drive backwards D: both A,B, and C		The other small shiny thing that is the same shape as the tiny yellow shiny object is what color? A: red B: cyan C: purple D: brown		How many apples are there in the image? And how many bananas are there? A: 0 apples and 5 bananas B: 1 apples and 4 bananas C: 0 apples and 4 bananas D: 1 apples and 5 bananas	
LLaVA-NeXT-7B	B	LLaVA-NeXT-7B	A	LLaVA-NeXT-7B	B
Chameleon-7B	B	Chameleon-7B	C	Chameleon-7B	C
CoCoVa-1.5B	A	CoCoVa-1.5B	B	CoCoVa-1.5B	D

Figure 11: Qualitative comparisons on MMBench [Liu et al., 2024c].

							
Can you identify the season in which the picture was taken? A: spring B: summer C: fall D: winter		How many people are visible in this picture? A: three B: six C: seven D: eight		How many tables are there in the room? A: Four tables B: Two tables C: Three tables D: One table		What is the total volume of the measuring cup? (Unit: g) A: 1000 B: 2000 C: 800 D: 1200	
LLaVA-NeXT-7B	B	LLaVA-NeXT-7B	C	LLaVA-NeXT-7B	A	LLaVA-NeXT-7B	D
Chameleon-7B	B	Chameleon-7B	B	Chameleon-7B	B	Chameleon-7B	C
CoCoVa-1.5B	A	CoCoVa-1.5B	D	CoCoVa-1.5B	D	CoCoVa-1.5B	A

Figure 12: Qualitative comparisons on MMStar [Chen et al., 2024a].

C Case Study

We conduct qualitative comparisons across a diverse set of representative examples to intuitively demonstrate the behavior and advantages of CoCoVa. Figures 11 and 12 present multiple-choice examples from MMBench and MMStar, Figure 13 shows ScienceQA examples covering both geometry and commonsense reasoning, and Figure 14 illustrates temporal and situational inference cases. For each example, we display the final predictions of both baseline models and CoCoVa, and visualize in Figure 15 the attention maps across four latent reasoning steps, revealing how the model progressively refines its visual focus and internal evidence alignment.

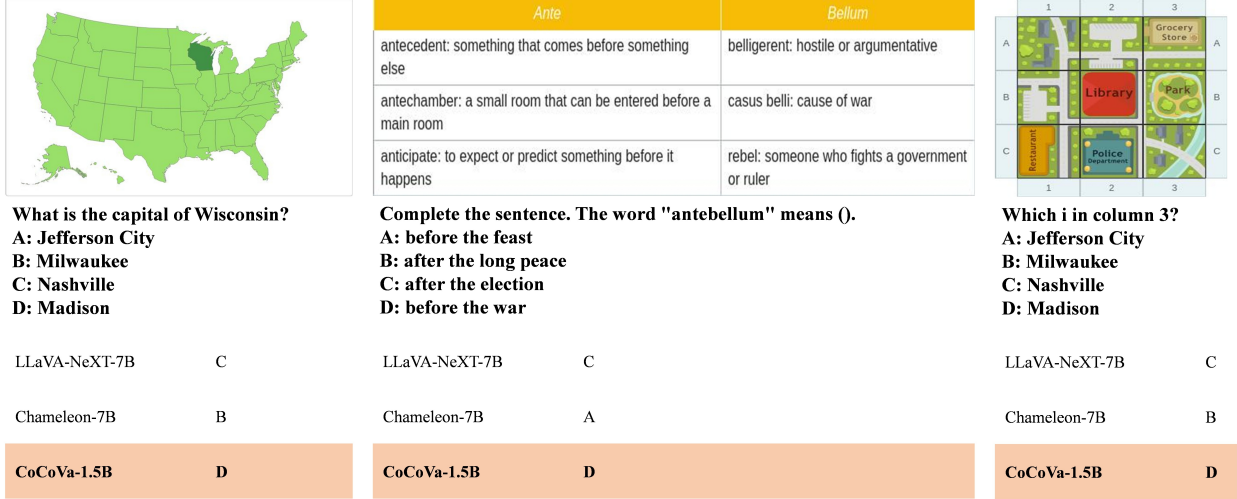


Figure 13: Qualitative comparisons on MMBench [Lu et al., 2022a].

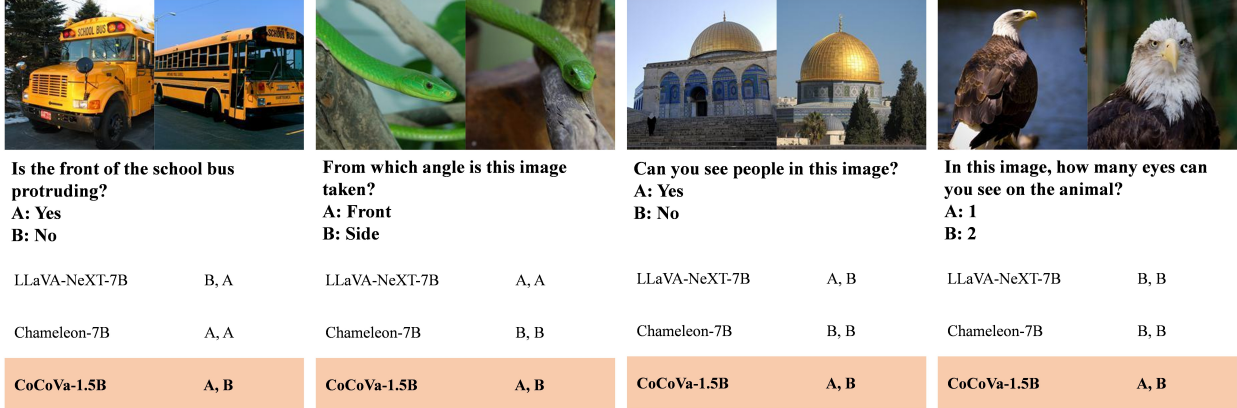


Figure 14: Qualitative comparisons on MMBench [Tong et al., 2024].

As shown in Figures 11 and 12, CoCoVa provides more consistent and accurate predictions compared to baseline models. This suggests that multi-step continuous latent reasoning compensates for the missing visual and semantic information inherent in single-pass projection. Moreover, the attention evolution in Figure 15 shows a clear progression from broad, dispersed focus to highly localized regions of interest: Step 1 generally covers wide candidate areas, Steps 2–3 strengthen attention on the most task-relevant details (such as a person’s face, held objects, or key vertices in a geometric diagram), and Step 4 culminates in concentrated activations over a few critical regions. This trajectory reflects the LQ-Former’s ability to perform a *global-to-local* information aggregation process over multiple reasoning steps, enhancing the quality of visual evidence that supports the final language decoding.

For visual counting and object recognition tasks (MMBench / MMStar; Figures 11, 12), CoCoVa’s dynamic token selection effectively suppresses background noise and concentrates on the relevant object clusters, reducing both undercounting and misclassification. In temporal and situational reasoning tasks (MMVP; Figure 14), the iterative latent reasoning enables the model to internally simulate motion patterns or scene progressions, thus improving accuracy in “next-step prediction” problems. For geometry-related ScienceQA questions (Figure 13), the evolving attention successfully highlights critical vertices, arcs, and chords related to angle relationships, allowing the language model to extract the necessary spatial information for correct reasoning.

We also observe a few failure cases where CoCoVa’s attention converges to secondary or misleading features, particularly under complex backgrounds or multiple competing salient regions (as shown in several examples of Figure 12). In such scenarios, the single-window selection strategy can occlude important contextual information, leading to incorrect predictions. These issues suggest two promising improvement directions: (1) incorporating multi-region selection or

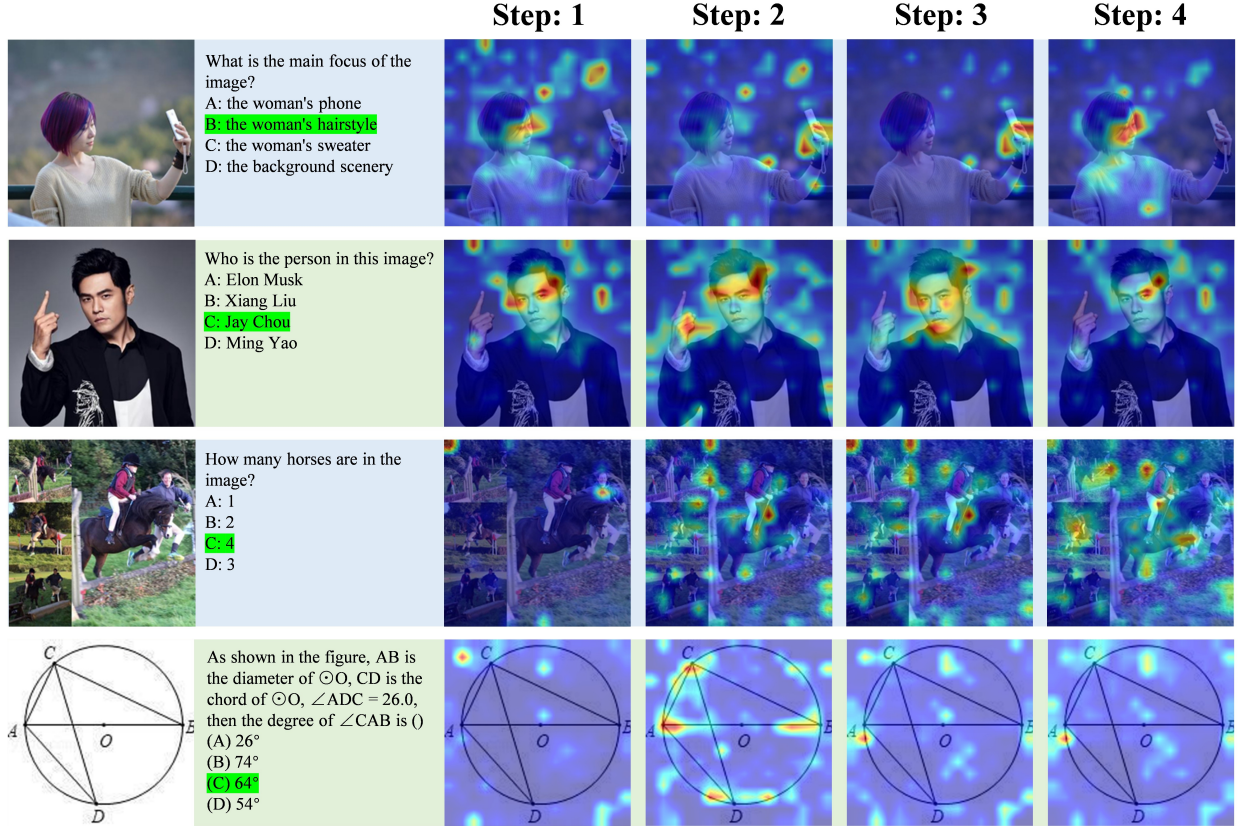


Figure 15: Attention visualization across four latent steps for different questions.

multimodal verification to prevent information loss, and (2) introducing an adaptive halting mechanism that allows the model to stop earlier upon convergence or extend its search when additional evidence is needed.

Overall, these qualitative examples provide intuitive evidence that CoCoVa’s continuous latent reasoning and dynamic visual selection mechanisms produce more focused and interpretable vision-language representations, leading to more robust performance in downstream reasoning and multi-choice decision tasks. Nonetheless, further refinements in multi-region attention and adaptive computation are needed to better handle complex scenes and multi-object settings.

D Discussion

This study demonstrates that operating in a continuous latent reasoning space provides significant advantages over discrete linguistic reasoning for vision-language models. Our experimental results consistently show that CoCoVa achieves superior performance while maintaining greater token efficiency across diverse benchmarks. These findings suggest that the conventional approach of verbalizing all reasoning steps creates an expressive bottleneck that limits model capabilities.

The emergent properties of the continuous reasoning space offer particular insight. The topological organization of latent thoughts into semantically meaningful clusters, combined with the convergence patterns observed in reasoning trajectories, indicates that our framework develops a structured representation of different reasoning types. This geometric organization provides a plausible explanation for both the performance improvements and computational efficiency of our approach.

Several practical implications emerge from our work. The ability of smaller models equipped with continuous reasoning to compete with significantly larger conventional models suggests an alternative pathway for scaling multimodal intelligence. Furthermore, the substantial reduction in output tokens makes advanced reasoning capabilities more feasible for deployment in resource constrained environments.

However, certain limitations warrant acknowledgment. Our current implementation uses a fixed number of reasoning steps in its default configuration, though our exploration of adaptive mechanisms represents a preliminary step toward more dynamic computation. Additionally, while our analyses provide substantial insight into the model’s internal processes, a more comprehensive theoretical understanding of continuous reasoning dynamics would strengthen the field.

Future research directions appear promising. Extending the continuous reasoning paradigm to temporal domains such as video understanding presents an interesting challenge. Developing more sophisticated adaptive mechanisms that can dynamically adjust both the number of steps and the focus of attention across modalities represents another valuable direction. Exploring how continuous reasoning interacts with other emerging capabilities such as tool use and world models could lead to more generally intelligent multimodal systems.

In conclusion, CoCoVa establishes continuous latent space reasoning as an effective alternative to discrete Chain of Thought approaches. By enabling iterative refinement of multimodal representations before verbalization, our framework bridges the representational gap between continuous visual perception and discrete linguistic expression. We believe this work contributes meaningful progress toward developing multimodal systems that better capture the fluid nature of human thought.