Data-Efficient Latent Reasoning Training

Joyce Lu

Nov. 5 2025

Abstract

This project proposes a data-efficient pretraining framework to enable multimodal large language models (MLLMs) to acquire latent reasoning, the ability to internally represent and manipulate visual concepts in a compact latent space. While existing MLLMs achieve strong perceptual understanding, they primarily focus on visual recognition and captioning rather than internal visual imagination or reasoning. Developing these capabilities typically requires expensive interleaved reasoning datasets that are hard to scale.

The objective of this work is to design a pretraining objective that leverages abundant image–text paired datasets to train models to predict visual latent embeddings from textual descriptions. By aligning predicted and target latents derived from a frozen vision encoder, the model learns to "imagine" latent visual features consistent with the underlying image. This pretraining acts as a latent reasoning stage, improving the model's multimodal grounding and reasoning ability before standard instruction fine-tuning.

The proposed work builds upon existing vision–language model backbones (e.g., Qwen2.5-VL) by introducing a lightweight latent prediction head and a multi-objective training loss combining text generation and latent prediction. Experiments will compare standard vision language models (VLMs) with the latent-pretrained model on various reasoning benchmarks, evaluating reasoning accuracy, alignment quality, and data efficiency.

The anticipated outcomes include (1) a novel latent prediction pretraining method that repurpose large-scale image-text data for learning latent reasoning, (2) analytical insights into how latent alignment affects multimodal reasoning, and (3) empirical evidence that latent pretraining improves reasoning and generalization without costly supervision.

1 Motivation

1.1 Problem Statement

The goal of this project is to develop a scalable and data-efficient pretraining strategy to train multimodal large language models (MLLMs) for latent reasoning, the ability to internally represent and manipulate visual concepts in a continous latent space, using readily available image-text datasets.

Current MLLMs show impressive visual understanding, but still struggle with reasoning tasks that require imagination, planning, or spatial abstraction. This limitation arises as training such reasoning abilities often require large amounts of interleaved reasoning datasets, where text and visual reasoning are interleaved. However, such data are expensive and difficult to collect and are often domain-specific.

However, we already possess abundant paired image-text data that encode rich visual-semantic information but lack detailed reasoning supervision. This proposal aims to bridge the gap between these abundant datasets and reasoning-oriented latent training by enabling models to predict and align visual latents from text, allowing the model to learn visual grounding in a latent space.

This raises a question: ”Can we repurpose existing large-scale image-text data to pretrain the models with latent reasoning capability?”

1.2 Background and Context

Existing MLLMs have achieved remarkable progress in visual understanding by combining pretrained vision encoders with large language models. Techniques such as Chain-of-Thought (CoT) [2] further enables models to perform step-by-step reasoning to solve complex problems. However, these models primarily excel at perception and captioning tasks, while their ability to perform visual imagination or latent reasoning remains limited.

Humans naturally complement verbal reasoning with mental imagery to aid understanding and problem solving in complicated visual reasoning tasks. Motivated by this, recent research has explored MLLMs with visual thinking to enhance reasoning. Recent frameworks such as Latent Sketchpad [1] and Mirage [3] attempt to enhance multimodal reasoning by training models to generate or predict visual latent features during reasoning. While demonstrating the benefits of predicting visual latents for reasoning, these frameworks typically rely on external tools or interleaved text-and-image reasoning datasets that are costly and hard to scale.

At the same time, massive collections of image–text pairs are abundant and easy to collect but remain underutilized for reasoning-centric pretraining. As language models scale in size and capability, there is a growing need for data-efficient objectives that can leverage such widely available text-image paired data to develop richer reasoning capabilities.

This project explores whether such data can be leveraged through a latent reasoning pretraining objective to instill internal visual imagination and improve downstream multimodal reasoning.

1.3 Limitations of Existing Approaches

Current approaches still suffer from several limitations:

- **Data Inefficiency and Dependence on Curated Reasoning Sets.** Most MLLMs depend on task-specific multimodal supervision, such as interleaved text-image reasoning traces or instruction-tuned datasets, which are costly to curate and difficult to scale across domains.
- **Lack of Latent Visual Reasoning.** Existing MLLMs like Qwen2.5-VL are trained to process and describe visual inputs but not to generate or predict visual representations themselves. As a result, they lack the ability to internally simulate visual context or perform reasoning in visual latent space.
- **Limited Generalization Across Reasoning Tasks.** Existing latent reasoning frameworks are often trained on smaller, domain-specific datasets (e.g., spatial puzzles, navigation tasks), limiting their ability to generalize to diverse visual tasks.

1.4 Objectives

The objectives of this proposal are:

- **Latent Reasoning Pretraining.** Introduce a scalable pretraining Objective for MLLMs, where the model learns to predict visual latent embeddings from textual inputs, enabling internal visual imagination.
- **Data-Efficient Multimodal Latent Reasoning Learning.** Leverage large-scale, image-text paired datasets to learn cross-modal structure in a cost-effective and scalable way, without requiring reasoning annotations.
- **Evaluate Reasoning Gains and Latent Alignment.** Compare the latent-pretrained model to baseline VLMs across visual reasoning benchmarks to determine whether latent reasoning pretraining improves multimodal reasoning performance. Analyze how latent space alignment and structure relate to downstream reasoning ability.

2 Approach

2.1 Overview

We propose to enhance a pretrained MLLM with an additional latent prediction head and a new latent prediction pretraining objective. The goal is to encourage the model to develop an internal latent reasoning capability that better aligns its textual and visual representations. The high-level pipeline is as follows:

1. **Image Encoding and Latent Target Extraction.** Each image from an image–text dataset is first encoded by a frozen, pretrained vision encoder to obtain a compact visual latent representation. These latents capture the semantic and compositional structure of the image and serve as supervision targets during pretraining.

2. **Latent Prediction.** The paired text (e.g. caption or description) is processed by the language model to produce hidden states h. A lightweight projection module – the latent prediction head $M_\phi(\cdot)$ – maps these hidden states into the same latent space as the vision encoder, which we denote as

$$\hat{z} = M_\phi(h).$$

   , representing the predicted latent. This trains the language model to predict how an image would be represented in visual latent space.

3. **Pretraining Objective.** During pretraining, the model jointly optimizes the standard next-token objective along with additional latent-level objectives that encourage the predicted latents to be consistent with those of the visual encoder. These objectives include both instance-level similarity between the predicted and target latents and broader semantic alignment across batches. This helps the model learn to organize its internal representations in a way that reflects visual semantics.

4. After latent pretraining, the model is fine-tuned on downstream multimodal reasoning benchmarks (e.g., GQA, ScienceQA).

2.2 Design Principle

- **Data Efficiency.** We leverage large-scale image–text datasets that are easy to collect to provide a scalable source of multimodal supervision without requiring expensive reasoning annotations.

- **Utilize Pretrained Visual Knowledge.** The vision encoder is frozen to retain its existing visual features, ensuring the model focuses on learning reasoning dynamics rather than low-level visual reconstruction.

- **Internal Visual Simulation.** A latent prediction head establishes an internal pathway for the model to simulate and predict latent visual representations consistent with the vision encoder, forming the basis for internal "mental imagery."

- **Multi-objective Loss.** The model is optimized using a standard next-token loss, a similarity-based latent prediction loss that aligns predicted and visual latents, and an alignment loss that promotes semantic consistency. Together, they encourage the model to learn representations that are both accurate and semantically aligned across vision and language.

3 Experiments

3.1 Experimental Setup

Datasets. For pretraining, we will use large-scale image–text datasets such as COCO Captions, Conceptual Captions (3M), and LAION-400M. These datasets provide diverse caption-style supervision without explicit reasoning traces, enabling scalable learning of latent visual representations. Downstream evaluation will be conducted on established multimodal reasoning benchmarks such as GQA (compositional visual QA), ScienceQA (multimodal science QA), and BLINK-Jigsaw (visual perception), which test visual grounding, compositional reasoning, and knowledge-based reasoning capabilities.

Models and Training. We will use Qwen2.5-VL (7B) as our backbone model, which combines a pretrained vision encoder and a large language model. The proposed latent prediction head will be trained jointly with the language backbone under the latent prediction objective, while the vision encoder remains frozen.

Evaluation Protocol. We will compare against three representative open-source VLMS: Qwen2.5-VL, LLaVA, InternVL2.5. All models are fine-tuned and evaluated on the same multimodal reasoning benchmarks (e.g., GQA, ScienceQA) under identical settings. Results from these standard VLMs serve as baselines for comparison with our latent-pretrained VLM, which applies the proposed latent reasoning pretraining stage to the Qwen2.5-VL backbone before fine-tuning. All models will be trained using 8×A100 GPUs.

Evaluation Metrics. We will measure task-specific reasoning accuracy and answer correctness (e.g., accuracy on GQA and ScienceQA).

Ablation Studies. Some experiments we will conduct include comparing different weighting for latent prediction loss and different similarity functions (e.g., cosine, L2, contrastive). For qualitative evaluation, latent space visualizations and clustering analyses will be used to study how alignment in latent space correlate with reasoning performance.

Timeline

Month 1: Dataset preparation, baseline LLM setup.
Month 2-3: Implementation of latent prediction pretraining, ablation studies.
Month 4: Downstream evaluation, analysis, and report writing.

Expected Contributions

- A latent prediction pretraining method that efficiently reuses image–text data to enable latent reasoning capability.
- Demonstration that large-scale, image-text paired datasets can provide efficient resource for latent reasoning, reducing dependence on costly interleaved reasoning datasets.

- Empirical evidence showing latent reasoning pretraining improves reasoning accuracy and generalization compared to standard VLM baselines.

References

[1] Anonymous. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in MLLMs. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[3] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens, 2025.