

Interleaved Latent Visual Reasoning with Selective Perceptual Modeling

Shuai Dong^{1,2}, Siyuan Wang^{3*}, Xingyu Liu¹, Zhongyu Wei^{2,4*}

¹China University of Geosciences, Wuhan

²Shanghai Innovation Institute

³University of Southern California ⁴Fudan University

dongshuai_iu@cug.edu.com, sw_641@usc.edu

liuxingyu@cug.edu.cn, zywei@fudan.edu.cn

Abstract

Interleaved reasoning paradigms enhance Multimodal Large Language Models (MLLMs) with visual feedback but are hindered by the prohibitive computational cost of repeatedly re-encoding pixel-dense images. A promising alternative, latent visual reasoning, circumvents this bottleneck yet currently forces a critical trade-off: methods either sacrifice precise perceptual modeling by over-compressing features or fail to model dynamic problems due to static, non-interleaved structures. We introduce Interleaved Latent Visual Reasoning (ILVR), a framework that unifies dynamic state evolution with precise perceptual modeling. ILVR interleaves textual generation with latent visual representations that act as specific, evolving cues for subsequent reasoning. To enable this, we employ a self-supervision strategy where a Momentum Teacher Model selectively distills relevant features from *helper images* into sparse supervision targets. This adaptive selection mechanism guides the model to autonomously generate context-aware visual signals. Extensive experiments on multimodal reasoning benchmarks demonstrate that ILVR significantly outperforms existing approaches, effectively bridging the gap between fine-grained perception and sequential multimodal reasoning. The code is available at <https://github.com/XD111ds/ILVR>.

1 Introduction

Multimodal Large Language Models (MLLMs) (Li et al., 2024; Bai et al., 2025a; Wang et al., 2025b) have demonstrated remarkable capabilities in bridging the gap between vision and language. Capitalizing on the reasoning prowess of Large Language Models, recent works have successfully adapted Chain-of-Thought (CoT) methodologies to the multimodal domain (Zhang et al., 2023; Bai et al., 2025b; Huang et al., 2025a; Wei et al., 2022). This

enables models to decompose complex visual tasks into sequential intermediate steps, achieving sophisticated reasoning grounded in visual content.

Recent work explores interleaved image-text reasoning by injecting intermediate visual images within textual CoTs, further enhancing multimodal understanding and planning (Shao et al., 2024; Deng et al., 2025; Chern et al., 2024). These approaches generally fall into two paradigms. The first uses external tools to statically manipulate the input image, e.g., highlighting key regions (Fu et al., 2025),

drawing auxiliary lines (Hu et al., 2024), or shifting image styles (Liu et al., 2025), to improve fine-grained perception. While remaining reliance on a single visual state, it is unable to model evolving scenarios or simulate action outcomes crucial for sequential tasks (Li et al., 2025a). The second paradigm addresses this by dynamically visualizing imagined intermediate or future states (Chern et al., 2024; Deng et al., 2025). However, integrating visual generation and reasoning into a unified model often degrades reasoning performance. More critically, both paradigms incur prohibitive computational cost from iteratively re-encoding pixel-dense images, severely hindering multi-step reasoning.

Inspired by latent reasoning in LLMs (Shen et al., 2025; Hao et al., 2024), the emerging paradigm of latent visual reasoning (Yang et al., 2025; Li et al., 2025b) replaces explicit images with implicit latent representations to circumvent high pixel-level processing costs. However, current methods face two major limitations, as illustrated in Fig. 1. First, existing approaches generally adopt a static, non-interleaved structure (Li et al., 2025b; Yang et al., 2025). For instance, LVR (Li et al., 2025b) focuses on enhancing perception by generating a latent representation of a specific region within the *static input image*. However, this design prevents dynamic state evolution. In the chess puzzle shown in Fig. 1(a), simply zooming in on the

* Corresponding authors.

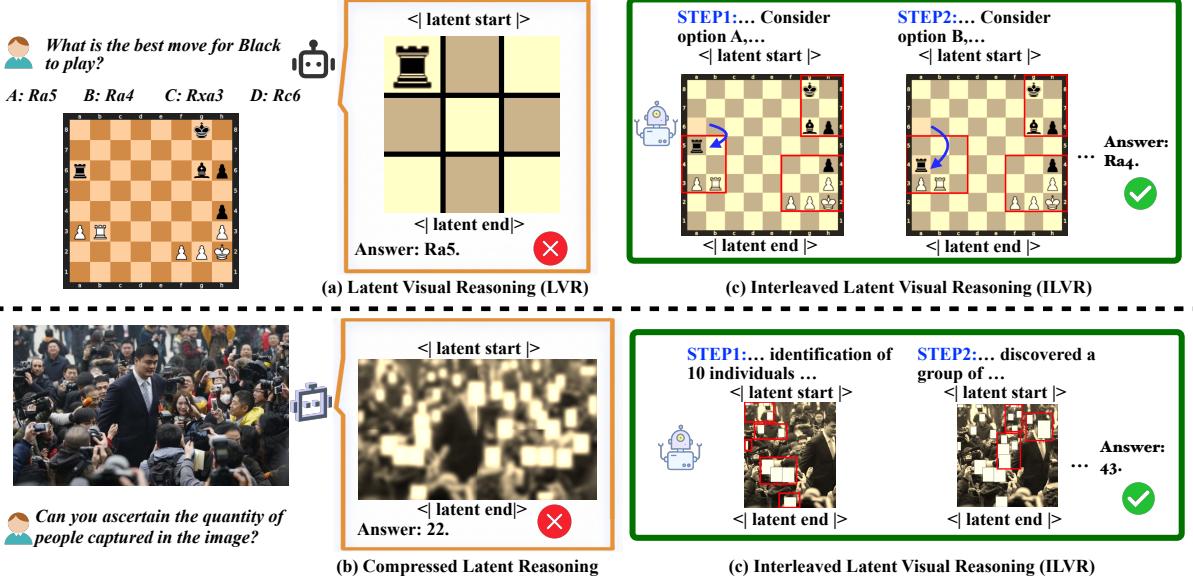


Figure 1: **Comparison of ILVR with prior latent visual reasoning methods.** On a multiple-choice chess puzzle (top row), prior approaches like LVR (a) are limited to representing static details of the initial input (e.g., a zoomed-in view of the rook), failing to model the hypothetical board states required to evaluate different options. On a dense counting task (bottom row), methods relying on heavily compressed latent representations (b) lose fine-grained details, resulting in a hallucinated count. In contrast, our proposed ILVR (c) successfully addresses both tasks by interleaving textual reasoning with dynamically updated latent states. Each latent representation provides specific visual cues essential for facilitating the subsequent reasoning step (visualized by red boxes), unifying dynamic evolution with precise perceptual modeling to arrive at the correct answer.

rook is insufficient; the model fails because it cannot simulate the hypothetical board states resulting from different candidate moves. Second, methods like Mirage (Yang et al., 2025) derive latent representations by highly compressing dense visual features to support textual generation. As shown in the dense counting task in Fig. 1(b), this over-compression inevitably discards crucial perceptual details and leads to hallucinated answers.

To this end, we propose Interleaved Latent Visual Reasoning (ILVR), a framework that combines dynamic latent visual reasoning with precise perceptual modeling. ILVR interleaves reasoning between explicit textual generation and dynamically updated latent visual representations that capture selectively filtered visual cues essential at each reasoning step. We train the model to learn such interleaved paradigm by approximating ground-truth interleaved image-text trajectories, with text outputs supervised using cross-entropy loss while the latent visual representations are aligned with the embedding of their corresponding image counterparts, which we refer to *helper images*. To enable precise perceptual modeling, we employ a Momentum Teacher Model (He et al., 2019), a temporally smoothed copy of the model being trained, to con-

struct latent supervision targets at each reasoning step. This mechanism selectively distills the most relevant features from *helper images* by aggregating highly attended patches alongside the reasoning process. By internalizing this capability, ILVR effectively unifies precise perceptual modeling with the dynamic evolution of latent visual states.

In summary, our contributions are threefold:

- We propose Interleaved Latent Visual Reasoning (ILVR), a reasoning framework that interleaves explicit token generation and iteratively updated latent visual representations, allowing dynamic state evolution.
- We introduce an adaptive selection mechanism to distill context-relevant visual signals from a *helper image* into latent representations at each reasoning step. This is achieved through a self-supervised strategy guided by a Momentum Teacher Model, eliminating the need for external supervision.
- Through extensive experiments on challenging multimodal reasoning benchmarks, we validate the effectiveness of our method in integrating fine-grained perception with the

interleaved latent reasoning required for complex, evolving tasks.

2 Related Work

2.1 Interleaved Image-Text Reasoning

Interleaved image-text reasoning refers to the capability of models to generate intermediate visual feedback (Chern et al., 2024; Li et al., 2025a; Deng et al., 2025), either directly or via external tools (Hu et al., 2024; Shao et al., 2024; Su et al., 2025), to enhance their reasoning abilities. Early approaches focused on augmenting fine-grained perception by invoking external tools to manipulate a static input image. These range from spatial utilities like cropping to specialized modules such as OCR (Huang et al., 2025b) or chart parsers, with tool invocation learned through supervised finetuning (Wang et al., 2025a; Zhang et al., 2025a) or reinforcement learning (Zhang et al., 2025b; Yu et al., 2025; Wu et al., 2025; Geng et al., 2025). While effective for perception, these methods are fundamentally unable to model problems where the visual state evolves. Recent work leverages foundational models that natively support multimodal generation (Chern et al., 2024; Deng et al., 2025), enabling them to create entirely new images depicting intermediate or future task states. This generative approach, however, introduces a new trade-off, as maintaining high-fidelity generation can compromise the model’s reasoning capabilities compared to text-centric MLLMs. More fundamentally, both tool-based manipulation and dynamic image generation are constrained by the same bottleneck: the prohibitive computational cost of repeatedly re-encoding pixel-dense visual feedback.

2.2 Latent Reasoning

Inspired by the representational richness of continuous hidden states, a line of work in natural language processing (Shen et al., 2025; Hao et al., 2024; Cheng and Durme, 2024) has explored reasoning directly in the latent space, moving beyond the constraints of discrete token generation. This concept has been extended to multimodal reasoning (Yang et al., 2025; Li et al., 2025b), where images are condensed into latent representations to guide text generation. For instance, Mirage (Yang et al., 2025) learns to precede its textual reasoning with a self-generated latent representation. This latent is formed by encoding a problem-specific *helper*

image and then aggressively pooling its patch embeddings into highly compressed vectors. While this provides a guiding signal, the heavy compression sacrifices crucial details, limiting the model’s capacity for fine-grained perception. LVR (Li et al., 2025b) adopts a similar strategy but isolates key visual cues within a bounding box, generating latent representations of only that targeted region. Despite these variations, a more fundamental limitation plagues both approaches. In their paradigm, a model generates latent representations of a *helper images* once, and all subsequent steps are confined to pure textual reasoning. This non-interleaved structure inherently renders the visual information static and detached from the evolving reasoning trajectory.

3 Method

In this section, we present Interleaved Latent Visual Reasoning (ILVR), a framework that performs reasoning by interleaving explicit textual generation with latent visual representations. We first introduce the interleaved generation paradigm. Then, we detail how we construct latent supervision targets by extracting key features from helper images using a momentum teacher. Finally, we describe the two-stage training strategy used to instill this interleaved capability.

3.1 Interleaved Latent-Text Paradigm

Our framework operates in an interleaved reasoning paradigm where the model autoregressively generates both text and latent representations. The reasoning process is structured as a unified sequence \mathcal{S} containing textual tokens interspersed with latent segments:

$$\mathcal{S} = [t_{1,1}, \dots, t_{1,M}, <|\text{latent_start}|>, z_{1,1}, \dots, z_{1,K}, <|\text{latent_end}|>, t_{2,1}, \dots, t_{2,N}, <|\text{latent_start}|>, z_{2,1}, \dots, z_{2,K}, <|\text{latent_end}|>, \dots] \quad (1)$$

where $t_{i,j}$ denotes discrete text tokens and $z_{i,k}$ represents continuous latent embeddings. The special tokens $<|\text{latent_start}|>$ and $<|\text{latent_end}|>$ explicitly delimit the boundaries of the latent visual reasoning phases.

During inference, the model generates text tokens normally. When the model produces a $<|\text{latent_start}|>$ token, it switches to a latent generation mode for a fixed number of steps K . In this phase, instead of projecting the hidden state to the vocabulary size to sample a discrete token,

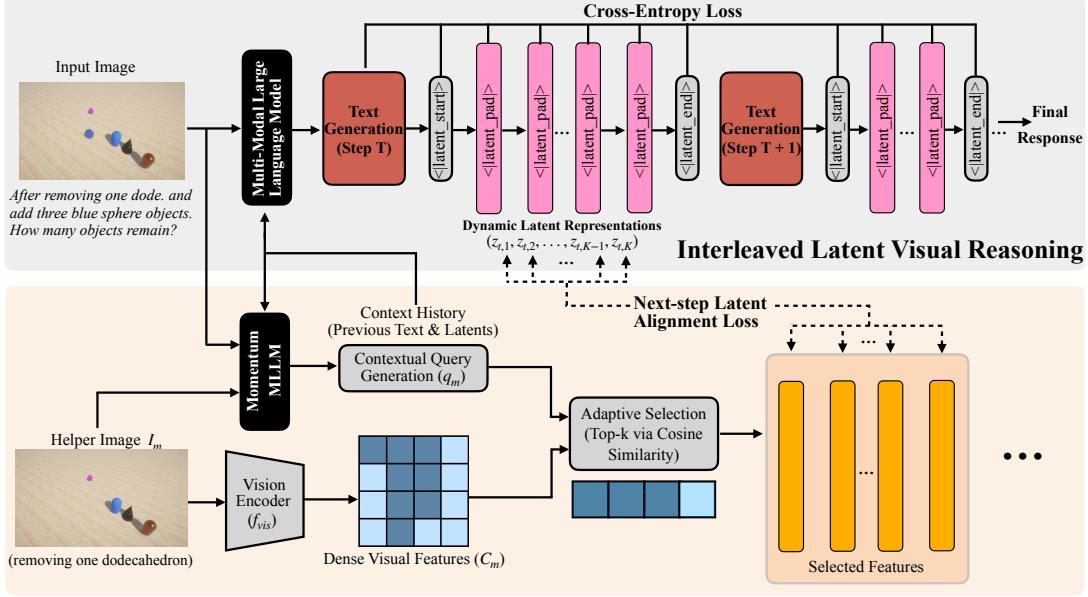


Figure 2: **The Interleaved Latent Visual Reasoning (ILVR) framework.** The model performs multi-step reasoning by interleaving textual generation with dynamic latent visual representations. Given a multimodal input, the Momentum Teacher Model (bottom) utilizes the current context and latent representations history to generate a Contextual Query (q_m), which selectively extracts the most relevant visual patches (yellow blocks) from a *helper image*. Simultaneously, the model being trained (top) generates a sequence of Latent Representations (pink blocks) interleaved with reasoning text. These generated latents are supervised via a *Next-step Latent Alignment* objective to match the Momentum-selected key visual features, enabling the model to ground its reasoning in precise, evolving visual evidence.

the hidden state from the previous step \mathbf{h}_t is used directly as the input embedding for the current timestep, effectively bypassing the discrete embedding lookup:

$$\mathbf{e}_{t+1} = \mathbf{h}_t. \quad (2)$$

The sequence of K hidden states produced in this loop constitutes the model’s self-generated latent representation. After K steps, the model generates $\langle|\text{latent_end}|>$ and resumes explicit textual reasoning, utilizing the accumulated latent information as context.

To train the model for this paradigm, we utilize pre-constructed interleaved sequences formatted as “reasoning text→helper image→reasoning text→helper image ...”. We convert this structure into a unified supervision sequence by replacing each *helper image* I_m with a latent segment: $\langle|\text{latent_start}|>$ followed by K instances of $\langle|\text{latent_pad}|>$ and terminated by $\langle|\text{latent_end}|>$. Within this sequence, the $\langle|\text{latent_pad}|>$ tokens serve as placeholders designated to reconstruct specific visual signals extracted from I_m . Thus, the core of our method lies in determining exactly which visual features from I_m should serve as the regression targets to

supervise the hidden states generated at these pad positions.

3.2 Interleaved Latent-Text Supervision Construction

To enable the model to generate meaningful latent representations, we need to construct high-quality supervision targets. We employ a teacher model to perform this task. Given the same reasoning context as the model being trained, the teacher analyzes the corresponding *helper image* I_m and selects the most relevant visual features to serve as the ground-truth latent supervision. The text parts are supervised using standard explicit textual supervision.

Momentum Teacher Model We opt for a self-supervision strategy where the teacher is a Momentum Model—a temporally smoothed copy of the model being trained (the *online model*). This ensures that the supervision signal remains stable and aligned with the representational space of the learning model. The parameters of the Momentum Model θ_m are updated as an Exponential Moving Average (EMA) of the online parameters θ with a

decay factor τ :

$$\theta_m \leftarrow \tau\theta_m + (1 - \tau)\theta. \quad (3)$$

Teacher-Guided Selective Perceptual Modeling

The goal of the momentum teacher is to distill the pixel-dense *helper image* into a sparse set of K feature vectors that are most critical for the current reasoning step.

First, the teacher encodes the *helper image* I_m using its frozen vision encoder f_{vis} to obtain a dense pool of visual features:

$$\mathbf{C}_m = f_{\text{vis}}(I_m) = \{\mathbf{c}_{m,j} \in \mathbb{R}^H\}_{j=1}^{P_m}, \quad (4)$$

where H is the hidden dimension and P_m is the number of patches.

However, raw patch features often suffer from varying information density depending on the image resolution. In high-resolution scenarios, individual patches may only capture local textures rather than semantic concepts. To address this, we introduce a spatial aggregation step to refine the candidate pool. We define a threshold L to regulate the feature density. If the number of raw patches P_m exceeds L , we aggregate the raw features by performing pooling over local spatial windows to form a refined candidate pool \mathbf{C}'_m . If the feature map is small ($P_m < L$), we retain the fine-grained features. This operation is formulated as:

$$\mathbf{C}'_m = \begin{cases} \text{GroupMean}(\mathbf{C}_m, L), & \text{if } P_m \geq L \\ \mathbf{C}_m, & \text{if } P_m < L \end{cases} \quad (5)$$

where *GroupMean* spatially aggregates the sequence into L semantic units, ensuring the subsequent selection operates on robust features regardless of input resolution.

Next, the teacher determines which candidates from \mathbf{C}'_m are most relevant. To do this, we compute an attention-weighted summary of the image based on the user's initial text query. Using weights (W_Q, W_K) from the model's attention module, we compute a text-to-image attention matrix \mathbf{A} :

$$\mathbf{A} = \frac{1}{\sqrt{H}} \mathbf{Q}_{\mathcal{P}}^{\text{pre}} (\mathbf{K}_{\mathcal{J}}^{\text{pre}})^{\top}. \quad (6)$$

This attention mechanism enables a differentiated pooling strategy. The user text is typically semantically dense, making a uniform average an effective summary. In contrast, image patches are spatially unstructured and often contain irrelevant

background; a simple average would dilute the important visual signals with noise. Therefore, attention allows the model to selectively focus on the most salient visual regions. Consequently, we form an attention-weighted visual summary \mathbf{r}_{img} and a mean-pooled textual summary \mathbf{r}_{txt} from the final-layer hidden states \mathbf{H}^{last} :

$$\mathbf{r}_{\text{img}} = \sum_{j \in \mathcal{J}} \bar{w}_j \mathbf{h}_j^{\text{last}}, \quad \mathbf{r}_{\text{txt}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbf{h}_p^{\text{last}}. \quad (7)$$

A global user intent vector is derived as $\mathbf{u} = \frac{1}{2}(\mathbf{r}_{\text{img}} + \mathbf{r}_{\text{txt}})$. Finally, for each specific reasoning step m , we construct a step-specific query \mathbf{q}_m by fusing the global intent, the local textual history $\mathbf{q}_m^{\text{text}}$, and the previous latent state $\bar{\mathbf{z}}_{m-1}$:

$$\mathbf{q}_m = \text{mean}(\mathbf{u}, \mathbf{q}_m^{\text{text}}, \mathbb{I}[m > 1] \cdot \bar{\mathbf{z}}_{m-1}). \quad (8)$$

The teacher then calculates the cosine similarity between \mathbf{q}_m and each feature in the refined pool \mathbf{C}'_m , selecting the top- K features to form the supervision set \mathbf{Z}_m .

3.3 Two-stage Learning

We employ a two-stage training pipeline to effectively train the online model using these constructed supervision signals.

Stage 1: Interleaved Latent-Text Joint Supervision

In the first stage, we enforce precise perceptual modeling. The features selected by the teacher in \mathbf{Z}_m are injected as inputs for the K tokens. We optimize the model using a joint loss: a standard cross-entropy loss \mathcal{L}_{CE} for text tokens, and a next-step latent alignment loss that forces the student's generated hidden state \mathbf{h}_{t-1} to match the teacher's selected feature \mathbf{z}_t .

$$\mathcal{L}_{S1} = \mathcal{L}_{\text{CE}} + \lambda_{\text{sim}} \cdot \frac{1}{\sum_m K} \sum_m \sum_{t \in \mathcal{T}_m} \left(1 - \cos(\mathbf{h}_{t-1}, \mathbf{z}_t) \right), \quad (9)$$

where \mathcal{T}_m denotes the indices of the latent tokens and λ_{sim} balances the two objectives.

Stage 2: Text-Only Supervision with Latent Relaxation

In the second stage, we relax the strict alignment constraint to allow the model to internalize the reasoning process. We remove the latent alignment loss and the teacher supervision, optimizing only for the final text generation.

$$\mathcal{L}_{S2} = \mathcal{L}_{\text{CE}}(\mathcal{X}_{\text{text}}), \quad (10)$$

Table 1: **Performance comparison on COMT and VSP benchmarks.** We compare distinct tasks in COMT and overall accuracy for VSP. “Direct Ans.” and “Text CoT” denote direct answer and text-only chain-of-thought fine-tuning. For latent methods, we contrast the static “Direct” paradigm against our dynamic “Interleaved” approach. **Bold** indicates the best result.

Model	Paradigm	COMT					VSP Acc.
		Creation	Deletion	Selection	Update	Avg.	
<i>Baselines</i>							
Qwen2.5-VL (Zero-shot)	Direct Ans.	68%	38%	35%	14%	38.8%	6%
Qwen2.5-VL (Direct-FT)	Direct Ans.	52%	60%	51%	49%	53%	72%
Qwen2.5-VL (CoT-FT)	Text CoT	80%	52%	45%	46%	55.8%	47%
<i>Stage 1: Latent Alignment</i>							
Mirage	Direct	53%	54%	45%	42%	48.5%	65.8%
ILVR (Ours)	Interleaved	69%	66%	46%	47%	57%	77.3%
<i>Stage 2: Latent Relaxation (Best Configurations)</i>							
Mirage	Direct	65%	62%	47%	50%	56%	76%
ILVR (Ours)	Interleaved	71%	68%	53%	51%	60.8%	81.5%

where $\mathcal{X}_{\text{text}}$ represents the textual tokens. This encourages the model to use the latent states as flexible internal priors optimized end-to-end for the reasoning task.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our framework on the COMT (Cheng et al., 2024) dataset containing 3,450 training and 400 test samples across Creation, Deletion, Selection, and Update tasks along with the VSP (Wu et al., 2024) dataset containing 1,000 training and 400 test samples. To assess generalization we train on a 10k-sample Zebra-CoT (Li et al., 2025a) subset spanning Scientific Reasoning tasks like Physics and Graph Algorithms plus Visual Logic tasks including Chess, Ciphers, Maze, Tetris, and RPM along with 3D Visual Reasoning tasks covering Counting, Planning, and Embodied CoT. Evaluation targets three unseen benchmarks including a held-out Zebra-CoT 2D set with 200 samples each for Visual Jigsaw and Visual Search plus EMMA BENCH (Hao et al., 2025) with 400 problems across Chemistry, Coding, Math, and Physics and VisualLogic (Xu et al., 2025) with 290 questions testing Positional, Quantitative, and Stylistic reasoning.

Compared Methods. We implement ILVR on the Qwen2.5-VL 7B (Bai et al., 2025a) architecture and benchmark it against three base variants: the native Zero-shot model, a Direct-FT version fine-tuned for direct answers, and a CoT-FT version fine-tuned on text-only chain-of-thought. To evaluate latent reasoning strategies, we contrast

the *direct reasoning* paradigm, represented by a re-implementation of Mirage (Yang et al., 2025) using a single initial image, against our *interleaved reasoning* paradigm that weaves visual updates into the textual chain. We exclude LVR (Li et al., 2025b) as its reliance on ground-truth bounding boxes renders it incompatible with these general-purpose benchmarks.

Implementation Details. All experiments use Qwen2.5-VL 7B (Bai et al., 2025a) as the base model where fine-tuning utilizes the AdamW optimizer with a 1e-5 learning rate and a cosine scheduler with the random seed fixed at 42. Unless varied in ablation studies, we set the latent token size K to 8, the alignment weight λ_{sim} to 1, and the EMA decay τ to 0.999. Additionally, we set the target group size L to 784, enabling adaptive feature grouping specifically for the Zebra-CoT dataset. Training lasts for 15 epochs on COMT (Cheng et al., 2024) and VSP (Wu et al., 2024), and for 2 epochs on the larger Zebra-CoT (Li et al., 2025a) dataset, while we use Qwen2.5-VL 72B as the judge for open-ended evaluations.

4.2 Main Results

Performance on COMT and VSP. Tab. 1 summarizes the evaluation on the COMT and VSP benchmarks. The results on VSP initially highlight the critical role of precise perceptual modeling, as text-only CoT lags significantly behind the Direct-FT baseline, dropping from 72% to 47% accuracy. Our approach effectively bridges this gap. By interleaving dynamic visual signals, ILVR significantly outperforms the static latent paradigm employed by Mirage. In the initial training stage, ILVR achieves

Table 2: **Generalization performance on unseen benchmarks.** We evaluate models trained on the curated Zebra-CoT dataset against three OOD benchmarks: EMMA BENCH (scientific), VisualLogic (fine-grained reasoning), and the held-out 2D tasks from Zebra-CoT. “Avg.” denotes the average accuracy within each benchmark. The final column reports the macro average across all 1090 test samples. Bold indicates the best result.

Model	Paradigm	EMMA BENCH (400)					VisualLogic (290)				Zebra-CoT (OOD 400)			Total
		Chem.	Code	Math	Phys.	Avg.	Pos.	Quant.	Style	Avg.	Jigsaw	Search	Avg.	Avg.
<i>Baselines</i>														
Qwen2.5-VL (Zero-shot)	Direct Ans.	18%	25%	28%	33%	26.0%	29%	24%	27%	26.6%	23%	65%	44%	32.8%
Qwen2.5-VL (Direct-FT)	Direct Ans.	16%	27%	28%	32%	25.8%	25%	23%	23%	23.8%	17%	73%	45%	32.3%
Qwen2.5-VL (CoT-FT)	Text CoT	21%	26%	33%	31%	27.8%	27%	23%	28%	25.9%	21.5%	68.5%	45%	33.6%
<i>Stage 1: Latent Alignment</i>														
Mirage	Direct	13%	21%	30%	37%	25.3%	25%	24%	21%	23.4%	16%	71%	43.5%	31.5%
ILVR (Ours)	Interleaved	23%	26%	34%	35%	29.5%	26%	23%	24%	24.5%	20.5%	74.5%	47.5%	34.8%
<i>Stage 2: Latent Relaxation</i>														
Mirage	Direct	15%	25%	35%	33%	27.0%	24%	26%	30%	26.6%	20%	74.5%	47.3%	34.3%
ILVR (Ours)	Interleaved	31%	35%	34%	33%	33.3%	27%	30%	31%	29.3%	22.5%	73%	47.8%	37.5%

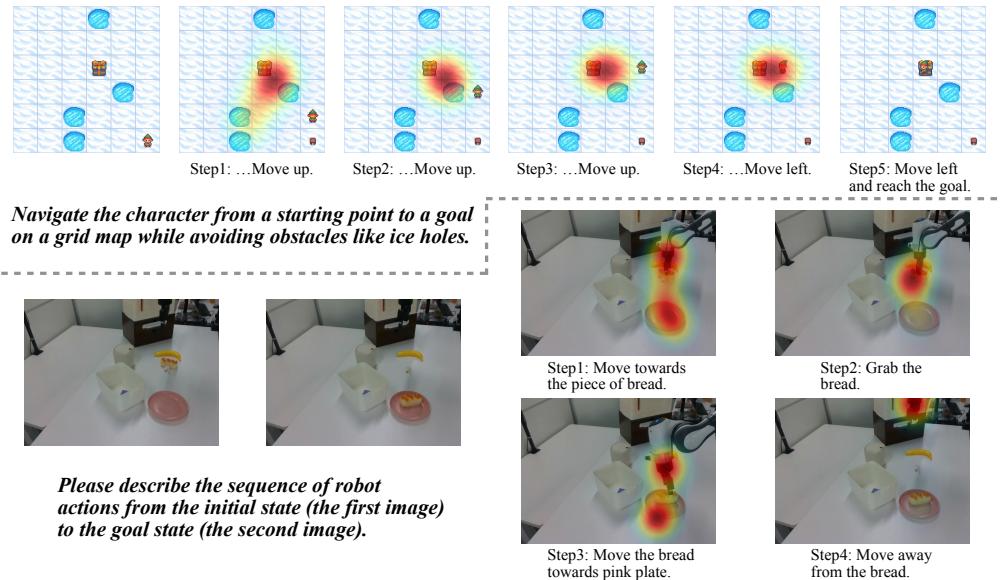


Figure 3: **Visualization of dynamic latent modeling.** Heatmaps depict the Gaussian-smoothed aggregation of relevant image patches for $K = 8$ generated latents. Top: Latents sequentially track the character’s path in navigation. Bottom: Visual attention shifts from the object (bread) to the target (plate) during robotic manipulation. This confirms precise alignment between generated latents and the step-wise reasoning context.

average accuracies of 57% on COMT and 77.3% on VSP, leading Mirage by over 8 percentage points on both benchmarks. This advantage indicates that evolving multi-step problems demand up-to-date visual cues rather than a single compressed representation. After the final latent relaxation stage, ILVR establishes a new state-of-the-art with 60.8% average accuracy on COMT and 81.5% on VSP, consistently exceeding the best competitor configurations.

Generalization Across Diverse Benchmarks. To assess the robustness of our framework, we evaluate models trained on the curated Zebra-CoT dataset against three challenging OOD benchmarks: EMMA BENCH, VisualLogic, and the held-out 2D

Visual Reasoning subset of Zebra-CoT. The results, summarized in Tab. 2, demonstrate the superior generalization capability of ILVR.

(1) Overall Performance. ILVR (Stage 2) achieves the highest overall average accuracy of 37.5% across all 1090 test samples, surpassing the Zero-shot baseline (32.8%), the best fine-tuned baseline (CoT-FT, 33.6%), and the Mirage competitor (34.3%). This consistent lead confirms that our interleaved latent reasoning paradigm learns transferable skills rather than merely overfitting to training patterns.

(2) Fine-Grained & Scientific Reasoning. On VisualLogic, which tests fine-grained perception across Positional, Quantitative, and Stylistic

Table 3: **Ablation of core components.** Comparison of the Adaptive Selection mechanism against the Average Pooling baseline (Mirage) and the dynamic Interleaved structure against the static Direct setup.

Method	Structure	Mechanism	Accuracy			
	(Int. vs. Dir.)	(Adapt. vs. Pool.)	VisLog	EMMA	Zebra	Total
Direct (Pooling)		×	23.4%	25.3%	43.5%	31.5%
Direct (Adaptive)	×	✓	24.1%	26.3%	44.5%	32.4%
ILVR (Ours)	✓	✓	24.5%	29.5%	47.5%	34.8%

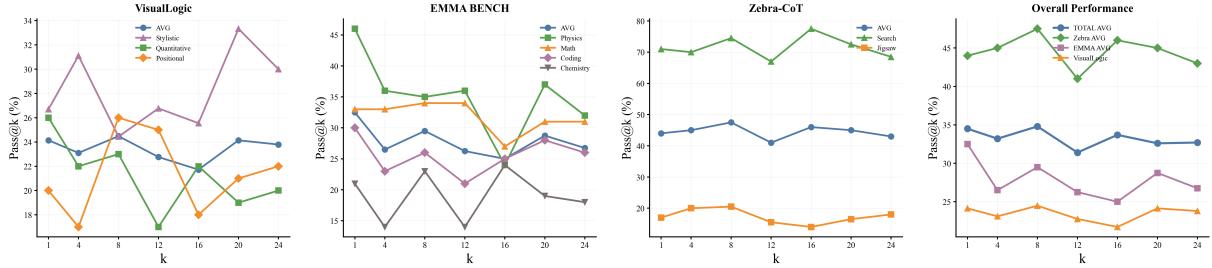


Figure 4: **Impact of Latent Size K .** Performance trends across VisualLogic, EMMA, and Zebra-CoT (and the overall average) as the number of latent tokens K varies. λ_{sim} is fixed at 1.0. $K = 8$ yields the most robust performance across metrics.

Table 4: **Sensitivity to alignment loss weight λ_{sim} .** We report the average accuracy across benchmarks. λ_{sim} represents the effective weight of the alignment loss relative to the text generation loss.

λ_{sim}	Accuracy			
	VisLog	EMMA	Zebra	Total
0.1	23.4%	25.8%	44%	31.8%
0.5	20%	30.5%	45.8%	33.3%
1	24.5%	29.5%	47.5%	34.8%
2	21.7%	27.8%	42.5%	31.6%
10	21.4%	27.5%	48.5%	33.6%

categories, ILVR achieves a dominant average of 29.3%, significantly outperforming Mirage (26.6%) and Zero-shot (26.6%). Notably, in Stylistic Reasoning, ILVR gains a substantial lead (31% vs. 27% for Zero-shot), indicating a deeper understanding of visual nuances. similarly, on EMMA BENCH, ILVR demonstrates strong versatility, particularly in Chemistry (31%) and Coding (35%), where it outperforms Mirage by margins of +16% and +10% respectively. This suggests that our dynamic latent updates effectively capture the evolving states crucial for multi-step scientific problem solving.

(3) 2D Visual Reasoning. On the held-out Zebra-CoT 2D tasks, ILVR maintains its advantage with an average accuracy of 47.8%. In the Visual Jigsaw task, which demands precise spatial assembly logic, ILVR (22.5%) outperforms both CoT-FT (21.5%) and Mirage (20%). This further validates that our method’s ability to ground reasoning in

precise visual latents generalizes well to unseen spatial tasks.

Qualitative Analysis. Fig. 3 visualizes the spatial focus of generated latents via aggregated attention heatmaps. In the navigation task (top), the model prioritizes task-critical features, specifically the goal (chest) and hazards (ice holes), demonstrating an awareness of global environmental constraints beyond simple agent tracking. Similarly, the robotic manipulation example (bottom) exhibits a clear semantic shift where attention transitions from the manipulation object (bread) to the target destination (plate) as the action sequence progresses. These patterns confirm that ILVR dynamically aligns its latent visual attention with the specific semantic requirements of each reasoning step.

4.3 Ablation Study

To dissect the contribution of each component and assess the robustness of our framework, we conduct extensive ablation studies. All models are trained on the curated Zebra-CoT dataset and evaluated on three unseen benchmarks: VisualLogic, EMMA BENCH, and the held-out Zebra-CoT 2D tasks.

Impact of Core Components. Tab. 3 dissects the impact of our core components. The Direct (Pooling) baseline mirrors Mirage’s static average pooling and yields the lowest accuracy of 31.5% as it indiscriminately discards details. Introducing momentum-guided selection in Direct (Adap-

tive) boosts performance to 32.4% confirming that context-aware extraction outperforms blind compression even in a static setup. Finally, ILVR (Ours) achieves a dominant 34.8% demonstrating that complex reasoning demands both precise feature modeling and the temporal flexibility to update visual focus dynamically.

Latent Representation Size (K). We analyze the impact of the latent token budget K , which controls the information capacity of the visual representation. For this experiment, we fix the alignment weight $\lambda_{\text{sim}} = 1.0$. As illustrated in Fig. 4, we vary K from 1 to 24. Performance initially improves as K increases, peaking at $K = 8$ with a total average accuracy of 34.8%. This suggests that $K = 8$ offers an optimal trade-off, providing sufficient capacity to encode rich visual details without introducing the noise or optimization difficulties associated with longer sequences (e.g., performance drops to 31.3% at $K = 12$). Consequently, we adopt $K = 8$ as the default configuration.

Alignment Loss Weight (λ_{sim}). Finally, we examine the sensitivity of ILVR to the strength of the alignment supervision. We vary the effective weight λ_{sim} relative to the cross-entropy loss. Tab. 4 summarizes the results with the latent size fixed at $K = 8$. We observe that $\lambda_{\text{sim}} = 1.0$ yields the best balance (34.8% total accuracy). When λ_{sim} is too low (e.g., 0.1), the model ignores the visual alignment objective, degrading performance to 31.8%. Conversely, an excessively high weight (e.g., 10.0) over-constrains the latent space to match the selected features at the expense of the reasoning flexibility, dropping accuracy to 33.6%. This confirms that the alignment loss should serve as a guide rather than a hard constraint.

5 Conclusion

In this paper, we introduce Interleaved Latent Visual Reasoning (ILVR), a framework that unifies dynamic state evolution with precise perceptual modeling. By interleaving textual generation with dynamically updated latent visual representations, ILVR effectively models evolving tasks while circumventing the prohibitive computational cost of pixel-level re-encoding. Our adaptive selection mechanism, guided by a Momentum Model, enables the system to distill context-relevant visual signals for each reasoning step, ensuring robust modeling without over-compression. Extensive ex-

periments across diverse benchmarks demonstrate that ILVR significantly outperforms existing static latent methods, validating that reasoning within a dynamic, interleaved latent space offers a scalable and effective path for enhancing multimodal intelligence.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. [Qwen2.5-vl technical report](#). *ArXiv*, abs/2502.13923.
- Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025b. [Univg-r1: Reasoning guided universal visual grounding with reinforcement learning](#). *ArXiv*, abs/2505.14231.
- Jeffrey Cheng and Benjamin Van Durme. 2024. [Compressed chain of thought: Efficient reasoning through dense representations](#). *ArXiv*, abs/2412.13171.
- Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2024. [Comt: A novel benchmark for chain of multi-modal thought on large vision-language models](#). *ArXiv*, abs/2412.12932.
- Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. [Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation](#). *ArXiv*, abs/2407.06135.
- Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Shi Guang, and Haoqi Fan. 2025. [Emerging properties in unified multimodal pretraining](#). *ArXiv*, abs/2505.14683.
- Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei A. F. Florêncio, and Cha Zhang. 2025. [Refocus: Visual editing as a chain of thought for structured image understanding](#). *ArXiv*, abs/2501.05452.
- Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Quchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. [Webwatcher: Breaking new frontier of vision-language deep research agent](#). *ArXiv*, abs/2508.05748.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E. Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *ArXiv*, abs/2412.06769.

- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. **Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark.** *ArXiv*, abs/2501.05444.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. **Momentum contrast for unsupervised visual representation learning.** *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke S. Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. **Visual sketchpad: Sketching as a visual chain of thought for multimodal language models.** *ArXiv*, abs/2406.09403.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaoshen Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025a. **Vision-r1: Incentivizing reasoning capability in multimodal large language models.** *ArXiv*, abs/2503.06749.
- Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Junjie Hu, and Yong Jae Lee. 2025b. **Visualtoolagent (vista): A reinforcement learning framework for visual tool selection.** *ArXiv*, abs/2505.20289.
- Ang Li, Charles L. Wang, Kaiyu Yue, Zikui Cai, Olle Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. 2025a. **Zebra-cot: A dataset for interleaved vision language reasoning.** *ArXiv*, abs/2507.16746.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muham Chen, and Zicheng Liu. 2025b. **Latent visual reasoning.** *ArXiv*, abs/2509.24251.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. **Llava-onevision: Easy visual task transfer.** *ArXiv*, abs/2408.03326.
- Dairu Liu, Ziyue Wang, Minyuan Ruan, Fuwen Luo, Chi Chen, Peng Li, and Yang Liu. 2025. **Visual abstract thinking empowers multimodal reasoning.** *ArXiv*, abs/2505.20164.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. **Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models.** *ArXiv*, abs/2403.16999.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. **Codi: Compressing chain-of-thought into continuous space via self-distillation.** *ArXiv*, abs/2502.21074.
- Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhua Chen. 2025. **Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning.** *ArXiv*, abs/2505.15966.
- Jiacong Wang, Zijiang Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. 2025a. **Vgr: Visual grounded reasoning.** *ArXiv*, abs/2506.11991.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xinguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 44 others. 2025b. **Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency.** *ArXiv*, abs/2508.18265.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. **Chain of thought prompting elicits reasoning in large language models.** *ArXiv*, abs/2201.11903.
- Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. 2025. **Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use.** *ArXiv*, abs/2505.19255.
- Qiucheng Wu, Handong Zhao, Michael Stephen Saxon, Trung M. Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. 2024. **Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms.** *ArXiv*, abs/2407.01863.
- Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wen gang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhui Wang, Jifeng Dai, and Jinguo Zhu. 2025. **Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models.** *ArXiv*, abs/2504.15279.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025. **Machine mental imagery: Empower multimodal reasoning with latent visual tokens.** *ArXiv*, abs/2506.17218.
- Zhao Yu Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 2025. **Openthinking: Learning to think with images via visual tool reinforcement learning.** *ArXiv*, abs/2505.08617.
- Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, and Hao Jiang. 2025a. **Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation.** *ArXiv*, abs/2503.05255.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2025b. **Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl.** *ArXiv*, abs/2505.15436.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao,
George Karypis, and Alexander J. Smola. 2023. **Multimodal chain-of-thought reasoning in language models**. *Trans. Mach. Learn. Res.*, 2024.