FUNCTIONALLY GROUNDED LATENT DISTILLATION FOR VISUAL LATENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual latent reasoning offers a promising alternative to explicit multimodal reasoning by performing intermediate computation in continuous representation space. Existing approaches commonly learn latent states by aligning them with perceptual features or privileged intermediate representations, yet representational similarity alone does not guarantee that the learned states preserve the reasoning computation they are intended to replace. We propose HCLR, a hierarchical latent distillation framework for visual latent reasoning that moves beyond representation alignment by learning latent states according to the downstream reasoning behavior they preserve. We formulate this principle as functional equivalence and progressively internalize explicit multimodal reasoning through hierarchical latent distillation. HCLR first learns compact latent substitutes for intermediate visual observations and then compresses the remaining textual and local-latent reasoning scaffold into a compact global latent trajectory. Privileged visual observations and reasoning scaffolds are used only during training and are subsequently removed through distillation, enabling the final model to perform intermediate reasoning entirely in continuous latent space using only the original image and question. Empirical results demonstrate that HCLR achieves strong performance across diverse real-world visual perception and reasoning settings, while exhibiting strong out-of-distribution generalization on challenging abstract visual reasoning tasks.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in visual understanding and reasoning. Chain-of-thought (CoT) reasoning further improves their ability to solve problems that require multiple intermediate decisions Wei et al. (2022); Zhang et al. (2024). Recent studies extend this paradigm beyond linguistic reasoning by grounding intermediate steps in visual regions, acquiring task-relevant evidence, or actively manipulating visual inputs Mitra et al. (2024); Shao et al. (2024); Xu et al. (2025); Wu & Xie (2024); Zheng et al. (2025). Such approaches allow visual information to participate directly in the reasoning process. However, the intermediate computation is still explicitly instantiated as textual rationales, auxiliary visual observations, or visual operations. As the reasoning process becomes longer or more visually involved, this explicit scaffold introduces additional autoregressive decoding and repeated visual processing during inference.

Visual latent reasoning provides a different direction by carrying intermediate computation in continuous representation space Li et al. (2025); Yang et al. (2025); Wang et al. (2025a); Jiang et al. (2026); Zhu et al. (2026). Instead of materializing every intermediate reasoning step, the model maintains latent states that participate directly in subsequent computation. This formulation removes the need to expose the complete reasoning process, but introduces a fundamental learning problem. Unlike textual rationales or visual observations, latent reasoning states have no natural supervision target. Existing methods therefore derive training signals from perceptual features, vision-encoder representations, privileged intermediate states, or reasoning traces constructed with additional multimodal context Li et al. (2025); Yang et al. (2025); Qin et al. (2025); Jeon et al. (2026); Wang et al. (2025a); Jiang et al. (2026); Zhu et al. (2026). The resulting latent states are typically encouraged to approach an intermediate representation chosen by the learning procedure.

These practices leave unresolved whether representational similarity is sufficient for functional substitution. A latent state may closely resemble a visual feature or a privileged hidden state without

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

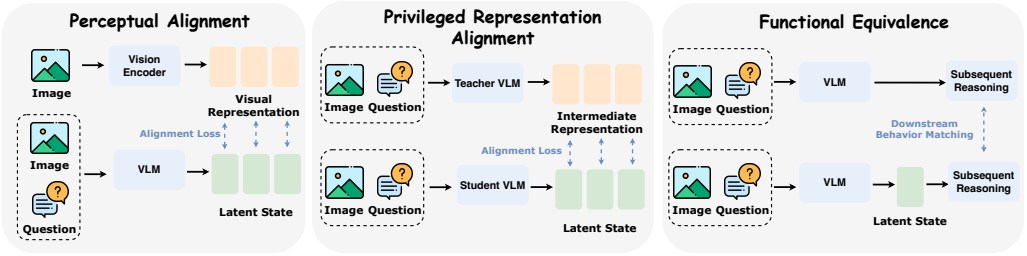

Figure 1: Comparison of perceptual alignment, privileged representation alignment, and functional equivalence for visual latent reasoning.

preserving how the corresponding intermediate component affects subsequent reasoning. We therefore formulate latent learning in terms of the computation that must remain after the original component is removed, defining *functional equivalence* as the preservation of downstream reasoning behavior rather than the reproduction of a prescribed intermediate representation.

Figure 1 summarizes this distinction between representation-oriented supervision and our functional criterion. Existing approaches commonly derive latent targets either from perceptual representations produced by visual encoders or from privileged intermediate states obtained with richer multimodal context. Although these two forms of supervision differ in the source of their targets, both primarily constrain the representation that a latent state should approximate. HCLR instead evaluates latent states through the downstream computation they support, using preservation of reasoning behavior as the learning criterion. The three panels therefore contrast perceptual alignment, privileged representation alignment, and functional equivalence, highlighting a shift from matching intermediate representations to preserving the reasoning function they are intended to replace.

Building on this principle, we propose HCLR, a hierarchical latent distillation framework for visual latent reasoning. The hierarchy is designed to progressively remove the explicit components of multimodal reasoning while retaining their computational contribution. At the local level, intermediate visual observations provide privileged evidence for learning compact latent substitutes. These states are trained to retain the role of the corresponding observations in subsequent reasoning, after which the visual evidence is removed through distillation. The remaining reasoning process still contains explicit textual computation together with local latent states. HCLR therefore performs a second transfer at the global level, where this variable-length scaffold is condensed into a compact global latent trajectory. Distillation then removes the scaffold itself. The final model consequently performs intermediate reasoning in continuous latent space using only the original image and question. In summary, our contributions are threefold:

- We introduce *functional equivalence* as a learning principle for visual latent reasoning. It characterizes latent states through the downstream reasoning behavior they preserve rather than solely through similarity to a predefined intermediate representation.
- HCLR provides a hierarchical latent distillation framework that progressively internalizes explicit multimodal reasoning through local functional latent learning followed by global latent reasoning. The resulting model removes privileged observations and explicit reasoning scaffolds from inference.
- Extensive evaluation demonstrates strong performance across diverse visual perception and reasoning benchmarks, together with out-of-distribution generalization on abstract visual reasoning. Additional analyses establish the functional contribution of the learned latent trajectory and validate the proposed hierarchical learning design.

2 Related Work

2.1 Chain-of-Thought and Thinking with Images

Chain-of-thought (CoT) reasoning has become a widely adopted paradigm for improving multi-step reasoning in multimodal large language models (MLLMs) Wei et al. (2022); Zhang et al. (2024).

Early multimodal CoT methods primarily formulate intermediate reasoning as linguistic rationales conditioned on both textual and visual inputs Zhang et al. (2024), while subsequent studies introduce more structured reasoning trajectories and strengthen visual grounding by associating intermediate steps with localized visual evidence or task-relevant regions Mitra et al. (2024); Shao et al. (2024); Xu et al. (2025). Despite these advances, the reasoning trajectory remains predominantly expressed through discrete linguistic or symbolic representations, requiring intermediate computation to be explicitly generated and consequently introducing additional autoregressive decoding overhead Wang et al. (2025a); Zhu et al. (2026).

Recent *thinking with images* approaches further incorporate visual computation into the reasoning process. Some methods actively acquire task-relevant evidence through grounding, visual search, cropping, zooming, or iterative visual inspection Wu & Xie (2024); Shao et al. (2024); Zheng et al. (2025), whereas others construct or transform intermediate visual states through drawing, editing, or executable visual operations Hu et al. (2024); Fu et al. (2025). Although these approaches strengthen the interaction between perception and reasoning, iterative visual operations can introduce additional execution and repeated visual-processing overhead during inference Hu et al. (2024); Fu et al. (2025). Learning such behaviors may further require specialized multimodal trajectories, tool-use supervision, or reinforcement learning Fu et al. (2025); Zheng et al. (2025), while the resulting reasoning process remains constrained by the form and granularity of the available visual operations Wang et al. (2025a); Zhu et al. (2026). In contrast, we internalize intermediate multimodal computation into continuous latent representations, enabling reasoning without explicitly materializing auxiliary visual observations or intermediate rationales.

2.2 Visual Latent Reasoning

Visual latent reasoning performs intermediate multimodal computation in continuous representation space Li et al. (2025); Yang et al. (2025); Wang et al. (2025a). Existing approaches vary in how extensively latent states replace explicit reasoning. Some interleave latent visual representations with textual reasoning Yang et al. (2025); Wang et al. (2025a); Tong et al. (2025), whereas others compress multimodal reasoning into latent trajectories and decode only the final answer Jiang et al. (2026); Zhu et al. (2026).

A central challenge is how to learn these latent states in the absence of ground-truth latent reasoning annotations. Existing methods employ markedly different forms of supervision. A common strategy derives latent supervision from perceptual representations produced by vision encoders or visual experts Li et al. (2025); Yang et al. (2025); Qin et al. (2025); Jeon et al. (2026). While such targets provide strong visual grounding, they primarily specify the perceptual content to be encoded rather than the reasoning transformations to be performed over it. Beyond purely perceptual targets, recent methods construct supervision from explicit reasoning scaffolds, distilling latent representations from auxiliary visual evidence, intermediate reasoning traces, or multimodal CoT Wang et al. (2025a); Jiang et al. (2026); Zhu et al. (2026). Such supervision incorporates richer task-dependent reasoning information, but remains tied to particular teacher representations or constructed reasoning trajectories. Other approaches reduce their reliance on explicit intermediate alignment and allow downstream task objectives to shape latent representations more directly Yang et al. (2025); Jiang et al. (2026), offering greater representational flexibility but leaving the computational role of individual latent states less explicitly constrained.

Across these designs, latent supervision largely differs in how latent states are learned and what representations are used as alignment targets. Existing approaches therefore focus primarily on constructing suitable intermediate targets and encouraging latent states to match them. However, representational alignment does not necessarily ensure that the learned latent state preserves the reasoning computation it is intended to replace. We instead formulate latent supervision in terms of *functional equivalence*, focusing on whether latent states preserve the reasoning behavior required for subsequent inference rather than only their similarity to a predefined intermediate representation.

3 Method

In this section, we present HCLR, a hierarchical latent distillation framework for visual latent reasoning that progressively internalizes explicit multimodal reasoning into compact continuous latent

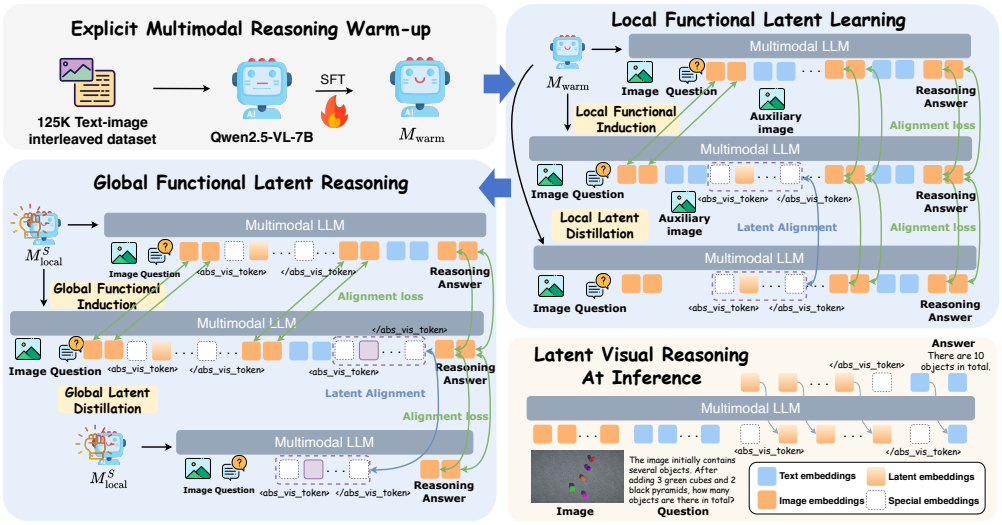

Figure 2: Overview of HCLR. Starting from explicit multimodal reasoning traces, HCLR progressively transfers intermediate reasoning into continuous latent computation through local and global functional latent learning. After training, the model performs latent visual reasoning using only the original image and question.

computation under the principle of functional equivalence. As illustrated in Fig. 2, HCLR first establishes explicit multimodal reasoning capability from interleaved visual and textual reasoning traces, and then progressively removes this intermediate reasoning scaffold through local and global latent transfer. The resulting model performs reasoning entirely in continuous latent space while requiring only the original image and question at inference time.

We first introduce the overall framework in Sec. 3.1. Sec. 3.2 describes the explicit multimodal reasoning warm-up, which provides the common initialization for subsequent latent learning. We then present local functional latent learning in Sec. 3.3, where privileged intermediate visual observations are first used to induce local latent states that preserve their function in downstream reasoning. Distillation subsequently removes access to these observations, enabling the model to generate the corresponding local latent states without auxiliary visual input. Sec. 3.4 extends the same principle to the complete reasoning scaffold, compressing the remaining textual and local latent computation into a compact global latent trajectory. Finally, Sec. 3.5 describes latent visual reasoning at inference, where the model directly generates the latent trajectory from the original image–question pair before producing the final answer.

3.1 Method Overview

Given an original image I_0 and a question Q, explicit multimodal reasoning can be represented as a sequence of textual reasoning segments $\{T_m\}$ interleaved with intermediate visual observations $\{I_m\}$. Such traces provide rich supervision for both linguistic reasoning and visual evidence, but retaining the complete scaffold at inference requires explicit intermediate observations and textual reasoning traces. HCLR learns latent states according to the reasoning function of the components they replace. We formulate this principle as *functional equivalence* and realize it through a hierarchical transfer from explicit to latent reasoning.

At the local level, each intermediate visual observation I_m serves as privileged information for inducing a latent block E_m that preserves its function in downstream reasoning. Access to I_m is subsequently removed through distillation, yielding an auxiliary-free local reasoning process composed of textual segments and local latent states.

At the global level, HCLR further compresses the remaining textual and local-latent reasoning scaffold into a compact global latent trajectory $R = \{r_k\}_{k=1}^{K}$. The trajectory is first learned to preserve the downstream predictive behavior of the complete scaffold. Distillation then removes the scaffold

itself, enabling the model to generate the global latent trajectory directly from (I_0, Q). Throughout latent reasoning, the hidden state produced at one latent position is directly propagated as the input embedding of the next position without intermediate vocabulary decoding. After training, the explicit reasoning scaffold is no longer required at inference: the model directly constructs the global latent trajectory from the original image and question and uses it to generate the final response.

3.2 Explicit Multimodal Reasoning Warm-up

To provide a reliable explicit reasoning scaffold for subsequent latent learning, we first perform supervised fine-tuning on image–text interleaved reasoning traces. We adopt the Monet-SFT-125K dataset introduced in Monet Wang et al. (2025a), which contains textual reasoning interleaved with intermediate visual observations. Each training example consists of an original image I_0, a question Q, and a variable-length assistant response composed of textual reasoning segments $\{T_m\}$ interleaved with intermediate visual observations $\{I_m\}$. Following the final visual observation, the response continues with textual reasoning that completes the reasoning process and produces the final response.

We optimize the model with standard next-token prediction over the entire assistant response, enabling it to acquire the explicit multimodal reasoning capability demonstrated by these interleaved traces. The resulting model, denoted as M_{warm}, serves as the common initialization for the subsequent local and global latent learning stages.

3.3 Local Functional Latent Learning

At the local level, we transfer the functional role of each intermediate visual observation into a compact sequence of continuous latent states. We first use privileged intermediate visual observations to induce local latent states that preserve their functional role in downstream reasoning. We then remove access to these observations through distillation, enabling an auxiliary-free student to generate the corresponding local latent states.

Local functional induction. Starting from M_{warm}, we construct a frozen direct path and a trainable latent path. The direct path retains the original interleaved multimodal reasoning trace and provides the reference predictive behavior of explicit reasoning. For the latent path, we insert a local latent block $E_m = \{e_{m,k}\}_{k=1}^{K_l}$ immediately after each intermediate visual observation I_m and before the subsequent textual reasoning segment T_{m+1}.

To ensure that the contribution of I_m is mediated through E_m, we introduce a bottleneck attention mask. Each local latent block can access its corresponding observation, whereas subsequent reasoning tokens cannot attend to that observation directly while retaining access to the original image, question, preceding textual context, as well as the local latent block.

Rather than assigning E_m a predefined representation target, we supervise it through the behavior it induces in downstream reasoning. Let p_{dir} and p_{lat} denote the downstream predictive distributions of the direct and latent paths, respectively. We optimize

$$\mathcal{L}_{\text{local}}^{\text{ind}} = \mathcal{L}_{\text{text}} + \lambda_{\text{KD}} D_{\text{KL}} \left(p_{\text{dir}} \,\|\, p_{\text{lat}} \right), \tag{1}$$

where $\mathcal{L}_{\text{text}}$ is the standard language modeling objective and D_{KL} denotes the grouped KL distillation used to match downstream predictive behavior. Unlike representation-level alignment, this objective does not constrain E_m to reproduce the hidden representation of I_m or a feature produced by an external visual encoder. Instead, E_m is optimized to preserve the functional role of the intermediate observation in downstream reasoning. We refer to this property as *local functional equivalence*.

Local latent distillation. The induced local reasoning process still relies on privileged intermediate visual observations, we therefore remove this dependency by training an auxiliary-free student. We freeze M_{local}^{T} as the teacher and initialize the student from M_{warm}. The teacher receives the complete interleaved reasoning trace and produces privileged local latent blocks $\{E_m^T\}_{m=1}^{M}$ together with the downstream predictive behavior induced by these states. The student retains the original

image I_0, the question Q, and the explicit textual reasoning context, but does not observe any intermediate visual observations. It instead produces the corresponding local latent blocks $\{E_m^S\}_{m=1}^M$ from the auxiliary-free context.

Let p_T^{local} and p_S^{local} denote the downstream predictive distributions of the teacher and student, respectively. The student is optimized with

$$\mathcal{L}_{\text{local}}^{\text{dist}} = \mathcal{L}_{\text{text}} + \lambda_E \mathcal{L}_{\text{align}}^{\text{local}} + \lambda_{\text{KD}} D_{\text{KL}} \left(p_T^{\text{local}} \,\|\, p_S^{\text{local}} \right), \tag{2}$$

where $\mathcal{L}_{\text{align}}^{\text{local}}$ is the mean position-wise cosine distance between corresponding teacher and student local latent states. The teacher states are treated as stop-gradient targets, while the KL term preserves the downstream predictive behavior realized through the privileged local states. After distillation, M_{local}^S realizes the local latent reasoning process without access to intermediate visual observations, leaving an interleaved scaffold of explicit textual reasoning and auxiliary-free local latent states for subsequent global latent reasoning.

3.4 Global Functional Latent Reasoning

The auxiliary-free local model still reasons through an explicit scaffold of textual segments and local latent blocks. We denote the complete local reasoning scaffold as $\mathcal{S}_{\text{local}}$ and the final response as Y_f. At the global level, we extend functional equivalence from individual visual observations to the complete intermediate reasoning process. Our goal is to replace $\mathcal{S}_{\text{local}}$ with a compact global latent trajectory that preserves the computation required to generate Y_f.

Global functional induction. Starting from M_{local}^S, we construct a frozen scaffold path and a trainable global-latent path. The scaffold path retains the complete local reasoning scaffold and provides the reference predictive behavior for Y_f. The global-latent path consumes the same scaffold and rolls out a fixed-length sequence of global latent states $R^T = \{r_k^T\}_{k=1}^K$.

To ensure that the contribution of the local scaffold is mediated through the global trajectory, we introduce a global bottleneck. When generating Y_f, direct attention to $\mathcal{S}_{\text{local}}$ is blocked, forcing information from the scaffold to pass through R^T. Let p_{scaf} and p_{glob} denote the predictive distributions over Y_f produced by the scaffold and global-latent paths, respectively. We optimize

$$\mathcal{L}_{\text{global}}^{\text{ind}} = \mathcal{L}_{\text{term}} + \lambda_{\text{KD}} D_{\text{KL}} \left(p_{\text{scaf}} \,\|\, p_{\text{glob}} \right), \tag{3}$$

where $\mathcal{L}_{\text{term}}$ is the language modeling objective over Y_f. As in local functional induction, the global latent states are not constrained to reproduce a predefined representation. Instead, R^T is optimized to preserve the downstream predictive behavior of the complete local reasoning scaffold. We refer to this property as *global functional equivalence*. The resulting model M_{global}^T therefore provides a compact latent realization of the computation carried by $\mathcal{S}_{\text{local}}$.

Global latent distillation. Constructing the global trajectory still requires the explicit local reasoning scaffold. We therefore remove this dependency through a scaffold-free student, we freeze M_{global}^T as the teacher and initialize the student from M_{local}^S. The teacher constructs R^T from the complete local scaffold, whereas the student receives only the original image and question (I_0, Q) and directly rolls out its own global latent trajectory $R^S = \{r_k^S\}_{k=1}^K$ before generating Y_f.

Let p_T^{global} and p_S^{global} denote the predictive distributions over Y_f produced by the teacher and student, respectively. The student is optimized with

$$\mathcal{L}_{\text{global}}^{\text{dist}} = \mathcal{L}_{\text{term}} + \lambda_R \mathcal{L}_{\text{align}}^{\text{global}} + \lambda_{\text{KD}} D_{\text{KL}} \left(p_T^{\text{global}} \,\|\, p_S^{\text{global}} \right), \tag{4}$$

where $\mathcal{L}_{\text{align}}^{\text{global}}$ aligns corresponding teacher and student global latent states, while the KL term preserves the predictive behavior supported by the teacher trajectory. After distillation, the student can construct the global latent reasoning trajectory directly from (I_0, Q) and generate the final response without requiring the explicit textual and local-latent scaffold.

3.5 Visual Latent Reasoning at Inference

At inference, the final model receives only the original image I_0 and question Q. It first rolls out a fixed-length sequence of K global latent states in continuous space, subsequently generating the final response Y_f conditioned on the resulting latent trajectory.

No privileged information or explicit intermediate reasoning scaffold is required during this process. The model does not rely on auxiliary visual observations, generate intermediate textual CoT or invoke any teacher model. Instead, the information required for intermediate reasoning is carried by the compact global latent trajectory constructed directly from the original multimodal input. The final model therefore performs visual latent reasoning without reconstructing the explicit multimodal reasoning trace used during training.

4 Experiments

4.1 Experimental setup

Training and Evaluation Setup. We adopt Qwen2.5-VL-7B-Instruct as the base model for HCLR. The vision encoder and visual merger remain frozen throughout training. During the multimodal reasoning warm-up, we fine-tune the language-model backbone, while the latent learning stages update only the last eight transformer blocks and the final normalization layer. We use $K_l = 8$ local latent states for each intermediate visual observation and set the global latent length to $K = 8$. The complete training, inference and evaluation setups are provided in Appendix A.3 and A.4.

Benchmarks & Metrics. We evaluate our method on V* Wu & Xie (2024), HRBench4K Wang et al. (2025b), HRBench8K Wang et al. (2025b), and MME-RealWorld-Lite Zhang et al. (2025), covering fine-grained perception and multimodal reasoning across high-resolution and real-world scenarios. All evaluations are conducted with VLMEvalKit Duan et al. (2024) for fair and consistent comparison. V* emphasizes attribute recognition and spatial reasoning, while HRBench4K and HRBench8K focus on fine-grained high-resolution perception. MME-RealWorld-Lite targets real-world multimodal perception and reasoning. We report the official overall and sub-category scores for these benchmarks. To assess out-of-distribution reasoning, we additionally evaluate on VisualPuzzles Song et al. (2025), which contains visual logic puzzles requiring abstract reasoning with limited reliance on prior knowledge.

Base Models & Baselines. We compare our method with representative baselines spanning three multimodal reasoning paradigms. Textual reasoning baselines include GPT-4o, Qwen2.5-VL-7B-Instruct, and a vanilla SFT variant trained on the same reasoning data, which serves as a data-matched control. Tool-based visual reasoning is represented by DeepEyes Zheng et al. (2025) and PixelReasoner Su et al. (2025), which enhance reasoning through explicit visual interaction or active perception. For visual latent reasoning, we include LVR Li et al. (2025), SkiLa Tong et al. (2025), and Monet Wang et al. (2025a), covering representative latent visual reasoning strategies with continuous intermediate states.

4.2 Main Results and Out-of-Distribution Generalization

As shown in Table 1, HCLR achieves the strongest overall performance among visual latent reasoning methods on V*, HRBench4K, and MME-RealWorld-Lite, while remaining competitive on HRBench8K. The improvements are not confined to a single capability. HCLR performs particularly well on spatial reasoning in V*, cross-instance perception in both HRBench settings, as well as the reasoning and perception subsets of MME-RealWorld-Lite. This pattern suggests that the proposed latent reasoning process remains effective across both fine-grained visual understanding and more complex multimodal reasoning. In comparison, vanilla SFT on the same interleaved reasoning traces provides limited or inconsistent improvements over the base model, whereas HCLR improves the overall score across all four benchmarks, suggesting that the observed gains cannot be explained by additional supervision alone.

Table 2 further evaluates out-of-distribution generalization on VisualPuzzles. HCLR achieves the highest overall score among the compared 7B models, with particularly strong performance on in-

Table 1: Main results on four multimodal reasoning benchmarks. Best results among visual latent reasoning methods are shown in **bold**. Results marked with * are taken from prior work.

Model	V*			HRBench4K			HRBench8K			MME-RealWorld-Lite		
	Overall	Attr.	Spa.	**Overall**	FSP	FCP	**Overall**	FSP	FCP	**Overall**	Rea.	Perc.
Textual Reasoning												
GPT-4o	67.5*	72.2*	60.5*	59.0*	70.0*	48.0*	55.5*	62.0*	49.0*	52.0*	48.3*	54.4*
Qwen2.5-VL-7B-Instruct	76.96	78.26	75.00	68.38	80.00	56.75	64.13	76.50	51.75	45.81	39.60	49.79
+ vanilla SFT	76.96	75.65	78.95	67.13	79.50	54.75	63.50	72.00	55.00	48.15	42.27	51.92
Tool-based Visual Reasoning												
DeepEyes	83.25*	84.35*	81.58*	71.25*	83.75*	58.75*	65.13*	77.00*	53.25*	54.28*	50.53*	56.63*
PixelReasoner	80.60*	83.50*	76.30*	72.90*	86.00*	60.30*	66.90*	80.00*	54.30*	49.70*	44.50*	53.10*
Visual Latent Reasoning												
LVR	80.63	81.74	78.95	70.25	83.00	57.50	66.50	78.25	54.75	50.34	43.73	54.58
SkiLa	79.58	82.61	75.00	68.88	85.00	52.75	60.38	77.25	43.50	48.88	43.87	52.10
Monet	80.63	81.74	78.95	70.50	**85.50**	55.50	**66.88**	**79.50**	54.25	49.82	44.27	53.38
HCLR	**83.25**	**84.35**	**81.58**	**71.63**	82.50	**60.75**	65.00	75.00	**55.00**	**51.80**	**46.27**	**55.35**

Table 2: Out-of-distribution evaluation on VisualPuzzles, which measures challenging abstract visual reasoning. Best results among comparable 7B models are shown in **bold**. Results marked with * are taken from prior work.

Model	VisualPuzzles					
	Overall	**Algorithmic**	**Analogical**	**Deductive**	**Inductive**	**Spatial**
Gemini 2.5 Pro	49.5*	60.0*	64.0*	60.0*	29.7*	36.4*
GPT-4o	41.3*	49.2*	58.3*	49.0*	27.3*	26.2*
Qwen2.5-VL-7B-Instruct	33.05	37.40	22.27	**47.50**	26.32	31.82
+ vanilla SFT	32.62	**37.79**	**27.49**	43.00	25.84	29.37
LVR	28.94	30.92	26.07	38.00	23.44	26.92
HCLR	**34.42**	**37.79**	27.01	**47.50**	**29.19**	**33.22**

ductive and spatial reasoning. It also matches the strongest results on algorithmic and deductive reasoning. The gains on these abstract reasoning categories suggest that the learned latent computation generalizes beyond the reasoning patterns encountered during training.

4.3 Effect of Latent Token Number

We investigate the effect of latent reasoning length by varying K from 0 to 24. As shown in Fig. 3, introducing latent computation consistently improves performance over $K = 0$ across all four benchmarks, confirming that the latent trajectory provides useful intermediate computation before answer generation. Performance is strongest with a moderate latent budget, with $K = 8$ achieving the best overall results on three of the four benchmarks. Further increasing the trajectory length brings little additional benefit and often results in mild fluctuations or degradation, indicating that additional recurrent steps may introduce redundant computation or accumulate deviations along the latent trajectory.

4.4 Latent Intervention Analysis

To assess whether the learned latent trajectory carries functionally relevant information, we perturb the latent states at inference time using Gaussian noise, zeroing, random permutation, and repeated copies of the first state. Figure 4 shows consistent performance degradation after intervention on both V* and MME-RealWorld-Lite. Gaussian perturbation and random permutation have relatively limited effects, indicating that the learned trajectory is not highly sensitive to moderate numerical perturbations or changes in latent ordering. More substantial degradation occurs when the latent states are zeroed or collapsed into repeated copies, with spatial reasoning on V* being particularly affected. This contrast suggests that effective latent reasoning depends more strongly on the information and variation preserved across the trajectory than on its exact state ordering.

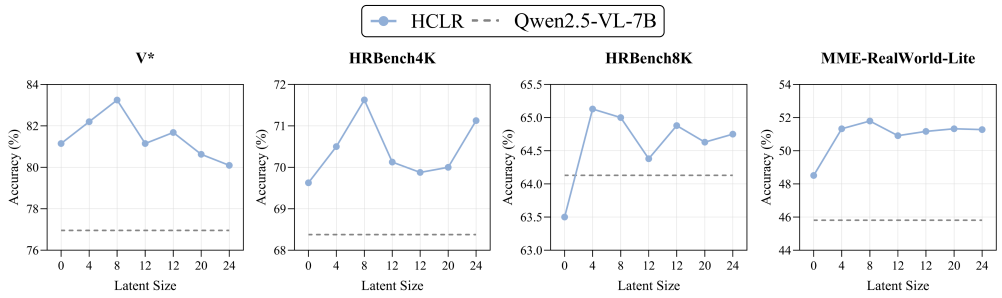

Figure 3: Effect of the latent reasoning length K on four multimodal reasoning benchmarks.

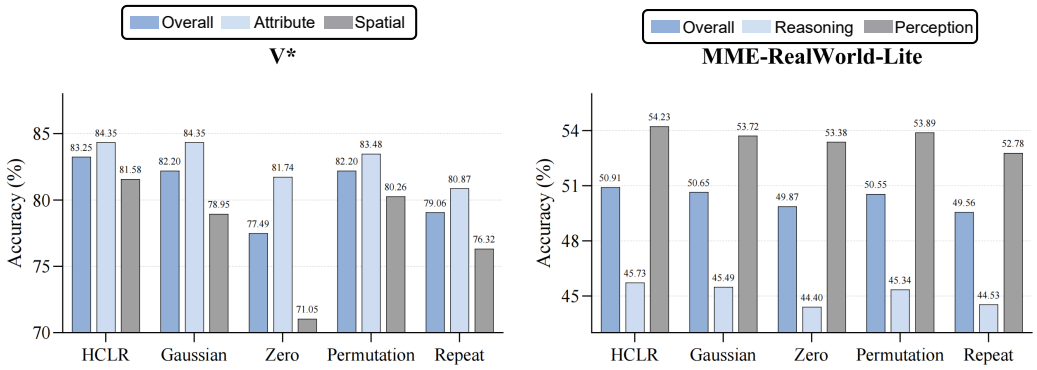

Figure 4: Latent intervention analysis on V* and MME-RealWorld-Lite. We perturb the latent trajectory with Gaussian noise, zeroing, permutation, and repeated copies of the first latent state.

4.5 Ablation Study

Table 3: Ablation study on the hierarchical latent learning design and latent distillation objectives. Best results are shown in **bold**.

Variant	V*	HRBench4K	MME-RealWorld-Lite
HCLR	**83.25**	**71.63**	**51.80**
w/o Local Functional Latent Learning	81.15	71.25	49.71
CE Only	76.96	69.38	49.09
w/o Latent Alignment	79.06	70.25	50.81
w/o Behavioral KD	81.68	70.38	50.29

We ablate the hierarchical latent learning design and distillation supervision in Table 3. Removing local functional latent learning reduces performance across all three benchmarks, suggesting that directly compressing the reasoning scaffold without first establishing local latent substitutes loses intermediate computation. Both distillation components contribute consistently to performance. Removing latent alignment or behavioral KD degrades performance across all benchmarks, while using CE alone produces the largest overall drop. These results demonstrate the complementary roles of latent-state alignment and behavioral preservation in effective latent reasoning.

5 Conclusion

We presented HCLR, a hierarchical latent distillation framework for visual latent reasoning that moves beyond representation alignment through *functional equivalence*. HCLR internalizes explicit multimodal reasoning into continuous latent computation while removing privileged reasoning scaffolds from inference. Experiments across diverse multimodal benchmarks demonstrate effectiveness and generalization. Important questions remain about the structure and behavior of latent reasoning, HCLR still relies on privileged intermediate observations during training, while the fixed latent reasoning length K may not be optimal across tasks. The computational role of individual latent states and the meaning of trajectory length also remain insufficiently understood, motivating future work on adaptive latent computation and more systematic analysis of latent reasoning dynamics.

AI USE STATEMENT

Generative AI tools were used for brainstorming during research development and for language polishing during manuscript preparation. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a general methodology for visual latent reasoning using existing training datasets, benchmarks, and pretrained multimodal models. It does not involve human-subject experiments or the collection of personal or sensitive data. The proposed method is not developed for a specific high-risk application.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce HCLR and the reported experimental results in the main paper and appendix. The learning procedure is described in Sec. 3.2, Sec. 3.3, and Sec. 3.4, covering explicit multimodal reasoning warm-up, local functional latent learning, as well as global functional latent reasoning. The inference procedure is specified in Sec. 3.5. Detailed training configurations, latent distillation settings, and hyperparameters are given in Appendix A.3. Inference settings and evaluation protocols are documented in Appendix A.4. Descriptions of the evaluation benchmarks and baseline methods are also included in the appendix to support reproduction of the reported results.

REFERENCES

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 11198–11201. ACM, 2024. doi: 10.1145/3664647.3685520. URL https://doi.org/10.1145/3664647.3685520.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhen-Yu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. URL https://api.semanticscholar.org/CorpusID:259243928.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei A. F. Florêncio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/fu25d.html.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *CoRR*, abs/2602.04476, 2026. doi: 10.48550/ARXIV.2602.04476. URL https://doi.org/10.48550/arXiv.2602.04476.

Houcheng Jiang, Jiajun Fu, Junfeng Fang, Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Univlr: Unifying text and vision in visual latent reasoning for multimodal llms. *CoRR*, abs/2605.11856, 2026. doi: 10.48550/ARXIV.2605.11856. URL https://doi.org/10.48550/arXiv.2605.11856.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *CoRR*, abs/2509.24251, 2025. doi: 10.48550/ARXIV.2509.24251. URL https://doi.org/10.48550/arXiv.2509.24251.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14420–14431. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01367. URL https://doi.org/10.1109/CVPR52733.2024.01367.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *CoRR*, abs/2511.19418, 2025. doi: 10.48550/ARXIV.2511.19418. URL https://doi.org/10.48550/arXiv.2511.19418.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *CoRR*, abs/2504.10342, 2025. doi: 10.48550/ARXIV.2504.10342. URL https://doi.org/10.48550/arXiv.2504.10342.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *CoRR*, abs/2505.15966, 2025. doi: 10.48550/ARXIV.2505.15966. URL https://doi.org/10.48550/arXiv.2505.15966.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *CoRR*, abs/2512.16584, 2025. doi: 10.48550/ARXIV.2512.16584. URL https://doi.org/10.48550/arXiv.2512.16584.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *CoRR*, abs/2511.21395, 2025a. doi: 10.48550/ARXIV.2511.21395. URL https://doi.org/10.48550/arXiv.2511.21395.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 7907–7915. AAAI Press, 2025b. doi: 10.1609/AAAI.V39I8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh

(eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13084–13094. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01243. URL https://doi.org/10.1109/CVPR52733.2024.01243.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 2087–2098. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00202. URL https://doi.org/10.1109/ICCV51701.2025.00202.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *CoRR*, abs/2506.17218, 2025. doi: 10.48550/ARXIV.2506.17218. URL https://doi.org/10.48550/arXiv.2506.17218.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=y1pPWFVfvR.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *CoRR*, abs/2505.14362, 2025. doi: 10.48550/ARXIV.2505.14362. URL https://doi.org/10.48550/arXiv.2505.14362.

Shoutai Zhu, Tianyang Xu, Bin Sun, Mingyuan Xu, Yu Liu, and Qinzhen Guo. OPLD: on-policy latent distillation for multimodal reasoning. *CoRR*, abs/2607.28154, 2026. doi: 10.48550/ARXIV.2607.28154. URL https://doi.org/10.48550/arXiv.2607.28154.

A Implementation Details

A.1 Benchmarks

We evaluate HCLR on benchmarks covering fine-grained perception, high-resolution understanding, real-world multimodal reasoning, and abstract visual reasoning. Unless otherwise specified, all evaluations are conducted with VLMEvalKit Duan et al. (2024) following its standard protocols.

V*. Wu & Xie (2024) evaluates detailed visual understanding in high-resolution images containing abundant and complex visual information. It is built on 191 high-resolution images and contains two subtasks, including 115 attribute-recognition samples and 76 spatial-relation samples. Attribute recognition focuses on fine-grained object properties, while spatial relation requires reasoning about the relative positions between objects.

HRBench4K and HRBench8K. Wang et al. (2025b) evaluates fine-grained perception of high-resolution images and provides separate 4K and 8K settings. Each setting contains 200 underlying QA samples, consisting of 100 Fine-grained Single-instance Perception (FSP) samples and 100 Fine-grained Cross-instance Perception (FCP) samples. Under the circular multiple-choice evaluation used by VLMEvalKit, each split contains 800 evaluation instances.

MME-RealWorld-Lite. Zhang et al. (2025) provides a lightweight evaluation of perception and reasoning in high-resolution real-world scenarios. The Lite version samples up to 50 examples from each task. Tasks with fewer available examples retain their full set, resulting in 1,919 examples in total. The benchmark covers five real-world domains, including monitoring, OCR with complex context, diagrams and tables, autonomous driving as well as remote sensing. The evaluation is further organized into perception and reasoning tasks.

VisualPuzzles. Song et al. (2025) targets visual reasoning while deliberately reducing dependence on domain-specific knowledge. It contains 1,168 puzzles spanning algorithmic, analogical, deductive, inductive, and spatial reasoning. The benchmark additionally provides easy, medium, and hard difficulty labels.

A.2 Baselines

We compare HCLR against representative baselines spanning general-purpose multimodal reasoning, tool-based visual reasoning, and visual latent reasoning. This selection covers both explicit visual-interaction methods and approaches that internalize intermediate computation into continuous latent representations, providing a broad comparison across different multimodal reasoning paradigms.

DeepEyes. Zheng et al. (2025) enables interleaved visual reasoning by allowing the model to actively inspect task-relevant regions during the reasoning process. It leverages the model's native grounding capability as a visual tool, rather than relying on separate specialized grounding models. The resulting tool-use behavior is learned through end-to-end reinforcement learning without requiring cold-start SFT on pre-collected reasoning trajectories.

PixelReasoner. Su et al. (2025) performs visual reasoning directly in pixel space by equipping the model with operations such as zoom-in and frame selection. These operations allow the model to actively acquire additional visual evidence while reasoning over images or videos. Its two-stage training first applies instruction tuning on synthesized reasoning traces and then uses curiosity-driven reinforcement learning to encourage effective pixel-space exploration.

LVR. Li et al. (2025) enables autoregressive reasoning directly in the visual embedding space. During supervised fine-tuning, the model generates latent states that reconstruct query-relevant visual tokens, jointly optimized with standard next-token prediction. It further adapts GRPO to latent reasoning, allowing latent visual computation to be interleaved with text generation.

SkiLa. Tong et al. (2025) introduces continuous latent sketch tokens as intermediate visual thoughts within a hybrid autoregressive reasoning process. The model dynamically switches between textual thinking and visual sketching modes, while a training-time sketch encoder provides visual-semantic targets for reconstructing the latent sketch states. This reconstruction objective grounds the generated latent tokens in the semantics of intermediate sketch images.

Monet. Wang et al. (2025a) performs reasoning in latent visual space by generating continuous embeddings as intermediate visual thoughts. Its training first uses a three-stage distillation-based SFT pipeline to learn latent visual representations from interleaved multimodal reasoning traces. It then introduces Visual-latent Policy Optimization (VLPO), which explicitly incorporates latent embeddings into policy-gradient optimization.

A.3 Training Configuration

Training setup. We initialize HCLR from Qwen2.5-VL-7B-Instruct and use Monet-SFT-125K Wang et al. (2025a) throughout training. The resulting training set is used to optimize the model for four epochs, while each subsequent latent-learning stage is trained for one epoch using the corresponding training view constructed from the same corpus. The vision encoder and visual merger are kept frozen throughout training. We optimize the language-model parameters during warm-up, while the latent-learning stages update only the last eight language transformer blocks and the final normalization layer.

Latent distillation. Both local and global reasoning use eight latent states in the 3584-dimensional hidden space of the language model. Behavioral distillation is implemented with grouped KL divergence, where the teacher-selected top-128 vocabulary entries are retained explicitly and the remaining probability mass is aggregated into a single tail bucket. We use $\mathrm{KL}(p_T \| p_S)$ with temperature $T = 2$. Local and global latent alignment are computed position-wise using cosine distance between corresponding teacher and student states, with the teacher states treated as stop-gradient targets. The corresponding objective weights are set to $\lambda_{\mathrm{KD}} = 1$, $\lambda_E = 1$, and $\lambda_R = 1$.

Hyperparameters. All experiments are conducted on eight NVIDIA H800 GPUs using DeepSpeed ZeRO-2 with BF16 precision. We use a per-device batch size of 1 and 16 gradient-accumulation steps, resulting in an effective global batch size of 128. The maximum sequence length is set to 4096. The complete training hyperparameters are summarized in Table 4.

Table 4: Training hyperparameters used for HCLR.

Hyperparameter	**Value**
Backbone	Qwen2.5-VL-7B-Instruct
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.01
LR scheduler	Cosine
Warm-up ratio	0.03
Maximum gradient norm	1.0
Per-device batch size	1
Gradient accumulation steps	16
Global batch size	128
Precision	BF16
Maximum sequence length	4096
DeepSpeed	ZeRO-2
GPUs	8× H800
Warm-up epochs	4
Latent-stage epochs	1
Local latent length K_l	8
Global latent length K_g	8
Latent dimension	3584
KD temperature T	2
KD top-k	128

A.4 Inference and Evaluation Settings

Inference configuration. We use a fixed global latent length of $K_g = 8$ for all evaluations. The latent reasoning span is delimited by `<abs_vis_token>` and `</abs_vis_token>`, with eight continuous latent states generated between the two boundary tokens. Upon completing the latent trajectory, the model transitions to autoregressive response generation. No auxiliary visual observations or explicit intermediate reasoning traces are required during inference.

System prompt. We use the following system prompt throughout evaluation: "*You are a helpful multimodal assistant. You are required to answer the question based on the image provided. Put your final answer in* $\backslash$*boxed*$\{\}$." Benchmark-specific questions and answer choices are then provided in their corresponding input formats.

Response generation. We adopt greedy decoding for all main experiments and use the same decoding strategy across benchmarks. This avoids stochastic variation during evaluation and ensures that each input produces a deterministic model response.

Evaluation and judging. All benchmark evaluations are conducted with VLMEvalKit Duan et al. (2024). For V*, HRBench4K, and HRBench8K, generated responses are first processed by the rule-based answer inference procedure provided by VLMEvalKit. Responses that cannot be resolved by this procedure are further evaluated using DeepSeek-V4-Flash as an auxiliary LLM judge. For MME-RealWorld-Lite, we follow the standard rule-based exact-match evaluation protocol without using an external LLM judge.

B Additional Experimental Results

We provide additional experimental results that complement the evaluations reported in the main paper. We first report the complete results on MME-RealWorld-Lite, extending the aggregated results in the main paper with performance across individual task domains. We then examine capability preservation beyond the target visual reasoning benchmarks.

B.1 Fine-Grained Results on MME-RealWorld-Lite

We report the complete results on MME-RealWorld-Lite in Table 5. The main paper reports the aggregated performance over perception and reasoning, whereas the results here are further separated across the individual task domains. The detailed evaluation reports category-level results for monitoring, autonomous driving, OCR with complex context, diagram and table understanding as well as remote sensing. The complete evaluation provides a more detailed view of HCLR across the diverse real-world scenarios covered by the benchmark.

Table 5: Detailed results on MME-RealWorld-Lite. Best results are shown in **bold**, and second-best results are underlined.

Model	Reasoning					Perception					
	Overall	MO	AD	OCR	DT	**Overall**	MO	AD	OCR	DT	RS
Base	39.60	31.33	27.00	<u>73.00</u>	<u>69.00</u>	49.79	30.09	32.29	**90.40**	**85.00**	41.33
SFT	42.27	<u>44.67</u>	29.75	<u>73.00</u>	58.00	51.92	<u>36.36</u>	32.57	<u>89.60</u>	<u>84.00</u>	46.00
Monet	<u>44.27</u>	40.67	**33.50**	67.00	**70.00**	<u>53.38</u>	36.05	**40.86**	84.00	82.00	<u>49.33</u>
Ours	**45.07**	**50.67**	<u>31.50</u>	**74.00**	62.00	**54.75**	**39.50**	<u>38.57</u>	86.80	<u>84.00</u>	**52.00**

B.2 General Capability Preservation

To examine whether latent reasoning training preserves capabilities beyond the target reasoning benchmarks, we further evaluate HCLR on MME translation and POPE. Table 6 reports the corresponding results. HCLR achieves a score of 192.5 on MME translationFu et al. (2023), matching the strongest result among the compared methods. On POPE, HCLR consistently achieves the highest performance across the overall score, three evaluation settings and accuracy. These results show that the proposed latent reasoning training maintains strong performance on complementary multimodal capabilities.

Table 6: General capability evaluation on MME translation and POPE. Best results are shown in **bold**.

Model	$\text{MME}_{\text{translation}}$	POPE				
		Overall	**Adversarial**	**Popular**	**Random**	**Accuracy**
Qwen2.5-VL-7B-Instruct	185.0	84.35	83.76	84.43	84.88	86.22
+ vanilla SFT	160.0	81.73	81.03	81.79	82.38	81.34
LVR	**192.5**	84.48	83.72	84.44	85.30	86.10
SkiLa	**192.5**	83.27	82.35	82.90	84.61	83.29
Monet	**192.5**	84.80	83.99	85.04	85.39	86.52
HCLR	**192.5**	**86.75**	**85.46**	**86.72**	**88.09**	**87.63**