ELVIS: EVIDENCE-GROUNDED LATENT VISUAL STATE FOR MULTI-STEP REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-step multimodal reasoning requires sustained access to visual evidence, yet recurrent methods, constructing latent visual states sequentially, show weak sensitivity to decisive evidence and provide limited support for subsequent reasoning in our diagnostics. We introduce ELVIS, a framework that constructs evidence-grounded latent visual states in parallel and invokes new states as reasoning progresses. Learnable queries interact bidirectionally within each state while preserving causality across reasoning steps. To supervise state content, we construct AVR-112K, an Action–Visual–Readout corpus whose explicit observations guide teacher-based evidence selection. Evidence alignment and visual anchoring train latent states to capture the selected evidence and attend to the original image. An Invocation Progressive Reward encourages additional invocations when they yield higher empirical success than smaller observed invocation counts. Experiments demonstrate strong performance across visual perception and spatial reasoning tasks, including a 22.04% gain over Qwen2.5-VL-7B on Zebra Maze. At 16 latent positions and matched invocation counts, ELVIS achieves an 11.65–17.40× speedup in latent-state computation over the fastest applicable recurrent baseline.

1 INTRODUCTION

Multi-step multimodal reasoning requires sustained access to visual evidence. Tool-assisted approaches revisit or manipulate images (Hu et al., 2024; Zheng et al., 2026), while latent visual reasoning represents intermediate visual feedback with continuous states (Li et al., 2026b; Wang et al., 2026b; Dong et al., 2026). A common approach generates *latent tokens* sequentially, feeding each decoder output back to produce the next token. Yet recent analyses question their visual content and report diminishing information gain in longer sequences (Li et al., 2026c; Wang et al., 2026a; Viveiros et al., 2026).

To better understand their limitations, we begin by probing existing recurrent latent methods using three diagnostics (Fig. 1a). (i) Masking decisive image regions while fixing the question and reasoning prefix produces small changes in the latents of LVR (Li et al., 2026b), Monet (Wang et al., 2026b), and ILVR (Dong et al., 2026), and generally smaller at later tokens. (ii) Transplanting latent states from original-image runs into masked-image runs restores only 0.0-5.8% accuracy. (iii) Decoder layer-wise trajectories projected into PCA space increasingly overlap across later tokens even for different samples, similarly found in (Li et al., 2026c; Cui et al., 2026; Guo et al., 2026). (Detailed diagnostic setups are provided in the Appendix A.1.) This limited evidence sensitivity, recovery, and representational diversity motivates us to revisit both state formation and visual supervision. We therefore ask: *how to construct a state that carries evidence needed at each reasoning step, and how to specify that evidence during training*.

Our first insight is to expand the *width* of a latent state through parallel construction (Fig. 1b). Recurrent methods require repeated decoder pass for each latent token to represent an intermediate visual state, we instead form a set of latent tokens, with bidirectional interaction within the state and causal ordering across steps and intervening text. Each state takes one parallel pass through the decoder stack, and further states can be invoked as reasoning progresses. This reduces the measured latency of state construction while preserving updates to visual evidence across reasoning steps.

The second challenge is to specify what each state should represent and how it should be supervised. Existing interleaved reasoning corpuses (Li et al., 2026a; Gu et al., 2026) unfortunately can weaken

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.

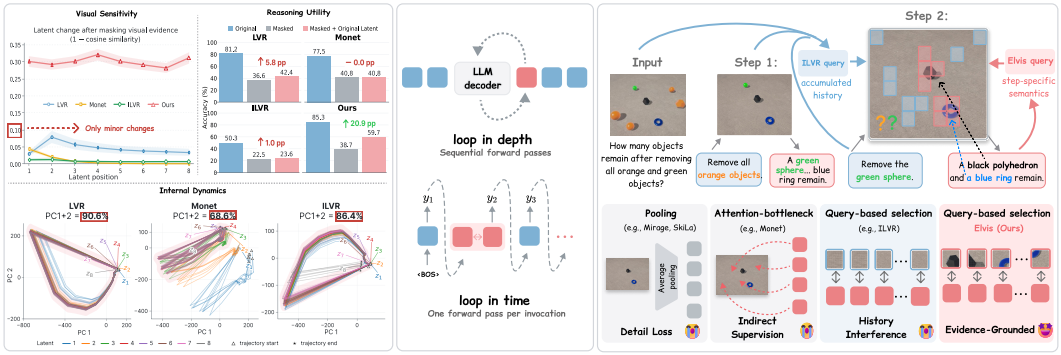


Figure 1: **From recurrent latents to evidence-grounded visual states.** (a) Masking, transplantation, and PCA diagnose recurrent states. ELVIS shows stronger evidence sensitivity and accuracy recovery. (b) Sequential latent feedback (*loop in depth*) gives way to jointly formed states interleaved with text (*loop in time*). Each state has bidirectional internal attention. (c) Readout-conditioned evidence selection targets the current observation, addressing limitations of pooling, indirect supervision, and history-conditioned selection.

this supervision in two ways: preceding text may already disclose an intermediate image's observation (*information leakage*), or subsequent text may omit that observation (*information omission*). In either case, subsequent predictions need not clearly depend on the image. Even with a perfect reasoning trajectory as supervision, broad pooling (e.g., Mirage (Yang et al., 2026), SkiLa (Tong et al., 2025)) may lose local detail, attention aggregation (e.g., Monet (Wang et al., 2026b)) may leave the relevant evidence implicit, and selection from preceding reasoning (e.g., ILVR (Dong et al., 2026)) may reflect what the model intended to inspect rather than what it observed (Fig. 1c). We therefore connect the continuation target to the evidence target through the observation made at each step.

We first construct **AVR-112K** to make this connection explicit, where each **Action–Visual–Readout** step specifies an operation without revealing its outcome, followed by an intermediate image and a readout of the relevant observation. Grounded in the image and not disclosed in preceding text, the readout serves both as a prediction target and as a training-time guide for selecting visual evidence. That said, this connects trajectory design to supervision of the image's latent replacement.

We then introduce ELVIS, our evidence-grounded latent visual state framework, which learns this replacement in two supervised stages. First, a visual-reasoning teacher learns from AVR trajectories, and readout-to-visual attention selects fine-grained evidence targets. The student then replaces intermediate images with shared learnable queries processed jointly within the decoder under a trajectory-causal, state-bidirectional attention mechanism. Moreover, evidence alignment matches the selected visual representations from the teacher, while visual anchoring further ensures attention to the original image. We further train ELVIS with GRPO (Shao et al., 2024b) and an additional *invocation progressive reward* so that it can learn how many latent-state invocations each problem needs. Experiments span visual perception, spatial reasoning, and out-of-distribution benchmarks show that ELVIS can achieve strong performance across these tasks. And ELVIS improves over Qwen2.5-VL-7B by 22% on Zebra Maze and achieves an 11.65 - 17.40× speedup in latent-state computation over the fastest applicable recurrent baseline at 16 latent positions and matched invocation counts.

Contributions. (i) We introduce AVR-112K and readout-conditioned evidence supervision, linking each step's visual observation to its continuation and latent target. (ii) We develop ELVIS, which jointly forms evidence-grounded states within the decoder and learns adaptive invocation. (iii) We thoroughly evaluate ELVIS through diverse benchmarks, ablations, and analyses of the correspondence between visual evidence and the learned latent state.

2 Related Work

Visual Feedback and Latent State Construction. Visual Sketchpad (Hu et al., 2024) and DeepEyes (Zheng et al., 2026) revisit visual evidence through drawing tools and reinforcement-learned

active perception, respectively. The execution and processing costs of explicit visual observations motivate internalizing such feedback as latent states. Recurrent approaches include LVR's autoregressive reconstruction of query-relevant visual tokens (Li et al., 2026b), Monet's staged distillation (Wang et al., 2026b), and ILVR's interleaving of recurrent latent segments with text (Dong et al., 2026). Parallel construction has also been explored: SCOLAR (Wang et al., 2026a) maps LLM hidden states to auxiliary visual tokens with a lightweight detransformer, while DeepLatent (Lu et al., 2026) uses a dedicated LatentFormer whose learnable spatial queries interact with image features under reasoning-conditioned modulation. We extend parallel construction within the language decoder, where a shared query bank acquires evidence from the preceding multimodal context through bidirectional within-state interactions while preserving causality along the reasoning trajectory.

Grounding and Functional Utility of Visual Latents. Recent analyses report weak input–latent and latent–answer connections under interventions, as well as limited visual information and high latent-token similarity (Li et al., 2026c), alongside diminishing information gain in longer recurrent sequences and limitations of pooled supervision (Wang et al., 2026a). To strengthen grounding and latent use, CrystaL transfers representations between intact- and corrupted-image runs with consistency objectives (Zhang et al., 2026a), while RIS combines region-level supervision with a progressive attention bottleneck (Cui et al., 2026). Monet filters useful auxiliary images and identifies key visual-observation tokens (Wang et al., 2026b), while DeepLatent filters for intermediate-image necessity (Lu et al., 2026). ILVR selects intermediate-image features using a momentum teacher conditioned on preceding reasoning (Dong et al., 2026). ELVIS connects trajectory design with evidence selection: AVR-112K separates actions, visual results, and subsequent readouts, making newly acquired visual information an explicit continuation target. Readout-to-visual attention then selects fine-grained evidence for each latent state. This couples step-level visual dependencies with representation targets. The readout guides supervision only and remains unavailable during latent construction at inference time.

3 Action–Visual–Readout Training Corpus

3.1 Visual-Dependency Shortcuts

Latent visual reasoning often learns to replace intermediate images in interleaved trajectories from multimodal generation datasets (Li et al., 2026a; Gu et al., 2026) or tool-based reasoning traces (Fu et al., 2025; Shao et al., 2024a). Yet an image's presence does not specify its contribution to subsequent reasoning due to two issues (Fig. 2): *information leakage*, where preceding text already reveals the intended observation, and *information omission*, where subsequent text leaves that observation unstated. These can enable textual shortcuts or leave visual supervision underspecified. So we pair each intermediate image with an explicit readout of its evidence for the current step.

3.2 Action–Visual–Readout Reformulation

Given an input image I, question q, and answer y, we represent an M-step trajectory in **Action–Visual–Readout** (AVR) format:

$$\tau = (I, q, S_1, \ldots, S_M, y), \qquad S_i = (A_i, V_i, R_i). \tag{1}$$

The action A_i specifies a visual operation, such as zooming into a region, drawing an auxiliary line, or predicting a subsequent scene state, without disclosing its outcome. The intermediate image V_i shows the result, and the immediately following readout R_i states the observation relevant to the action. This observation must be grounded in V_i and not already stated in the preceding text, although its underlying information may be present in I. The readout provides both a textual continuation target and a guide for selecting visual evidence to supervise the latent state (Sec. 4.1).

3.3 Corpus Construction

We curate samples from Zebra-CoT (Li et al., 2026a), Monet-SFT-125K (Wang et al., 2026b), Visual-CoT (Shao et al., 2024a), and ThinkMorph (Gu et al., 2026), covering visual perception, spatial reasoning, and embodied planning. Each intermediate image defines one AVR step. Text between consecutive images is split into the earlier image's readout and the next image's action.

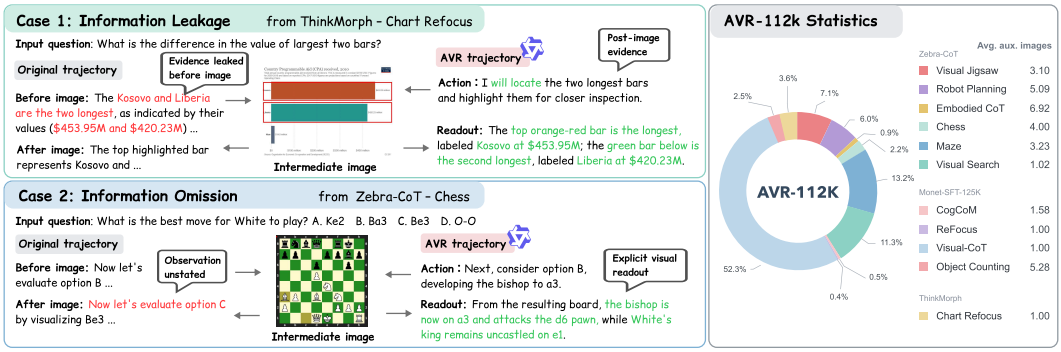


Figure 2: **AVR-112K construction and composition. Left**: examples of information leakage and omission, with AVR reformulations that place image-grounded observations in post-image readouts. **Right**: task distribution and average number of auxiliary images per trajectory.

To further refine the corpus, we use Qwen2-VL-72B-Instruct (Wang et al., 2024) to reexamine the images and restructure the text, i.e., add missing action-relevant observations, rewrite readouts that do not match their images, and move prematurely disclosed observations from actions to readouts. This preserves prospective actions while making each image's contribution explicit. To better expose latent visual reasoning, we also extend some multi-object spatial-relation examples originally containing a single intermediate image into multi-step trajectories, such that successive images localize different regions, each with a paired action and readout. The resulting **AVR-112K** corpus combines single-step and multi-step supervision, with 112,034 trajectories and 1.92 intermediate visual states per trajectory (Fig. 2). More detailed statistics are provided in the Appendix A.2.

4 METHOD

ELVIS forms an evidence-grounded latent visual state in parallel and reconstructs it as reasoning progresses. The Action–Visual–Readout (AVR) corpus pairs each intermediate image with a readout that serves both as a text prediction target and as a guide for selecting visual supervision. Training proceeds in three stages: extracting teacher evidence targets (Sec. 4.1), learning parallel latent states through evidence alignment and visual anchoring (Sec. 4.2), and optimizing when to invoke them through reinforcement learning (Sec. 4.3). Figure 3 illustrates the two supervised stages.

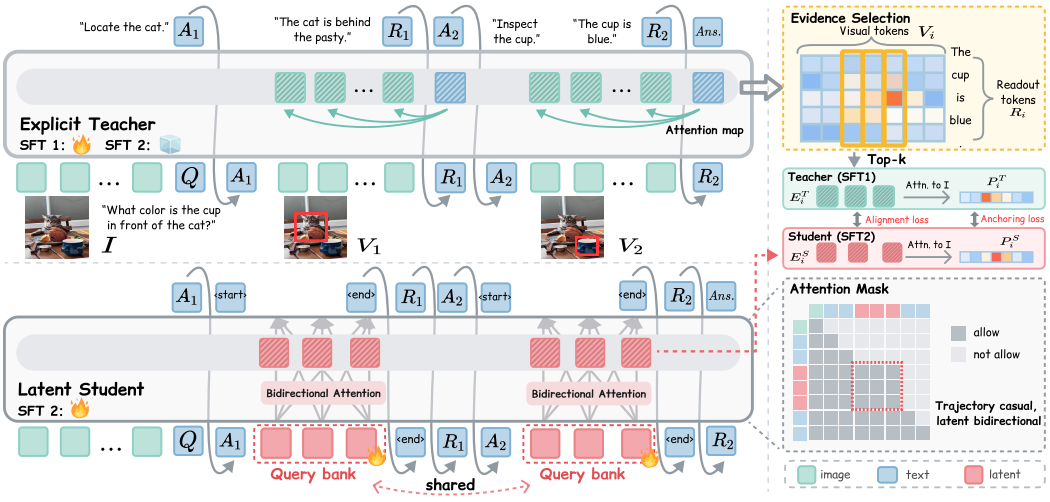


Figure 3: **Supervised learning of evidence-grounded latent visual states.** An explicit teacher learns from AVR trajectories and selects visual evidence through readout-to-visual attention (SFT1). The student replaces intermediate images with a shared query bank, forming each latent state through bidirectional within-state attention while preserving trajectory causality (SFT2). Evidence alignment matches teacher representations, and visual anchoring aligns attention to the original image.

4.1 READOUT-CONDITIONED VISUAL EVIDENCE EXTRACTION

Explicit visual-reasoning teacher. We fine-tune Qwen2.5-VL-7B on the AVR trajectories with cross-entropy loss, retaining the intermediate images as inputs. At step i, the model predicts readout R_i conditioned on action A_i, intermediate image V_i, and the preceding multimodal context. This trains a teacher whose visual representations can supply targets for the latent replacement of V_i.

Readout-conditioned evidence selection. The preceding action describes what to inspect, while the subsequent readout states the observation obtained from the image. We therefore use readout-to-visual (**r2v**) attention as a training-time signal for selecting evidence relevant to that observation. At decoder layer $\ell^\star$, we average attention over the readout tokens and H heads:

$$\mathbf{A}_i^T = \frac{1}{H|R_i|} \sum_{h=1}^{H} \sum_{t \in R_i} \boldsymbol{\alpha}_{t,V_i}^{\ell^\star,h} \in \mathbb{R}^{|V_i|}, \tag{2}$$

where $\boldsymbol{\alpha}_{t,V_i}^{\ell^\star,h}$ is the attention from readout token t to the visual tokens of V_i. We select the K highest-scoring visual positions and arrange them in spatial sequence order, $p_1 < \cdots < p_K$. Hence, their corresponding output vectors at the same decoder layer form the evidence target:

$$\mathbf{E}_i^T = [\mathbf{h}_{i,p_1}^{\ell^\star}, \ldots, \mathbf{h}_{i,p_K}^{\ell^\star}] \in \mathbb{R}^{K \times d}, \tag{3}$$

where d is the hidden dimension. For images with more than 256 visual tokens, we first spatially average-pool both scores and output vectors to an aspect-ratio-preserving grid of at most 256 positions, then apply the same selection. Thus, each target vector represents a selected token or local region. Note that the teacher is frozen during later student training, and the current readout is used to construct supervision and is unavailable to the student when forming its latent state.

4.2 PARALLEL LATENT VISUAL STATE FORMATION

To train the student, we replace each intermediate image with a latent state whose positions are formed jointly. These positions expand representational width without imposing a generation order within the state. Note that the successive states still follow the causal reasoning trajectory.

Shared latent queries. We initialize the student from the teacher and replace each V_i with a latent block Z_i. After action A_i, the token $\langle$`latent_start`$\rangle$ triggers K learnable query embeddings:

$$\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_K] \in \mathbb{R}^{K \times d}. \tag{4}$$

This query bank is shared across steps. Each invocation processes all queries together within the language decoder, retrieving information from the current multimodal context.

Trajectory-causal, state-bidirectional attention. Let C_i contain all positions preceding Z_i, including the original image I, question q, previous reasoning and latent states, current action, and start token. Let $\mathcal{Z}_i$ denote the positions of the current latent block. For query position u and key position v, we have a mask

$$\mathsf{M}_{uv} = \begin{cases} 0, & u \in \mathcal{Z}_i,\ v \in C_i \cup \mathcal{Z}_i, \\ 0, & u \notin \bigcup_i \mathcal{Z}_i,\ v \le u, \\ -\infty, & \text{otherwise.} \end{cases} \tag{5}$$

Thus, latent positions attend to the preceding context and all positions in their own state, while non-latent positions retain causal attention. No position accesses subsequent reasoning. At the supervision layer $\ell^\star$, the student representation is

$$\mathbf{E}_i^S = \mathbf{Z}_i^{\ell^\star} = F_\theta^{\ell^\star}(\mathbf{Q}; C_i, \mathsf{M}) = [\mathbf{z}_{i,1}^{\ell^\star}, \ldots, \mathbf{z}_{i,K}^{\ell^\star}], \tag{6}$$

where $F_\theta^{\ell^\star}$ returns the decoder outputs at the latent positions. The last latent position is trained to predict $\langle$`latent_end`$\rangle$. At inference, this token is inserted after the fixed-size block, and autoregressive text generation resumes.

Evidence alignment. We align the student positions with the spatially ordered teacher targets using mean squared error:

$$\mathcal{L}_{\text{evi},i} = \frac{1}{Kd} \left\| \mathbf{E}_i^S - \text{sg}(\mathbf{E}_i^T) \right\|_F^2, \tag{7}$$

where sg stops gradients through the teacher target.

Visual anchoring. To preserve a direct connection to the original image, we also align teacher T and student S attention over its token positions $\mathcal{I}$. At layer $\ell^\star$, we normalize each attention map over $\mathcal{I}$, then average over evidence positions and heads:

$$\mathbf{P}_i^m = \frac{1}{KH} \sum_{j=1}^{K} \sum_{h=1}^{H} \bar{\boldsymbol{\alpha}}_{i,j\to\mathcal{I}}^{m,\ell^\star,h}, \qquad m \in \{T, S\}. \tag{8}$$

Here, j indexes a selected teacher position or its corresponding student slot, and $\bar{\boldsymbol{\alpha}}$ denotes attention normalized over the input-image tokens. For a pooled teacher region, we average these normalized maps over its constituent tokens. The anchoring loss is

$$\mathcal{L}_{\text{anchor},i} = D_{\text{KL}}\big(\text{sg}(\mathbf{P}_i^T) \,\|\, \mathbf{P}_i^S\big)\,. \tag{9}$$

Evidence alignment specifies the state content, while visual anchoring supervises its attention to the original image. If fewer than K teacher candidates are available, both losses use only valid target slots and replace K in their averages with the valid count.

Training objective. Each AVR step (A_i, V_i, R_i) becomes (A_i, Z_i, R_i). We combine cross-entropy on the supervised output tokens with the two auxiliary losses, averaged over latent blocks:

$$\mathcal{L}_{\text{SFT2}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{evi}}\mathcal{L}_{\text{evi}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}}. \tag{10}$$

4.3 Adaptive Latent Invocation via Reinforcement Learning

An *invocation* constructs one complete latent-state block. Supervised trajectories teach repeated latent state construction. Here, RL trains the model to decide when to initiate another block. Our **Invocation Progressive Reward** adds a bonus to correct trajectories only when their invocation count achieves higher empirical success than all smaller observed counts for the same prompt.

For each prompt, we sample G trajectories and group them by invocation count. Let $\mathcal{B}$ be the set of observed counts. For $r \in \mathcal{B}$, the empirical success rate is $p_r = \frac{c_r}{N_r}$, where N_r and c_r count all sampled trajectories and correct trajectories with r invocations, respectively. We define the gain over smaller observed counts as

$$g_r = \begin{cases} 0, & \{s \in \mathcal{B} : s < r\} = \varnothing, \\ \Big[p_r - \max\limits_{s\in\mathcal{B}:s<r} p_s\Big]_+, & \text{otherwise}, \end{cases} \tag{11}$$

where $[x]_+ = \max(x, 0)$. The smallest observed count receives no gain. For rollout b with invocation count n_b, the final reward is

$$\mathcal{R}_b = a_b + \lambda_f f_b + \lambda_p a_b g_{n_b}, \tag{12}$$

where $a_b \in \{0, 1\}$ indicates answer correctness and $f_b \in \{0, 1\}$ indicates a boxed answer with well-formed latent markers and at least one invocation. We set $\lambda_f = 0.1$ and $\lambda_p = 0.3$, and optimize the combined reward with GRPO (Shao et al., 2024b). The progressive term supplies a within-prompt signal based on observed success across invocation counts, and additional calls alone earn no progressive bonus.

5 Experiments

We evaluate ELVIS across visual perception, spatial reasoning, and out-of-distribution tasks, then analyze its visual evidence and invocation behavior. Ablations examine evidence supervision, latent-state formation, state size, and training components. Further analyses characterize task-dependent representations and the cost of latent-state computation.

5.1 Experimental Setup

Baselines. We compare ELVIS against three categories of baselines: (1) textual reasoning models, including GPT-4o (Hurst et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and vanilla SFT, where

Table 1: Accuracy (%) across visual perception, spatial reasoning, and OOD benchmarks. Bold indicates the best result among latent visual reasoning models, including ties. The final row reports the improvement (%) of our full model over the base model Qwen2.5-VL-7B.

Model	Latent Invocation	Visual Perception												Spatial Reasoning			OOD
		V*			HRBench4K			HRBench8K			MME-RealWorldLite			Blink		Zebra	VisuLogic
		Overall	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall	Reason.	Percept.	Jigsaw	Multi-v.	Maze	
Textual Reasoning																	
GPT-4o	–	67.54	72.17	60.53	59.00	70.00	48.00	55.50	62.00	49.00	52.01	48.27	54.41	55.33	55.64	22.03	26.30
Qwen2.5-VL-7B	–	76.44	77.39	75.00	67.88	77.25	58.50	65.25	78.75	51.75	45.75	39.73	49.62	62.67	42.86	15.25	25.60
+ vanilla SFT	–	79.06	80.87	76.32	68.13	86.75	49.50	65.50	81.25	49.75	47.79	43.87	50.30	48.67	42.11	35.59	23.90
Tool-based Visual Reasoning																	
PixelReasoner	–	80.63	83.48	76.32	72.88	86.00	60.25	66.88	80.00	54.25	49.71	44.53	53.12	–	–	–	–
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	–	–	–	–
Thyme	–	82.20	83.48	80.26	77.00	91.00	63.00	72.00	86.50	57.50	55.18	49.07	59.11	–	–	–	–
Latent Visual Reasoning Models																	
LVR	single	81.68	84.35	77.63	70.75	83.75	57.75	63.00	74.50	51.50	50.23	43.33	54.66	56.67	44.36	22.03	25.30
Monet	single	83.25	83.48	82.89	71.00	85.25	56.75	68.00	**79.75**	56.25	55.50	51.07	58.34	40.67	50.38	10.17	24.20
COVT	single	82.72	86.09	77.63	72.00	**85.50**	58.50	66.13	78.00	54.25	48.62	42.53	52.52	57.33	44.36	22.88	25.60
ILVR	adaptive	72.77	75.65	68.42	58.13	65.75	50.50	53.00	60.00	46.00	39.55	34.53	42.77	41.33	45.86	21.19	21.40
ELVIS (SFT)	adaptive	84.82	85.22	**84.21**	71.88	84.75	59.00	66.13	76.00	56.25	52.42	49.60	54.23	54.00	**57.89**	36.44	24.90
ELVIS (SFT + GRPO)	adaptive	83.25	83.48	82.89	71.13	82.75	59.50	66.13	76.50	55.75	**56.23**	**51.73**	**59.11**	56.67	54.14	34.75	**25.90**
ELVIS (SFT + GRPO + Prog.)	adaptive	**85.34**	**86.09**	**84.21**	**73.75**	**85.50**	**62.00**	**69.38**	79.00	**59.75**	55.50	49.87	**59.11**	**58.00**	50.38	**37.29**	25.70
Relative improvement		+8.90	+8.68	+9.21	+5.87	+8.25	+3.50	+4.13	+0.25	+8.00	+9.33	+10.94	+8.29	-4.67	+7.52	+22.00	+0.10

vanilla SFT is obtained by fine-tuning Qwen2.5-VL-7B on AVR-112K. Qwen2.5-VL-7B also serves as the backbone for the subsequent baselines. (2) Tool-based visual reasoning models, including PixelReasoner (Su et al., 2025), DeepEyes (Zheng et al., 2026), and Thyme (Zhang et al., 2026b), which enhance visual reasoning via explicit tool operations. (3) Latent visual reasoning models, including LVR (Li et al., 2026b), CoVT (Qin et al., 2026), Monet (Wang et al., 2026b), and ILVR (Dong et al., 2026). Among them, LVR, Monet, and CoVT use a single latent invocation, whereas ILVR supports multiple latent invocations.

Benchmarks. We evaluate ELVIS on seven benchmarks covering diverse scenarios. For Visual Perception, we adopt V* (Wu & Xie, 2024), HRBench4K/8K (Wang et al., 2025), and MME-RealWorld-Lite (Zhang et al., 2025), covering fine-grained visual understanding, high-resolution perception, and complex real-world scenes. For Spatial Reasoning, we evaluate on the Jigsaw and Multi-view subsets of BLINK (Fu et al., 2024), which require reasoning over visual transformations and cross-view relations, and Zebra Maze (Li et al., 2026a), which tests visual planning. VisuLogic (Xu et al., 2026) is used to measure out-of-distribution generalization on logical reasoning.

Training details. We use Qwen2.5-VL-7B as the base model and conduct all training on 8 NVIDIA A800-80G GPUs. For SFT, we train on AVR-112K with a per-device batch size of 1 and a gradient accumulation step of 8. SFT Stage 1 is trained for 2 epochs, while SFT Stage 2 is trained for 1 epoch. For RL training, we use a 5K subset of HyLaR-RL (Cheng et al., 2026) for 1 epoch with 16 rollouts per prompt. We use $K = 16$ latent positions and decoder layer $\ell^\star = 22$ (zero-based indexing), selected empirically, for evidence supervision. Details are provided in Appendix A.3.

5.2 Main Results

Table 1 reports accuracies. ELVIS with the progressive reward obtains the strongest latent-method result on V*, HRBench4K, HRBench8K, BLINK Jigsaw, and Zebra Maze. Vanilla GRPO has the highest latent-method MME-RealWorld-Lite and VisuLogic scores, while the SFT model performs best on BLINK Multi-view. Thus, the full model performs well across distinct task families, although the progressive reward does not uniformly improve every task. On HRBench's cross-instance FCP split, the full model exceeds the best prior latent baseline by 3.5% at both 4K and 8K. On Zebra Maze, ELVIS reaches 37.29%, compared with 15.25% for the base model, a gain of 22.04%. This extends the improvements to a task requiring multi-step spatial planning.

5.3 Visual Evidence and Adaptive Invocation

Beyond overall performance, we examine how attended visual evidence evolves, how evidence concentration relates to reasoning success, and how invocation behavior changes across training stages.

Evolving visual evidence. To inspect the evidence attended by the learned latent states, we visualize the attention from latent state to the input image. As shown in Fig. 4(a), ELVIS exhibits structured evidence patterns that evolve with reasoning: on V*, successive invocations shift attention from the white statue to the green statue, while on Zebra-Maze, different invocations progressively form route-like patterns over the image. In contrast, ILVR (Dong et al., 2026) exhibits little change in its

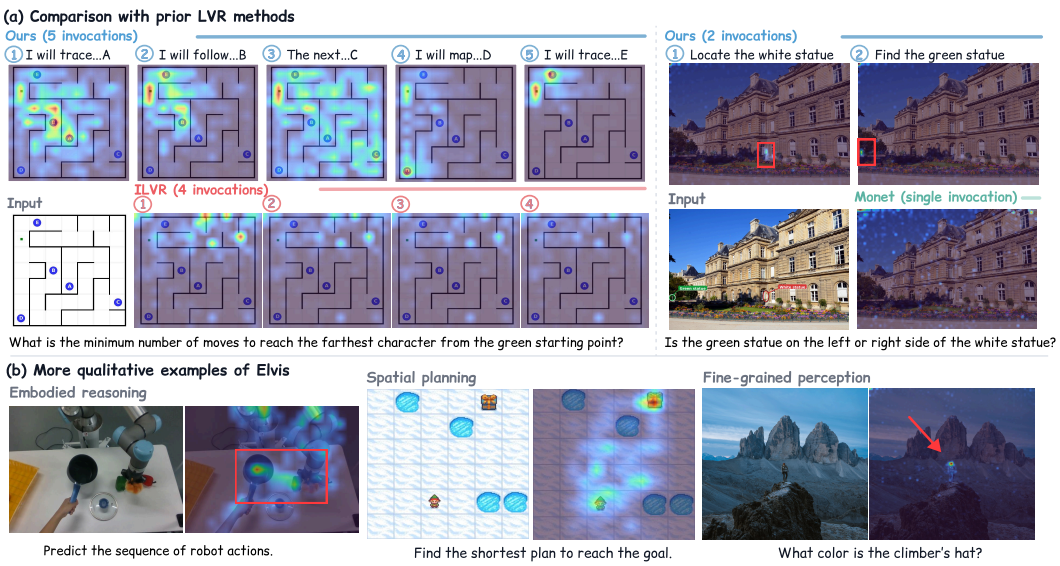


Figure 4: **Evolving visual evidence across latent invocations.** The comparison examples in (a) are from Zebra-Maze and V*. The examples in (b) are from Zebra-Robot-Planning, Zebra-Maze, and HRBench4K, respectively. The Zebra-Robot-Planning visualization overlays attention from four invocations, while the latter two show a single invocation.

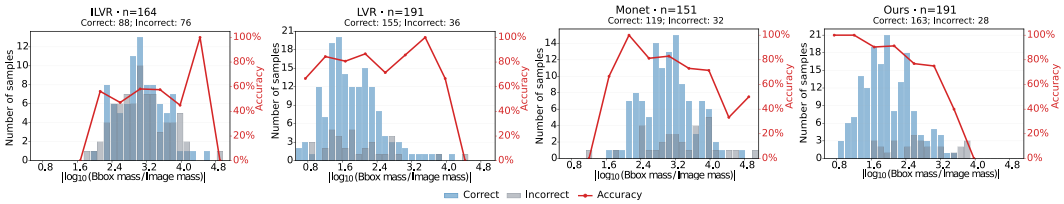


Figure 5: **Latent evidence concentration and reasoning outcomes on V*.** The x-axis shows the negative log of $\frac{\text{AttentionMass}(\text{latent}\to\text{BBox})}{\text{AttentionMass}(\text{latent}\to\text{Image})}$, larger values indicate weaker evidence concentration.

attention patterns across steps, while Monet's single invocation (Wang et al., 2026b) provides weak evidence concentration. Fig. 4(b) further shows task-relevant attention across different benchmarks. These qualitative examples illustrate changes in attended evidence as reasoning proceeds.

Evidence concentration and reasoning success. To assess the relationship between the visual evidence carried by latent states and reasoning outcomes, we quantify evidence concentration as the proportion of latent-to-image attention mass falling within the annotated evidence regions and examine how it relates to answer accuracy. We group samples by the negative log of this ratio, and compare accuracy together with the distributions of correct and incorrect predictions. As shown in Fig. 5, ELVIS exhibits a clearer trend: as evidence concentration weakens, accuracy steadily decreases, with correct samples concentrated in stronger-evidence regions and errors shifting toward weaker ones. In contrast, LVR, Monet, and ILVR exhibit less consistent or noisier trends.

Adaptive invocation across training stages. To examine whether ELVIS learns an adaptive invocation policy, we compare per-sample invocation counts between SFT2 and GRPO, and between GRPO and GRPO with the Invocation Progressive Reward. For each comparison, we report the proportion of samples whose invocation count changes and evaluate accuracy before and after on the same changed subset. As shown in Tab. 2, the Invocation Progressive Reward induces more invocation changes and brings consistent accuracy gains on the changed subsets, with improvements of up to 20%. In contrast, vanilla GRPO yields fewer changes and smaller gains. The changed proportions vary across datasets, indicating task-dependent adjustments in invocation behavior.

5.4 Ablation Studies and Further Analysis

Training component. We assess the contribution of each component in our model in Tab. 3. Replacing pooled features with selected visual evidence improves all four benchmarks, confirming

Table 2: Invocation changes across training stages.

Dataset	Comparison	Changed (%)	Before Acc. (%)	After Acc. (%)	ΔAcc. (pp)
V*	SFT2 → GRPO	2.09	25.00	25.00	0.00
	GRPO → + Prog.	**12.04**	73.91	86.96	**+13.04**
HR4K	SFT2 → GRPO	2.13	58.82	70.59	+11.76
	GRPO → + Prog.	**2.50**	35.00	55.00	**+20.00**
Jigsaw	SFT2 → GRPO	18.00	48.15	55.56	+7.41
	GRPO → + Prog.	**48.67**	56.16	72.60	**+16.44**

Table 3: Component ablation of ELVIS.

Mode	V*	HR4K	Jigsaw	Maze
ELVIS-SFT (Pooling Sup.)	82.20	69.75	47.33	34.75
ELVIS-SFT (Evidence Sup.)	83.77	70.75	52.00	35.59
+ Anchor KL	84.82	71.88	54.00	36.44
+ GRPO	83.25	71.13	56.67	34.75
+ Prog. Reward (Full)	**85.34**	**73.75**	**58.00**	**37.29**

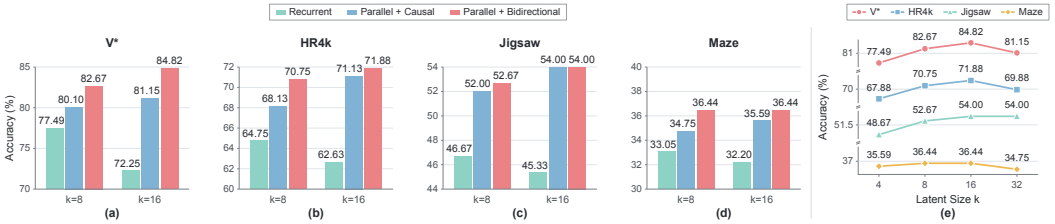


Figure 6: (a–d) Ablation of latent-state formation mechanisms. (e) Ablation of latent-state size k.

the benefit of fine-grained evidence supervision. Visual anchoring brings further consistent gains. Vanilla GRPO yields mixed changes, whereas adding the progressive reward consistently improves over GRPO across all four reported benchmarks, confirming the benefit of the progressive reward.

Latent-state. We compare three latent formation mechanisms in SFT2 at $K = 8$ and $K = 16$: recurrent sequential generation, parallel insertion with causal attention, and our parallel-bidirectional design. As shown in Fig. 6(a)-(d), parallel formation consistently outperforms recurrence, with the bidirectional design achieving the strongest overall performance. As K increases, recurrent formation degrades across all benchmarks, whereas parallel formation remains stable or improves, supporting our collaborative state design. We further ablate latent-state size K in Fig. 6(e), which jointly determines the number of selected teacher evidence targets and the number of student latent components. We vary $K \in \{4, 8, 16, 32\}$ while keeping all other training settings fixed. At $K = 16$, V* and HRBench4K reach their best, Jigsaw ties with $K = 32$, and Maze with $K = 8$. We also report latent t-SNE projections (van der Maaten & Hinton, 2008) in Fig. 7. We can see that ELVIS adapts both its representations and latent-state usage to task demands.

Latent-state computation latency. We compare the cumulative latency of latent-state computation in Tab. 4. ELVIS substantially reduces latency by constructing all latent components in parallel. At $K = 16$, ELVIS achieves $11.65\times$–$17.40\times$ speedup over the fastest applicable recurrent baseline under the same number of invocations. Moreover, its latency remains nearly unchanged when increasing K from 8 to 16 while recurrent LVR methods scale increasingly costly with latent size.

Table 4: Latent-state latency (ms).

Model	Invocations	V*			Maze		
		$K = 8$	$K = 16$	Speedup	$K = 8$	$K = 16$	Speedup
LVR	1	183.45	345.99	–	182.68	344.85	–
Monet	1	164.73	329.10	–	160.86	321.36	–
ILVR	1	184.88	348.77	–	179.90	340.17	–
ELVIS	1	**28.38**	**28.24**	**11.65×**	**20.01**	**19.86**	**16.18×**
ILVR	2	370.87	700.18	–	360.82	680.86	–
ELVIS	2	**57.45**	**57.08**	**12.27×**	**39.97**	**39.14**	**17.40×**
ILVR	4	733.22	1385.91	–	717.78	1354.92	–
ELVIS	4	–	–	–	**81.64**	**81.03**	**16.72×**

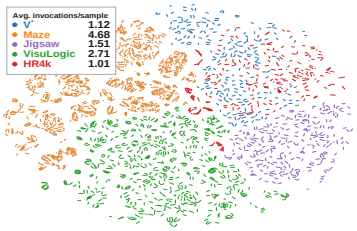


Figure 7: Latent state t-SNE.

6 Conclusion

We presented ELVIS, a framework for constructing evidence-grounded latent visual states for multi-step reasoning. Motivated by the limited evidence sensitivity and downstream utility of recurrent latents, ELVIS jointly forms each state through parallel, bidirectional interactions and invokes new states as reasoning progresses. AVR-112K connects explicit visual observations to readout-conditioned evidence supervision, while evidence alignment and visual anchoring ground the learned states. An Invocation Progressive Reward further supports adaptive state invocation. Experiments and ablations demonstrate the benefits of this design for visual perception and spatial reasoning, alongside substantially reduced latent-state computation latency. These findings support coupling latent-state construction with explicit, step-specific visual evidence supervision.

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tao Cheng, Shi-Zhe Chen, Hao Zhang, Yixin Qin, Jinwen Luo, and Zheng Wei. Hylar: Hybrid latent reasoning with decoupled policy optimization, 2026. URL https://arxiv.org/abs/2604.20328.

Jin Cui, Xinyue Long, Xunyong Zhang, Yadong Zhang, Chuanchang Su, Jingye Gan, Boran Zhao, and Pengju Ren. Retrieve, Integrate, and Synthesize: Spatial-Semantic Grounded Latent Visual Reasoning. arXiv preprint arXiv:2605.07106, 2026. URL https://arxiv.org/abs/2605.07106.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved Latent Visual Reasoning with Selective Perceptual Modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1351. URL https://aclanthology.org/2026.acl-long.1351/.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024*, pp. 148–166. Springer, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://link.springer.com/chapter/10.1007/978-3-031-73337-6_9.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. ReFocus: Visual Editing as a Chain of Thought for Structured Image Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17783–17805. PMLR, 2025. URL https://proceedings.mlr.press/v267/fu25d.html.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. ThinkMorph: Emergent Properties in Multimodal Interleaved Chain-of-Thought Reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e5095602ad1c6a835e2b643ec4ed97d0-Abstract-Conference.html.

Jiawei Guo, Yu Chen, Xiang Wang, Shuai Li, Xinpei Zhao, Huaxing Liu, Shuai Dong, Feifei Zhai, and Yu Zhou. Beyond Visual Memory: Mechanistic Diagnostics of Latent Visual Reasoning. arXiv preprint arXiv:2606.01287, 2026. URL https://arxiv.org/abs/2606.01287.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024. doi: 10.52202/079017-4423. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-CoT: A Dataset for Interleaved Vision-Language Reasoning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6087a60306544be7ba0d0cf34aa93c8f-Abstract-Conference.html.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent Visual Reasoning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/efa0939caa8a906fd0a61a9c60e46e8b-Paper-Conference.pdf.

You Li, Chi Chen, Yanghao Li, Fanhu Zeng, Kaiyu Huang, Jinan Xu, and Maosong Sun. Imagination Helps Visual Reasoning, But Not Yet in Latent Space. In *International Conference on Machine Learning*, 2026c. URL https://arxiv.org/abs/2602.22766.

Dongchen Lu, Zhimo Li, Mao Shu, and Huo Cao. DeepLatent: Think with Images via Parallel Latent Visual Reasoning. arXiv preprint arXiv:2606.00562, 2026. URL https://arxiv.org/abs/2606.00562.

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. In *European Conference on Computer Vision*, 2026. URL https://wakalsprojectpage.github.io/covt-website/.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024a. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 9130–9159, 2025. doi: 10.52202/085713-0280. URL https://papers.nips.cc/paper_files/paper/2025/file/0c38f54740062529aa4117a04b583f3c-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2512.16584.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://jmlr.org/papers/v9/vandermaaten08a.html.

André G. Viveiros, Nuno Gonçalves, André F. T. Martins, and Matthias Lindemann. What's Holding Back Latent Visual Reasoning? arXiv preprint arXiv:2605.18445, 2026. URL https://arxiv.org/abs/2605.18445.

Chenfeng Wang, Wei He, Xuhan Zhu, Chunpeng Zhou, Qizhen Li, Song Yan, Yufei Zheng, Chengjun Yu, Fan Lu, Wei Zhai, Yang Cao, Pengfei Yu, and Zheng-Jun Zha. Self-Consistent Latent Reasoning: Long Latent Sequence Reasoning for Vision-Language Model. arXiv preprint arXiv:2605.12163, 2026a. URL https://arxiv.org/abs/2605.12163.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in Latent Visual Space Beyond Image and Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Monet_Reasoning_in_Latent_Visual_Space_Beyond_Image_and_Language_CVPR_2026_paper.html.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025. doi: 10.1609/aaai.v39i8.32852. URL https://ojs.aaai.org/index.php/AAAI/article/view/32852.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=mXuzDDVXxi.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine Mental Imagery: Empower Multimodal Reasoning with Latent Visual Tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Yang_Machine_Mental_Imagery_Empower_Multimodal_Reasoning_with_Latent_Visual_Tokens_CVPR_2026_paper.pdf.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. CrystaL: Spontaneous Emergence of Visual Latents in MLLMs. arXiv preprint arXiv:2602.20980, 2026a. URL https://arxiv.org/abs/2602.20980.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think Beyond Images. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8080912b919ad0344c0d34e68983dfcb-Abstract-Conference.html.

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*, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. EasyR1: An efficient, scalable, multi-modality RL training framework. https://github.com/hiyouga/EasyR1, 2025.

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. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

A Appendix

A.1 Diagnosing Recurrent Latent Visual Reasoning Methods

We detail the diagnostic experiments summarized in Fig. 1(a), examining recurrent LVR methods from three perspectives: visual sensitivity, reasoning utility, and internal dynamics.

Visual Sensitivity. We examine whether recurrent latent states are sensitive to the visual evidence. We conduct the experiment on V^*, which provides annotated bounding boxes for target regions. For each sample, we mask only the annotated target regions while leaving the rest of the image unchanged. The reasoning text tokens preceding the latent state are fixed to those from the original run, thereby isolating the effect of removing decisive visual information. For each latent position, we compare the layer-averaged hidden states produced from the original and masked images using cosine distance ($1 -$ cosine similarity). As shown in Fig. 1(a), LVR, Monet, and ILVR exhibit only limited representational changes after the decisive evidence is removed, with the sensitivity further weakening toward later recurrent steps. These results reveal *a weak and progressively diminishing connection between their latent states and the decisive visual evidence.*

Reasoning Utility. Sensitivity alone does not establish whether the information carried by latent states can actually support reasoning. We therefore perform a latent-transplantation experiment. For each sample, we compare three conditions: inference on the original image, inference after masking the evidence region, and inference on the masked image while replacing the latent hidden states at each decoder layer with those obtained from the original-image run. The visual input remains masked throughout the transplantation condition, allowing us to directly test whether the original latents contain information that can compensate for the removed evidence. If the latent states preserve usable task-relevant visual information, transplanting them should recover part of the accuracy lost after masking. As shown in Fig. 1(a), LVR, Monet, and ILVR show negligible gains after transplantation, with improvements of 5.8, 0.0, and 1.1 points. In contrast, ELVIS recovers 21.0 points. *These results indicate that existing recurrent latents provide little functional support to downstream reasoning.*

Internal Dynamics. We further examine the internal dynamics of recurrence using 10 samples from V^*. For each sample, we track the hidden state of each latent position across decoder layers and project the resulting trajectories into PCA space. As shown in Fig. 1(a), later latent positions increasingly follow overlapping trajectories, with trajectories from different samples also converging to similar paths. Meanwhile, the first two principal components explain 90.6%, 68.6%, and 86.4% of the variance for LVR, Monet, and ILVR, respectively, indicating that their variation is concentrated in a few dominant directions. For clearer inspection of these dynamics, Fig. 9 further visualizes the latent trajectories of individual samples separately. Taken together, these patterns suggest a tendency toward latent-state collapse under recurrence, characterized by *intra-sample, inter-position latent homogenization and feature-level collapse into a few dominant directions*.

A.2 Dataset Construction

We summarize the composition of AVR-112K in Tab. 5. Each training instance contains an image–text interleaved CoT with intermediate auxiliary images, which is reformulated into an Action–Visual–Readout (AVR) trajectory. Representative AVR training samples are provided in Figs. 10 and 11.

A.3 Implementation Details

A.3.1 Training Details

We adopt Qwen2.5-VL-7B as the base model and perform supervised fine-tuning in two stages. In Stage 1, we freeze the vision encoder and train only the language model backbone. In Stage 2, the vision encoder remains frozen, while the language model backbone and the learnable latent query embeddings are jointly optimized. We conduct SFT on AVR-112K and perform RL training on a 5K subset of HyLaR-RL. RL is implemented using EasyR1 (Zheng et al., 2025), an open-source reinforcement learning framework for multimodal LLMs. Detailed training hyperparameters are provided in Tab 6, Tab 7.

Table 5: Composition of AVR-112K. Avg. Aux. Images denotes the average number of intermediate images per sample.

Source	Subset	Task Type	# Samples	Avg. Aux. Images
Zebra-CoT	Visual Jigsaw	2D spatial reasoning	8,000	3.10
	Robot Planning	3D robot planning	6,744	5.09
	Embodied CoT	3D embodied reasoning / navigation	1,000	6.92
	Chess	Chess planning	2,483	4.00
	Maze	Spatial reasoning / maze pathfinding	14,800	3.23
	Visual Search	Visual perception / grounding	12,687	1.02
Monet-SFT-125K	CogCoM	Visual reasoning / charts grounding	567	1.58
	ReFocus	Chart reasoning / visual refocusing	426	1.00
	Visual-CoT	Visual perception / grounding	58,561	1.00
	Object Counting	3D multi-hop object counting	2,766	5.28
ThinkMorph	Chart Refocus	Chart reasoning / visual refocusing	4,000	1.00
Total / Average			**112,034**	**1.92**

Evidence Supervision Layer. We ablate the decoder layer used to extract teacher evidence for SFT2. For each intermediate image, we average readout-to-visual attention over reasoning tokens and heads. As shown in Fig. 8, layer 22 yields the strongest text-visual semantic alignment, with attention most concentrated on task-relevant regions, whereas others exhibit less accurate localization or pronounced attention sinks in irrelevant regions.

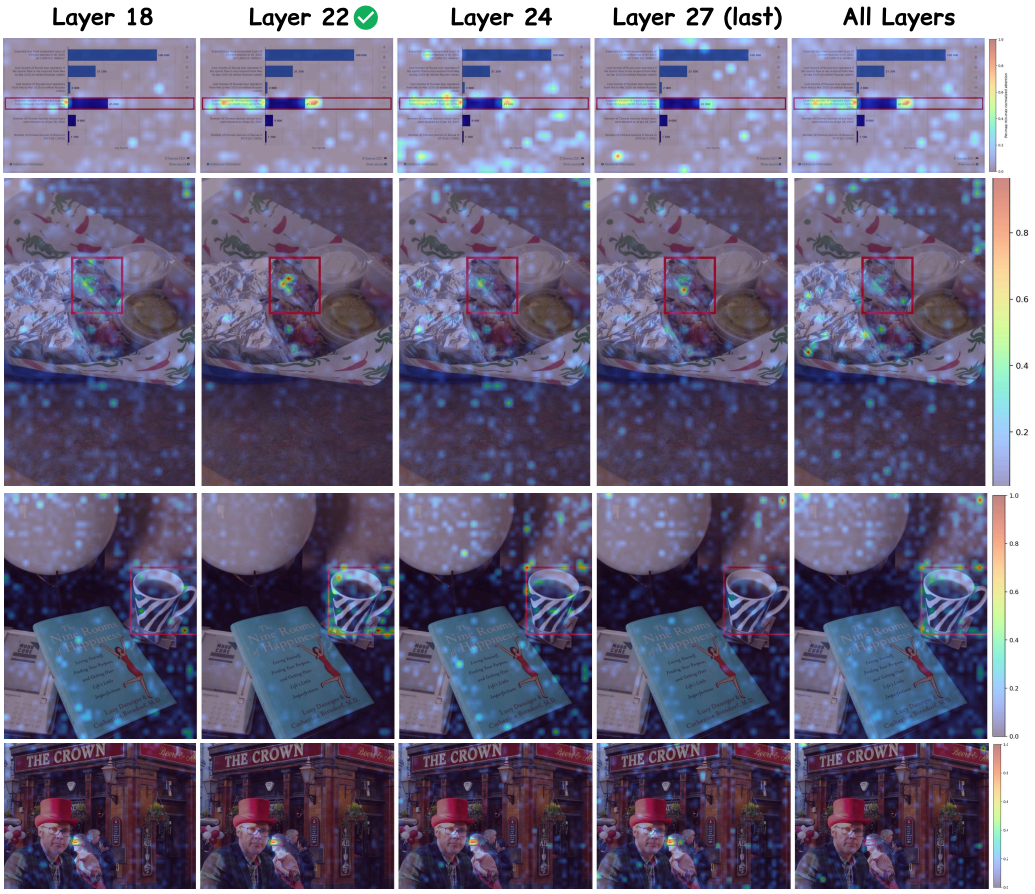


Figure 8: Readout-to-visual attention across decoder layers for teacher evidence extraction. Red boxes denote task-relevant regions; Layer 22 shows the most consistent concentration.

Table 6: Training hyperparameters for the two-stage supervised fine-tuning.

Hyperparameter	**SFT1**	**SFT2**
Initial model	Qwen2.5-VL-7B-Instruct	SFT1 checkpoint-3500
Learning rate	1×10^{-5}	2×10^{-6}
Per-device batch size	1	1
Gradient accumulation steps	8	8
Number of GPUs	8	8
Effective global batch size	64	64
Optimizer	AdamW	AdamW
Weight decay	0.01	0.01
LR scheduler	Cosine	Cosine
Warmup steps	10	100
Training steps	3,500 ($\sim$2 epochs)	1,766 (1 epoch)
Latent size	–	16
Maximum latent blocks	–	No fixed limit (max. 19 in training data)
Teacher feature selection	–	Attention Top-K
Alignment loss weight (MSE)	–	0.1
Anchoring loss weight (KL)	–	0.1
Max. total image pixels per sample	$4096 \times 28 \times 28$	–
Max. pixels per image	$2048 \times 28 \times 28$	$2048 \times 28 \times 28$
Distributed strategy	DeepSpeed ZeRO-2	DeepSpeed ZeRO-2

Table 7: Training hyperparameters for reinforcement learning.

Hyperparameter	**Value**
Algorithm	GRPO
Initial model	SFT2 final checkpoint
Learning rate	1×10^{-6}
Optimizer	AdamW
Weight decay	0.01
LR scheduler	Constant with warmup
Warmup ratio	0.03
Training epochs	1
Rollout batch size	64
Rollouts per prompt	16
Temperature	1.0
Top-p / Top-k	1.0 / -1
Maximum response length	2048 tokens
Latent size	16
Maximum latent blocks	No fixed limit
Max. pixels per image	$2048 \times 28 \times 28$
Accuracy / format / progress reward weights	1.0 / 0.1 / 0.3
Number of GPUs	8

Latent Query Initialization. We initialize the learnable latent queries using the average RMS of the visual embeddings rather than the standard $\mathcal{N}(0, 0.02^2)$ PyTorch initialization. Specifically, we sample from $\mathcal{N}(0, \mathrm{rms}_{\mathrm{img}}^2)$, where $\mathrm{rms}_{\mathrm{img}} \approx 0.9$. We find this scale-matched initialization improves training stability, preventing the MSE alignment loss from being dominated by magnitude differences.

A.3.2 Evaluation Details

We use VLMEvalKit (Duan et al., 2024) for evaluation on all supported benchmarks. For Zebra Maze, we evaluate on a held-out set of 118 samples from Zebra-CoT (Li et al., 2026a) and implement

a corresponding evaluation script. Inference outputs from all models are evaluated through an API-based LLM-as-a-Judge using GPT-5-2025-08-07 as the judge model.

A.4 Additional Qualitative Results

We present additional complete inference examples generated by ELVIS in Figs. 12–15. For each sample, we visualize the attention from the latent states to the input image, following the visual-evidence analysis in Sec. 5.3. These examples further illustrate how ELVIS attends to task-relevant visual evidence throughout the reasoning process.

A.5 AI Use Statement

We used generative AI tools to polish the writing of this paper, including improving grammar, wording, and readability. The authors reviewed and edited all AI-assisted text to ensure that it accurately reflects our work. We take full responsibility for the final content of this paper, including its text, claims, and results.

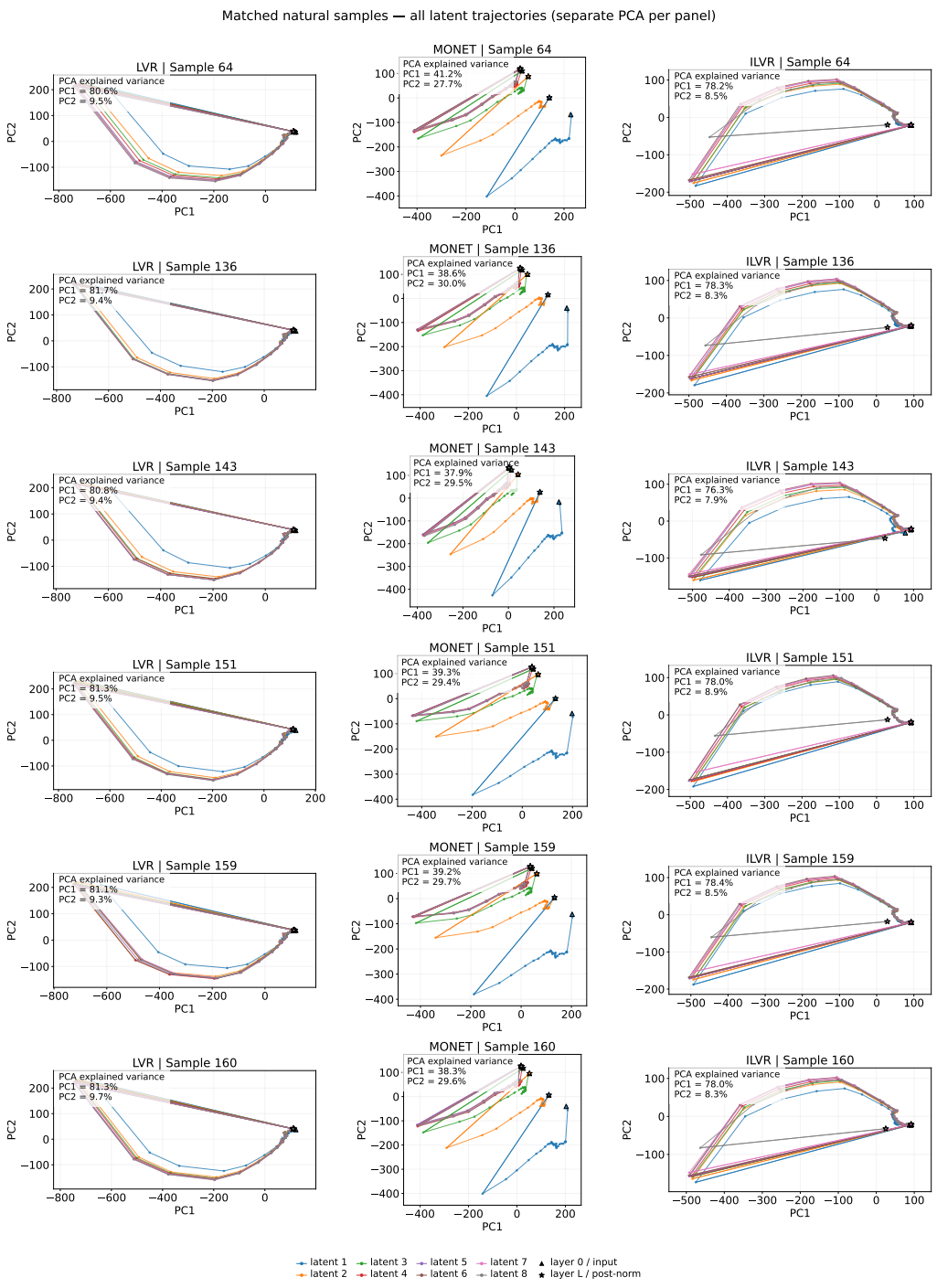


Figure 9: Layer-Wise Latent Trajectories in LVR, MONET, and ILVR. Each color denotes a latent position. Each point in that path represents its PCA-projected hidden state at a decoder layer.

AVR-112k example 1: multi-object spatial reasoning (from Zebra-CoT Visual Search)

Question:
Based on the image provided, is the dining table situated a considerable distance away from the truck? Put your final answer within \\boxed{}.

Question Image:



AVR-112 Reasoning Trajectory:

STEP 1

Action:
I need to locate both the dining table and the truck before comparing their separation. First, I will zoom in on the dining-table area in the foreground.

Intermediate image:



Readout:
The cropped view shows the dining table in the foreground, with plates, glasses, and meals clearly visible on its surface.

STEP 2

Action:
The table has been located in the foreground. I now need to locate the truck in the street area to assess how far apart they are.

Intermediate image:



Readout:
The second crop shows a red pickup truck parked across the sidewalk in the background, with other vehicles and the sidewalk separating it from the foreground table area.

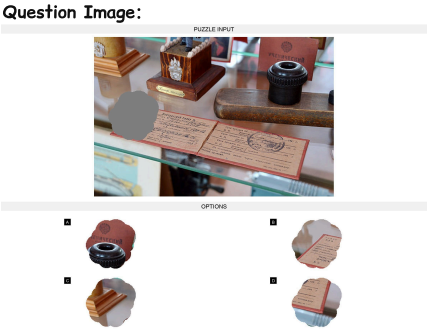
Final Answer:
Therefore, the final answer is $\boxed{Yes}$.

Figure 10: AVR-112k example from Zebra-CoT Visual Search.

AVR-112k example 2: Visual manipulation reasoning (from Zebra-CoT Jigsaw)

Question:
A jigsaw puzzle piece is missing from the main image. Identify which of the provided options correctly fills the empty space. Put your final answer within \\boxed{}.

Question Image:



AVR-112 Reasoning Trajectory:

STEP 1

Action:
I will place the pieces from Option A into the gray gaps and inspect their shape and image alignment.

Intermediate image:



Readout:
Option A inserts a red cover and a black ridged rim over the document's left side, interrupting its handwriting and red border.

STEP 2

Action:
Next, I will test Option B in the missing areas and check whether its visual content continues the surrounding image.

Intermediate image:



Readout:
Option B aligns the document's slanted red border and handwritten entries with the surrounding page while smoothly continuing the tabletop reflection above.

STEP 4

Action:
The next candidate to try is Option D; I will compare its placed pieces with the gap boundaries and nearby details.

Intermediate image:



Readout:
Option D inserts a lower-right page corner and blue reflection into the left-side gap, misaligning both the red border and the tabletop background.

STEP 3

Action:
I will now insert Option C's pieces into the puzzle to evaluate the fit.

Intermediate image:



Readout:
Option C replaces part of the document with a stepped wooden base and gray tabletop, breaking the continuity of the page and its red outline.

Final Answer:
The final answer is $\boxed{B}$.

Figure 11: AVR-112k example from Zebra-CoT Jigsaw.

ELVIS inference example 1: Fine-grained Perception (from V*)

Question:

What is the color of the bicycle? Options:A. blue B. white C. silver D. red

Question Image:



ELVIS Response:

I need to locate the bicycle in the provided image and determine its color. To accurately identify the bicycle's color, I will focus on the specific area where it is visible.

<latent state>

The zoomed-in view clearly shows the bicycle's frame and wheels, allowing for a precise color identification. Based on the visual evidence from the detailed view, the bicycle is distinctly red. Therefore, the final answer is D.

Figure 12: ELVIS inference example from V*.

ELVIS inference example 2: Ultra-high Resolution Grounding (from HR8K)

Question:

What is the color of the flag? Options: A. Blue and white B. Yellow and black C. Green and red D. Red and white

Question Image:



ELVIS Response:

I need to locate the flag within the provided image and identify its colors. To accurately determine the flag's colors, a closer view of the flag is necessary.

<latent state>

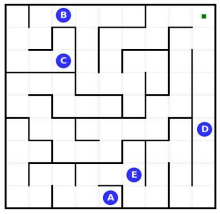
The zoomed-in image clearly shows the flag's distinct colors. Based on the visual evidence from the detailed view, the flag is red and white. Therefore, the final answer is D.

Figure 13: ELVIS inference example from HRBench8k.

ELVIS inference example 3: Spatial Reasoning (from Zebra-CoT Maze)

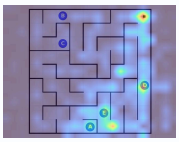
Question:
Determine the shortest possible sequence of steps to travel from the green marker to the farthest letter, moving only in up/down/left/right directions around obstacles. Options: A. 19 B. 18 C. 21 D. 20

Question Image:



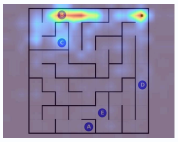
ELVIS Response:

STEP 1
To identify the farthest letter, I will compare the shortest legal route from the green square to each letter. I will trace the shortest legal path from the green square to A and count its orthogonal moves.
<latent state>



The traced path reaches A from the green square in 20 orthogonal moves.

STEP 2
I will now follow the most direct open corridor from the green square to B.
<latent state>



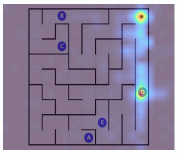
The traced path reaches B from the green square in 12 orthogonal moves.

STEP 3
The next distance to check is the shortest path from the green square to C.
<latent state>



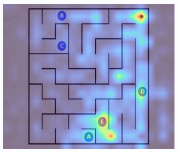
The traced path reaches C from the green square in 18 orthogonal moves.

STEP 4
I will map the minimum-length route connecting the green start to D.
<latent state>



The traced path reaches D from the green square in 5 orthogonal moves.

STEP 5
I will trace the shortest legal path from the green square to E and count its orthogonal moves.
<latent state>



The traced path reaches E from the green square in 16 orthogonal moves.

Final Answer
Comparing the five traced distances, the maximum is 20 moves, reached by A.
The final answer is \\boxed{D. 20}.

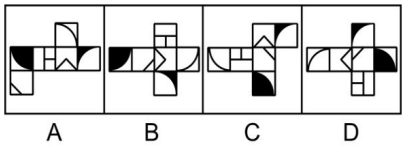
Figure 14: ELVIS inference example from Zebra-CoT Maze.

ELVIS inference example 4: Out-of-Distribution Performance (from VisuLogic)

Question:

Which of the following unfolded views of the outer surface of four cube boxes will form a box that is different from the other three when folded? Options: A. A B. B C. C D. D

Question Image:

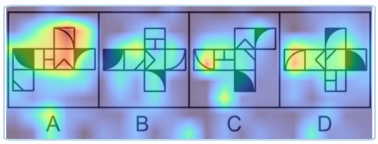


ELVIS Response:

STEP 1

I will examine each provided net to determine which, when folded, does not form a valid cube. To identify the correct answer, I will first visualize the complete cube for comparison.

<latent state>

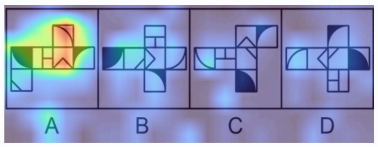


The visual thought shows the complete cube.

STEP 2

Now, I will test Option A by attempting to fold its net into a cube.

<latent state>

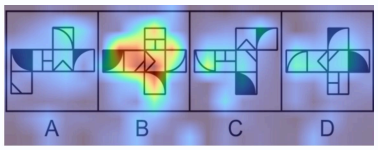


The visual shows that Option A does not form a closed cube.

STEP 3

Next, I will test Option B in the same manner.

<latent state>

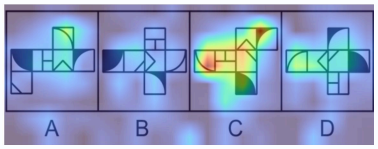


The visual shows that Option B does not form a closed cube.

STEP 4

I will now test Option C.

<latent state>

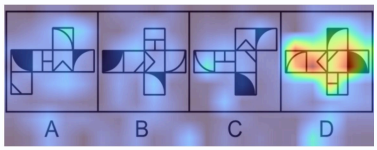


The visual shows that Option C does not form a closed cube.

STEP 5

Finally, I will test Option D.

<latent state>



The visual shows that Option D forms a closed cube, unlike A, B, and C.

Final Answer

The final answer is \\boxed{D}.

Figure 15: ELVIS inference example from VisuLogic.