REASON THROUGH THE LATENT! MAKING LATENT VISUAL REASONING NECESSARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent visual reasoning aims to perform multimodal reasoning through hidden-state computation rather than explicit textual chains of thought. However, visual information being present in a latent state does not imply that the model actually relies on that state when producing its answer, especially when alternative image-conditioned paths remain available. We introduce **C**ausal **V**isual **R**ecurrent **R**easoning (CVRR), which preserves pretrained visual competence while making recurrent computation the required image-conditioned path to prediction. CVRR initializes recurrence from the question hidden state after the pretrained vision-language model has incorporated the image, then repeatedly updates this state while re-reading the same fixed visual evidence. Before decoding, visual states and the original multimodal KV cache are removed so that only the final recurrent state carries image-conditioned information to the answer. Across the V^*, MMVP, BLINK, and MME-RealWorld-Lite benchmarks, CVRR retains strong performance under this strict interface, while compatible latent reasoners fail to recover comparable visual competence even when retrained under the same constraint. Causal interventions further show that predictions remain sensitive to recurrent content when the question is held fixed, and that persistent visual evidence causally revises the recurrent trajectory. These results distinguish latent informativeness from latent computation that is actually used for prediction.[1]

1 INTRODUCTION

Latent visual reasoning promises to let multimodal models reason through hidden states without verbalizing every intermediate step. Continuous latent reasoning has shown that intermediate computation need not be expressed as text (Hao et al., 2025; Shen et al., 2025), and recent multimodal methods extend this idea to visual inputs through latent visual tokens or internal hidden states (Yang et al., 2026; Li et al., 2026a). These methods differ in how latent states are constructed, updated, and coupled with visual evidence (Zhang et al., 2025a), with recent work exploring recurrent updates, hybrid text–latent trajectories, and repeated interaction with the image. Such approaches are particularly appealing for visual reasoning, where relevant evidence may be spatially distributed and difficult to express faithfully through intermediate text. Latent visual reasoning therefore offers a natural substrate for maintaining and refining visual information across multi-step inference.

However, a latent state containing task-relevant visual information is not necessarily one that the model actually uses to determine its answer. To test whether existing latent visual reasoners depend on their proposed latent computation, we intervene on their latent states while leaving the original multimodal answer context intact. For several methods, predictions remain largely unchanged, showing that the proposed latent content is not behaviorally necessary when alternative image-conditioned routes remain available. This exposes a central ambiguity in latent visual reasoning: an internal state can be informative without being causally required for prediction. We do not argue that every useful latent reasoner must eliminate all redundant visual paths. Rather, when the scientific claim is that prediction depends on a proposed latent computation, latent informativeness alone is insufficient evidence that the model actually uses that computation. Our goal is therefore not simply to make latent states informative, but to make prediction depend on computation through them without sacrificing the visual competence formed by pretraining.

[1] Our code is publicly available at https://anonymous.4open.science/r/CVRR-FC3F.

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.

This perspective suggests two design requirements. The latent computation should begin from a representation that preserves pretrained visual competence, and it should become the required route through which that competence reaches the answer. We introduce **C**ausal **V**isual **R**ecurrent **R**easoning (CVRR) to satisfy both. CVRR initializes its latent state from the question-token hidden state after the pretrained vision-language model (VLM) has incorporated the image. It then reuses a single decoder layer as a shared recurrent transition, repeatedly updating the question state while re-reading fixed visual evidence. After the final recurrent step, visual rows and the original multimodal KV cache are removed, leaving the recurrent question state as the sole image-conditioned path to prediction. Path necessity is deliberately enforced by construction rather than discovered post hoc. The nontrivial question is whether this constraint can be imposed without destroying pretrained visual competence and whether computation along the required path performs prediction-relevant visual refinement. CVRR is designed to address both.

Experiments across V^* (Wu & Xie, 2024), MMVP (Tong et al., 2024), BLINK (Fu et al., 2024), and MME-RealWorld-Lite (Zhang et al., 2025b) show that this path constraint can be imposed while retaining strong visual reasoning performance. CVRR reaches 81.2% on V^*, 52.7% MMVP pair accuracy, and 55.2% on BLINK overall under the strict no-bypass interface, while compatible latent visual reasoners fail to recover comparable competence even when retrained under the same constraint. Our analyses show that this performance does not arise from simply routing a pretrained representation through a required bottleneck. Retraining without recurrent visual re-reading reduces V^* accuracy by 12.1 points and MMVP pair accuracy by 34.7 points despite retaining native h_1 and the learned transition, while removing the learned transition also degrades performance. Causal interventions show that persistent visual evidence reshapes the recurrent trajectory and that predictions remain tied to image-conditioned recurrent content when the question is controlled. Together, these results show that CVRR makes latent computation necessary for prediction while performing learned, prediction-relevant visual refinement along that required path. Our contributions are:

1. We recast latent-state reliance in visual reasoning as a causal path-design problem, distinguishing latent informativeness from path necessity.
2. We introduce CVRR, which preserves pretrained competence through image-conditioned initialization while making recurrent latent computation necessary for prediction.
3. We show through ablations and causal interventions that CVRR performs prediction-relevant visual refinement through learned updates and persistent visual re-reading.

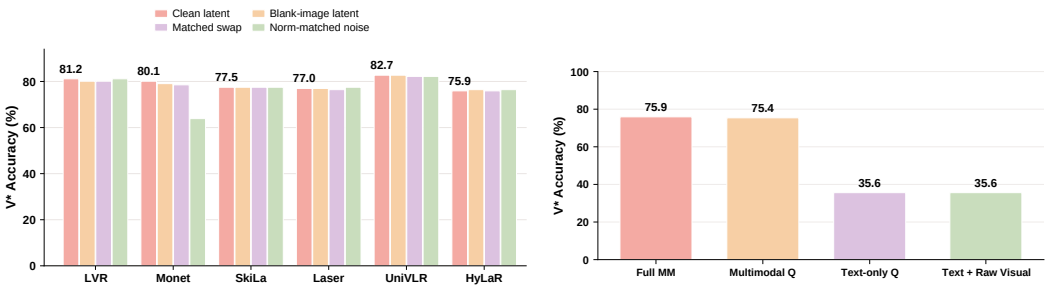
2 Preliminary Experiments

We first establish two observations that motivate CVRR: existing latent visual reasoners need not depend strongly on their proposed latent content, and pretrained visual competence is already integrated into the native multimodal question state.

2.1 Weak Behavioral Reliance on Latent Content

We first ask whether existing latent visual reasoners actually depend on their proposed latent states. For each method introduced in Appendix A.1, we replace the latent content with a blank-image latent, a matched latent from an example with a different answer, or row-norm-matched noise, while leaving the original multimodal answer context intact. We compare each intervention with the paired clean prediction from the unaltered latent state. Small changes in accuracy therefore indicate weak behavioral dependence on the proposed latent content.

As shown in Figure 1a, several methods remain close to their clean accuracy even after substantial changes to the latent state. For example, UniVLR changes by at most 0.5 percentage point across all three replacements. These interventions are not strict no-bypass evaluations because the original multimodal answer context remains available. Instead, they ask whether the standard inference path actually requires the proposed latent content. Stable predictions under replacement show that the answer can be sustained without strong dependence on that content. Thus, strong task performance does not by itself establish that prediction depends on the proposed latent computation.



(a) *Latent content is often nonessential* (b) *Multimodal question states preserve competence*

Figure 1: Preliminary diagnostics motivating CVRR. (a) Accuracy under blank-image, matched-state, and norm-matched-noise replacements relative to clean latent states. (b) Accuracy from the full multimodal state (*Full MM*), its question rows (*Multimodal Q*), text-only question rows (*Text-only Q*), and text-only rows with raw visual features (*Text+Raw Visual*). Panel (b) uses the frozen pretrained Qwen2.5-VL-7B-Instruct before CVRR adaptation.

2.2 Multimodal Question States Preserve Visual Competence

Removing alternative answer paths raises a complementary question: which representation should carry pretrained visual competence once those paths are unavailable? We compare four states formed at an intermediate layer of the frozen pretrained VLM: the full multimodal hidden state containing visual and question tokens (*Full MM*), the question rows from the same multimodal forward pass (*Multimodal Q*), the question rows from a text-only pass (*Text-only Q*), and the text-only question rows concatenated with raw visual features (*Text+Raw Visual*). Each state is passed alone to the same frozen upper decoder. As shown in Figure 1b, *Full MM* reaches 75.9% and *Multimodal Q* reaches 75.4%, whereas both text-only conditions achieve only 35.6%.

The 0.5 percentage point gap between *Full MM* and *Multimodal Q* indicates that nearly all of the visual competence available from the full multimodal state is already carried by the question representation at this layer. In contrast, attaching raw visual features to a text-only state does not recover this competence. Together with the weak latent reliance above, these findings motivate two requirements for latent visual reasoning: the latent path should begin from the native multimodal question representation, and alternative image-conditioned answer paths should be removed so that prediction actually depends on computation through that state.

3 Method

The preliminary results suggest that latent visual reasoning requires two properties simultaneously. The latent path should preserve the visual competence already formed by the pretrained VLM, and prediction should actually depend on computation through that path. CVRR is designed around these requirements. It begins recurrence from a native image-conditioned question representation, repeatedly refines this state while re-reading persistent visual evidence, and removes all alternative image-conditioned routes before answer decoding. The resulting recurrent state therefore carries pretrained visual competence while becoming the required visual path to prediction. Figure 2 summarizes the architecture, and Appendix C provides the complete forward procedure.

3.1 Causal Visual-Read Boundary

The recurrent path should not begin from an arbitrary hidden state. It should start after substantial visual information has entered the question representation, while visual states still retain downstream influence that recurrence can re-read. We locate this transition using layer-wise activation patching on the frozen base VLM (Geiger et al., 2021; Meng et al., 2022). We construct contrastive pairs $x = (I, q, y)$ and $x' = (I', q, y')$ that share the same question but contain different images and yield different answers. At each candidate layer, we replace the visual-token activations of one example

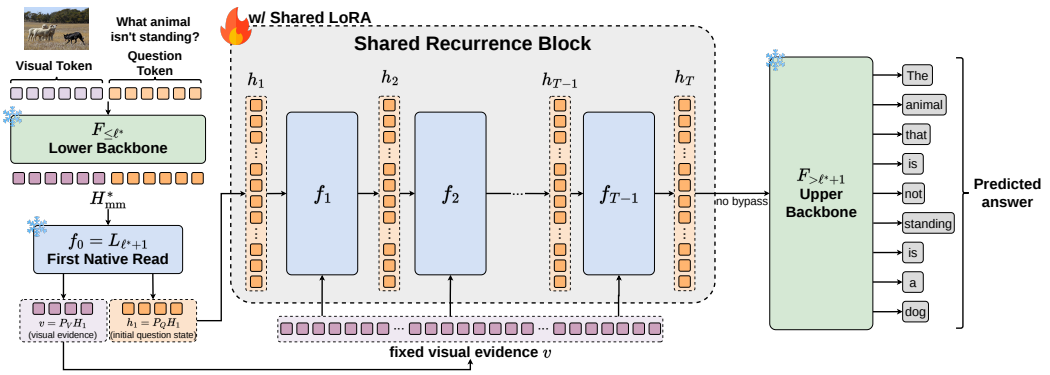


Figure 2: Overview of CVRR. The native multimodal forward forms the image-conditioned question state h_1 and persistent visual evidence v at the causal visual-read boundary. A shared adapted layer repeatedly updates h_t while re-reading the same fixed v. After T steps, visual states and multimodal caches are discarded, and only h_T enters the frozen upper backbone for answer prediction.

with those from its paired example and measure how much this reduces the preference for its original answer. Let M_x denote the log-probability margin favoring the answer of x over that of x', with $M_{x'}$ defined in the opposite direction, and let $P_\ell(x \leftarrow x')$ denote patching the visual activations of x with those of x' at layer ℓ. We measure the bidirectional effect as

$$m_i(\ell) = \frac{1}{2}\Big[M_x(x) - M_x(P_\ell(x \leftarrow x')) + M_{x'}(x') - M_{x'}(P_\ell(x' \leftarrow x))\Big], \qquad \widehat{m}_i(\ell) = \frac{m_i(\ell)}{m_i(0)}, \tag{1}$$

where larger normalized effects indicate that visual states at that layer still exert substantial downstream influence. We aggregate these profiles across contrastive pairs and use the layer preceding the first sustained decline as the candidate recurrent boundary.

For Qwen2.5-VL-7B (Bai et al., 2025b), this procedure yields $\ell^* = 20$, and the subsequent layer $f_0 = L_{21}$ initializes recurrence. At this boundary, the frozen lower backbone produces

$$H^*_{\text{mm}} = F^{\text{mm}}_{\le \ell^*}(I, q), \qquad Q^*_{\text{txt}} = F^{\text{txt}}_{\le \ell^*}(q). \tag{2}$$

The multimodal state supplies the native image-conditioned representation from which recurrence begins, whereas the text-only state is reserved for image-free decoder context. Their question rows preserve the same token identity, ordering, and masking structure. The multimodal forward is terminated after f_0, and no later multimodal hidden state or prefix KV cache is retained for answer decoding. Appendix D further shows that this region jointly provides strong question-state sufficiency and residual visual influence.

3.2 Persistent Visual Recurrence

A competent image-conditioned initialization is necessary for preserving pretrained visual ability, but initialization alone does not provide further visual computation. CVRR therefore keeps visual evidence available throughout recurrence while allowing the question state to evolve. Reusing shared Transformer parameters across depth follows prior work on universal, looped, and recurrent-depth Transformers (Dehghani et al., 2019; Kohli et al., 2026; Fan et al., 2026). Unlike these approaches, CVRR uses recurrence to repeatedly re-read persistent visual evidence from a native multimodal state while making the resulting recurrent state the required image-conditioned path to prediction. Given the multimodal boundary state, we first apply the subsequent decoder layer $f_0 = L_{\ell^*+1}$ once in its pretrained form, without the low-rank adapter used in later recurrent steps:

$$H_1 = f_0(H^*_{\text{mm}}), \qquad h_1 = P_Q H_1, \qquad v = P_V H_1. \tag{3}$$

Here, P_Q and P_V select the question-token and visual-token rows. The question rows h_1 form the initial recurrent state, while the visual rows v provide persistent evidence throughout recurrence.

Because both originate from the same native multimodal forward pass, recurrence begins from the visual computation already established by the pretrained model rather than attempting to reconstruct it from raw visual features.

At each subsequent step, we reconstruct the native multimodal scaffold by replacing only the question rows with the previous recurrent state while keeping the visual rows fixed. Each transition therefore receives an evolving question state together with the same visual evidence:

$$\begin{aligned} H_t &= \mathrm{Replace}_Q(H_1, h_{t-1}), \qquad P_V H_t = v, \\ \widetilde{h}_t &= P_Q f_\phi(H_t), \\ h_t &= (1-\beta)h_{t-1} + \beta\widetilde{h}_t, \qquad t = 2, \ldots, T. \end{aligned} \tag{4}$$

Visual re-reading occurs through the native causal self-attention of the shared decoder layer without introducing a separate cross-attention module. The original token ordering, causal mask, and positional structure are preserved, and all recurrent steps share the same layer and low-rank parameters. Thus, h_t is repeatedly revised against persistent visual evidence rather than transformed in isolation.

3.3 Strict Causal Decoder Interface and Training Objective

Even a visually grounded recurrent state need not be behaviorally necessary if the decoder can still access the original multimodal context. CVRR therefore removes every alternative image-conditioned route before answer decoding. Below the recurrent boundary, autoregressive answer tokens use only image-free prefix caches $\mathcal{K}_{\mathrm{txt}}$ constructed from the text-only question path. Above the boundary, the final recurrent state h_T serves as the question-prefix representation for the frozen upper decoder. All visual rows and original multimodal KV caches are discarded before decoding. The text-only caches provide the lower-layer linguistic context required for autoregressive processing but contain no image-conditioned information. Consequently, h_T is the sole source of image-conditioned information available to the answer decoder.

The backbone and upper decoder remain frozen, and only the shared low-rank parameters ϕ are optimized. We use no latent targets, teacher trajectories, or reconstruction objectives to encourage reliance on the recurrent state. Instead, path reliance follows from the decoder interface itself. For a minibatch $\mathcal{B}$ and valid answer-token set $\mathcal{A}_i$, we minimize answer-token cross-entropy,

$$\mathcal{L}_{\mathrm{ans}} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{A}_i|}\sum_{j\in\mathcal{A}_i}\log p_\phi\left(y_{i,j} \mid y_{i,<j}, h_{T,i}, \mathcal{K}_{\mathrm{txt},i}\right). \tag{5}$$

Prompt and padding positions are excluded from the loss, and token losses are averaged within each example and then across the minibatch. The same recurrent trajectory and strict decoder interface are used at inference time.

4 Experiments

4.1 Datasets & Metrics

We train on Visual CoT (Shao et al., 2024), containing 438K QA pairs with answer-relevant bounding-boxes. We evaluate on four VQA benchmarks: V^* (Wu & Xie, 2024), MMVP (Tong et al., 2024), BLINK (Fu et al., 2024), and MME-RealWorld-Lite[2] (Zhang et al., 2025b). V^* contains 191 high-resolution examples, while MMVP contains 300 questions over 150 visually similar image pairs. From BLINK, we report the Counting, IQ-Test, Jigsaw, Relative Reflectance, and Spatial Relation subsets following LVR (Li et al., 2026a), together with overall accuracy across the benchmark. We report accuracy for all benchmarks: overall accuracy and the Direct Attributes (D.A.) and Relative Position (R.P.) splits for V^*; item and pair accuracy for MMVP, where both questions in a pair must be correct; per-subset and overall accuracy for BLINK; and overall accuracy for MME-RealWorld-Lite.

[2] https://huggingface.co/datasets/yifanzhang114/MME-RealWorld-Lite

Table 1: Results across V^*, MMVP, BLINK, and MME-RealWorld-Lite. **Bold** and underline indicate the best and second-best within each model group. Colored arrows show point changes from the corresponding pretrained backbone. Indented *w/ SFT* rows denote full-multimodal controls trained on the same data without recurrent reasoning. The shaded row denotes CVRR.

Method	V^*			MMVP		BLINK						MME-RW-Lite
	Overall	D.A.	R.P.	Item	Pair	Overall	Counting	IQ-Test	Jigsaw	Rel. Reflect	Spatial Rel.	Overall
General-Purpose VLMs w/ CoT												
GPT-5	73.3	70.4	77.6	**87.7**	**77.3**	**74.0**	**78.3**	**35.3**	**78.0**	**68.7**	89.5	**57.8**
Qwen3-VL-8B-Thinking	78.5	81.7	73.7	52.3	20.7	52.8	65.8	18.7	44.7	57.5	55.9	46.7
Qwen2.5-VL-7B-Instruct	79.6	79.1	**80.3**	70.3	47.3	55.0	74.2	30.0	67.3	41.0	84.6	50.8
w/ SFT	80.6 ↑1.0	80.9 ↑1.8	**80.3**	74.7 ↑4.4	52.0 ↑4.7	53.0 ↓2.0	70.8 ↓3.4	24.0 ↓6.0	59.3 ↓8.0	34.3 ↓6.7	**90.9** ↑6.3	44.7 ↓6.1
Visual Reasoning												
VL-Rethinker	81.7 ↑2.1	80.9 ↑1.8	82.9 ↑2.6	**74.3** ↑4.0	**52.0** ↑4.7	**54.5** ↓0.5	68.3 ↓5.9	20.7 ↓9.3	**66.0** ↓1.3	41.0	**90.2** ↑5.6	50.9 ↑0.1
DeepEyes	81.2 ↑1.6	**81.7** ↑2.6	80.3	69.7 ↓0.6	43.3 ↓4.0	52.7 ↓2.3	**70.8** ↓3.4	**26.7** ↓3.3	54.7 ↓12.6	37.3 ↓3.7	81.8 ↓2.8	**54.1** ↑3.3
PixelReasoner	**82.7** ↑3.1	**81.7** ↑2.6	**84.2** ↑3.9	73.3 ↑3.0	48.0 ↑0.7	51.1 ↓3.9	67.5 ↓6.7	**26.7** ↓3.3	64.0 ↓3.3	**44.8** ↑3.8	83.2 ↓1.4	50.6 ↓0.2
Latent Visual Reasoning												
LVR-7B	81.2 ↑1.6	83.5 ↑4.4	**77.6** ↓2.7	71.0 ↑0.7	44.7 ↓2.6	51.8 ↓3.2	**70.8** ↓3.4	**28.0** ↓2.0	58.0 ↓9.3	38.8 ↓2.2	89.5 ↑4.9	51.3 ↑0.5
Monet-7B	79.1 ↓0.5	80.9 ↑1.8	76.3 ↓4.0	70.7 ↑0.4	46.0 ↓1.3	46.3 ↓8.7	60.8 ↓13.4	17.3 ↓12.7	50.7 ↓16.6	41.8 ↑0.8	79.0 ↓5.6	50.1 ↓0.7
SkiLa-7B	77.5 ↓2.1	79.1	75.0 ↓5.3	71.7 ↑1.4	47.3	52.8 ↓2.2	67.5 ↓6.7	18.7 ↓11.3	64.0 ↓3.3	38.8 ↓2.2	89.5 ↑4.9	44.7 ↓6.1
Laser-7B	77.5 ↓2.1	82.6 ↑3.5	69.7 ↓10.6	64.7 ↓5.6	32.7 ↓14.6	53.4 ↓1.6	63.3 ↓10.9	24.7 ↓5.3	63.3 ↓4.0	44.8 ↑3.8	83.2 ↓1.4	50.8
UniVLR-7B	**82.7** ↑3.1	**87.8** ↑8.7	75.0 ↓5.3	70.0 ↓0.3	42.7 ↓4.6	50.2 ↓4.8	66.7 ↓7.5	22.0 ↓8.0	42.0 ↓25.3	42.5 ↑1.5	84.6	**51.8** ↑1.0
HyLaR-7B	75.9 ↓3.7	77.4 ↓1.7	73.7 ↓6.6	74.3 ↑4.0	51.3 ↑4.0	53.4 ↓1.6	68.3 ↓5.9	20.0 ↓10.0	64.7 ↓2.6	43.3 ↑2.3	**90.2** ↑5.6	44.9 ↓5.9
PEARL-7B	80.6 ↑1.0	82.6 ↑3.5	**77.6** ↓2.7	**75.7** ↑5.4	**54.7** ↑7.4	**56.7** ↑1.7	68.3 ↓5.9	24.7 ↓5.3	**67.3**	**50.0** ↑9.0	88.8 ↑4.2	47.6 ↓3.2
Latent Visual Reasoning w/o Bypass												
LVR-7B	35.6 ↓44.0	24.4 ↓54.7	52.6 ↓27.7	50.0 ↓20.3	0.7 ↓46.6	37.9 ↓17.1	39.2 ↓35.0	20.0 ↓10.0	52.7 ↓14.6	26.9 ↓14.1	53.9 ↓30.7	26.2 ↓24.6
Monet-7B	36.1 ↓43.5	25.2 ↓53.9	52.6 ↓27.7	50.7 ↓19.6	2.0 ↓45.3	38.4 ↓16.6	45.0 ↓29.2	24.7 ↓5.3	52.7 ↓14.6	26.9 ↓14.1	55.9 ↓28.7	33.1 ↓17.7
SkiLa-7B	38.7 ↓40.9	28.7 ↓50.4	54.0 ↓26.3	29.3 ↓41.0	0.0 ↓47.3	38.8 ↓16.2	40.0 ↓34.2	24.0 ↓6.0	47.3 ↓20.0	40.3 ↓0.7	51.1 ↓33.5	15.6 ↓35.2
Laser-7B	37.2 ↓42.4	27.0 ↓52.1	52.6 ↓27.7	50.0 ↓20.3	0.7 ↓46.6	37.8 ↓17.2	40.8 ↓33.4	19.3 ↓10.7	47.3 ↓20.0	26.9 ↓14.1	57.3 ↓27.3	31.6 ↓19.2
UniVLR-7B	37.7 ↓41.9	27.8 ↓51.3	52.6 ↓27.7	49.3 ↓21.0	0.0 ↓47.3	38.4 ↓16.6	40.0 ↓34.2	20.0 ↓10.0	52.7 ↓14.6	26.9 ↓14.1	58.7 ↓25.9	27.3 ↓23.5
HyLaR-7B	39.8 ↓39.8	26.1 ↓53.0	60.5 ↓19.8	42.3 ↓28.0	2.7 ↓44.6	33.7 ↓21.3	36.7 ↓37.5	6.7 ↓23.3	53.3 ↓14.0	32.8 ↓8.2	37.8 ↓46.8	32.6 ↓18.2
CVRR-7B	**81.2** ↑1.6	**82.6** ↑3.5	**79.0** ↓1.3	**74.7** ↑4.4	**52.7** ↑5.4	**55.2** ↑0.2	**70.0** ↓4.2	**28.0** ↓2.0	**62.7** ↓4.6	**42.5** ↑1.5	**90.9** ↑6.3	**46.7** ↓4.1

4.2 Baselines

Our comparisons are designed to separate visual competence from causal reliance on the latent path. As general-purpose reference models, we evaluate GPT-5 (Singh et al., 2025), Qwen2.5-VL-7B-Instruct, and Qwen3-VL-8B-Thinking (Bai et al., 2025a) with chain-of-thought (CoT) reasoning. Visual reasoning models include VL-Rethinker (Wang et al., 2025), DeepEyes (Zheng et al., 2026), and PixelReasoner (Su et al., 2025), while latent visual reasoning methods include LVR, Monet (Wang et al., 2026a), SkiLa (Tong et al., 2025), Laser (Wang et al., 2026b), UniVLR (Jiang et al., 2026), HyLaR (Cheng et al., 2026), and PEARL (Adhikari & Lapata, 2026). Qwen2.5-VL-7B-Instruct serves as the pretrained backbone reference for CVRR. To distinguish recurrent computation from ordinary supervised adaptation, we additionally train a full-multimodal SFT control on the same Visual CoT data while retaining the standard multimodal answer path.

The most direct comparison tests whether removing bypasses alone is sufficient. We retrain compatible latent reasoning baselines under the same strict decoder interface while preserving their original latent mechanisms and training objectives. These no-bypass variants use the same Visual CoT data and common SFT settings as CVRR, but visual states and multimodal prefix caches are removed before answer decoding. This comparison isolates the central design question: whether a method can make its latent path necessary without losing the visual competence of the pretrained model. Implementation and evaluation details are provided in Appendix B.

4.3 Main Results

The central comparison in Table 1 is not absolute benchmark rank, but whether visual competence survives once the answer is required to depend on the latent path. CVRR retains this competence under the strict no-bypass interface, reaching 81.2% on V^*, 52.7% MMVP pair accuracy, and 55.2% on BLINK overall. In contrast, latent reasoning baselines retrained under the same interface reach at most 39.8% on V^*, 2.7% MMVP pair accuracy, 38.8% on BLINK, and 33.1% on MME-RealWorld-Lite. The contrast shows that simply forcing prediction through a latent state is not enough. The path must also begin from a representation that preserves the visual computation established by pretraining, which CVRR obtains through native multimodal initialization.

The full-multimodal SFT control provides a complementary check that these results are not merely an effect of training on Visual CoT. CVRR matches or exceeds this control on V^*, MMVP, and BLINK despite removing the standard multimodal answer path. On MME-RealWorld-Lite, CVRR drops 4.1 points from its pretrained backbone, compared with a 6.1 point drop for the SFT control. Thus, the strict recurrent interface does not introduce a larger degradation than ordinary supervised adaptation on this benchmark. Taken together, the results support the intended claim: latent computation can be made behaviorally necessary while retaining much of the visual competence that makes it useful. Appendix E further evaluates CVRR across additional backbones and larger models.

5 Analysis

The results show that CVRR preserves visual competence while making the recurrent state the required image-conditioned path to prediction. Path necessity alone, however, does not reveal what computation occurs along that path. We therefore examine whether learned recurrence improves on native initialization, whether persistent visual evidence causally shapes the trajectory, whether sensitivity to h_T reflects image-conditioned content, and how visual influence changes across steps.

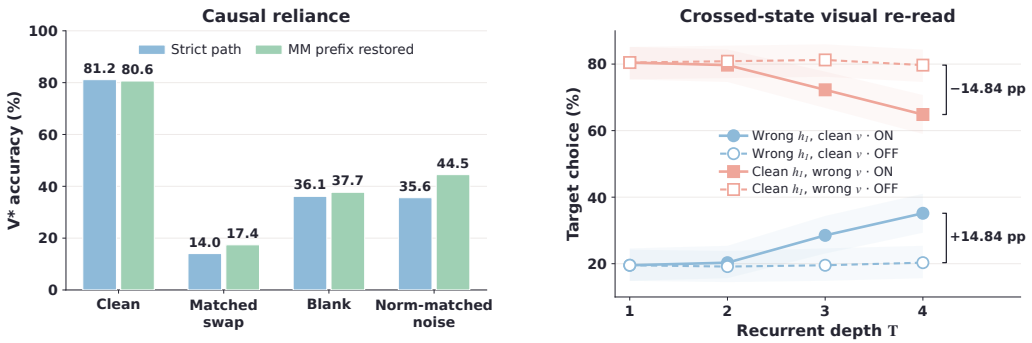
5.1 What Does Recurrence Add Beyond h_1?

Table 2: Component analysis on V^* ($n = 191$) and MMVP ($n = 300$). Δ denotes the percentage-point change from Full CVRR. Variants without native initialization replace h_1 with the text-only anchor B. Component ablations use $T = 4$ unless otherwise indicated.

	Components			V^*				MMVP			
Variant	Native h_1	Visual Re-read v	Transition f_ϕ	Overall	D.A.	R.P.	Δ	Item	Pair	Δ Item	Δ Pair
Full CVRR ($T = 4$)	✓	✓	✓	**81.2**	**82.6**	**79.0**	–	**74.7**	**52.7**	–	–
Recurrent depth controls											
$T = 2$	✓	✓	✓	78.5	80.0	76.3	-2.7	74.0	50.7	-0.7	-2.0
$T = 1$	✓	✗	✗	75.4	79.1	69.7	-5.8	60.7	22.7	-14.0	-30.0
Component ablations											
w/o Visual Re-read	✓	✗	✓	69.1	71.3	65.8	-12.1	56.7	18.0	-18.0	-34.7
w/o Native MM Init ($h_1 \to B$)	✗	✓	✓	37.2	27.0	52.6	-44.0	50.7	2.0	-24.0	-50.7
w/o Learned Transition (f_ϕ)	✓	✓	✗	76.4	80.0	71.1	-4.8	65.7	32.7	-9.0	-20.0
w/o Entire Visual Path	✗	✗	✓	37.2	27.0	52.6	-44.0	50.0	0.0	-24.7	-52.7

Table 2 separates the roles of native multimodal initialization and recurrent refinement. Here, B denotes the text-only question state obtained by passing the same question without the image through the frozen lower backbone and native layer f_0, aligning it with h_1 in depth and token structure while removing image-conditioned content. Replacing h_1 with B reduces V^* accuracy from 81.2% to 37.2% and MMVP pair accuracy from 52.7% to 2.0% despite retaining visual evidence v and a retrained f_ϕ. On V^*, this matches removing the entire visual path and falls within the 35–40% range of the no-bypass baselines in Table 1. Thus, persistent visual access does not reconstruct the competence already integrated into the native multimodal question state. The role of h_1 is therefore to carry the pretrained model's native image-conditioned representation into the required path.

Given this starting state, we isolate what recurrent computation contributes. All trainable component ablations are independently retrained under the same recipe as Full CVRR. Decoding h_1 at $T = 1$ reaches 75.4% on V^* but only 22.7% MMVP pair accuracy, showing that native initialization alone is insufficient. Removing visual re-reading while retaining native h_1, the learned transition, and $T = 4$ reduces V^* accuracy by 12.1 points and MMVP pair accuracy by 34.7 points. Thus, CVRR does not merely pass a competent h_1 through trainable latent transformations; persistent visual evidence provides prediction-relevant information during recurrence. Conversely, removing the learned transition reduces performance by 4.8 and 20.0 points, showing that visual access alone is also insufficient. The trained $T = 2$ control reaches 78.5% on V^* and 50.7% MMVP pair accuracy, while full CVRR reaches 81.2% and 52.7%. Together, these results support CVRR's central design: pretrained visual competence enters through h_1 and is actively refined by learned computation that repeatedly re-reads visual evidence.



(a) (b)

Figure 3: Causal tests of recurrent computation. (a) Corrupting h_T degrades prediction under the strict path, while restoring the multimodal prefix partially compensates for the corrupted state. (b) Clean v corrects an incorrect h_1, while incorrect v degrades a clean h_1 when question-to-visual KV connections are enabled (ON); blocking these connections (OFF) largely removes these effects.

5.2 Causal Visual Re-reading

We first verify the claim enforced by the architecture: prediction under the strict interface should depend on recurrent content. Figure 3a shows that accuracy falls from 81.2% to 14.0% under matched recurrent-state replacement and to 36.1% and 35.6% under blank and norm-matched noise interventions. This sensitivity verifies path necessity, but it does not by itself establish visually grounded recurrent computation. Restoring the original multimodal prefix raises matched-swap and noise accuracy to 17.4% and 44.5%. An alternative image-conditioned route can therefore compensate for corrupted recurrent content, directly illustrating why an informative latent state can become behaviorally nonessential when another visual path remains available.

We then test whether recurrence actually uses the persistent visual evidence. We construct crossed states that pair the question state from one image with visual evidence from its matched alternative. If v causally guides the recurrent update, clean evidence should pull an incorrect h_1 toward the correct prediction, while incorrect evidence should pull a correct h_1 away from it. Figure 3b shows this bidirectional pattern when question-to-visual KV connections remain available, whereas blocking those connections nearly removes both effects, yielding difference-in-differences effects of $+14.8$ and -14.8 points. Thus, the recurrent trajectory is not merely a repeated transformation of h_1. It is causally revised by the visual evidence that remains available at each step.

5.3 Image-Conditioned Content of the Recurrent State

A further alternative explanation remains. Replacing h_T across examples changes both image-conditioned content and the question representation, so the degradation could reflect question disruption rather than visual content. We therefore use MMVP pairs in which two images I and I' share the same question q. For a target (I, q), we replace $h_T(I, q)$ with $h_T(I', q)$ at decoding time. Because the question and token alignment remain fixed, this intervention isolates changes in image-conditioned content.

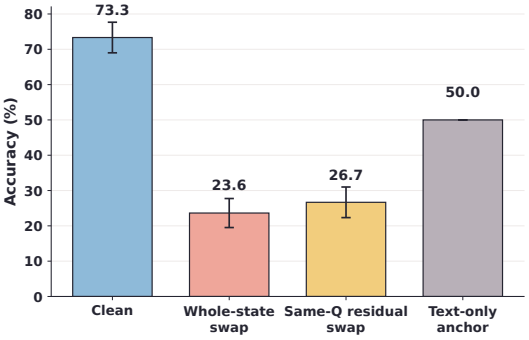


Figure 4: Same-question MMVP intervention with a fixed question and text-only anchor.

As shown in Figure 4, clean accuracy is 73.3%, while whole-state and same-question replacement reduce it to 23.6% and 26.7%. Similar sensitivity after fixing the question rules out question disruption as the primary explanation. The text-only anchor reaches 50.0%, above the 26.7% obtained with content from the competing image, suggesting that incompatible image-

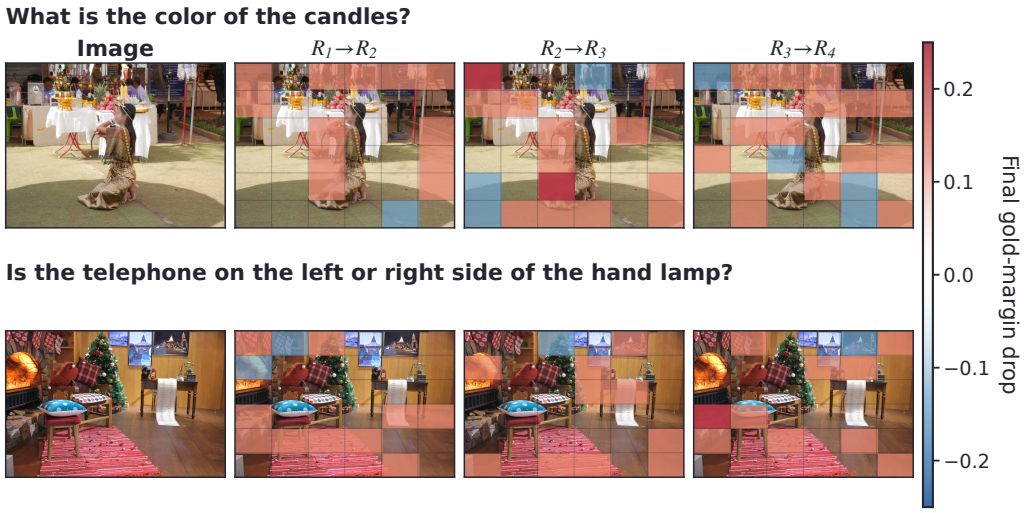


Figure 5: Step-varying causal visual influence across recurrence. Blocking local question-to-visual connections reveals image regions that affect the final gold-answer margin at each step.

conditioned content is more disruptive than removing such content alone. Together, these controls show that prediction follows the image-conditioned content carried by h_T rather than merely requiring an intact hidden state.

5.4 Step-Varying Visual Influence

The preceding interventions establish that recurrent updates use visual evidence, but not whether successive steps rely on the same evidence or shift which regions matter to prediction. We therefore block question-to-visual connections within each 6×6 image region and measure the resulting change in the final gold-answer margin. Figure 5 shows that causally influential regions shift across recurrent steps and, in the visualized examples, later transitions qualitatively emphasize objects implicated by the query. This variation also appears across the full V^* set, where consecutive causal maps have mean cosine similarities of 0.37 and 0.35. Recurrence therefore does not treat v as a static packet of visual information whose contribution remains fixed. Instead, as the question state evolves, different parts of the persistent evidence become causally relevant to the prediction. Full population statistics are reported in Appendix I. Taken together, the analyses support a coherent computational picture: h_1 carries pretrained visual competence into the required path, learned recurrence improves on that starting state, persistent visual evidence causally redirects the trajectory, and the visual evidence relevant to prediction changes across recurrent steps.

6 Conclusion & Limitations

We address a central ambiguity in latent visual reasoning: a latent state can contain task-relevant visual information without being necessary for prediction. We introduce CVRR to align these properties by preserving pretrained visual competence while requiring image-conditioned prediction to pass through recurrent latent computation. CVRR starts from the native multimodal question state, refines it against persistent visual evidence, and removes alternative multimodal answer routes before decoding. Under this strict interface, CVRR retains strong visual reasoning performance, while causal analyses show that predictions follow the image-conditioned recurrent state and that persistent visual evidence revises its trajectory. These results support latent reasoning in which informativeness alone is insufficient and the proposed computation should also be a path that prediction actually uses. The visual-read boundary remains backbone-specific, our mechanistic analyses mainly use multiple-choice settings, and the current design relies on persistent rather than adaptively acquired visual evidence. Extending causal path necessity to more adaptive and open-ended multimodal reasoning remains an important direction.

Reproducibility statement

To support reproducibility, we provide the implementation code and execution instructions through the anonymous repository and document the key experimental settings in the appendix. These include the training and inference procedures for CVRR, recurrent reasoning settings, causal interventions, boundary localization, and the evaluation of the main baselines. Model and data configurations, training hyperparameters, inference settings, evaluation protocols, hardware, and other experimental details are documented in the appendix. We also distinguish the settings used for the main benchmarks from those used in the analysis experiments and describe the implementation conditions for the compared methods, including the bypass-free baselines. We verified the implementation and checked that the reported results are consistent with the corresponding experimental settings.

AI Use Statement

Generative AI tools were used solely to assist with language editing, including wording and grammar, and with code debugging during implementation. All AI-assisted suggestions were independently reviewed and verified by the authors. Generative AI tools were not used to formulate research ideas, design the methodology or experiments, interpret the results, or develop scientific claims. These aspects of the work were developed and verified independently by the authors. The authors take full responsibility for the content of the paper.

References

Ashutosh Adhikari and Mirella Lapata. Multimodal latent reasoning via predictive embeddings. *arXiv preprint arXiv:2604.08065*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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*, 2025b.

Tao Cheng, Shi-Zhe Chen, Hao Zhang, Yixin Qin, Jinwen Luo, and Zheng Wei. Hylar: Hybrid latent reasoning with decoupled policy optimization. *arXiv preprint arXiv:2604.20328*, 2026.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyzdRiR9Y7`.

Ying Fan, Anej Svete, and Kangwook Lee. Bridging the gap between latent and explicit reasoning with looped transformers. *arXiv preprint arXiv:2606.31779*, 2026.

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 *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

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. *arXiv preprint arXiv:2605.11856*, 2026.

Harsh Kohli, srinivasan parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=8fz7WRThKL.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=j84WR5ORsC.

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. *arXiv preprint arXiv:2602.22766*, 2026b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ff38d72a2e0aa6dbe42de83a17b2223-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VeZkY3JjWV.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

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. *arXiv preprint arXiv:2512.16584*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

André G Viveiros, Nuno Gonçalves, André FT Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=4oYxzssbVg.

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*, pp. 12030–12040, 2026a.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. Forest before trees: Latent superposition for efficient visual reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11272–11288, 2026b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

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*, pp. 33510–33520, 2026.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, et al. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*, 2025a.

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

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xUyMXkI958.

Dongyao Zhu, Zhen Wang, Xi Xiao, Han Jiang, Saeed Vahidian, Wei-Lun Chao, Tanya Berger-Wolf, Yu Su, Raju Vatsavai, and Jianyang Gu. Leveraging latent visual reasoning in silence. *arXiv preprint arXiv:2605.18641*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Related Works

A.1 Visual and Latent Visual Reasoning

Recent visual reasoning methods improve inference by explicitly acquiring, revisiting, or manipulating visual evidence. VL-Rethinker encourages multimodal self-reflection through reinforcement learning (Wang et al., 2025), DeepEyes repeatedly searches for image regions relevant to the current reasoning step (Zheng et al., 2026), and PixelReasoner incorporates pixel-space operations into the reasoning trajectory (Su et al., 2025). In parallel, latent visual reasoning moves intermediate computation from explicit text into continuous hidden representations. LVR iteratively develops latent visual states (Li et al., 2026a), Monet uses continuous visual embeddings as intermediate reasoning representations (Wang et al., 2026a), and SkiLa combines textual and latent computation (Tong et al., 2025). Laser explores richer latent representations (Wang et al., 2026b), UniVLR integrates textual and visual evidence in a shared latent space (Jiang et al., 2026), HyLaR develops a hybrid latent reasoning framework (Cheng et al., 2026), and PEARL learns predictive latent representations from visual reasoning trajectories (Adhikari & Lapata, 2026). Despite their architectural differences, these methods primarily study how latent states are constructed, updated, or coupled with visual evidence. Importantly, introducing a latent reasoning state does not by itself make that state necessary for prediction. When answer decoding retains access to the original image-conditioned hidden states, visual tokens, or multimodal prefix KV caches, image information can reach the answer through both the proposed latent state and the original multimodal pathway. The latter forms a bypass around the latent computation. Such a bypass need not harm predictive performance, but it makes strong benchmark accuracy insufficient to establish that the model actually reasons through the proposed latent state. CVRR instead focuses on controlling this computational interface while preserving the visual competence of the pretrained model.

A.2 Causal Utilization in Multimodal and Latent Reasoning

A related line of work distinguishes information encoded in an internal representation from information that actually influences model behavior. Causal mediation analyses show that changes in visual or intermediate representations need not propagate to the final prediction (Li et al., 2026b), while latent replacement and removal experiments find that predictions can remain stable even when intermediate states are substantially altered (Viveiros et al., 2026; Zhu et al., 2026). Mechanistic studies further separate the contribution of latent content from interface, boundary, or formatting effects (Guo et al., 2026). Related concerns arise in multimodal question answering, where strong predictions can partly reflect linguistic priors, answer-choice regularities, or other signals that reduce dependence on the intended visual evidence (Goyal et al., 2017; Leng et al., 2024). In latent visual reasoning, an analogous problem arises when the decoder can still attend to image-conditioned representations that were created before or alongside the latent trajectory. In this case, the latent state may remain highly informative or even improve prediction while being behaviorally nonessential because the original multimodal context provides a parallel route to the answer. Causal interventions are therefore useful for distinguishing latent informativeness from latent utilization, but post-hoc diagnosis alone does not guarantee that the intended path is used. CVRR turns this distinction into an architectural constraint by retaining the native image-conditioned initialization, allowing recurrent states to re-read persistent visual evidence, and removing alternative image-conditioned routes before decoding so that the final recurrent state becomes the required visual path to prediction.

B Implementation Details

B.1 CVRR Training and Recurrent Depth

Unless otherwise specified, we use Qwen2.5-VL-7B-Instruct as the base multimodal backbone. We localize the causal boundary using training examples only, without access to any evaluation benchmark examples. This selects $\ell^* = 20$, and we reuse layer L_{21} as the shared recurrent transition with $T = 4$ fixed as the default recurrent depth before downstream benchmark evaluation. The recurrent state is initialized from the native image-conditioned question representation h_1, while the visual state v is retained as fixed evidence throughout recurrence. For text-only controls, we define B by passing the same question without the image through the frozen lower backbone and the na-

tive recurrent layer f_0. This places B at the same depth and token alignment as h_1 while removing image-conditioned content. Before decoding, visual rows and the original multimodal KV cache are removed. The answer decoder receives h_T together with image-free text-only prefix caches, making h_T the sole source of image-conditioned information. We cap the visual input at 8,192 visual tokens. The pretrained backbone remains frozen, and we adapt only the recurrent transition using LoRA with rank 32, scaling factor $\alpha = 16$, and dropout 0.01. This results in $2{,}883{,}584$ trainable parameters out of approximately 8.295B total parameters, or 0.0348%.

CVRR training minimizes answer-token cross entropy only. We use $\beta = 0.50$ during both training and inference. We optimize with AdamW using a learning rate of 2×10^{-5}, a cosine learning-rate schedule, a warmup ratio of 0.05, and gradient clipping at 1.0. Models are trained for three epochs with a per-device batch size of 32, gradient accumulation of 1, bfloat16 precision, and random seed 0 on four NVIDIA B200 180GB GPUs, giving an effective global batch size of 128. We select the checkpoint with the lowest validation loss for evaluation. For recurrent-depth experiments, we independently train models for $T \in \{2, 3, 4, 6, 8\}$ under the same settings. The $T = 1$ condition has no recurrent transition or trainable recurrent LoRA and directly decodes h_1 through the strict interface. All trainable component ablations in Table 2 are independently retrained under the same recipe as Full CVRR, changing only the indicated component.

B.2 No-Bypass Baselines

For the no-bypass comparison, we use the six latent reasoning baselines for which the strict answer interface can be imposed while preserving the method's original latent pathway. For each compatible baseline, we retain its architecture, latent mechanism, and method-specific training objectives, including auxiliary losses when applicable. We standardize the Visual CoT data and common SFT settings, modifying only the answer interface to remove visual states and multimodal prefix KV caches before decoding. PEARL is excluded from this comparison because imposing the same strict decoder interface would require altering its predictive latent pathway rather than only removing an alternative multimodal answer route. Thus, the comparison tests whether existing latent mechanisms can retain visual competence when alternative image-conditioned routes are removed without redefining those mechanisms to fit CVRR's interface.

B.3 Evaluation Protocol

The main benchmark results use generation-based evaluation with greedy decoding for CVRR and all baselines. We otherwise retain each implementation's default generation configuration, including its maximum-new-token setting. Preliminary latent-replacement experiments and, unless otherwise specified, causal and mechanistic analyses use restricted first-option-token readout. Each trained model is evaluated from a single training run with random seed 0. Reported 95% confidence intervals are computed by nonparametric bootstrap over evaluation examples and quantify evaluation-sample uncertainty rather than decoding or training-run variability. For MMVP, bootstrap resampling is clustered by image pair to preserve within-pair dependence.

C Algorithm

Algorithm 1 summarizes the training and inference procedure in three stages: native multimodal initialization, persistent visual re-reading, and strict answer decoding. The frozen lower backbone first processes the image–question pair and the question alone. With the recurrent adapter disabled, the native layer F_0 produces the multimodal state H_1, whose question and visual rows initialize h_1 and v, respectively. The text-only path produces the depth-aligned anchor B and image-free prefix caches $\mathcal{K}_{\text{txt}}$. The shared adapter ϕ is then enabled for $T - 1$ recurrent updates. At each step, the previous question state replaces the question rows of H_1, while the visual evidence, token ordering, mask, and positions remain fixed. The adapted layer F_ϕ computes a candidate question state, which is mixed with the previous state using β. Thus, recurrence updates the question representation against persistent visual evidence without carrying updated visual states between steps.

Before answer decoding, all visual states and multimodal/recurrent caches are discarded. The decoder receives only h_T and the image-free text-only caches, making the final recurrent state its sole image-conditioned input. During training, the negative log-likelihood is averaged over valid answer

tokens within each example and then across the minibatch. Gradients update only the shared LoRA parameters ϕ, while the pretrained backbone remains frozen. At inference, the same initialization, recurrence, and strict decoding stages are executed without parameter updates.

Algorithm 1 Training and inference with CVRR

Require: Frozen VLM $M = \{L_1, \ldots, L_N\}$ and causal boundary $\ell^\star$
Require: Image–question pair (I, q), recurrent depth T, mixing coefficient β
Require: Shared LoRA parameters ϕ on recurrent layer $F = L_{\ell^\star+1}$
Ensure: Answer distribution $p_\phi(y \mid I, q)$

Native multimodal initialization
$H^\star_{\mathrm{mm}} \leftarrow M^{\mathrm{mm}}_{\le \ell^\star}(I, q)$
$(Q^\star_{\mathrm{txt}}, \mathcal{K}^{\le \ell^\star}_{\mathrm{txt}}) \leftarrow M^{\mathrm{txt}}_{\le \ell^\star}(q; \texttt{cache} = \mathrm{on})$
Disable ϕ and denote the native recurrent layer by F_0
$H_1 \leftarrow F_0(H^\star_{\mathrm{mm}}; \texttt{cache} = \mathrm{off})$
$(B, \mathcal{K}^{\ell^\star+1}_{\mathrm{txt}}) \leftarrow F_0(Q^\star_{\mathrm{txt}}; \texttt{cache} = \mathrm{on})$ ▷ text-only anchor at the h_1 depth
$\mathcal{K}_{\mathrm{txt}} \leftarrow \left(\mathcal{K}^{\le \ell^\star}_{\mathrm{txt}}, \mathcal{K}^{\ell^\star+1}_{\mathrm{txt}}\right)$
$h_1 \leftarrow P_Q H_1$
$v \leftarrow P_V H_1$ ▷ persistent visual evidence
Enable the shared recurrent adapter ϕ

Persistent visual re-reading
for $t = 2, \ldots, T$ **do**
  $X_t \leftarrow \mathrm{Replace}_Q(H_1, h_{t-1})$
  ▷ Keep $P_V X_t = v$ and preserve the native mask, token ordering, and positions
  $\widetilde{h}_t \leftarrow P_Q F_\phi(X_t; \texttt{cache} = \mathrm{off})$
  $h_t \leftarrow (1 - \beta) h_{t-1} + \beta \widetilde{h}_t$
end for

Strict answer decoding
Discard v, $H^\star_{\mathrm{mm}}$, and all multimodal/recurrent caches
$p_\phi(y \mid I, q) \leftarrow \mathrm{AnswerDecode}(h_T, \mathcal{K}_{\mathrm{txt}})$
  ▷ h_T is the only image-conditioned state available to the answer decoder
if training with gold answer y **then**
  $\ell_i \leftarrow -\dfrac{1}{|\mathcal{A}_i|} \displaystyle\sum_{j \in \mathcal{A}_i} \log p_\phi(y_{i,j} \mid y_{i,<j}, h_{T,i}, \mathcal{K}_{\mathrm{txt},i})$
  $\mathcal{L}_{\mathrm{ans}} \leftarrow \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{i \in \mathcal{B}} \ell_i$
  Update only ϕ using $\nabla_\phi \mathcal{L}_{\mathrm{ans}}$
end if
return $p_\phi(y \mid I, q)$

D Localizing the Causal Visual-Read Interface

Selecting a recurrent boundary requires more than locating where visual mediation begins to decline. The boundary is selected from training-pair activation-patching profiles using the first sustained decline criterion described in the main text. To interpret and validate this choice, we additionally examine whether the selected region already provides a question state with substantial pretrained visual competence. Question-state sufficiency evaluates strict decoding from the native post-layer question state after removing visual rows and multimodal caches, while residual visual influence measures the effect of visual-row intervention on the downstream answer.

Figure 6 shows why the selected transition is suitable for recurrence. Early layers retain strong residual visual influence but provide insufficient question states for strict decoding. With increasing depth, question-state sufficiency rises while residual visual influence declines, creating a region

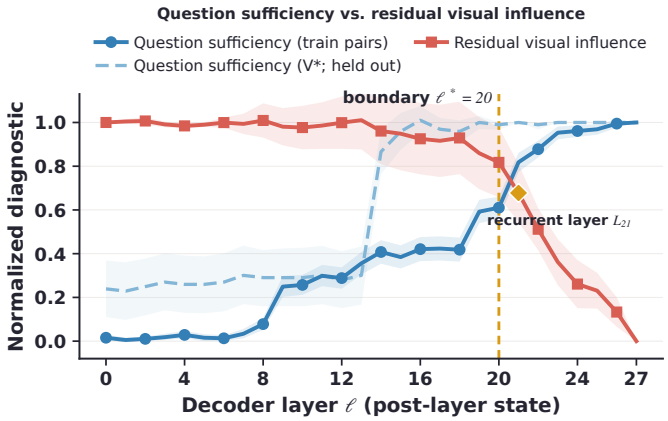


Figure 6: Localizing the causal visual-read interface. Question-state sufficiency and residual visual influence characterize the selected transition near $\ell^* = 20$; the dashed V^* curve is held out from boundary selection.

where a competent question state coexists with visual signal that remains available for further refinement. The activation-patching criterion selects $\ell^* = 20$, immediately before recurrent layer L_{21}. The training-pair sufficiency curve supports this choice, while the dashed V^* curve is held out from selection and independently confirms that question-state sufficiency is already high in this region.

E Generalization Across Backbones and Larger Models

E.1 Cross-Backbone Competence and Reliance

We first examine whether the competence–reliance pattern observed with Qwen2.5-VL transfers to other multimodal backbones. For each backbone, we independently localize its visual-read boundary and train CVRR under the same strict decoder interface. Table 3 compares strict and full-multimodal decoding and reports sensitivity to recurrent-content corruption. InternVL3-9B (Zhu et al., 2025) reaches 75.9% under the strict path versus 75.4% for the full multimodal continuation, while matched-state replacement causes a 57.1-point drop. Gemma-3-4B and Gemma-3-12B (Team et al., 2025) also match or exceed their full-multimodal references in observed accuracy, whereas Gemma-4-12B (Team et al., 2026) shows a 4.7-point decrease. All evaluated backbones exhibit sensitivity to recurrent-content corruption, although the effects are smaller for Gemma-3-4B than for InternVL3-9B. These results support the applicability of the strict interface beyond Qwen2.5-VL, while showing that competence retention and recurrent-state sensitivity vary across backbones.

Table 3: Cross-backbone generalization on V^*. Strict–Full reports the paired accuracy difference, and intervention columns report drops under recurrent-content corruption.

Backbone	ℓ^*	Full MM	Strict	Strict − Full (95% CI)	Swap drop	Blank drop	Noise drop
Gemma-3-4B	21	36.1	40.3	+4.2 [-3.1, +11.5]	7.3	9.4	9.4
Gemma-3-12B	25	43.5	44.0	+0.5 [-5.2, +6.3]	12.4	11.5	10.5
Gemma-4-12B	32	52.9	48.2	-4.7 [-10.5, +0.5]	20.2	14.7	26.7
InternVL3-9B	35	75.4	75.9	+0.5 [-1.1, +2.6]	57.1	42.9	42.4

E.2 Performance on Larger Backbones

We further evaluate CVRR on four instruction-tuned multimodal backbones ranging from 27B to 38B parameters to examine its applicability beyond the default 7B setting. Table 4 reports

Table 4: Generation-based evaluation on larger instruction-tuned multimodal backbones. Zero-shot and full-multimodal SFT controls use CoT, while CVRR uses the strict no-bypass interface. **Bold** and underline indicate the best and second-best scores within each backbone group, including ties. Shaded rows denote CVRR.

Backbone	Method	V^*			MMVP		BLINK						MME-RW-Lite
		Overall	D.A.	R.P.	Item	Pair	Overall	Counting	IQ-Test	Jigsaw	Rel. Reflect	Spatial Rel.	Overall
Qwen2.5-VL-32B	Zero-shot	**84.3**	**82.6**	**86.8**	**77.0**	**56.7**	**58.3**	**70.0**	**26.7**	**72.7**	**47.0**	**84.6**	45.3
	SFT	73.8	80.0	64.5	67.7	40.0	52.8	63.3	19.3	54.7	43.3	78.3	46.4
	CVRR	80.6	80.9	80.3	71.7	47.3	56.0	65.8	22.0	57.3	44.0	81.8	**46.6**
Gemma-4-31B	Zero-shot	61.3	65.2	55.3	**85.7**	**72.7**	**70.7**	**70.8**	19.3	**84.0**	**61.2**	**88.1**	48.4
	SFT	**64.4**	64.4	**64.5**	79.7	60.7	60.1	67.5	28.7	66.0	33.6	79.7	**51.8**
	CVRR	63.9	**67.0**	59.2	80.3	62.0	66.9	62.5	**30.7**	82.0	51.5	81.1	51.2
Gemma-3-27B	Zero-shot	48.2	44.4	54.0	**75.0**	**51.3**	**53.6**	**60.8**	23.3	70.7	35.8	**76.2**	**44.4**
	SFT	49.7	43.5	59.2	60.3	22.0	51.7	59.2	25.3	70.7	**42.5**	73.4	43.9
	CVRR	**51.8**	**45.2**	**61.8**	63.3	27.3	51.7	56.7	**27.3**	**72.7**	37.3	72.7	44.2
InternVL3-38B	Zero-shot	**79.6**	76.5	**84.2**	80.7	64.0	61.8	**78.3**	26.7	49.3	**51.5**	88.1	54.4
	SFT	79.1	76.5	82.9	81.0	64.0	**62.3**	76.7	**28.0**	47.3	50.8	**90.2**	**55.6**
	CVRR	78.5	**77.4**	80.3	**82.7**	**66.7**	61.2	72.5	24.7	**53.3**	**51.5**	86.7	54.8

generation-based results for zero-shot and full-multimodal SFT controls using CoT, alongside CVRR under the strict no-bypass interface. CVRR exceeds the zero-shot reference in several settings. On Gemma-3-27B, V^* overall accuracy increases from 48.2% to 51.8%, while InternVL3-38B reaches 66.7% MMVP pair accuracy compared with 64.0% for the zero-shot reference. CVRR also exceeds the zero-shot reference on MME-RealWorld-Lite for three of the four backbones.

Compared with full-multimodal SFT, CVRR remains competitive across the evaluated backbones. In particular, CVRR achieves higher MMVP pair accuracy on all four backbones, with gains ranging from 1.3 to 7.3 percentage points. On Qwen2.5-VL-32B, CVRR outperforms the SFT control on every reported metric and retains substantially more of the zero-shot performance on V^* overall, decreasing by 3.7 percentage points compared with 10.5 percentage points for SFT. Similarly, on Gemma-4-31B, the BLINK overall decrease is 3.8 percentage points for CVRR versus 10.6 percentage points for SFT. These results show that the strict interface can be applied to substantially larger backbones while remaining competitive with full-multimodal adaptation, although the degree of performance preservation varies across backbones and tasks.

F Efficiency and Computational Cost

Table 5: Accuracy and inference cost across recurrent depths. Latency is measured on a single NVIDIA B200 over 32 examples with three repeated runs, excluding image preprocessing. Relative latency is normalized to $T = 1$. FLOPs report decoder-side computation, and peak Δ memory reports incremental GPU memory during inference.

T	V^* Acc. (%)	Latency (s)	Rel. Latency	TFLOPs	Peak Δ Mem. (GiB)
1	75.4	1.044	1.000	47.58	1.04
2	78.5	1.099	1.053	49.70	1.12
3	79.6	1.163	1.114	51.83	1.16
4	**81.2**	1.146	1.098	53.96	1.16
6	78.5	1.178	1.128	58.22	1.16
8	80.1	1.154	1.105	62.48	1.16

Table 5 reports how inference cost and accuracy change with recurrent depth. The $T = 1$ diagnostic reaches 75.4% on V^*, while independently trained recurrent models reach 78.5% at $T = 2$ and peak at 81.2% at the default $T = 4$. The trend is not monotonic: accuracy decreases to 78.5% at $T = 6$ and recovers to 80.1% at $T = 8$. Because models with $T \geq 2$ are independently trained at their respective depths, this sweep should not be interpreted as a monotonic test-time scaling law. Instead, it shows that additional recurrent depth is useful only up to a task- and training-dependent regime,

with $T = 4$ providing the best accuracy among the tested settings. From $T = 1$ to $T = 4$, measured latency increases by 9.8% and decoder-side computation by 13.4%, while peak incremental memory increases by only 0.12 GiB.

Table 6: Inference cost compared with six latent visual reasoning baselines under a matched four-step budget. Latency is measured on a single NVIDIA B200 over 32 examples with three repeated runs, excluding image preprocessing. Peak and Δ report total and incremental GPU memory during inference.

Model	Latency (s)	Decoder TFLOPs	Peak / Δ (GiB)
CVRR	1.146 ± 0.114	**53.96**	**16.62 / 1.16**
LVR-7B	1.397 ± 0.091	59.72	17.38 / 1.92
Monet-7B	1.396 ± 0.092	59.85	17.30 / 1.85
SkiLa-7B	1.377 ± 0.090	59.70	17.38 / 1.92
Laser-7B	1.329 ± 0.085	59.70	17.31 / 1.86
UniVLR-7B	1.516 ± 0.090	59.70	18.38 / 2.88
HyLaR-7B	1.378 ± 0.090	59.70	17.38 / 1.92

Table 6 compares CVRR with six latent visual reasoning baselines under the same four-step reasoning budget. CVRR has the lowest measured latency, decoder-side computation, and incremental memory among these models. Its latency is 1.146 seconds compared with 1.329–1.516 seconds for the baselines, while decoder computation is 53.96 TFLOPs compared with approximately 59.7–59.9 TFLOPs. CVRR also uses 1.16 GiB of incremental memory compared with 1.85–2.88 GiB for the baselines. This efficiency follows from reusing a single native decoder layer across recurrent steps rather than repeatedly invoking a larger portion of the multimodal backbone. We use $T = 4$ throughout the main experiments unless otherwise specified.

G Query-Channel Corrections Across Recurrence

We further examine how recurrent computation modifies the question representation across successive visual re-reads. We report the relative query correction $\|\Delta q\|_2/\|q^{(0)}\|_2$, where $q^{(0)}$ denotes the corresponding reference query representation and Δq denotes the recurrent correction. Figure 7 visualizes this quantity across question-token positions and hidden channels for each recurrent transition. The aggregate correction increases from 8.9% at $R_1 \to R_2$ to 10.6% at $R_2 \to R_3$ and 12.0% at $R_3 \to R_4$. Later recurrent steps therefore continue to modify the question state rather than simply reproducing the initial multimodal representation.

The correction is non-uniform across both token positions and hidden channels. The two-dimensional maps show this structured variation directly, while the three-dimensional surfaces provide a complementary view of its magnitude and concentration. These quantities characterize functional changes induced by recurrent updates and should not be interpreted as attention weights or semantic feature importance. We avoid strong conclusions from the farthest token-distance regions because the number of available examples decreases substantially in the tail.

H Projection-Wise Functional Corrections

Table 7: Projection-wise functional correction magnitudes (%).

Projection	$R_1 \to R_2$	$R_2 \to R_3$	$R_3 \to R_4$
q_{proj}	8.9	10.6	12.0
k_{proj}	2.9	2.9	2.9
v_{proj}	4.3	4.3	4.3
o_{proj}	**48.3**	**53.0**	**54.9**
$\text{gate}_{\text{proj}}$	10.6	12.4	13.6
up_{proj}	11.0	12.3	12.9
$\text{down}_{\text{proj}}$	14.8	14.2	12.5

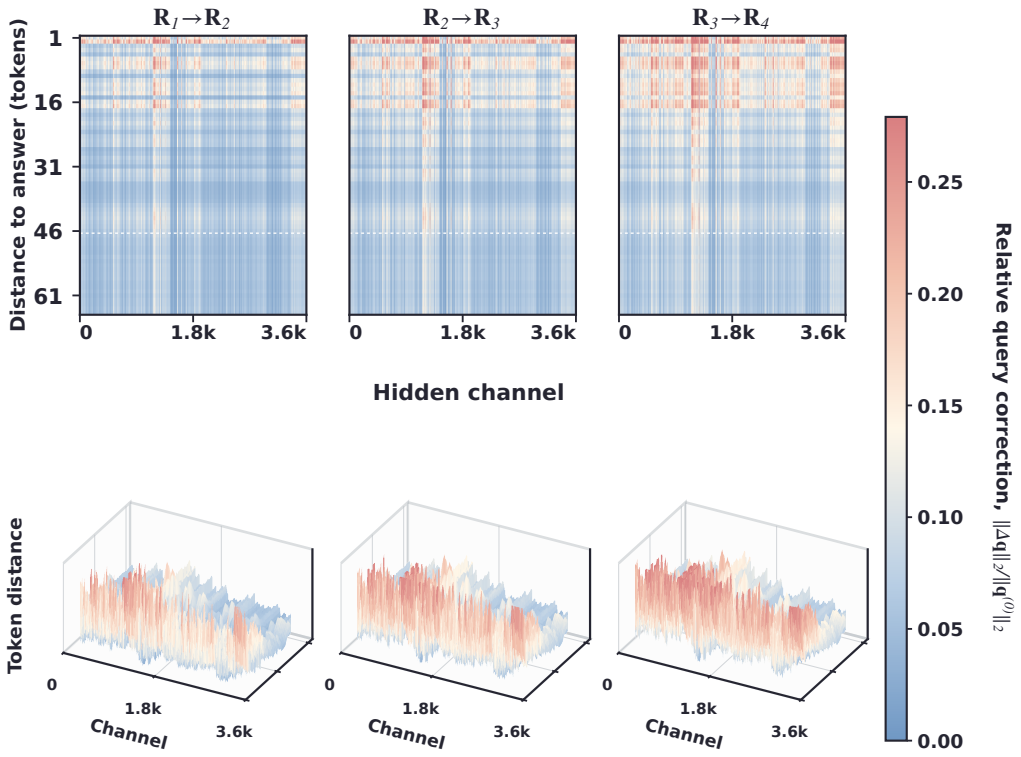


Figure 7: Query-channel corrections across recurrent transitions. Top panels show relative correction magnitude across question-token positions and hidden channels for $R_1 \to R_2$, $R_2 \to R_3$, and $R_3 \to R_4$. Bottom panels show the corresponding three-dimensional surfaces.

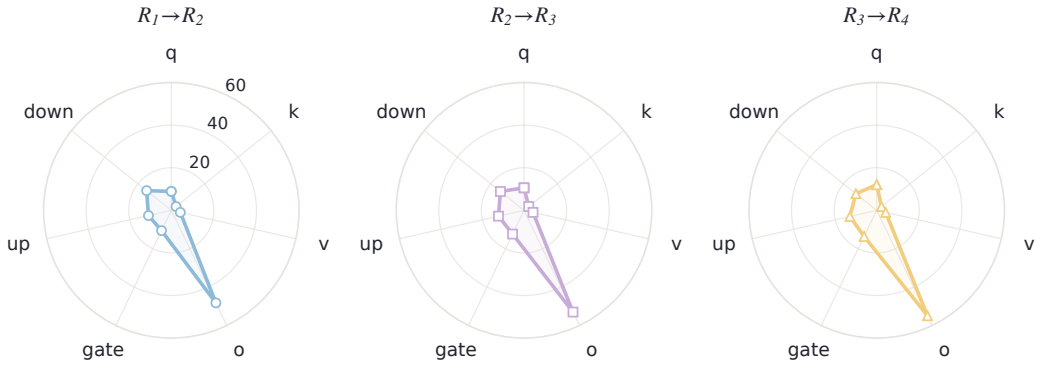


Figure 8: Projection-wise functional corrections across recurrence. Polar plots show relative correction magnitude for transformer projections at $R_1 \to R_2$, $R_2 \to R_3$, and $R_3 \to R_4$. All panels use the same projection ordering and radial scale.

We further decompose the recurrent update by transformer projection to examine where functional corrections are expressed within the shared recurrent layer. Figure 8 and Table 7 show that the output projection receives the largest relative correction throughout recurrence, increasing from 48.3% to 53.0% and 54.9% across the three transitions. Corrections in the query, gate, and up projections also increase across recurrent steps, while the key and value projections remain comparatively stable.

This pattern is consistent with the CVRR design, where visual evidence remains fixed while the question representation evolves across recurrent steps. The query- and output-side corrections show that substantial modification continues to be expressed in the evolving recurrent state, while key and value corrections vary little across transitions. These quantities characterize where functional corrections are expressed within the shared layer rather than the semantic importance of individual projections, and should not be interpreted as showing that any single projection uniquely performs visual reasoning.

I Population Statistics for Spatial Visual Re-reading

Table 8 reports population-level statistics over the full V^* evaluation set to complement the qualitative examples in Figure 5. Causal influence is spatially concentrated across all recurrent transitions, while consecutive region maps show moderate similarity. These results support the conclusion that the step-varying visual re-reading observed in the examples is not driven by a small set of selected cases.

Table 8: Population statistics for spatial causal re-reading on V^*. Intervals denote 95% confidence intervals.

Statistic	$R_1 \to R_2$	$R_2 \to R_3$	$R_3 \to R_4$
Mean signed effect	0.008 [−0.004, 0.020]	0.006 [−0.006, 0.018]	0.005 [−0.007, 0.017]
Max positive effect	0.113 [0.099, 0.127]	0.120 [0.107, 0.133]	0.111 [0.098, 0.125]
Positive regions (%)	24.6 [20.8, 28.7]	24.3 [20.4, 28.2]	24.0 [20.3, 27.9]
Top-quartile positive mass (%)	48.4 [42.6, 54.2]	52.9 [47.1, 58.8]	46.6 [40.9, 52.5]
Next-map cosine similarity	0.367 [0.327, 0.407]	0.350 [0.311, 0.390]	–