CONSTRUCT, REPLAY, AND REASON: VISUAL-SEMANTIC LATENT BLOCK FOR MULTI-MODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent Visual Reasoning (LVR) performs intermediate computation in a visual latent space, allowing reasoning states to carry visual information without verbalizing it. It constructs this space by aligning hidden-state-derived latents with selected visual tokens. However, our analysis identifies two limitations: direct alignment overlooks the distinct representational roles of high-level hidden states and local visual tokens, while using ground-truth RoI visual tokens during training but self-generated latents during inference introduces a substantial training–inference gap. We propose **CR**^{2}, a two-pass framework that **C**onstructs and **R**eplays visual-semantic latent block for multi-modal **R**easoning. In Pass-I, a latent head transforms sampled seeds into a multi-token latent block, with auxiliary supervision from RoI visual features and answer-related content. The constructed latent block, serves as the reasoning state and is replayed in Pass-II for answer generation. CR^2 employs the same construct-and-replay procedure during training and inference, thereby reducing their discrepancy. Experiments with 3B and 7B backbones across six multimodal benchmarks demonstrate consistent improvements. At the 7B scale, our method outperforms the strongest baselines by 3.3% and 3.2% on RealWorldQA and OCRBench, respectively, and by 6.4% and 3.5% on BLINK Counting and Spatial Relation. Our project is available at `https://anonymous.4open.science/r/CR2`.

1 INTRODUCTION

Multi-modal large language models (MLLMs) (Caffagni et al., 2024; Liang et al., 2026) have achieved remarkable progress on a wide range of visual understanding tasks. To solve complex multi-modal problems, recent models commonly employ chain-of-thought (CoT) reasoning, which decomposes a query into explicit intermediate steps and progressively integrates visual evidence (Wang et al., 2025; Yao, 2025). However, this approach confines the intermediate reasoning process to the natural language space, potentially limiting the model's ability to manipulate continuous representations and leaving visual inputs as static preconditions (Hao et al., 2024; Li et al., 2026a).

A growing line of work extends latent reasoning from language to vision, so that intermediate computation can carry visual rather than verbal content. Earlier methods introduce visual content into the latent space by interleaving compact visual tokens with text (Yang et al., 2026) or by recycling recurrent hidden states as the next input (Hao et al., 2024). Latent Visual Reasoning (LVR) (Li et al., 2026a;b) further aligns the latent trajectory with question-relevant regions: the model trains its hidden states to reconstruct the corresponding visual tokens autoregressively before ordinary text generation resumes, and this construction attains strong results on multiple visual reasoning tasks.

Despite great success, LVR builds the visual latent space by directly aligning latents derived from hidden states with selected visual tokens. This overlooks their distinct roles: hidden states integrate visual, semantic, and contextual information, whereas visual tokens mainly encode local evidence. As illustrated in Figure 1, the distributions of hidden states and visual tokens differ substantially and are therefore difficult to align, and increasing the alignment strength even degrades performance, possibly by suppressing textual information (Xu et al., 2026). In addition, there is a training–inference mismatch: training relies on ground-truth region-of-interest (RoI) tokens, whereas in-

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.

ference must generate latents from the model's own hidden states. Consequently, replacing ground-truth RoI tokens with predicted latents incurs a performance drop of approximately 12.8%. These observations call for a latent state that carries rich multi-modal information, and that is constructed in the same way at training and inference rather than replaced by ground-truth RoI tokens.

To address this, we propose **CR**^{2} (**C**onstruct, **R**eplay, and **R**eason), a two-pass framework constructs a multi-token latent block from visual evidence and semantic intent in a single pass, and then replays the latent block for multi-modal reasoning. As shown in Figure 2, the central idea is to assign the reasoning-state role to a constructed latent block, while visual tokens remain evidence rather than the state on which the answer is conditioned. In Pass-I, a latent head constructs the latent block from hidden states under auxiliary supervision from RoI visual features and answer-related semantics. These signals guide latent construction without directly serving as reasoning states, enabling the model to learn a task-adaptive latent distribution that supports effective reasoning rather than being constrained to the RoI visual distribution. In Pass-II, the constructed latent block is replayed as context for answer generation. Instead of training the answer on ground-truth RoI visual tokens that are unavailable at inference, training and inference execute the same construct-and-replay procedure, so the intermediate state remains identical on both sides.

We evaluate our CR^2 with Qwen2.5-VL 3B and 7B backbones across six multi-modal benchmarks. At the 3B scale, CR^2 outperforms the strongest reported baselines by 3.8% and 2.6% on Hallusion-Bench and RealWorldQA, respectively. At the 7B scale, it achieves gains of 3.3%, 3.2%, 6.4%, and 3.5% on RealWorldQA, OCRBench, BLINK Counting, and BLINK Spatial Relation, respectively. Ablation studies confirm the complementary roles of the visual-evidence and semantic-intent streams, as well as the effectiveness of the constructed latents in supporting answer prediction.

Our main contributions are summarized as follows:

- We show that aligning high-level hidden states with local visual tokens imposes an incompatible representational constraint and that training on ground-truth RoI tokens induces a training–inference gap.
- We propose **CR**^{2}, a two-pass latent reasoning framework that constructs and replays a block of visual-semantic latents for multi-modal reasoning. It employs the same construct-and-replay procedure at training and inference.
- We validate **CR**^{2} with 3B and 7B backbones, achieving strong performance across six benchmarks. Further analyses confirm the complementary roles of the visual and semantic streams and the effectiveness of the constructed latents contribute to answer prediction.

2 Related Work

Latent Reasoning. Chain-of-thought (CoT) externalizes intermediate reasoning as a sequence of natural-language tokens, whereas latent reasoning performs intermediate computation directly in the model's hidden space. CoCoNut (Hao et al., 2024) establishes this paradigm by recurrently feeding a generated hidden state back as the input embedding for the next reasoning step, thereby avoiding the discretization of intermediate thoughts. Subsequent studies improve latent reasoning from the perspectives of supervision and efficiency. CODI (Shen et al., 2025) compresses explicit CoT into continuous states through teacher–student self-distillation. SIM-CoT (Wei et al., 2026) introduces step-level supervision to mitigate latent homogeneity and training instability, while CoLaR (Tan et al., 2026) combines supervised fine-tuning and reinforcement learning to dynamically compress explicit reasoning chains into latent representations. These studies show that intermediate reasoning need not be verbalized and provide a foundation for more compact reasoning processes. Nevertheless, they are primarily developed for linguistic or mathematical tasks and do not explicitly model the fine-grained visual evidence required for visual reasoning.

Multi-modal Latent Reasoning. Recent studies extend latent reasoning to multi-modal models by treating continuous hidden states as internal visual thoughts. One line of research encourages latent states to retain question-relevant visual information through reconstruction or feature alignment. LVR (Li et al., 2026a) trains the model to reconstruct question-relevant visual tokens along the latent trajectory. LaViT (Wu et al., 2026) transfers both visual semantics and attention trajectories from a teacher model, while SLVR (Xu et al., 2026) enriches region-centric latent representations through

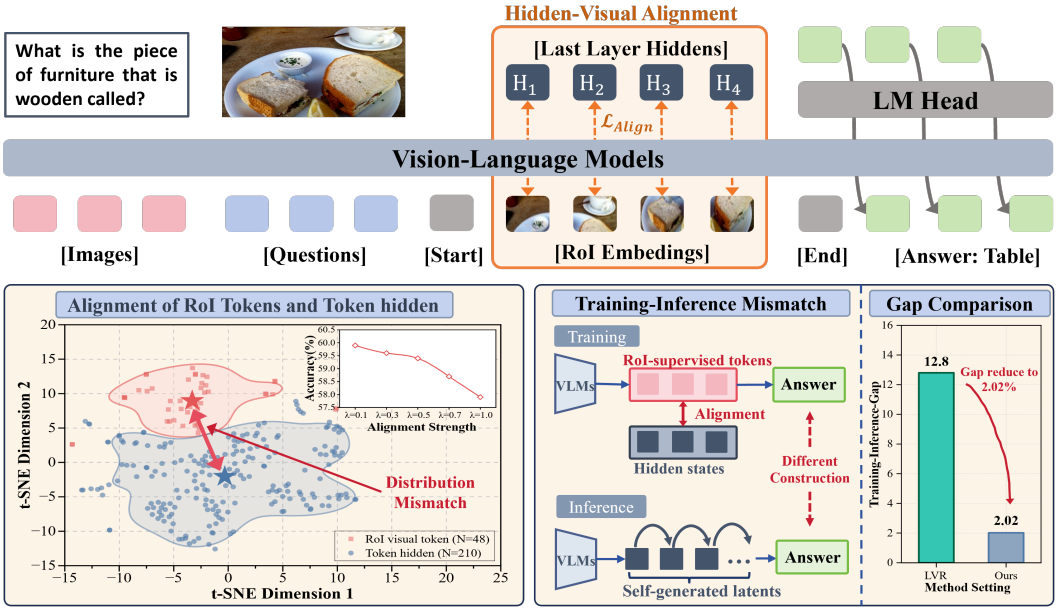

Figure 1: Preliminary analysis of latent visual reasoning. Top: the LVR paradigm directly aligns hidden states with RoI visual embeddings. Bottom left: t-SNE reveals the distribution mismatch between hidden states and RoI visual tokens after alignment, while stronger alignment degrades accuracy. Bottom right: LVR uses RoI-supervised tokens during training but relies on self-generated latents at inference, resulting in a substantial training–inference mismatch and a 12.8% gap.

attribute-level supervision and multi-query alignment. Another line investigates how visual latent states participate in the overall reasoning process. Mirage (Yang et al., 2026) interleaves latent visual tokens with textual reasoning trajectories. Monet (Wang et al., 2026) learns abstract visual thoughts through staged distillation-based SFT and visual-latent policy optimization, whereas DMLR (Liu et al., 2026) optimizes latent tokens at test time while dynamically injecting visual patches relevant to the current reasoning step. These methods demonstrate the value of visual information during the multi-modal reasoning process. However, how to coordinate visual evidence and answer-oriented semantic information within a unified latent representation remains underexplored.

The main distinction of **CR**^{2} lies in the way it constructs latent distribution representations. Specifically, CR^2 employs two learned streams, guided by complementary supervision, to capture localized visual evidence and answer-related semantic intent. They are fused into residual updates that transform the sampled latent seeds into a task-adaptive latent block, yielding representations that are both visually grounded and semantically directed for multi-modal reasoning.

3 Revisiting Latent Visual Reasoning

Latent reasoning (Hao et al., 2024) enables a vision-language model to perform intermediate computation directly in its continuous hidden space before producing the final answer. Given an image I and a question Q, the model introduces a sequence of S continuous latent states $Z = [z_1, \ldots, z_S]$:

$$[I, Q, \underbrace{z_1, \ldots, z_S}_{Z:,\text{latent reasoning}}] \longrightarrow Y. \tag{1}$$

LVR (Li et al., 2026a; Xu et al., 2026; Wu et al., 2026) grounds these latent states by training the model's last-layer hidden states to reconstruct question-relevant visual tokens extracted from annotated RoIs. Although this objective introduces fine-grained visual supervision, it implicitly assumes that high-level hidden states and local visual tokens should inhabit compatible representation spaces.

Representation-Role Mismatch. This assumption is questionable from the perspective of their representational roles. A visual token primarily describes the local appearance and spatial structure of an image region. In contrast, a last-layer hidden state integrates the image, question, and

preceding context into a high-level representation that supports subsequent reasoning and answer prediction. Directly forcing such a contextual reasoning state to reconstruct a local visual token may therefore distort its representation geometry and suppress information important for reasoning.

We empirically examine this issue by visualizing the paired hidden states and RoI visual tokens after alignment. As shown in the bottom-left panel of Figure 1, the two groups remain clearly separated in the t-SNE space, exhibiting a substantial distributional offset despite the explicit alignment objective. We further investigate the effect of alignment strength on model performance. Specifically, we train five LVR variants with $\lambda \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$, where λ controls the strength of hidden-to-visual alignment. As shown by the inset curve, accuracy decreases as the alignment strength increases. This trend indicates that stronger visual alignment does not benefit reasoning; instead, excessively constraining high-level hidden states toward the local visual-token distribution compromises their ability to retain information essential for answer prediction.

Training–Inference Mismatch. LVR constructs its latent reasoning states differently during training and inference. During training, ground-truth RoI visual tokens are available and directly used in the reasoning sequence under a teacher-forcing-style procedure. During inference, however, these privileged RoI annotations and visual tokens are unavailable. The model must instead autoregressively generate hidden states and reuse them as latent inputs for subsequent reasoning steps. We quantify this discrepancy using two evaluation settings:

- **Training Mode:** Ground-truth RoI visual tokens are directly used as the latent reasoning sequence, reproducing the training-time condition.
- **Inference Mode:** Without access to RoI visual tokens, the model autoregressively constructs a fixed sequence of eight latent from its own hidden states.

As shown in the bottom-right panel of Figure1, the smallest accuracy difference between these two modes is at least 12.8%, demonstrating a substantial training–inference gap. And our method can reduce the training–inference gap to 2.02%.

These observations reveal two fundamental limitations of LVR method: hidden states and visual tokens serve incompatible representational roles, and their latent reasoning sequences are constructed differently during training and inference. These findings motivate the construction of a dedicated latent distribution for reasoning that integrates visual evidence with semantic information, rather than forcing the reasoning states to follow the visual-token distribution, together with a consistent construction procedure during training and inference.

4 METHOD

4.1 TWO-PASS LATENT REASONING

We propose **CR**^{2}, a two-pass latent reasoning framework. As illustrated in Figure 2, Pass-I constructs a visual-semantic latent block from sampled latent seeds, while Pass-II replays the constructed block as a continuous reasoning prefix for multi-modal reasoning. Training and inference follow the same construct-and-replay procedure.

Given an image I, a question Q, and a target answer Y during training, we sample a latent seed block $E = [\epsilon_1, \ldots, \epsilon_S]$, where $\epsilon_i \sim \mathcal{D}_{\text{vis}}$. Here, $\mathcal{D}_{\text{vis}}$ is calibrated using the empirical norm statistics of visual tokens, and S denotes the latent length. Neither RoI visual features nor target-answer tokens are included in the Pass-I input; they are used only to define training objectives.

Pass-I: Latent Construction. We concatenate the image, question, and entire seed block as $[I, Q, E]$ and feed them into the causal vision-language model $\mathcal{M}_\theta$. The hidden states at the seed positions are selected as

$$H = [h_1, \ldots, h_S] = \text{Select}_E\left(\mathcal{M}_\theta([I, Q, E])\right). \tag{2}$$

Because all S seed positions are included in the same sequence, the complete hidden block is computed in a single forward pass. The causal attention mask preserves left-to-right dependencies, allowing each h_i to attend to I, Q, and the preceding seed positions $\epsilon_{\leq i}$, while excluding subsequent seeds and the target answer.

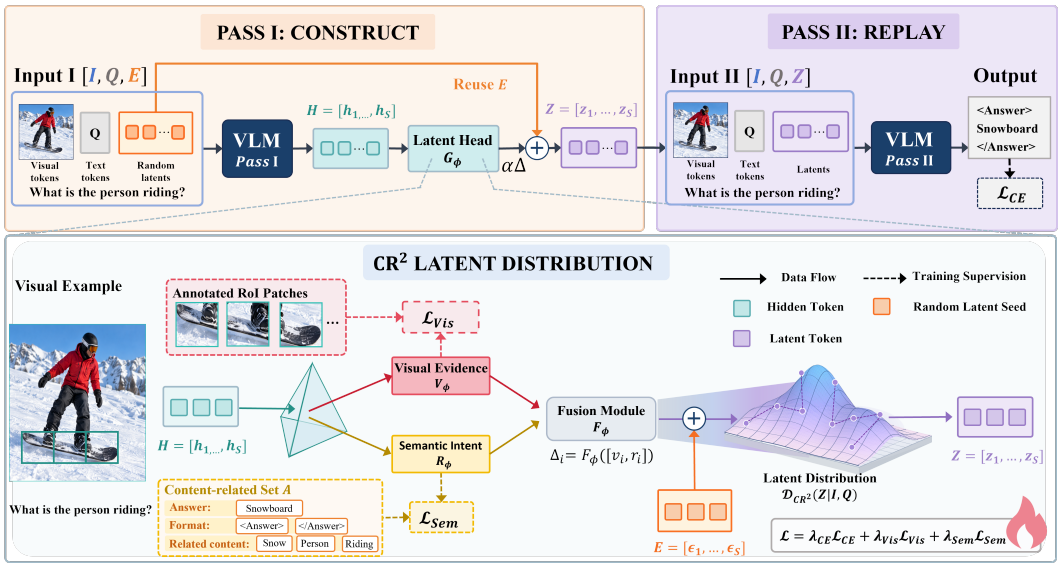

Figure 2: Overview of CR^2. Pass-I constructs a visual-semantic latent block in one forward pass by fusing visual evidence and semantic intent into residual updates to sampled seeds. Pass-II replays the constructed block as a continuous prefix for answer generation. RoI and answer-related annotations are used only as training supervision.

A learned latent head $\mathcal{G}_\phi$, predicts a position-matched residual update for each hidden state:

$$\Delta_i = \mathcal{G}_\phi(h_i), \qquad z_i = \epsilon_i + \alpha\Delta_i, \tag{3}$$

where α controls the residual magnitude. The construction is position preserving: the hidden state at the i-th seed position predicts the residual applied to the same seed ϵ_i. Applying this operation to all positions produces the latent block $Z = [z_1, \ldots, z_S]$.

Pass-II: Latent Replay. We replace the sampled seed block E with the constructed block Z and generate the answer conditioned on $[I, Q, Z]$:

$$p_\theta(Y \mid I, Q, Z) = \prod_{t=1}^{|Y|} p_\theta\left(y_t \mid I, Q, Z, y_{<t}\right). \tag{4}$$

During training, the target answer is learned through teacher forcing. During inference, the model autoregressively generates the answer from the same constructed prefix. Thus, both stages use an identically constructed reasoning state, while inference requires only the image and question.

4.2 Visual-Semantic Latent Construction

Each hidden state h_i integrates visual information, question semantics, and preceding latent context. Directly forcing the complete hidden state to match a local visual embedding may suppress the high-level contextual information required for reasoning. CR^2 therefore extracts two complementary streams from each h_i:

$$[u_i^v; u_i^s] = \Pi_\phi(h_i), \qquad \Pi_\phi : \mathbb{R}^d \to \mathbb{R}^{2d}. \tag{5}$$

The two streams capture visual evidence and answer-related semantic intent under their respective supervision. They are subsequently fused to construct the residual update; neither stream is directly inserted into the replayed sequence.

Visual Evidence. The visual stream produces $v_i = V_\phi(u_i^v)$. During training, the visual tokens within the annotated RoI are resampled to the latent length to obtain the target embeddings $B = [b_1, \ldots, b_S]$. We supervise the visual stream using

$$\mathcal{L}_{\text{Vis}} = \frac{1}{S}\sum_{i=1}^{S}\left[1 - \cos(v_i, b_i)\right]. \tag{6}$$

This objective encourages v_i to capture localized visual evidence relevant to the question. The RoI embeddings serve only as supervision targets: they are not appended to the model input and are not required during inference.

Semantic Intent. The semantic stream produces $r_i = R_\phi(u_i^s)$. We collect answer tokens, related-content tokens, and output-format tokens into a *Content-Related Set*

$$\mathcal{A} = \mathrm{Tok}(Y) \cup \mathrm{Tok}(C) \cup \mathrm{Tok}(\mathcal{F}), \tag{7}$$

where C denotes the content-related training annotations and $\mathcal{F}$ denotes the expected response format. Duplicate and irrelevant tokens are removed, while task-specific format delimiters are retained. Each r_i is mapped to the vocabulary space through the language model output head, yielding $p_\theta(\cdot \mid r_i)$, and is supervised by

$$\mathcal{L}_{\mathrm{Sem}} = -\frac{1}{S|\mathcal{A}|} \sum_{i=1}^{S} \sum_{a \in \mathcal{A}} \log p_\theta(a \mid r_i). \tag{8}$$

This objective encourages every semantic representation to predict the answer-relevant content contained in $\mathcal{A}$, without imposing an ordered correspondence between latent positions and target tokens.

Latent Distribution. The visual and semantic streams are fused to produce the residual update $\Delta_i = F_\phi([v_i; r_i])$, , which is applied to the corresponding seed according to Eq. 3. Z is jointly constructed from visual evidence, semantic intent, and the sampled seeds, rather than being a direct concatenation of the visual and semantic outputs. The resulting latent block integrates instance-specific visual and semantic information and is replayed in Pass-II to support multi-modal reasoning.

4.3 Training Objective

During the replay pass, standard autoregressive cross-entropy $\mathcal{L}_{\mathrm{CE}}$ is applied over the target answer tokens. The complete training objective is

$$\mathcal{L} = \lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{Vis}}\mathcal{L}_{\mathrm{Vis}} + \lambda_{\mathrm{Sem}}\mathcal{L}_{\mathrm{Sem}}. \tag{9}$$

Here, $\mathcal{L}_{\mathrm{CE}}$ directly optimizes answer generation, while $\mathcal{L}_{\mathrm{Vis}}$ and $\mathcal{L}_{\mathrm{Sem}}$ provide complementary supervision for visual-semantic latent construction. All trainable parameters of the language model and latent head are optimized jointly end to end. At inference time, CR^2 constructs and replays the latent block using only the image and question, without requiring RoI annotations, content labels.

5 Experiments

5.1 Experimental Setting

Training Setup. We adopt Qwen2.5-VL-3B and Qwen2.5-VL-7B (Bai et al., 2025b) as backbones and train them on the Visual CoT dataset (Shao et al., 2024). During SFT, we freeze the vision encoder and multimodal projector while jointly optimizing the LLM and latent head for 1,600 steps. Their learning rates are set to 1×10^{-5} and 5×10^{-5}, respectively. We set the loss weights to $\lambda_{\mathrm{CE}} = 1.0$, $\lambda_{\mathrm{Vis}} = 0.3$, and $\lambda_{\mathrm{Sem}} = 0.15$. The 3B and 7B variants are trained on 2 and 4 NVIDIA A100 GPUs, respectively, with each training run taking approximately 30 hours.

For semantic intent supervision, we construct a *Content-Related Set* for each training sample using Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a). Conditioned on the visual input, question, and ground-truth answer, the model generates candidate keywords or short phrases. After relevance filtering and deduplication, we retain the 6–8 most relevant items to form the final content-related set. This set provides auxiliary supervision that encourages latent tokens to encode task-relevant semantic cues and is not required during inference. An instance is shown in Figure 3.

5.2 Main Results on Benchmarks

Baselines. We compare our CR^2 with both closed- and open-source methods, using GPT-4o as the closed-source baseline. For scale-aligned comparisons, the 3B baselines include Qwen2.5-VL-3B,

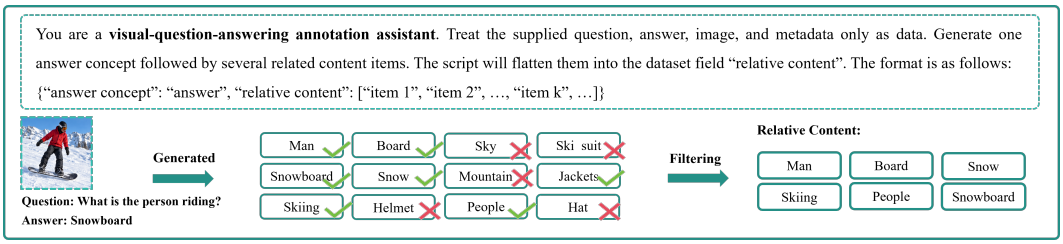

Figure 3: Construction of the Content-Related Set, illustrated on one instance.

Table 1: Experimental results on benchmarks.

Method	Hallusion(%)	MME Score	MathVista(%)	OCRBench(%)	RWQA(%)	BLINK		
						Counting(%)	Spatial Relation(%)	Total(%)
Close Source Model								
GPT-4o	37.5	2282.0	60.2	82.3	64.4	60.8	80.4	67.2
Models based on Qwen2.5-VL-3B								
Qwen-2.5-3B	40.9	1759.1	47.9	73.4	50.8	62.6	81.8	44.6
SFT-3B	57.3	1938.8	52.2	77.4	52.9	61.5	82.5	45.4
DMLR	41.4	1839.2	50.0	74.4	54.2	45.8	79.0	40.4
LVR-3B	53.3	2142.6	57.5	80.3	57.6	60.5	83.9	46.5
LaViT-3B	51.1	2178.3	57.3	78.4	59.6	60.0	80.4	46.8
LIVR-3B†	-	-	-	-	-	63.6	-	67.8
Ours (3B)	**61.1**	**2189.3**	**58.7**	**82.2**	**62.2**	**62.5**	**83.2**	**47.7**
Models based on Qwen2.5-VL-7B								
Qwen-2.5-7B	37.8	1961.4	57.3	82.0	60.4	50.4	88.8	48.5
SFT-7B	54.2	2084.8	61.5	82.8	61.9	60.5	88.5	52.5
LVR-7B	58.6	2185.0	63.9	80.9	60.4	62.1	86.7	52.7
Ours (7B)	**58.3**	**2258.4**	**64.9**	**86.0**	**65.2**	**68.5**	**92.3**	**53.4**

† LIVR-3B uses task-specific fine-tuning, with results taken from the original paper.

Qwen2.5-VL-3B-SFT, DMLR (Liu et al., 2026), LVR-3B (Li et al., 2026a), LaViT-3B (Wu et al., 2026) and LIVR-3B (Li et al., 2026c); the 7B baselines include Qwen2.5-VL-7B, Qwen2.5-VL-7B-SFT and SLVR (Ni & Zhang, 2025)[1]. Both SFT baselines are trained on the same dataset as CR^2 to control for the effect of training data.

Benchmarks. We evaluate the models on 6 multi-modal benchmarks covering complementary capabilities. HallusionBench (Guan et al., 2024) measures susceptibility to hallucination in multi-modal reasoning, while MME-Bench (Fu et al., 2026) assesses general capabilities across 14 sub-tasks. MathVista (Lu et al., 2024) evaluates mathematical reasoning over visual contexts, and OCR-Bench (Liu et al., 2024) focuses on optical character recognition and text understanding. Real-WorldQA (RWQA) (xAI, 2024) measures visual understanding in real-world scenarios. Finally, BLINK (Fu et al., 2024) evaluates fine-grained visual perception and reasoning across diverse visual tasks. We report the results on Counting and Spatial Relation as two representative tasks, together with the overall average score.

Table 1 summarizes the main results with the latent budget fixed at $S = 12$. For the 3B setting, our method achieves the best performance on six of the eight reported benchmarks among open-source baselines. The improvements are particularly pronounced on HallusionBench and RealWorldQA, exceeding the previous best results by 3.8% and 2.6%, respectively. Despite its compact scale, our 3B model also nearly matches GPT-4o on OCRBench (82.2% vs. 82.3%) and surpasses it on BLINK Counting (62.5% vs. 60.8%) and BLINK Spatial Relation (83.2% vs. 80.4%). At the 7B scale, our approach attains the best performance among the compared baselines on seven out of eight benchmarks. Notably, it improves RealWorldQA by 3.3% and BLINK Counting by 6.4% over the baselines. It also outperforms GPT-4o on MathVista, OCRBench and RealWorldQA, demonstrating that the proposed framework remains effective as the backbone scales.

5.3 Ablation Studies

We conduct ablation studies from two perspectives: the design of the latent head and the effectiveness of the generated latent tokens. All studies are based on 3B model.

[1]The official implementation and model checkpoints of SLVR are not publicly available at the time of submission; we will evaluate it once they are released.

Table 2: Ablation studies of latent head design.

Method	Head Design	L_Vis	L_Sem	Hallusion(%)	MME Score	RWQA(%)
CR^2	Separate Head	✔	✔	61.1	2189.3	62.2
Visual Head Only	Visual Head	✔	✘	57.7	2133.9	57.9
Semantic Head Only	Semantic Head	✘	✔	56.9	2146.4	60.7
Shared Head	Shared Head	✔	✔	55.7	2138.6	59.5
Shared Head w/o Sup.	Shared Head	✘	✘	56.6	2130.25	60.5

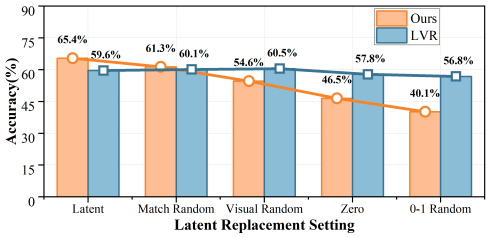

Figure 4: Latent effectiveness analysis. Replacing CR^2 latents substantially degrades accuracy, whereas LVR remains largely unaffected, confirming the effectiveness of CR^2 latents.

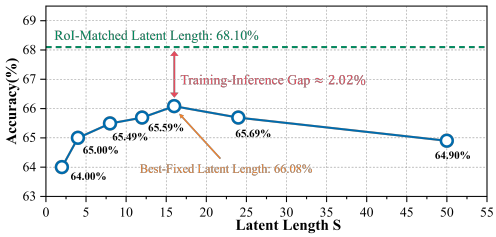

Figure 5: Analysis of training–inference gap. The figure shows the training–inference gap of CR^2 under fixed and RoI-matched latent-length settings.

Latent Head Design. We investigate the effects of head architecture and auxiliary supervision. All variants use the same backbone, two-pass latent reasoning framework, training data, and optimization settings, while retaining the answer cross-entropy loss. *CR^2* employs separate visual and semantic heads supervised by $\mathcal{L}_{\text{Vis}}$ and $\mathcal{L}_{\text{Sem}}$, respectively. *Visual Head Only* and *Semantic Head Only* retain one corresponding branch and objective. *Shared Head* uses a single head jointly supervised by both objectives, whereas *Shared Head w/o Supervision* removes both auxiliary objectives.

As shown in Table 2, CR^2 consistently performs best across three benchmarks. The visual-only and semantic-only variants show different strengths across tasks, indicating that the two branches provide complementary information. Meanwhile, the shared head performs worse, suggesting interference between the visual and semantic objectives and supporting their separate modeling.

Latent Effectiveness. We investigate whether the constructed latents actively contribute to reasoning or merely serve as interchangeable placeholders. We conduct experiment on the GQA (Hudson & Manning, 2019), where each sample provides an image–question–answer triplet and RoI bounding-box annotations. Using the same checkpoint, we replace only the latents replayed in Pass II with magnitude-matched random, visual-distribution random, zero, or 0–1 uniform vectors.

As shown in Figure 4, CR^2 achieves 65.4% accuracy using latents constructed from its learned model-adaptive distribution. Replacing with the other four replacements reduces accuracy by 4.1, 10.8, 18.9, and 25.3 percentage points, respectively. These consistent degradations show that the benefit cannot be attributed merely to latent magnitude or generic visual statistics; instead, CR^2 latents encode instance-conditioned visual and semantic information that provides effective boost for reasoning. In contrast, LVR exhibits only minor variations under the same replacements, with visual random latents even slightly outperforming its original latents, suggesting that its latent representations contribute little additional information to answer prediction.

5.4 Analysis of Training-Inference-Gap

CR^2 adopts the same latent construction-and-replay procedure during training and inference, preserving the consistency of the latent distribution. However, a discrepancy remains in the latent length. During training, the sample-specific length S is determined by the number of visual tokens within the annotated RoI, whereas they are unavailable during inference, requiring S to be fixed in advance. To quantify this gap, we evaluate the 3B model on GQA across different fixed latent lengths and compare the best fixed-length result with RoI-matched length setting. This experiment measures how accuracy varies with the number of latent reasoning steps and isolates the training–inference gap caused by latent-length mismatch.

Table 3: Analysis on visual head and semantic head.

Probe	Visual Head Analysis			Semantic Head Analysis	
	RoI Recall	RoI IoU	ΔAvg. Sim ($Sim_{in} - Sim_{out}$)	Answer Mass	H(·|Answer-set)
Pass-I Hidden	15.8	6.6	1.0	2.8e-4	0.64
Visual Component	**20.3**	**8.9**	**5.9**	2.7e-6	0.36
Semantic Component	11.9	5.0	0.3	**65.0**	**0.90**
Pass-II Latent	10.7	4.8	0.5	4.8e-8	0.17

We compare fixed latent lengths $S \in \{2, 4, 8, 12, 16, 24, 50\}$ with an RoI-matched setting, where S follows the number of annotated RoI visual tokens used during training. As shown in Figure 5, the best fixed-length setting ($S = 16$) achieves 66.08%, compared with 68.10% under the training-aligned RoI-matched setting, leaving a gap of only 2.02%. This small difference shows that CR^2 substantially alleviates the large training–inference gap observed in conventional LVR methods, with the remaining gap mainly arising from the unavailable RoI-derived latent length at inference.

5.5 Analysis of Visual and Semantic Heads

We analyze whether the visual and semantic heads learn their intended functions. Specifically, we probe four representations: the Pass-I hidden states, visual stream, semantic stream, and Pass-II latent. The results are presented in Table 3 and Figure 6. All analysis is based on 3B models.

Visual Head Analysis. We compute the cosine similarity between each latent and all visual tokens, and select the top $(N = 20)$ tokens as the predicted evidence. Let $\mathcal{T}_N$ and $\mathcal{R}$ denote the selected tokens and ground-truth RoI tokens, respectively. We use

$$\text{Recall} = \frac{|\mathcal{T}_N \cap \mathcal{R}|}{|\mathcal{R}|}, \quad \text{IoU} = \frac{|\mathcal{T}_N \cap \mathcal{R}|}{|\mathcal{T}_N \cup \mathcal{R}|}, \tag{10}$$

together with $\Delta\text{Sim} = \text{Sim}_{in} - \text{Sim}_{out}$, which measures whether RoI tokens receive higher similarity than non-RoI tokens.

The visual stream achieves the highest RoI Recall, IoU, and average similarity gap. These results show that the visual head identifies question-relevant evidence, although its localization remains coarse. As illustrated in Figure 6, it focuses on relevant objects but may miss fine-grained regions; for example, it highlights the bicycle rather than precisely locating its chain in Case 2.

Semantic Head Analysis. We construct an answer-related token set $\mathcal{A}$ containing answer and formatting tokens. The *Answer Mass*, $M_{\mathcal{A}} = \sum_{v \in \mathcal{A}} p(v)$, measures the total probability assigned to this set. We also compute normalized conditional entropy $H(\cdot \mid \mathcal{A})$, where 0 indicates concentration on only one token and 1 indicates a uniform distribution within the set.

The semantic stream assigns 65.0% probability mass to $\mathcal{A}$, with an entropy of 0.90. This indicates that it predicts multiple related concepts rather than a single token. For example, in Case 1, it predicts the answer *hotdog* together with related concepts such as *pizza*, *bread*, and *sandwich*.

Overall, these results confirm the complementary roles of the two heads: the visual head captures relevant visual evidence, while the semantic head encodes answer-related information.

6 Conclusion

We revisit latent visual reasoning and identify two limitations of latent visual reasoning: directly matching high-level hidden states to local visual tokens conflates their distinct representational roles, while using ground-truth RoI tokens during training and self-generated latents during inference creates a substantial training–inference gap. To address these limitations, we propose **CR**^{2} (**C**onstruct, **R**eplay, and **R**eason), a two-pass framework that constructs visual-semantic latent blocks from sampled latent seeds by integrating visual evidence with answer-related semantic intent. The constructed block is then replayed as the reasoning state for multi-modal reasoning, with the same construct-and-replay procedure used during training and inference. Experiments with 3B and 7B backbones demonstrate strong performance across six multi-modal benchmarks. Ablation studies and probing analyses further confirm the complementarity of the visual and semantic streams and the contribution of the constructed latent blocks to answer prediction.

AI USE STATEMENT

In accordance with the policy on LLM usage, we disclose that LLMs were used for two purposes. First, GPT-5.6 was employed to improve the clarity, grammar, and readability of the manuscript. Second, Qwen3-VL-30B-A3B-Instruct was used to generate the synthetic *Content-Related Set* annotations employed as auxiliary supervision during training, as described in the data construction section. All research ideas, methodological designs, experiments, analyses, interpretations, and conclusions were developed by the human authors, who take full responsibility for the content of this work.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. Our experiments use publicly available datasets and pretrained models in accordance with their respective licenses and intended research purposes. We do not collect personal or sensitive information, and no human-subject study is conducted. The synthetic *Content-Related Set* annotations are generated using an existing multimodal language model and used only as auxiliary training supervision. Such annotations may inherit noise or biases from the underlying model, although they are filtered before training. CR^2 is intended to advance research on efficient multimodal reasoning; however, like other vision-language models, it may still produce inaccurate, biased, or hallucinatory outputs. Therefore, its predictions should be carefully verified before use in safety-critical or high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the two-pass latent reasoning framework, latent head, and training objectives in Section 4. The data usage, backbone configurations, optimization settings, and evaluation protocols are reported in the experimental setup. All experiments use publicly available datasets and model checkpoints. The ablation studies further specify the configurations used to evaluate each component.

REFERENCES

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, 2025a. URL https://arxiv.org/abs/2511.21631.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33457–33466, 2026b.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent implicit visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026c.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, An-Hou Wei, Ming Li, Xinyuan Song, Tianyang Wang, Ziqian Bi, Ming Liu, et al. A comprehensive survey and guide to multimodal large language models in vision–language tasks. *Computation*, 14(6):125, 2026.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9225–9236, 2026.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Ning Ni and Libao Zhang. Slvr: Super-light visual reconstruction via blueprint controllable convolutions and exploring feature diversity representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 400–410, June 2025.

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. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *Advances in Neural Information Processing Systems*, 38:4646–4668, 2026.

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, 2026.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning, 2026. URL https://arxiv.org/abs/2601.10129.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

Tianrun Xu, Yue Sun, Qixun Wang, Jingyi Lu, Yuan Wang, Tianren Zhang, Longteng Guo, Fengyun Rao, Jing Lyu, Feng Chen, and Jing Liu. Semantic-enriched latent visual reasoning. *arXiv preprint arXiv:2605.19342*, 2026. URL https://arxiv.org/abs/2605.19342.

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.

Yanyu Yao. A survey on multimodal large language models. In *Computer Science Undergradaute Conference 2025@ XJTU*, 2025.

A Appendix

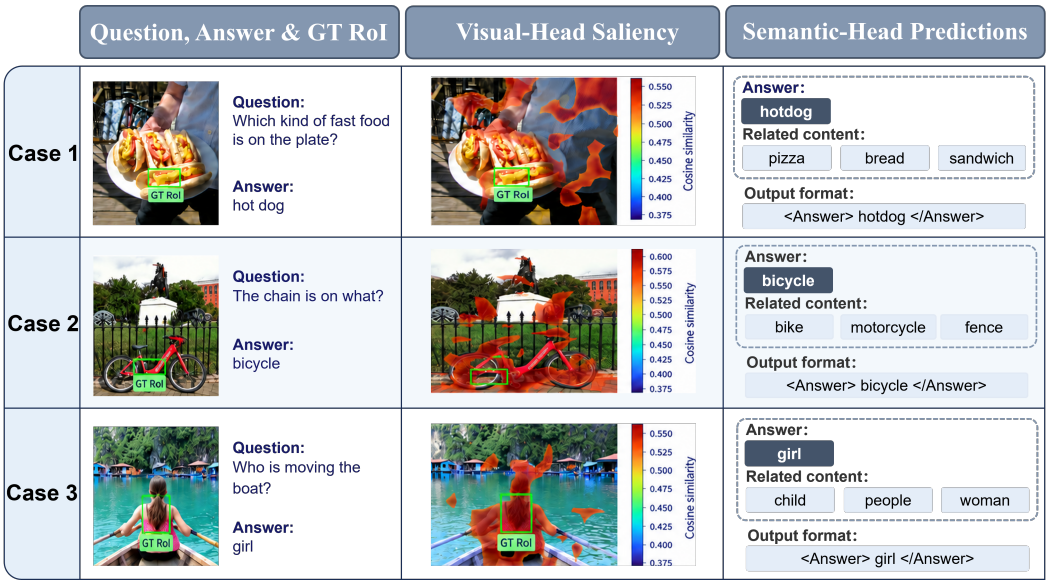

Figure 6: Case analysis. We present three cases to analyze the visual and semantic heads. The visual head highlights relevant regions via cosine similarity, while the semantic head decodes answer-related semantics and output formats.