Leveraging Latent Visual Reasoning in Silence

Dongyao Zhu[1†] **Zhen Wang**[3] **Xi Xiao**[4] **Han Jiang**[5] **Saeed Vahidian**[6]
Wei-Lun Chao[7] **Tanya Berger-Wolf**[2] **Yu Su**[2] **Raju Vatsavai**[1] **Jianyang Gu**[2†]

[1]North Carolina State University [2]The Ohio State University [3]UC, San Diego
[4]UAB [5]JHU [6]Duke University [7]Boston University
† dzhu6@ncsu.edu gu.1220@osu.edu

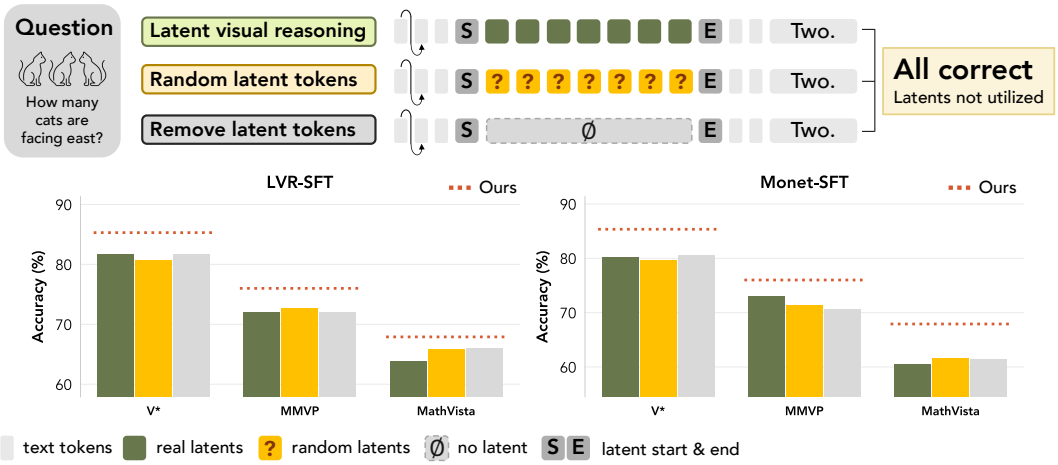


Figure 1: When standard latent tokens (green) are replaced with random values (yellow) or completely removed (grey), latent visual reasoning models retain comparable or achieve even higher performance across benchmarks. We demonstrate that latent reasoning may have limited value as a mandatory inference-time format, but can still be leveraged **in silence** to improve learning (red).

Abstract

Latent visual reasoning involves visual evidence more directly in multimodal reasoning by inserting continuous latent tokens before textual generation. However, the necessity of these latent tokens at inference remains ambiguous. We show that replacing latent tokens with random noise or removing them completely causes little performance degradation across spatial reasoning benchmarks. Reinforcement learning further diminishes the latent generation behavior after post-training. These observations raise a central question: *Is latent visual reasoning still meaningful*? We argue that its value should be measured by how effectively latent tokens guide learning, rather than whether they persist as an inference-time format. Our analysis shows that latent reasoning is unevenly favorable across question types, yet hard task-level routing for applying latent generation is brittle. Motivated by these findings, we propose an attention-based reward that encourages generated latent tokens to interact with later text tokens during RL. This reward promotes latent utilization when the latent mode is activated while preserving the flexibility to use pure-text reasoning. Experiments show that our method improves performance across perception and visual reasoning benchmarks, even when latent tokens are rarely generated after post-training. Our results highlight that, without explicit expression at inference, latent visual reasoning can shape better visual grounding and more accurate textual reasoning **in silence**. Code & model: .

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Latent visual reasoning offers an appealing alternative to purely text-based reasoning in multimodal models [14, 34, 31]. This paradigm allows models to generate continuous latent embeddings before discrete textual generation. These embeddings are usually optimized to capture visual information from prompt-related regions or auxiliary visual inputs such as depth maps [14, 34]. Many visual reasoning tasks depend on spatial evidence that can be difficult to express fully in language [18, 30, 17, 7, 39, 38]. If properly leveraged, latent tokens could summarize visual information more directly and improve the grounding of multimodal reasoning. This work seeks an opposite direction and asks whether latent visual reasoning can be useful even when it is not preserved as an explicit inference-time behavior. We refer to this possibility as leveraging latent visual reasoning "**in silence**."

This view is motivated by a basic ambiguity in current latent visual reasoning models: latent generation alone does not guarantee latent usage [34]. Building on this observation, we revisit latent dependence across representative models such as Monet [34] and LVR [14] under perturbations to the latent tokens. Applied perturbations include replacing the generated latent tokens with random noise before subsequent text generation or removing them completely. Across popular visual reasoning benchmarks [4, 17, 18], Figure 1 shows that these perturbations cause little performance degradation. The stable performance suggests that latent visual reasoning models may learn the format of latent generation, but rely only weakly on the generated tokens for the final answer.

This ambiguity becomes more pronounced during post-training through reinforcement learning (RL). We apply standard correctness-rewarded GRPO [9] to latent visual reasoning models and find that RL tends to diminish latent reasoning. As training proceeds, models increasingly avoid generating latent tokens and eventually converge toward pure-text reasoning. This behavior is counterintuitive, yet aligned with our earlier latent-dependence analysis. If latent tokens do not consistently improve the final answer, an outcome-oriented reward will provide little incentive to preserve them. To test whether this collapse is due to the lack of an explicit latent-related objective, we design a diagnostic latent necessity reward that is positive only when the response with latent is more accurate than its no-latent counterpart. During RL, this reward quickly drives the model to skip latent generation for most questions while maintaining response quality (§3.2). This further suggests that latent visual tokens may have limited value as a mandatory inference-time format.

These findings raise the central question of our work: *if latent tokens can often be removed without hurting performance, and RL tends to eliminate them, what is the value of latent visual reasoning*? We argue that the focus should shift from latent presence to latent utilization, meaning latent tokens should influence subsequent textual reasoning. Even when the final model does not always emit latent tokens, latent visual reasoning may still guide the learning process **in silence**. To further investigate the underlying mechanism, we compare the performance of reasoning with latents and pure text across different question types. We find that latent reasoning and pure-text responses are unevenly favorable for different questions and model families (Figure 3). While this makes identifying task type and applying corresponding objectives challenging, it motivates a more practical approach: improve the utilization of latent tokens without forcing latent generation on every example.

To this end, we propose an attention-based reward for latent visual reasoning during RL. Our intuition is that a useful latent token should effectively influence the tokens generated after it. We therefore use the attention from subsequent text tokens to latent tokens as an efficient proxy for latent influence over next-token prediction (§4). When combined with RL, this reward strengthens the interaction between latent tokens and textual reasoning. Empirically, models trained with this reward attend substantially more often to generated latent tokens and achieve improved final performance, even if latent generation still decreases later in training (§5.2). This outcome echoes our hypothesis that latent reasoning can improve learning in silence. We further study model behavior through attribution analyses (§5.3) and qualitative examples (§5.4) to interpret why our model performs better without generating latents. Results suggest that latent reasoning is useful less as a persistent response format and more as a training-time scaffold that shapes the learning process.

Through this study, we transition our focus from whether latent reasoning is necessary to whether latents are appropriately leveraged. While the presence of latents is not a reliable indicator of better performance, the attention reward can be used as an effective post-training signal. Our results suggest that the future of latent visual reasoning depends on training objectives that make latent computation influential during learning, even when its benefits are ultimately expressed in silence.

2 Related Work

Explicit reasoning in multimodal models. A prominent line of work improves multimodal reasoning by introducing explicit, interpretable intermediate steps. Originally introduced in the text domain, chain-of-thought prompting [36] demonstrates that eliciting step-by-step natural language reasoning substantially improves complex problem solving. In the visual domain, explicit reasoning takes several forms: textual verbalization of visual content [2, 11], active tool use such as grounding, zooming, cropping, and contouring [45, 3, 19, 44, 1], and generation of new visual information via code execution or external vision models such as depth estimators and segmentation networks [12, 15, 37]. A separate but related trend develops unified architectures that natively interleave text, image, and audio within a single model, typically featuring unimodal tokenizers [27, 29, 15]. In contrast to these approaches that externalize reasoning into human-interpretable modalities, our work studies reasoning in latent space, where intermediate computation is not directly observable.

Latent visual reasoning and reinforcement learning. Recent work also explores reasoning in continuous latent space. Some approaches feed predicted embeddings directly as inputs for subsequent steps [10]. Others generate free-form latent representations before decoding [24, 31]. A complementary direction constrains latent computation to a finite set of abstract tokens [21]. In the multimodal setting, the output hidden states are usually trained to carry related visual information [14, 34] before interleaving with text. These representations are often referred to as visual latents. Subsequent work improves the helpfulness of visual latents via task-agnostic training [14] or by coupling visual latents with external vision foundation models [20]. Despite these efforts, these methods implicitly assume that latent representations are both generated and meaningfully utilized during reasoning, especially in model post-training, which is typically based on RL [9]. Standard correctness/format rewards supervise final answers but not intermediate steps. While Monet [34] addresses this issue with their VLPO variant, we show that optimizing for latent generation does not guarantee its actual utilization during reasoning. We note concurrent work [16] that addresses underutilized visual latents by removing them entirely and reverting to explicit intermediate representations such as captioning. In contrast, we show that visual latents are not necessary at inference but can still shape learning via utilization signals **in silence**.

3 Latent Generation is Unnecessary as an Inference-Time Behavior

Latent visual reasoning augments multimodal language models with continuous latent embeddings for intermediate visual computation [14, 34, 31]. Instead of expressing the entire reasoning process through discrete language tokens, a latent visual reasoning model can switch into a latent mode by generating a special signal token, such as `<latent_start>`, producing a span of continuous `<latent>` tokens, and then return to textual generation after `<latent_end>`. These latent tokens are usually designed to encode task-relevant visual information, either from prompt-related regions in the original image or from auxiliary visual inputs such as depth maps [14, 34]. The resulting response contains both ordinary text tokens and latent tokens. Previous methods produce latent tokens that improve performance on visual reasoning tasks over text-only baselines. This section investigates the behavior of these models during inference and after RL post-training.

3.1 Latent Dependence During Inference

We first examine whether the latent visual reasoning models actually rely on their generated tokens during inference. Specifically, we apply perturbations to the generated `<latent>` tokens and evaluate the answer correctness. This question is central to understanding the role of these latents, which are expected to provide grounded visual information for subsequent textual reasoning. To isolate the effect from post-training dynamics, we conduct this analysis on latent reasoning models trained with supervised fine-tuning (SFT) before RL. All experiments use a temperature ≈ 0.0, following common evaluation settings [6]. We apply the following two perturbations:

Random replacement. Models can generate k `<latent>` tokens before `<latent_end>`. However, before the subsequent generation step, we replace latent tokens with random values drawn from a unit Gaussian distribution. Positional embeddings are kept unchanged. We set k to 8 for LVR and 10 for Monet, following the settings used in their best-reported performance.

Table 1: The ratio (%) of latent generation triggered during inference, and the accuracy (%) of responses with latent generated or with pure text on the ViRL39K and Thyme-RL datasets. RL significantly reduces the ratio where latent mode is triggered.

Model	Temp.	ViRL39K			Thyme-RL		
		Latent Ratio	Latent Acc	Text Acc	Latent Ratio	Latent Acc	Text Acc
LVR-SFT*	0.5	34.3	40.2	41.2	33.4	36.6	40.1
	1.0	37.5	31.4	33.4	37.5	30.5	35.2
Monet-SFT	0.5	36.1	42.3	41.1	69.2	36.4	40.0
	1.0	18.9	42.6	34.8	40.7	34.4	34.3
Monet-RL	0.5	10.0	60.8	41.1	55.0	30.9	36.7
	1.0	5.4	53.5	35.7	32.4	28.7	32.2

Removing latent tokens. In this setting, models can still predict `<latent_start>`, but we force the token `<latent_end>` as the next token, skipping the continuous latent input. Compared with random replacement, this setting tests whether the presence of `<latent>` tokens is needed at all.

As shown in Figure 1, we observe that random replacement performs on par with model-generated latent tokens. Even though removing latent tokens is out of the training distribution, it is better than using them in many cases. These results suggest that `<latent>` tokens can often be removed without hurting textual reasoning. In other words, the generated tokens are not consistently necessary for producing the final answer. However, these observations do not imply that latent visual reasoning should be discarded. First, in certain cases, such as MMVP [40], responses still benefit from using visual latent tokens. Second, prior work reports that models trained under the latent visual reasoning paradigm outperform those trained with pure textual reasoning on the same data [34, 14]. Although not universally required at inference, latent visual reasoning may still provide value in specific contexts or during training. Motivated by this distinction, we further investigate the latent visual reasoning behavior after different training stages.

3.2 Latent Generation Behavior

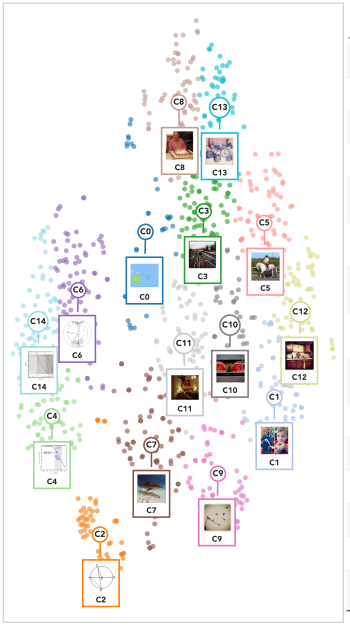
We examine whether models prefer latent generation to pure-text reasoning across different training stages, particularly after RL post-training. This analysis complements our previous perturbation study. If latent tokens are not consistently necessary for final-answer accuracy, an outcome-oriented RL objective may gradually reduce the use of latent mode. We therefore measure how often latent generation is triggered in SFT models and RL-post-trained models.

For Monet [34], we directly use the released SFT and RL models. For LVR [14], we use a retrained variant as the released LVR-SFT model is strongly constrained to follow the latent-reasoning format and always enters latent mode. To reduce this format bias, we add an additional SFT stage for 200 steps using the Visual-COT [23] dataset with latent targets randomly dropped. This exposes the LVR model to both latent and pure-text formats and allows it to choose between them during inference while maintaining performance. We denote the resulting model as LVR-SFT* hereafter. The detailed training information and model performance of LVR-SFT* are listed in §C.

We randomly sample $1,000$ questions from Thyme-RL [43] and ViRL39K [33]. For each question, we collect 8 responses from each model at temperatures 0.5 and 1.0, and summarize the ratio of triggering latent generation and the accuracy. As shown in Table 1, 10 out of 12 settings yield a latent-mode trigger rate below 50%, and the RL model shows a clear trend toward pure-text reasoning. We hypothesize that RL optimization encourages models to move away from the latent-reasoning format when latent tokens do not consistently improve the final answer. As a result, Monet-RL retains only a small amount of latent generation that leads to improved accuracy.

Training dynamics. We further reproduce RL post-training to study how this latent diminishing behavior emerges. Specifically, we train Monet-SFT on the Thyme-RL dataset using VLPO [34]. Due to computational constraints, this empirical study uses lighter configurations than those officially reported, as detailed in §D.

As shown in the first column of Figure 2, the ratio of responses that trigger latent mode decreases as RL training progresses. The reasonably high initial trigger ratio indicates that the model can enter latent mode at the start of training. The subsequent decrease therefore cannot be explained solely by



Cluster & sample question	Monet-SFT Acc. % (Latent)	Monet-SFT Acc. % (Text)	LVR-SFT Acc. % (Latent)	LVR-SFT Acc. % (Text)
C0: Which country is highlighted?	**40.2** (n=229)	32.0 (n=75)	52.3 (n=109)	**52.8** (n=195)
C1: Is the chair to the left or right of the remote?	**43.1** (n=408)	23.2 (n=112)	37.5 (n=192)	**42.1** (n=328)
C2: Choices: (A) 60° (B) 50° (C) 70° (D) 80°	21.1 (n=190)	**23.4** (n=346)	30.0 (n=200)	**32.7** (n=336)
C3: What is the color of the cup?	**36.2** (n=616)	30.4 (n=56)	53.6 (n=239)	**53.8** (n=433)
C4: What is the minimum number of edges that must be removed to disconnect all the red nodes from each other?A...	**42.9** (n=343)	39.1 (n=225)	42.0 (n=150)	**45.0** (n=418)
C5: Is the green partial on the left side of the woman in blue?	**26.5** (n=521)	21.5 (n=79)	13.5 (n=222)	**14.6** (n=378)
C6: The term multi tasking refers to which two activities seen here?	**63.2** (n=261)	63.0 (n=427)	61.6 (n=185)	**66.4** (n=503)
C7: What can Nathan and Reggie trade to each get what they want?	**46.5** (n=419)	40.4 (n=245)	34.3 (n=213)	**34.4** (n=451)
C8: Does the bag near the woman seem to have a brown color?	29.3 (n=505)	**38.2** (n=55)	24.9 (n=209)	**27.9** (n=351)
C9: What kind of birds are these?	**20.9** (n=282)	16.4 (n=134)	32.3 (n=155)	**33.7** (n=261)
C10: Who sponsors this player?	**31.5** (n=184)	19.1 (n=136)	25.9 (n=116)	**27.0** (n=204)
C11: What is the notebook on?	**37.4** (n=462)	36.0 (n=50)	35.7 (n=140)	**43.0** (n=372)
C12: On which side are the glasses?	35.4 (n=463)	**44.1** (n=145)	**30.7** (n=228)	28.2 (n=380)
C13: What is the color of the bars?	**25.1** (n=398)	23.5 (n=34)	**22.2** (n=162)	18.9 (n=270)
C14: What operation is performed at the end of the qubit lines |s_2>, |s_3>, and |s_4>?A. Pauli-X gate B. Quantum...	53.3 (n=255)	**59.7** (n=345)	**61.7** (n=154)	58.1 (n=446)
Overall	36.4 (n=5536)	**40.0** (n=2464)	36.6 (n=2674)	**40.1** (n=5326)

Figure 3: Accuracy comparison between using latent reasoning and text only on various question types. **Bold**: the better result between these two modes.

insufficient exploration of the latent mode. A more plausible possibility is that latent-mode responses are less favorable under the RL objective.

To make this interpretation more explicit and to avoid the potential reason of lacking latent-specific rewards, we further evaluate latent necessity during RL. We define latent tokens as *necessary* when keeping them yields a more accurate final response than skipping them from the same prefix before `<latent_start>`. Based on this definition, we introduce a diagnostic latent necessity reward that is positive only when the response with generated latents is more accurate. More details on the latent necessity calculation and resulting model performance can be found in §E. As shown in Figure 2 (smoothed every 10 steps), this reward drives the model to skip latent generation even faster than vanilla RL while maintaining response quality. This result supports the conclusion that latent tokens are not consistently useful as a mandatory format at inference time. It also motivates the next step of our study: rather than preserving latent generation itself, post-training should focus on whether the latent tokens are effectively utilized.

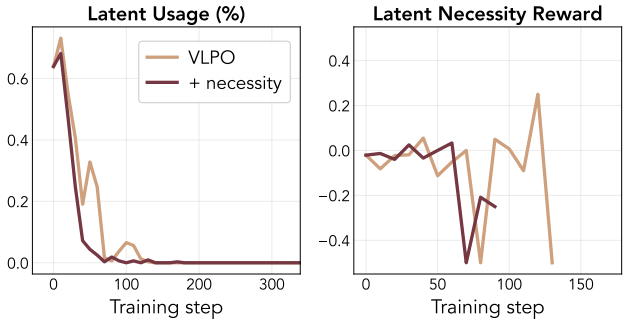


Figure 2: After adding the necessity reward, the model discards latent generation even faster than the VLPO baseline, without an increase in the reward value.

3.3 Where Does Latent Reasoning Help?

The previous section shows that RL tends to reduce latent generation, suggesting that latent tokens are not consistently useful as a mandatory inference-time format. However, this does not imply that latent reasoning is never useful. We therefore study how responses with and without latent tokens behave across different types of questions. We construct approximate groups by clustering question embeddings using Sentence-BERT [22]. The left panel of Figure 3 visualizes the resulting clusters for Thyme-RL, together with representative questions from each cluster in the middle panel. These clusters cover a mixture of question patterns, including spatial relations, object attributes, chart-like reasoning, mathematical reasoning, and general visual recognition.

The cluster-level results in Figure 3 reveal a heterogeneous pattern. For Monet-SFT, latent responses outperform pure-text responses in several clusters involving local visual evidence or spatial/object attributes, whereas pure-text responses are stronger in other clusters. For LVR-SFT*, the trend is different: pure-text responses are more competitive in most clusters. Therefore, a single coarse question type cannot explain latent preference, and it is not consistent across model families.

This analysis refines our interpretation of latent visual reasoning. Latent tokens can be useful, but their usefulness is conditional on various factors, including the model, the dataset, and the specific structure of the question. Some examples benefit from latent computation, while others are better handled by direct textual reasoning. This makes a hard routing of question type difficult, as deciding in advance which task category should use latent reasoning would be brittle. However, they motivate a softer post-training objective. Rather than forcing latent generation on every example or manually assigning question types to reasoning modes, we should encourage generated latent tokens to interact with subsequent textual reasoning whenever the latent mode is activated.

4 Leveraging Latent Visual Reasoning in Silence

Compared with the presence of latent generation, we identify that the key issue is whether generated latent tokens are actually used by subsequent textual reasoning. We aim to leverage latent visual reasoning as a post-training signal without forcing every response to contain latent tokens. The method design is based on a simple intuition from decoder-only transformers, where later tokens interact with earlier tokens through multi-head attention [32]. If the generated latents contribute to the final answer, the text generated after `<latent_end>` should attend to the `<latent>` tokens. We use this interaction as an efficient proxy for latent utilization. Without requiring a task-level decision about when latent reasoning should be used, this reward encourages latent tokens to be influential when the model chooses to generate them.

Latent attention reward. Let $\mathbf{o}_i = \pi_\theta(q)$ be a response produced by a latent visual reasoning policy π_θ given prompts q. When the response triggers latent mode, it has the format where textual reasoning tokens follow `<latent_end>`. Let T_L denote the indices of the `<latent>` tokens, and T_t denote the indices of subsequent text tokens. Let $A_{j,k}$ denote the attention weight from token j to token k in the first decoder layer, averaged over all attention heads, where $k < j$ under causal attention. We define the latent attention reward as the average attention mass assigned by tokens in T_t to tokens in T_L, normalized by the number of post-latent text tokens:

$$R_A^i = \frac{1}{|T_t|} \sum_{j \in T_t} \sum_{k \in T_L} A_{j,k}. \tag{1}$$

We compute the attention score at the first decoder layer, where text tokens attend directly to latent tokens before information is further mixed in deeper layers. If a response does not contain generated latents, we set the latent attention reward to zero. Empirically, we do not observe excessive attention on latent tokens (Figure 4 right), so no clipping is applied to R_A.

For VLPO [34], we add the latent attention reward to the standard outcome rewards r_i before calculating group advantage $\hat{A}_i$. Specifically, the scalar reward for response $\mathbf{o}_i$ is:

$$r_i = w_1 R_{\text{acc}}(\mathbf{o}_i) + w_2 R_{\text{format}}(\mathbf{o}_i) + w_3 R_A^i, \tag{2}$$

where R_{acc} is the correctness reward, R_{format} is the format reward to make sure the final answer is wrapped in "`\boxed{}`" [34], and R_A^i is the latent attention reward. The group-normalized advantage is then computed as: $\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}$. The final objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{q,o \sim \pi_{\text{old}}} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \min \left[r_{i,t}(\theta) \hat{A}_i, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right) \hat{A}_i \right] - \beta \, \text{KL}\left(\pi_\theta \| \pi_{\text{ref}}\right). \tag{3}$$

Following VLPO, the token-level ratio $r_{i,t}(\theta)$ is defined differently for continuous latent tokens and discrete text tokens:

$$r_{i,t}(\theta) = \begin{cases} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{h}_{i,t}^{\text{old}} - \mathbf{h}_{i,t}^{\theta}\|^2\right), & \text{if } t \in T_L, \\ \frac{\pi_\theta(\mathbf{o}_{i,t} | q, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(\mathbf{o}_{i,t} | q, \mathbf{o}_{i,<t})}, & \text{if } t \notin T_L. \end{cases} \tag{4}$$

Here, $\mathbf{h}_{i,t}^{\text{old}}$ and $\mathbf{h}_{i,t}^{\theta}$ denote the latent representations produced by the old and current policies, respectively, and $\mathbf{o}_{i,<t}$ denotes the prefix before position t. In practice, we use $w_1 = 0.9$, $w_2 = 0.1$, and $w_3 = 1.0$ to ensure that all terms are of comparable magnitude within the interval $[0, 1]$.

This design supports the central goal of leveraging latent visual reasoning "**in silence**." The attention reward encourages generated latent tokens to interact with subsequent textual reasoning during RL, allowing the model to extract useful visual grounding signals from the latent mode. It places no requirement on the final model to preserve latent generation as a fixed inference-time format.

5 Experiment Results

5.1 Settings

We evaluate whether the latent attention reward can improve post-training when added to RL objectives. We continue training Monet-SFT and LVR-SFT* on Thyme-RL with VLPO. Due to computational constraints, the models are trained with a batch size of 4 for 600 and 400 steps, respectively. More training hyperparameters, including learning rate and KL coefficient, are provided in §D.

We evaluate the trained models on various multimodal benchmarks, including MathVista-Mini (MathVista) [18], MathVerse-Mini (MathVerse) [41], MMVP [40], MMBench-V11-Mini (MMBench) [17], MME-RealWorld-Lite (MME-RW) [42], V* [4], and HRBench4K [35], comprehensively covering math reasoning, spatial reasoning, and fine-grained visual perception tasks. We use VLMEvalKit [6] with Gemini-2.5-Pro [5] / Gemini-3.1-Pro [8] as a judge following prior work [34].

5.2 Main Results

We first analyze how the attention reward changes latent behavior during RL. The right panel of Figure 4 shows the attention from subsequent text tokens to generated latent tokens for samples that trigger the latent mode, where adding R_A substantially increases this attention mass. At the end of training, this interaction is approximately $7\times$ higher than vanilla VLPO. This indicates that the reward successfully encourages generated latent tokens to participate in textual reasoning. The left panel of Figure 4 shows that the model also triggers the latent mode more frequently under the attention reward, allowing extended time to exploit potential knowledge from visual latents. As training progresses, the model gradually prunes low-value latent usages and retains only highly interactive latent responses before eventually internalizing the learned behavior into pure-text generation. Eventually, the latent usage ratio converges to zero.

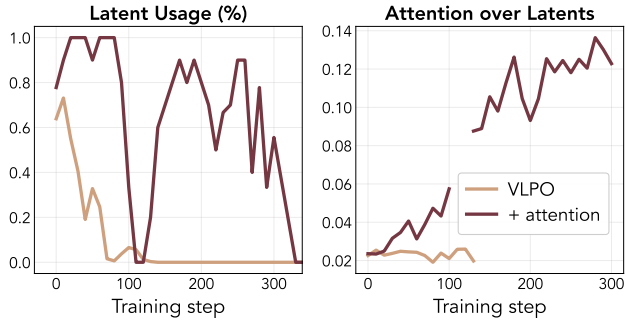


Figure 4: Attention reward encourages visual latents more attended to by subsequent textual reasoning. The gap in attention corresponds to the lack of latents around step 100.

Table 2 reports the main benchmark results. For both Monet and LVR, adding the attention reward improves the SFT model on most metrics. The largest gains are on visually grounded benchmarks, especially V* spatial reasoning (+9.2, +6.6) and HRBench4K-FSP (+6.0, +3.7). The reward also improves the performance of Monet on MathVista by 7.4. These results suggest that the attention reward helps the base model better exploit visual evidence during post-training. Our method shows even larger gains on MathVista (+8.0) and MMVP (+10.3) than VLPO for the Monet baseline, demonstrating the effectiveness of the attention reward in improving visual reasoning capabilities.

Overall, the results support the claim that latent generation alone is not a reliable indicator of better performance, and standard RL may suppress latent tokens when they are not consistently necessary. The attention reward encourages latent tokens to influence textual reasoning more softly than forcing the latent generation for every sample. While the latent mode is still diminished at the end of training, the RL training helps the model shape more accurate visual-related textual reasoning in silence.

Table 2: **Performance on perception and reasoning benchmarks**. Results with "†" are reported by other papers, otherwise evaluated by us. **Bold** entries denote that our method yields the best performance among the model family. ↑ denote our improvement compared to the SFT base model.

Model	MathVista	MathVerse	MMVP	MMBench	MME-RW	V*			HRBench4K		
	Overall	Overall	Overall	Overall	Overall	Overall	Attr.	Spat.	Overall	FSP	FCP
GPT-4o	–	–	–	–	52.0†	67.5†	72.2†	60.5†	59.0†	70.0†	48.0†
Qwen2.5-VL-7B	68.2	48.0	76.7	84.3	45.8	76.4	77.4	75.0	69.3	85.3	53.3
Deepeyes	–	–	–	–	54.3†	83.3†	84.4†	81.6†	71.3†	83.8†	58.8†
SkiLa	–	–	75.3†	83.3†	56.6†	84.3†	–	–	72.0†	–	–
CoVT	–	–	–	–	63.3†	78.5†	–	–	72.5†	–	–
LVR-SFT w/o latents	–	–	72.0†	–	–	79.1†	82.6†	73.7†	–	–	–
LVR-SFT*	63.0	36.8	75.7	77.8	50.4	82.7	85.2	78.9	70.5	80.8	60.3
LVR-SFT* + VLPO	61.5	37.3	76.7	83.3	50.8	81.7	82.6	80.3	71.5	83.3	59.8
LVR-SFT* + VLPO + Attn (Ours)	**63.7**	36.3	**77.7**	82.4	**50.8**	**85.9**	**86.1**	**85.5**	**71.9**	**84.5**	59.3
Δ vs. SFT	↑ 0.7	↓ 0.5	↑ 2.0	↑ 4.6	↑ 0.4	↑ 3.2	↑ 0.9	↑ 6.6	↑ 1.4	↑ 3.7	↓ 1.0
Monet-SFT w/o latents	–	–	–	–	51.3†	81.7†	83.5†	79.0†	68.4†	78.3†	58.5†
Monet-SFT w/o latents + GRPO	–	–	–	–	52.4†	78.5†	78.3†	79.0†	70.0†	83.3†	56.8†
Monet-SFT	60.5	35.1	73.0	74.1	52.7	80.1	82.6	76.3	68.5	79.5	56.5
Monet-SFT + VLPO	59.9	36.5	65.7	75.9	55.0	81.2	81.7	80.3	70.4	81.3	59.3
Monet-SFT + VLPO + Attn (Ours)	**67.9**	**39.9**	**76.0**	**77.8**	52.8	**85.3**	**85.2**	**85.5**	**71.1**	**85.5**	56.8
Δ vs. SFT	↑ 7.4	↑ 4.8	↑ 3.0	↑ 3.7	↑ 0.1	↑ 5.2	↑ 2.6	↑ 9.2	↑ 2.6	↑ 6.0	↑ 0.3

Table 3: **Ablation on latent type in attention reward**. Using random latents in VLPO degrades performance compared with real latents. Both models are trained for 400 steps.

Reward Type	MathVista	MMVP	MMBench	V*	HRBench4K
Rand. Lat. + R_A	60.9	73.0	75.0	80.6	65.6
Real Lat. + R_A	**62.9**	**74.3**	**80.6**	**85.3**	67.9

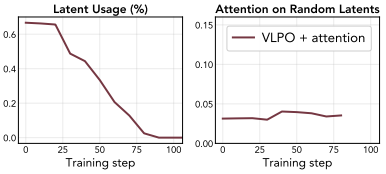


Figure 5: **Attention on random latents during VLPO training**.

5.3 Analysis

Necessity of latent tokens for attention reward. To verify that the benefit of R_A comes from meaningful visual latents, we conduct an ablation that replaces real latent embeddings with random values during RL rollout. Following §3.1, each generated latent embedding is replaced with Gaussian noise before subsequent token generation. We then use VLPO + R_A to post-train Monet-SFT. As shown in Figure 5, the latent mode collapses within the first 100 steps. Meanwhile, the attention assigned to random latents remains low, indicating that the model does not exploit R_A by attending to uninformative latent tokens. This addresses a potential reward-hacking concern, where the reward could be satisfied by arbitrary latent attention.

We also report the performance comparison of using random latents vs. real latents in Table 3. The results show that training with real latents consistently outperforms training with random latents across all five benchmarks, with an average gain of +3.2 points. The attention reward gradually improves visual grounding during post-training. The effect cannot be equally achieved with random latents. Through this ablation study, we confirm that the effectiveness of the attention reward comes from interaction with meaningful visual latent representations.

Attribution analysis for the attention reward. The proposed attention reward uses attention to latent tokens as a signal to indicate the use of generated latents. We use two complementary gradient-based attribution analyses to examine whether latent representations with higher attention values also influence the probability of the final answer more strongly.

Specifically, we save checkpoints at training steps 0, 30, and 60, before latent generation disappears under VLPO. We evaluate responses containing latent tokens on 300 held-out questions sampled from Thyme-RL, MMVP, and V*. For each response, we define the target as the length-normalized log-probability of the final answer span. We then compute attribution from the latent tokens using Gradient x Input (GxI, [26]) and Integrated Gradients (IG, [28]). Larger values indicate that the answer likelihood is more sensitive to the generated latent representations. Table 4 shows that as training progresses, our model generates latents that are more important for producing final answers than the original SFT model (Step 0) and the VLPO model. This trend aligns with the attention change in Figure 4 for both VLPO and our model.

Table 4: Attribution scores across training steps. GxI are scaled by 10^3 and IG by 10^2.

Model	Step 0		Step 30		Step 60	
	GxI	IG	GxI	IG	GxI	IG
Ours	0.62	3.63	**−0.35**	**7.57**	**10.56**	**20.95**
VLPO	0.62	3.63	−7.19	7.22	0.45	3.31

Table 5: Correlation between attention scores and attribution scores.

Model	Step 0		Step 30		Step 60	
	GxI	IG	GxI	IG	GxI	IG
Ours	−0.073	0.093	**0.160**	**−0.050**	**0.247**	**0.301**
VLPO	−0.073	0.093	0.032	−0.082	0.062	−0.026

We next compute the correlation between attention scores on latents and the attribution scores above. The results in Table 5 indicate that the correlation between attention scores and attribution scores is increasingly positive for our method. This supports the idea that the proposed reward drives training to identify latent representations that help derive the correct answers.

Source of gain for latent-free final models. We also probe the final model's behavior to understand its better performance. Because the final model emits no latents, we measure what its pure-text inference actually uses, applying the same gradient-attribution analysis as above to image tokens with Gradient $\times$ Input and Integrated Gradients. An image contributes hundreds of tokens, but most are irrelevant to any given query. We expect the model to focus on discriminative regions to derive the answer. Accordingly, we report the normalized entropy of per-image-token attribution, where lower values mean a smaller subset of the image tokens influence the output more significantly.

Table 6: Entropy of per-image-token gradient attribution before (Step 0) and after training (Step 600).

Model	Step 0		Step 600	
	GxI	IG	GxI	IG
Ours	0.859	0.898	**0.849**	**0.880**
VLPO	0.859	0.898	0.854	0.891

Table 7: Alignment of attribution with ground-truth important regions (AUC).

	AUC
SFT	0.654
VLPO	0.639
Ours	**0.669**

As shown in Table 6, image token attribution in our model is more concentrated (lower entropy) than both the final VLPO model and the initial SFT model (Step 0), and the reduction from initialization is roughly twice VLPO's (-0.010/-0.018 vs -0.005/-0.007). Crucially, total attention mass over image tokens is matched between the two models (≈ 0.025). Therefore, the two models differ mainly in how selectively they use image content to derive the answers. We hypothesize that the attention reward enables the model to focus on discriminative image regions during text reasoning.

To test whether attribution is focusing on the relevant part of the image rather than being randomly concentrated, we evaluate it against the ground-truth important regions annotated in 200 held-out samples from Visual-CoT [23]. We treat the per-image-token attribution as a score for the binary label "this token lies inside the annotated region" and report scale-invariant ROC-AUC, *i.e.*, the probability that a randomly drawn in-region token receives higher attribution than a randomly drawn out-of-region token, as the metric for alignment of attribution with relevant image regions.

Table 7 shows that our model ranks the annotated regions highest, while VLPO yields worse alignment than the shared SFT initialization. This is consistent with the qualitative pattern in §G in the main paper, where Monet-VLPO produces fluent reasoning that drifts from the image. Together, these results show that our final model reads the same amount of the image but relies more selectively on the regions that matter, a property that is beyond latent generation and plausibly explains the retained gain after latent disappears.

Why is latent token generation discarded during RL training? During training, latent responses are no more accurate than pure-text responses (Table 1), and since the correctness of the answer still dominates the objective (Equation 2). Thus, latent generation eventually diminishes.

5.4 Case Study

The case studies in Figure 6 illustrate how our method improves visual grounding during inference. In the HRBench4K spatial reasoning example, Monet-VLPO produces a long reasoning chain but fails to maintain a consistent spatial relation, leading to a wrong answer. Our model directly identifies the relation between the black car and the tram and predicts the correct answer. In the counting

Inference Example: spatial reasoning (from HRBench4K)

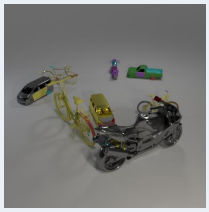


Question Prompt:
What is the position of the black car relative to the tram? Options: A. To the left of the tram B. In front of the tram C. Behind the tram. D. To the right of the tram (Answer: C)

Response by Monet-VLPO:
First, locate the black car in the image. Then, identify the tram and its position relative to the black car. The black car is positioned to the left of the tram. The tram is in front of the black car. The black car is behind the tram. The black car is to the right of the tram. The final answer is \\boxed{A}.

Response by Ours:
The black car is positioned behind the tram in the image. Therefore, the final answer is \boxed{C}.

Inference Example: counting (from MathVista)



Question Prompt:
Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Are there fewer large matte utility bikes than small yellow bicycles? Options: A. Yes B. No (Answer: A)

Response by Monet-VLPO:
To determine if there are fewer large matte utility bikes than small yellow bicycles, I will count the number of each type of bicycle present in the image.1. I will identify and count the large matte utility bikes.2. I will identify and count the small yellow bicycles.3. I will compare the two counts to determine the relationship between the quantities. After counting the bicycles in the image:- There is 1 large matte utility bike (the yellow one in the foreground).- There are 2 small yellow bicycles (one in the background and one in the foreground, slightly smaller than the matte bike).The count shows that there is 1 large matte utility bike and 2 small yellow bicycles. Therefore, there are not fewer large matte utility bikes than small yellow bicycles. The final answer is \boxed{B}.

Response by Ours:
The image clearly shows one large matte utility bike and two small yellow bicycles. The answer is \boxed{A}.

Figure 6: Qualitative comparison on reasoning examples. Monet-VLPO reasons longer but drifts from the image evidence. Our model responds concisely and stays visually grounded.

example, Monet-VLPO decomposes the problem into several counting steps, yet its final comparison contradicts the visual evidence. Our model gives a shorter response and correctly counts one large matte utility bike and two small yellow bicycles. These examples suggest that stronger visual grounding can reduce reasoning drift and lead to more concise answers aligned with the image. The conciseness may also relate to the design of the attention reward. In transformers, attention from later text tokens to a fixed early prefix can naturally decrease as the context grows. Longer continuations therefore reduce average attention to latent tokens. This formulation may encourage the model to generate shorter post-latent textual reasoning while maintaining interaction with visual latents. As a result, our model has improved response time as reported in §F.2 (more analyses in §F.3). Overall, these examples support the central finding of our paper: latent visual reasoning does not need to persist as an explicit inference-time format to be useful. Even when the final response is expressed mostly in text, the training signal from latent-token interaction can help the model shape more accurate visual grounding behavior. More inference examples are listed in §G.

6 Conclusion

This paper studies how latent visual reasoning behaves during inference and RL post-training. We show that generated latent tokens are often used only weakly to obtain the correct answer, and that RL drives models to skip the latent mode for most questions. We therefore shift the focus from latent necessity to latent utilization. We propose a latent attention reward that encourages generated latents to influence subsequent textual reasoning during RL. Experiments show that this reward improves performance on visual reasoning benchmarks, especially those requiring stronger visual grounding. Even when latent generation decreases after post-training, the learned policy can still benefit from the training signal provided by latent tokens in silence. We hope this perspective encourages a broader view of latent reasoning as a mechanism for improving visual grounding during learning, even when its benefits are ultimately expressed through ordinary textual responses.

Acknowledgments and Disclosure of Funding

This research is supported in part by grants from the National Science Foundation (OAC-2118240, HDR Institute: Imageomics). We also thank Prof. Ziang Xiao from Johns Hopkins University for his support.

References

[1] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025. 3

[2] Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1947–1963, 2022. 3

[3] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36:70115–70140, 2023. 3

[4] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025. 2, 7

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7

[6] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 3, 7

[7] 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*, pages 148–166. Springer, 2024. 2

[8] Google DeepMind. Gemini 3.1 pro model card, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf. 7

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 3

[10] 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. 3

[11] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2951–2963, 2023. doi: 10.1109/ICCV51070.2023.00277. 3

[12] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024. 3

[13] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 17

[14] Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *ICLR*, 2026. 2, 3, 4, 15, 16

[15] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: multimodal visualization-of-thought. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025. 3

[16] 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*, 2026. 3

[17] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 2, 7

[18] 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. *arXiv preprint arXiv:2310.02255*, 2023. 2, 7

[19] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*, 2024. 3

[20] Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025. 3

[21] Keshav Ramji, Tahira Naseem, and Ramón Fernandez Astudillo. Thinking without words: Efficient latent reasoning with abstract chain-of-thought. *arXiv preprint arXiv:2604.22709*, 2026. 3

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084. 5

[23] 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. 4, 9

[24] 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*, pages 677–693, 2025. 3

[25] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 16

[26] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017. 8

[27] Yutao Sun, Hangbo Bao, Wenhui Wang, Zhiliang Peng, Li Dong, Shaohan Huang, Jianyong Wang, and Furu Wei. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024. 3

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 8

[29] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. *URL https://arxiv. org/abs/2405.09818*, 9(8), 2024. 3

[30] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 2

[31] 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. 2, 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[33] 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. *arXiv preprint arXiv:2504.08837*, 2025. 4

[34] Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*, 2025. 2, 3, 4, 6, 7, 16, 17

[35] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915, 2025. 7

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[37] Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. *arXiv preprint arXiv:2505.19255*, 2025. (Accepted, ICLR 2026). 3

[38] Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, et al. Staying vigil ant: Mitigating visual laziness via counterfactual visual alignment in mllms. In *European Conference on Computer Vision*, pages 287–305. Springer, 2026. 2

[39] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025. (Accepted, ICLR 2026). 2

[40] He Zhang, Shenghao Ren, Haolei Yuan, Jianhui Zhao, Fan Li, Shuangpeng Sun, Zhenghao Liang, Tao Yu, Qiu Shen, and Xun Cao. Mmvp: A multimodal mocap dataset with vision and pressure sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21842–21852, 2024. 4, 7

[41] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 7

[42] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 7

[43] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025. (Accepted, ICLR 2026). 4, 16

[44] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. (Accepted, ICLR 2026). 3

[45] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025. 3

Appendix

The appendix is organized as follows.

- In §A, we discuss the societal impact of this research.
- In §B, we discuss the current limitations of this research as well as promising future directions.
- In §C, we describe the construction of LVR-SFT* with random latent dropping.
- In §D, we provide implementation details for RL post-training, including the lightweight training setup and hyperparameters used for the attention reward.
- In §E, we present the full formulation of the diagnostic latent necessity reward. app-length
- In §F, we report additional empirical results, including the comparison between zero-latent and standard latent rollouts and the inference speed analysis. We also show how controlling response length itself is not necessarily beneficial, while combining with our method can partly mitigate its effect.
- In §G, we provide additional qualitative examples across benchmarks that further illustrate the behavior of our method.

A Broader Impacts

This work improves visual grounding in multimodal reasoning models by better leveraging latent visual reasoning during post-training. Stronger visual grounding can benefit applications that require reliable interpretation of visual evidence. By encouraging models to make better use of intermediate visual representations, our method may reduce unsupported textual reasoning and improve alignment between generated answers and image content.

At the same time, our method is built on existing multimodal models and inherits their general risks, including potential hallucination and dataset bias. Improved visual reasoning could also increase user trust in model outputs, so deployment in high-stakes settings should still involve careful validation and appropriate human oversight. We do not anticipate misuse risks beyond those associated with the base models, although downstream applications should follow the same safety and privacy considerations required for multimodal systems.

B Limitations and Future Work

Our study focuses on two representative latent visual reasoning frameworks, Monet and LVR. These models cover common designs for injecting continuous visual latents into multimodal reasoning, yet future methods may use different latent construction mechanisms. It would be valuable to examine whether the observed RL dynamics and the benefits of the attention reward hold across a broader range of latent reasoning architectures.

Our attention reward uses attention mass as a proxy for latent utilization. This provides an efficient training signal. Future work could combine attention-based rewards with stronger counterfactual measurements of latent contribution. In addition, our experiments are conducted under limited computational budgets, so larger-scale post-training and broader hyperparameter sweeps may further clarify when latent reasoning is most beneficial.

Finally, our results suggest that latent reasoning can improve learning even when latent generation becomes rare at inference time. This raises new questions about how models improve latent-guided behavior and when explicit latent generation should be preserved. Future work may explore adaptive trigger policies, example-dependent latent budgets, and reward designs that better balance latent utilization, response quality, and inference efficiency.

C Details of LVR-SFT* Training

As stated in §3.2, the training samples used in the officially-released LVR-SFT model strictly follow the latent-generation format [14]. Therefore, the SFT model is not flexible in choosing whether to generate latents or not. Therefore, we perform an additional SFT stage with the latent target randomly disabled at a probability of 0.4.

Here we report the training details of the model. It is trained using 2 H100 GPUs, each of 80GB VRAM for 200 steps, with data packing of 4 instances per batch and longest single instance of 4096 tokens. We keep the vision head frozen, visual token range of min=128 and max=5120. We set the learning rate=1e−5, weight decay=0.1, warmup ratio=0.03, and LVR λ=0.1. Part of the performance comparison between the official LVR-SFT and our LVR-SFT* is listed in Table 2. We also report more comparisons on LVR's adopted benchmarks in Table 8. The results show that LVR-SFT* has a comparable performance to the original LVR-SFT [14].

Table 8: Performance comparison between LVR-SFT* compared and original LVR-SFT [14] (Continued). During inference, both latent models use 8 latent tokens, and the temperature is set to 0.

Method	V^*	$V^*_{D.A.}$	$V^*_{R.P.}$	MMVP	BLINK(overall)	Counting	IQ-Test	JigSaw	Relative Depth	Relative Reflect	Spatial Relation
Qwen2.5-VL-7B-Instruct	79.6	81.7	76.3	67.0	58.3	67.5	27.3	51.3	79.0	41.8	**88.1**
LVR-SFT	81.7	84.4	77.6	**72.0**	58.8	**70.8**	29.3	**52.7**	77.4	41.8	86.0
LVR-SFT*	**82.2**	**85.2**	**77.6**	71.3	**59.1**	68.3	**30.7**	52.0	**81.5**	**44.0**	83.2

D Details of RL Post-Training

Light training in §3.2. The official Monet-VLPO [34] is trained on 3.2K samples from Thyme-RL [43] for 1 epoch using a batch size of 64. Due to limited GPU availability (at most 2 GPUs for 8 hours per run), we train with a batch size of 4. This does not affect the size of the rollout group per sample.

Training with attention reward. We use the officially released model and code from Monet [34] and train the Monet-SFT-stage3 model and LVR-SFT* model on Thyme-RL dataset using verl [25], with the following hyperparameters:

Table 9: **Hyperparameters for RL.**

Hyperparameter	Value	Hyperparameter	Value
learning rate	1e−6	batch size	4
weight decay	0.01	rollout size	8
temperature	0.5	max response length	4096
VLPO σ	10.0	latent size	10
max pixels per img	$2000 \times 28 \times 28$	accuracy threshold	0.6
training steps	600 (Monet) 400 (LVR)	accelerator	2 NVIDIA H100-80GB
number of visual latents	10	data type	bfloat16

The reward itself introduces no additional trainable module. It only extracts attention from the first decoder layer during the log-probability computation of the old policy. The practical runtime and peak vram for our method is 82.2 H100 hours and 78.90 GB, compared to VLPO's 82.0 H100 hours and 76.11 GB. The small runtime and memory increases indicate that computing the reward will not be a scaling bottleneck. This supports computational compatibility with larger models.

Due to the substantial compute and API cost of latent-space RL, fully repeating both VLPO and our method across several seeds and architectures exceeds our available budget. The current one-run protocol follows the previous works, where LVR and Monet dont report standard deviation either. Moreover, our conclusion does not depend on a single evaluation set. The models are evaluated on seven benchmarks, and an overall average improvement is observed across two backbones.

E Details of Latent Necessity Reward

Let $\mathbf{o}_i = \pi_\theta(q)$ be a response containing latent tokens produced by a latent visual reasoning policy π_θ given prompts q. $\mathbf{o}_i$ has the format $\mathbf{o}_i$ = `<text1>` ∘ `<latent_start>` ∘ `<latents>` ∘ `<latent_end>` ∘ `<text2>` where `<latents>` denotes latent visual tokens and `<text1>`, `<text2>` denotes text tokens. We let π_θ continue to generate another response given the same prefix before `<latents>`: $\mathbf{o}_j = \pi_\theta(q \circ$ `<text1>` $\circ$ `<latent_start>` $\circ$ `<latent_end>`. $\mathbf{o}_j$ is different from

Table 10: **Performance of Monet-SFT + R_N on perception and reasoning benchmarks**. **Bold** entries denote the best performance.

Model	MathVista	MMVP	MMBench	MME-RW	V*			HRBench4K		
	Overall	Overall	Overall	Overall	Overall	Attr.	Spat.	Overall	FSP	FCP
Monet-SFT	60.5	**73.0**	74.1	**52.7**	**80.1**	**82.6**	**76.3**	68.5	79.5	56.5
Monet-SFT + VLPO + R_N	**71.3**	65.0	**76.9**	51.4	**80.1**	**82.6**	**76.3**	**70.5**	**84.3**	**56.8**

spontaneously generated textual responses in that a pure-text response with the same prefix is highly likely to still complete as $\mathbf{o}_i$. The necessity reward R_N for o_i is then defined as:

$$R_N(o_i) = \begin{cases} n, & r_i = 1,\ r_j = 0 \\ -\frac{n}{2}, & r_i = 1,\ r_j = 1 \\ -\frac{n}{2}, & r_i = 0,\ r_j = 1 \\ 0, & r_i = 0,\ r_j = 0 \end{cases} \tag{5}$$

where r_i, r_j denote the accuracies of o_i, o_j. Intuitively, R_N is positive only if a regeneration without `<latents>` has lower accuracy than the original latent response. We set $N = 1$ and train Monet-SFT model using VLPO + R_N for 600 steps. As Table 10 shows, this model performs on par with the SFT base model, suggesting that the necessity reward alone is insufficient to incentivize training process to utilize visual latents.

F More Empirical Results

F.1 Necessity of Generated Latents in RL

As shown in Table 2 and the original Monet paper [34], using latent visual tokens in SFT and later in RL post-training yields better performance than vanilla training without latents. We further investigate the importance of latent visual tokens solely at the RL stage. As the attention reward cannot be applied when there are no latent visual tokens, we conduct this experiment with VLPO. We discard the generated latent tokens when the model triggers the latent mode. The model is trained for 400 steps and compared with training with 10 latent tokens (default setting for Monet-VLPO). As reported in Table 11, training with 0 latent tokens when triggering latent mode is generally worse than training with 10 latent tokens, suggesting that such a shortcut has inferior potential.

Table 11: Performance of the Monet model trained with 0 latent tokens is worse than that trained with 10 latent tokens.

Model	MathVista	MMVP	V*			HRBench4K		
	Overall	Overall	Overall	Attribute	Spatial	Overall	FSP	FCP
Monet-7B (SFT)	60.5	73.0	80.1	82.6	76.3	68.0	79.5	56.5
VLPO w/ 0 latents	64.0	72.7	76.4	80.0	71.1	71.6	85.0	58.3
VLPO w/ 10 latents	65.8	73.0	79.1	79.1	78.9	68.6	83.8	53.5

F.2 Inference Speed and Response Length

Here, we report the improved inference efficiency of our model due to the concise nature of its responses in Table 12. We use vLLM [13] as our inference engine with a max model length of 16,384 on a single H100-80GB GPU. Our model responds with $\approx 38\%$ less time, $\approx 20\%$ to 40% fewer tokens than the Monet-VLPO model on V* and MathVista, while achieving better performance, as detailed in Table 2.

F.3 Controlling effects of response length

Observing that our model responds with fewer tokens (150 tokens on average) than the VLPO model, we investigate whether penalizing longer response length alone can improve model performance.

Table 12: **Inference speed comparison. Bold entries denote that our model responds with fastest speed and least tokens.**

Model	V*		MathVista	
	avg. response time	avg. response length	avg. response time	avg. response length
Monet-VLPO	1.91s	52.85 tokens	4.43s	184.02 tokens
Ours	**1.64**s	**41.31** tokens	**2.74**s	**111.02** tokens

We experiment with penalizing 0.1 to the responses with more than 150 tokens. As shown in Table 13, this yields inferior performance, but can be partly mitigated by adding our attention reward, suggesting that controlling response length itself is not enough compared to our method design.

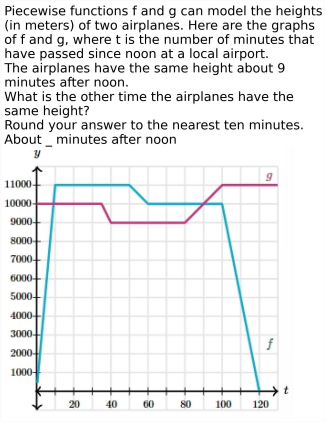
Table 13: Using restricted response length with attention reward degrades performance compared with using attention reward alone. All models are trained for 400 steps.

Reward Type	MathVista	MMVP	MMBench	V*	HRBench4K
-0.1 if len > 150	62.5	73.3	49.1	80.6	68.0
-0.1 if len $> 150 + R_A$	63.6	76.7	78.7	81.2	72.8

G Inference Examples

We show more inference examples generated by the Monet-VLPO model and our model below.

Inference Example 1: Chart Math Reasoning (MathVerse)



Input image

Question Prompt. Please directly answer the question and provide the final value, e.g., 1, 2.5, 300.

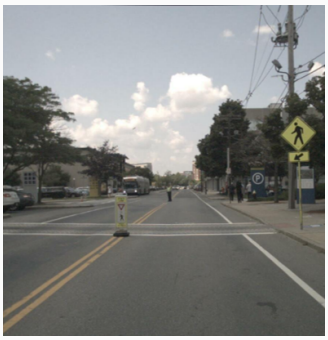
Question: Piecewise functions f and g can model the heights (in meters) of two airplanes. Here are the graphs of f and g, where t is the number of minutes that have passed since noon at a local airport. The airplanes have the same height about 9 minutes after noon. What is the other time the airplanes have the same height? Round your answer to the nearest ten minutes. About _ minutes after noon **(Answer: 90)**

Response by Monet-VLPO. Find the other time the functions have the same height. The functions have the same height at the points where the functions intersect. The functions intersect at the points (9, 11000) and (119, 11000). 119 rounded to the nearest ten is 120. The other time the functions have the same height is 120 minutes after noon.

Response by Ours. The problem asks for the other time the airplanes have the same height, given that they have the same height about 9 minutes after noon. The provided image clearly shows that the two functions intersect again around the 99 mark on the horizontal axis. The problem specifies to round the answer to the nearest ten minutes. The final answer is $\boxed{90}$.

Figure 7: Monet-VLPO fails to locate the intersection of the two lines, while our model precisely locates the coordinates of the intersection.

Inference Example 2: Visual Perception and Spatial Reasoning (MME-RealWorld-Lite)



Input image

Question Prompt. This image shows the front view of the ego car. What is the status of the buses that are to the front of the ego car? The choices are listed below. **(Answer: D)**

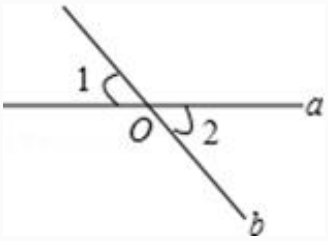
(A) One of the buses is moving, and two are stopped.
(B) Three buses are stopped.
(C) Many buses are stopped.
(D) One of the buses is moving, and one is stopped.
(E) The image does not feature the object.

Response by Monet-VLPO. First, I need to identify the buses in the image and determine their current status, moving or stopped. I will focus on the area in front of the ego car to get a clear view of the buses. `<latent_start><latent><latent_end>` The zoomed-in image clearly shows three buses in the distance. All three buses appear to be stationary and not in motion. The double yellow lines on the road also suggest that the area is a no-passing zone, which is consistent with the buses being stopped at a traffic light or a similar stationary point. **ANSWER:** The best answer is $\boxed{\text{B}}$.

Response by Ours. The image clearly shows one of the buses is moving, and one is stopped. The final answer is $\boxed{\text{D}}$.

Figure 8: Monet-VLPO responds with seemingly fluent reasoning but ends with hallucinations, while ours correctly identifies the one stopped bus.

Inference Example 3: Geometry Math Reasoning (MathVerse)



Input image

Question Prompt. Please directly answer the question and provide the correct option letter, e.g., A, B, C, D.

Question: As shown in the figure, if angle 1 is equal to $50.0°$, then angle 2 is equal to __. **(Answer: A)**

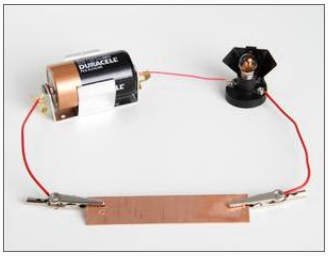
A: $50°$ **B:** $40°$ **C:** $140°$ **D:** $130°$

Response by Monet-VLPO. To find the measure of angle 2, we can use the property that angles on a straight line sum up to $180°$. Given that angle 1 is $50°$, we calculate angle 2 as $180° - 50° = 130°$. Hence, the measure of angle 2 is $130°$. The correct answer is $\boxed{\text{D}}$.

Response by Ours. The problem asks for the measure of angle 2 given that angle 1 is $50°$. The angles 1 and 2 are vertical angles, and vertical angles are always congruent. This means that angle 2 is also $50°$. The final answer is $\boxed{\text{A}}$.

Figure 9: Monet-VLPO fails to capture the position of angle 2 and continue the calculation based on the wrong logic, while our model correctly identifies that angles 1 and 2 are congruent.

Inference Example 4: Physical Attribute Reasoning (MMBench)



Input image

Question Prompt. Identify the question that Gina's experiment can best answer.

Question: The passage below describes an experiment. Read the passage and then follow the instructions below.
Gina built an electric circuit: she used wires to connect a battery to a light bulb, the light bulb to a small piece of copper, and the copper back to the battery. When the circuit was complete, the light turned on. Gina observed the brightness of the light for five seconds. She then replaced the copper with a piece of iron of equal size and noted whether the light became brighter or dimmer. Gina built three more of the same type of circuit. She repeated the tests with each circuit. Gina recorded whether the circuits produced brighter light when the circuit included copper or when the circuit included iron.
Figure: a circuit with a battery, a light bulb, and a piece of copper.
(Answer: B)
A: Can light bulbs stay lit longer when circuits include copper or when circuits include iron?
B: Do circuits that include iron produce dimmer light than circuits that include copper?

Response by Monet-VLPO. The question asks what Gina's experiment can best answer. The passage and the hint indicate that Gina's experiment involves comparing the brightness of a light bulb in a circuit with copper versus a circuit with iron. The hint specifically mentions that Gina observed the brightness of the light for five seconds and noted whether the light became brighter or dimmer when the copper was replaced with iron. Therefore, the experiment can best answer the question of whether circuits that include iron produce dimmer light than circuits that include copper. The final answer is $\boxed{\text{B}}$.

Response by Ours. The problem asks for the question that Gina's experiment can best answer based on the provided passage and the figure. The passage clearly states that Gina observed the brightness of the light for five seconds and then replaced the copper with iron to note the change in brightness. The passage also mentions that she recorded whether the circuits produced brighter light when the circuit included copper or when the circuit included iron. This directly addresses the second option. ANSWER: The final answer is $\boxed{\text{B}}$.

Figure 10: Both Monet-VLPO and our model can reason correctly when the image does not provide relevant information for the question.

Inference Example 5: Coarse Perception (MMVP)



Input image

Question Prompt. n/a

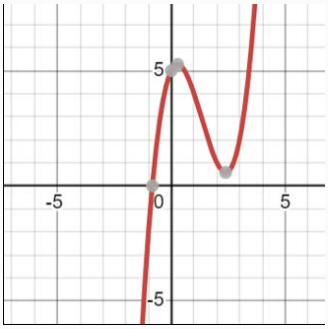
Question: Is the flame of the match more round or thin?
A. Round
B. Thin **(Answer: B)**

Response by Monet-VLPO. The flame of the match is more round, as it appears wider and less elongated. Therefore, the final answer is $\boxed{\text{A}}$.

Response by Ours. ANSWER: The flame of the match is thin, so the correct answer is $\boxed{\text{B}}$.

Figure 11: Monet-VLPO fails to capture the exact shape of the flame in the image using pure text reasoning, while our model correctly recognizes the flames thin shape.

Inference Example 6: Algebraic Reasoning (MathVista)



Input image

Question Prompt. Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.

Question:
Question: is f(3) > 0?
Choices:
A. Yes
B. No
(Answer: A)

Response by Monet-VLPO. To determine if f(3) is greater than 0, we need to locate the point where x = 3 on the given graph and observe its position relative to the x-axis. The point where x = 3 is located on the graph. Observing the position of this point, we can see that it is above the x-axis, indicating a positive value for f(3). THOUGHT N: The point where x = 3 is above the x-axis, indicating a positive value for f(3). ANSWER: The final answer is $\boxed{\text{A}}$.

Response by Ours. ANSWER: The problem asks for the value of the function f at x=3 and whether it is greater than 0. The provided image is a graph of the function f. I need to locate the point on the graph where x=3 and then observe the corresponding y value to answer the question.Therefore, the final answer is $\boxed{\text{No}}$.

Figure 12: Our model may fail to produce the correct final answer despite demonstrating correct reasoning toward a solution, which could be a hallucination inherited from the base model.

Inference Example 7: Spatial Reasoning (HRBench4K)



Input image

Question Prompt. n/a

Question:
Is the mailbox on the left or right side of the woman?
A. Directly above
B. In front
C. Left
D. Right
(Answer: D)

Response by Monet-VLPO. To determine the relative position of the mailbox to the woman, I will analyze the image. `<latent_start><latent><latent_end>`
The image clearly shows the mailbox is located to the right of the woman, next to the pillar she is leaning on. The woman is on the left side of the mailbox. THOUGHT N: The mailbox is on the right side of the woman.Therefore, the final answer is $\boxed{\text{D}}$.

Response by Ours. ANSWER: The image clearly shows the mailbox is positioned directly above the woman's head, relative to the viewer's perspective. The options provided are A. Directly above, B. In front, C. Left, and D. Right. The correct answer is $\boxed{\text{A}}$.

Figure 13: Our model may fail to perceive the correct spatial relationship between objects, likely due to conflating the target object's detection with the taller box behind the mailbox.