Beyond Format: Prompt Semantics Influence Visual Attention in Vision-Language RLVR

Anonymous ACL submission

Abstract

Training prompts in reinforcement learning with verifiable rewards (RLVR) are often treated as simple formatting instructions. This paper studies a different possibility: prompt semantics can act as a key influence on how vision-language models allocate attention to task evidence during RLVR. We conduct a controlled comparison of prompt conditions that share the same reward, data, optimizer, and final-answer format, but differ in whether they expose, suppress, or only semantically permit reasoning. In particular, we use `LatentThink` as a controlled prompt condition that permits internal reasoning while still requiring final-answer-only outputs, allowing us to isolate a semantic change without adding observable reasoning tokens. Across vision-centric VQA benchmarks, this intervention is associated with consistent performance gains over `NoThink`. More importantly, attention diagnostics reveal systematic prompt-induced influence: compared with `NoThink`, `LatentThink` shifts more answer-conditioned attention mass toward evidence image and question tokens, and exhibits a layer-wise entropy pattern that is broader in early layers but sharper near answer generation. Qualitative attention maps show the same depth-wise sharpening toward task-relevant visual evidence. These results suggest that prompt semantics should be treated as a crucial variable in vision-language RLVR, not merely as surface formatting.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a widely used post-training paradigm for improving vision-language models (VLMs) on challenging visual question answering (VQA) tasks (Bai et al., 2025b,a; Liu et al., 2024). Unlike supervised fine-tuning, RLVR optimizes sampled trajectories under reward feedback rather than token-level likelihood, meaning that training prompts can influence not only output format but also how models utilize task evidence during generation (Schulman et al., 2017; Christiano et al., 2017; Ouyang et al., 2022). Despite this, prompt templates are often treated as superficial formatting choices, even though prior work shows that seemingly minor prompt changes can substantially alter model behavior (Leviathan et al., 2025; Xu et al., 2024; Sclar et al., 2023). This raises a central question: *how do prompt semantics influence attention to task-relevant evidence during training?*

This question is particularly important because different prompting strategies impose different assumptions about where reasoning should occur. Explicit `Think` prompting encourages the model to generate textual reasoning traces (Shao et al., 2024; Wei et al., 2022), while final-answer-only `NoThink` prompting suppresses intermediate reasoning tokens and has recently been shown to outperform explicit reasoning on perception-heavy tasks (Li et al., 2025; Wang et al., 2025; Xu et al., 2025). These findings suggest that the key issue may not be whether reasoning itself is useful, but whether prompt semantics alter how model allocates attention to evidence-bearing visual and textual inputs.

To study this question, we conduct controlled comparisons between `Think`, `NoThink`, and `LatentThink` under identical data, reward functions, optimization settings, and final-answer evaluation. `LatentThink` serves as an analysis condition: it semantically permits internal reasoning while preserving the same final-answer-only output constraint as `NoThink`, thereby isolating semantic effects without introducing additional reasoning tokens into the observable response.

Our main finding is that prompt semantics systematically influences evidence-focused attention. As shown in Figure 1 and Figure 2, `LatentThink` shifts more answer-conditioned attention mass toward image and question tokens compared with `NoThink`. Since both modalities carry task evidence, this should be interpreted as a redis-

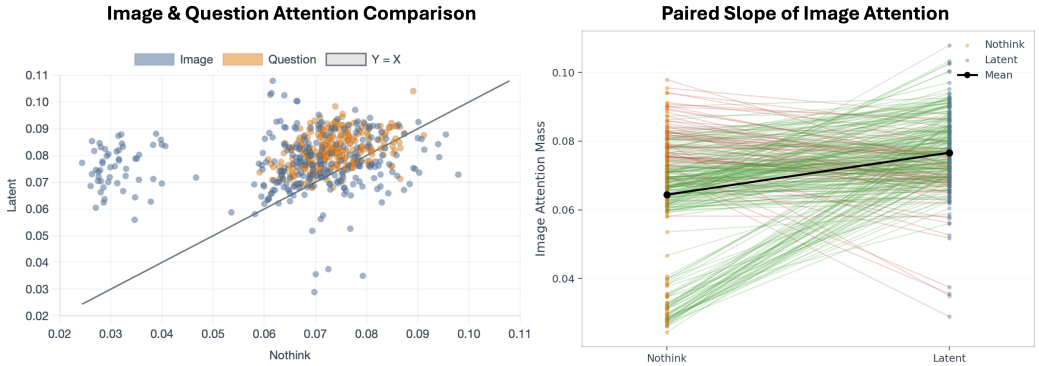

Figure 1: Prompt-induced attention influence on PuzzleVQA (Chia et al., 2024). **Left:** Per-example attention mass assigned to *image* tokens (blue) and *question* tokens (orange) under `NoThink` (x-axis) and the `LatentThink` controlled prompt condition (y-axis); the diagonal indicates parity. **Right:** Paired changes in image-attention mass for each example, with the mean shown in black.

tribution toward evidence-bearing inputs rather than an image-only effect. Layer-wise entropy further reveals a distinct processing trajectory: `LatentThink` exhibits broader attention allocation in lower and middle layers, followed by sharper concentration in deeper layers near answer generation (Figure 3). Qualitative visualizations show a consistent depth-wise sharpening toward task-relevant regions (Figure 4). Across counting, puzzle solving, spatial understanding, medical VQA, and vision-centric math benchmarks, the same prompt condition is also associated with consistent accuracy gains, suggesting that these attention influences are linked to downstream reasoning behavior rather than superficial formatting effects.

We summarize our contributions as follows:

- We present a controlled study of prompt semantics in vision-centric RLVR, treating prompt templates as experimental interventions rather than neutral formatting choices.
- We show that prompt semantics systematically influence answer-conditioned attention allocation and layer-wise attention entropy, producing distinct evidence-focused attention behaviors across transformer depth.
- We connect these attention influences with consistent improvements across diverse vision-centric benchmarks, demonstrating robustness across prompt paraphrases, out-of-domain evaluation, and model scales.

2 Related Works

RLVR and Reasoning in Vision-Language Models. Reinforcement Learning with Verifiable Rewards (RLVR) has become a dominant paradigm for post-training reasoning models, as demonstrated by DeepSeek-Math (Shao et al., 2024) and DeepSeek-R1 (Guo et al., 2025). Recent works extend RLVR to vision-language models for tasks such as spatial reasoning, counting, and medical VQA. However, several studies report that explicit chain-of-thought reasoning (`Think`) does not consistently benefit perception-heavy vision-centric tasks, where strict final-answer-only prompting (`NoThink`) can achieve better performance (Li et al., 2025; Wang et al., 2025; Lai et al., 2026). These findings suggest a tension between linguistic reasoning and visual evidence utilization.

Prompt Semantics and Formatting Sensitivity. A growing body of work shows that seemingly minor prompt formatting differences can significantly affect model behavior. FormatSpread (Sclar et al., 2023) demonstrates large performance variance caused solely by template changes, while prompt repetition and re-reading strategies further reveal strong sensitivity to input structure (Xu et al., 2024; Leviathan et al., 2025). Motivated by these findings, our work studies how prompt semantics reshape attention allocation and attention entropy under identical RLVR training settings.

3 Analysis Framework

We study prompt semantics as a controlled intervention in RLVR rather than as a new model architecture or reward design. For each multimodal input $x = (I, q)$, the VLM samples an answer trajectory y and receives a task verifier reward $R(x, y)$. In all comparisons, the reward function, training data, optimizer, model family, and final-answer evaluation remain fixed. The only intended inter-

vention is the prompt condition p. This lets us ask whether changing prompt semantics changes the effective response distribution $\pi_{\theta,p}(y \mid x)$ and the answer-conditioned attention patterns observed after training.

3.1 Controlled Prompt Conditions

We compare three prompt conditions that differ in how they position reasoning while keeping the downstream task and verifier unchanged. The explicit `Think` condition permits observable reasoning tokens. The `NoThink` condition enforces a strict final-answer-only output. The `LatentThink` condition is used as a controlled semantic intervention: it permits internal reasoning in the prompt wording, but preserves the same final-answer-only output constraint as `NoThink`. Thus, `LatentThink` changes the semantic instruction without adding observable reasoning traces to the response. The distinction is semantic rather than structural: `NoThink` explicitly prohibits reasoning-related behavior in the instruction, whereas `LatentThink` removes that prohibition while keeping the response space final-answer-only. The prompt templates are:

`Think Strategy:`

```
{Question} Output the thinking process in <think>
</think> and final answer from the options (text) in
<answer> </answer> tags.
```

`NoThink Strategy:`

```
{Question} Return ONLY the final answer from the options.
No other text.
```

`LatentThink Strategy:`

```
{Question} You may reason internally. Return ONLY the
final answer from the options. No other text.
```

These conditions can be viewed as different ways of shaping the policy support under the same terminal reward. The analysis does not assume access to hidden reasoning trajectories. Instead, it asks whether a semantic prompt change that is not directly rewarded is associated with systematic changes in attention and behavior. This framing is intended as an empirical probe of prompt semantics, not as a claim that `LatentThink` introduces a new RL algorithm.

3.2 Attention Diagnostics

For each evaluated example, we construct the full input-response sequence and extract self-attention maps from the trained VLM. We average attention over heads and answer-token query positions to obtain an answer-conditioned distribution over input tokens. We then aggregate attention mass into three groups: image tokens, question tokens, and other tokens. Image tokens are identified via vision-specific special tokens, while question tokens correspond to the formatted question span; all masses are normalized to sum to one.

We also compute layer-wise Shannon entropy over the same answer-conditioned attention distribution,

$$H = -\sum_{i=1}^{N} p_i \log p_i, \tag{1}$$

where p_i is the normalized attention assigned to token i and N is the number of input tokens. Higher entropy indicates a more diffuse allocation, while lower entropy indicates a sharper concentration. Together with qualitative layer-wise attention maps, these metrics let us characterize prompt-induced influences in both token-group allocation and depth-wise attention dynamics.

4 Experiments

4.1 Setup

Datasets: We evaluate prompt conditions on six vision-centric visual question answering (VQA) benchmarks spanning diverse reasoning types and modalities: SAT (Ray et al., 2024), CV-Bench (Tong et al., 2024), PuzzleVQA (Chia et al., 2024), OmniMedVQA (Hu et al., 2024), Super-CLEVR (Li et al., 2023), and MathVista (Lu et al., 2023). These datasets cover counting, spatial understanding, compositional puzzle solving, medical, and math VQA. All tasks require selecting a final answer from options, enabling verifiable reward during RLVR training.

Implementation Details: We perform RLVR post-training with GRPO implemented in PyTorch (Paszke et al., 2019). Unless otherwise specified, Qwen3-VL-4B (Bai et al., 2025a) is trained on 2× NVIDIA H100 GPUs, while Qwen3-VL-8B (Bai et al., 2025a) is trained on 2× NVIDIA H200 GPUs. We use bfloat16 precision and FlashAttention-2 (Dao, 2023) for efficient attention computation. The maximum prompt length is set to 1024 tokens, and the maximum image resolution budget is limited to 401,408 pixels. Training is performed with per-device batch size 1 and gradient accumulation over 2 steps, resulting in an effective batch size of 4. For GRPO updates, we sample $K = 4$ responses per input. All prompt strategies are trained under identical optimization hyperparameters and sampling configurations to

Table 1: PuzzleVQA (Chia et al., 2024) per-type accuracy (%) under different RLVR prompt conditions on Qwen-2.5-VL and Qwen-3-VL. The `LatentThink` condition is associated with higher overall accuracy than `NoThink`, with gains observed across most puzzle types.

Puzzle Type	**Qwen2.5-VL-3B** (Bai et al., 2025b)			**Qwen3-VL-4B** (Bai et al., 2025a)			
	Zero-shot	Nothink	Latent	Zero-shot	Think	Nothink	Latent
Circle Size Number	28.00	99.00	100.0	38.00	85.00	100.0	100.0
Color Grid	19.00	67.00	95.00	30.00	75.00	24.00	93.00
Color Hexagon	49.00	88.00	100.0	22.00	83.00	100.0	100.0
Color Number Hexagon	27.00	47.00	54.00	39.00	81.00	39.00	55.00
Color Overlap Squares	12.00	100.0	98.00	19.00	59.00	99.00	100.0
Color Size Circle	67.00	100.0	100.0	97.00	99.00	100.0	100.0
Grid Number	17.00	77.00	93.00	20.00	97.00	100.0	97.00
Grid Number Color	61.00	99.00	100.0	95.00	100.0	100.0	100.0
Polygon Sides Color	39.00	93.00	98.00	65.00	97.00	97.00	100.0
Polygon Sides Number	70.00	95.00	92.00	84.00	72.00	91.00	95.00
Rectangle Height Color	47.00	78.00	94.00	60.00	43.00	86.00	84.00
Rectangle Height Number	32.00	64.00	99.00	41.00	39.00	80.00	98.00
Shape Morph	53.00	83.00	85.00	72.00	63.00	100.0	93.00
Shape Reflect	50.00	99.00	100.0	94.00	99.00	100.0	100.0
Shape Size Grid	47.00	61.00	87.00	75.00	81.00	83.00	82.00
Shape Size Hexagon	60.00	68.00	84.00	47.00	76.00	100.0	100.0
Size Cycle	21.00	35.00	60.00	23.00	36.00	42.00	67.00
Size Grid	32.00	44.00	46.00	24.00	41.00	37.00	60.00
Triangle	35.00	98.00	99.00	63.00	75.00	92.00	100.0
Venn	43.00	96.00	100.0	39.00	100.0	100.0	100.0
Overall	40.45	79.55	89.20	51.75	76.20	83.50	91.20

ensure fair comparison.

4.2 Performance Effects of Prompt Semantics

Result 1: Puzzle Solving. Following official implementation of PuzzleVQA (Chia et al., 2024), we generate 4,000 training VQA pairs (200 per puzzle type) using the provided data generation code. Each model is trained for two epochs under identical optimization settings. Evaluation is conducted on the official PuzzleVQA test set (Chia et al., 2024), which contains 2,000 VQA pairs (100 per puzzle type).

As shown in Table 1, the `LatentThink` condition consistently achieves the highest overall accuracy across both model scales. On Qwen2.5-VL-3B, the overall accuracy increases from 79.55% (`NoThink`) to 89.20%. On Qwen3-VL-4B, performance improves from 83.50% to 91.20%. In contrast, the explicit `Think` strategy on Qwen3-VL-4B suffers a severe performance drop, achieving only 76.20% overall accuracy, with low scores on categories such as `Rectangle Height Number` (39.00%) and `Color Overlap Squares` (59.00%).

A per-category comparison further reveals that `LatentThink` is associated with improvements over `NoThink` across the vast majority of puzzle types. On Qwen3-VL-4B, notable improvements are observed on `Color Number Hexagon` (55.00% vs. 39.00%), `Rectangle Height Number` (98.00% vs. 80.00%), and `Size Cycle` (67.00% vs. 42.00%). Similar gains are evident on Qwen2.5-VL-3B, including `Rectangle Height Number` (99.00% vs. 64.00%) and `Shape Size Grid` (87.00% vs. 61.00%). Across all categories, the most pronounced absolute improvements are consistently observed in size- and scale-related tasks (e.g., `Rectangle Height Number`, and `Shape Size Grid`), which require detailed comparison of visual evidence.

Result 2: Medical VQA. As medical VQA represents another major category of vision-centric VQA task, we evaluate the same prompt conditions on OmniMedVQA (Hu et al., 2024). Following the setting of Med-R1 (Lai et al., 2026), we perform RLVR post-training using only the X-Ray training split (6,332 VQA pairs) and evaluate on the test sets of X-Ray and seven other imaging modalities (CT, MRI, Ultrasound, Dermoscopy, Fundus, OCT, and Microscopy).

Results are summarized in Table 2. The `LatentThink` condition consistently achieves the highest overall accuracy across both model scales. On Qwen2.5-VL-3B, the overall accuracy improves from 73.11% (`NoThink`) to 76.49%. On Qwen3-VL-4B, performance increases from 79.01% to 83.70%. Similar to the trends observed in puzzle solving, the explicit `Think` strategy empirically struggles in this setting, underperforming the `NoThink` baseline in overall accuracy on both Qwen2.5-VL-3B (69.54%) and Qwen3-VL-

Table 2: OmniMedVQA (Hu et al., 2024) modality-wise accuracy (%) under different RLVR prompt conditions on Qwen2.5-VL-3B and Qwen3-VL-4B. "Latent" denotes the `LatentThink` analysis condition and is associated with higher overall scores across most imaging modalities compared with Zero-shot, Think, and NoThink baselines.

Strategy	Modality								Overall
	CT	MRI	X-Ray	UltraS	Der	Fundus	OCT	Micro	
Qwen2.5-VL-3B (Bai et al., 2025b)									
Zero-shot	53.87	54.23	61.84	32.69	52.94	62.47	56.23	59.64	54.24
Think	78.90	59.73	93.50	38.28	65.85	76.05	78.89	65.14	69.54
NoThink	80.10	69.78	95.79	34.33	71.29	81.79	83.14	68.65	73.11
Latent	79.94	75.81	97.62	47.85	72.89	84.61	81.89	71.34	76.49
Qwen3-VL-4B (Bai et al., 2025a)									
Zero-shot	65.38	56.66	82.48	56.36	66.77	60.38	54.13	77.03	64.90
Think	70.63	68.38	94.30	63.16	68.84	74.41	64.98	80.00	73.09
NoThink	75.22	77.24	93.25	70.44	73.43	82.70	79.95	79.82	79.01
Latent	80.31	85.59	96.72	78.50	75.57	88.16	81.25	83.51	83.70

Modality: CT – Computed Tomography; MRI – Magnetic Resonance Imaging; UltraS – Ultrasound; Der – Dermoscopy; Fundus – Fundus Photography; OCT – Optical Coherence Tomography; Micro – Microscopy; X-Ray – X-ray Imaging.

Table 3: Performance comparison on additional vision-centric benchmarks.

Strategy	Dataset			
	SAT (Ray et al., 2024)	CV-Bench (Tong et al., 2024)	Super-CLEVR (Li et al., 2023)	MathVista (Lu et al., 2023)
NoThink	97.33	82.57	80.50	60.19
Latent	100.0	85.74	85.50	62.31

4B (73.09%). We interpret this drop cautiously: it may reflect a mismatch between explicit reasoning-token generation and final-answer verification under this RLVR setup, rather than an inherent weakness of explicit reasoning.

A per-modality breakdown reveals that `LatentThink` is associated with gains over `NoThink`. Despite being trained only on X-Ray data, the same prompt condition maintains improvements on unseen imaging types. For example, on Qwen3-VL-4B, significant improvements over `NoThink` are observed on MRI (85.59% vs. 77.24%), Ultrasound (78.50% vs. 70.44%), CT (80.31% vs. 75.22%), and Fundus (88.16% vs. 82.70%). Similar cross-modality improvements are evident on Qwen2.5 (e.g., Ultrasound improving from 34.33% to 47.85%), suggesting that prompt effect is not limited to training modality.

Result 3: Additional Vision-Centric Benchmarks. As shown in Table 3, we further evaluate prompt effects on additional spatial, counting, and mathematical VQA benchmarks. For spatial VQA, following Zhou et al. (Zhou et al., 2025), we fine-tune the model under the `NoThink` and `LatentThink` strategies on the static subset of SAT (Ray et al., 2024) for one epoch and evaluate on the official SAT and CV-Bench (Tong et al., 2024) test sets. `LatentThink` improves over `NoThink` by 2.67% on SAT and 3.17% on CV-Bench. For counting tasks, following Chen et al. (Chen et al., 2025), we fine-tune on the counting subset of Super-CLEVR (Li et al., 2023) for 200 training steps and evaluate on the official counting test set, where `LatentThink` is associated with a 5.00% absolute improvement. For math VQA, following Li et al. (Li et al., 2025), we fine-tune on Math-40K (Shi et al., 2024) for one epoch and evaluate on MathVista (Lu et al., 2023). `LatentThink` yields a 2.12% performance gain compared with `NoThink`. These results indicate that prompt semantics consistently changes downstream behavior across diverse vision-centric benchmarks.

4.3 Generalization

We further evaluate whether the prompt effect persists under out-of-domain evaluation. As shown in Table 4, following Li et al. (Li et al., 2025), we test Qwen3-VL-4B models fine-tuned on PuzzleVQA (Chia et al., 2024) under `NoThink` and `LatentThink` on the out-of-domain benchmark AlgoPuzzleVQA (Ghosal et al., 2025). Although overall accuracy decreases under domain shift, `LatentThink` outperforms `NoThink` on both the source dataset (91.20% vs. 83.50%) and the target dataset (31.44% vs. 28.83%), yielding a +2.61% gain on AlgoPuzzleVQA.

Similarly, in Table 2, models fine-tuned on the X-Ray training split of OmniMedVQA (Hu

Table 4: Cross-dataset generalization from PuzzleVQA to AlgoPuzzleVQA (Ghosal et al., 2025).

Dataset	Qwen3-VL-4B		
	Zero-shot	Nothink	Latent
PuzzleVQA	51.75	83.50	91.20
AlgoPuzzleVQA	27.94	28.83	31.44

Table 5: Scaling analysis of Qwen3-VL-4B and Qwen3-VL-8B on PuzzleVQA.

Model	Strategy		
	Zero-shot	Nothink	Latent
Qwen3-VL-4B	51.75	83.50	91.20
Qwen3-VL-8B	51.10	80.85	86.20

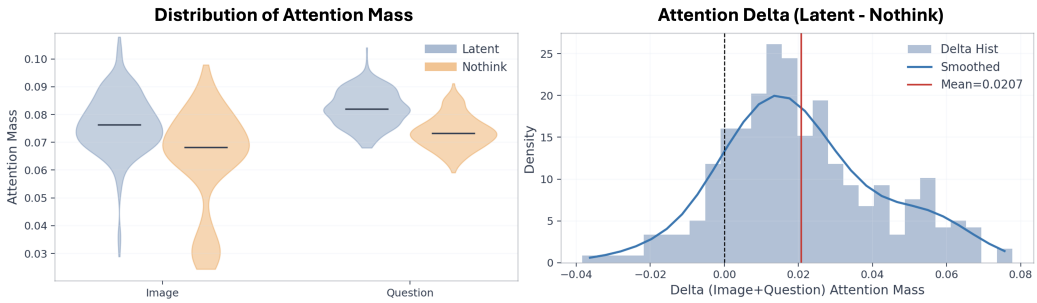

Figure 2: Attention mass allocation under different RLVR prompt conditions. **Left:** distributions of attention mass on *Image* and *Question* tokens for `LatentThink` (blue) and `NoThink` (orange). **Right:** histogram with smoothed density of per-sample attention-mass change Δ (`LatentThink` $-$ `NoThink`) over *Image+Question* tokens; the dashed black line marks $\Delta = 0$, and the red line indicates a positive mean shift.

et al., 2024) are evaluated across seven additional unseen imaging modalities. Across both model scales, `LatentThink` obtains higher overall accuracy than `NoThink` and maintains improvements on most unseen modalities, including CT, MRI, Ultrasound, Fundus, and Microscopy. These results indicate that performance differences associated with prompt semantics are preserved under cross-dataset evaluation rather than being confined to training domain.

4.4 Scale-up

We further examine whether the observed prompt effect persists under model scaling. In addition to Qwen3-VL-4B, we conduct RLVR post-training on Qwen3-VL-8B (Bai et al., 2025a) using the same 4,000 PuzzleVQA training pairs and identical optimization settings. Notably, the 8B model does not exhibit a clear zero-shot advantage over the 4B model on this benchmark (Table 5), suggesting that PuzzleVQA performance is not primarily driven by parameter scale under the default prompting setup. As shown in Table 5, `LatentThink` outperforms `NoThink` at both scales. On Qwen3-VL-4B, accuracy improves from 83.50% to 91.20%, and on Qwen3-VL-8B from 80.85% to 86.20%. Although the 8B model obtains slightly lower absolute accuracy under the same RLVR training budget, the relative gains remain stable, indicating that prompt semantics remains a meaningful variable across model sizes.

5 Evidence-Attention Influence

The performance results show that prompt semantics changes downstream behavior. We now analyze whether those behavioral differences are accompanied by systematic influence in attention to task evidence.

5.1 Prompt Semantics Shift Attention

As shown in Figure 2, the `LatentThink` condition allocates more answer-conditioned attention mass to both Image ($+19.0\%$) and Question ($+11.5\%$) tokens than `NoThink`. The left panel illustrates a distributional shift toward evidence-bearing input tokens under `LatentThink`. The right panel reports the per-sample attention difference $\Delta =$ `LatentThink` $-$ `NoThink` over Image+Question tokens. The distribution is shifted to the right, with a positive mean increase ($\mu = 0.0207$), indicating that the semantic prompt intervention reallocates attention toward task-relevant input tokens. This matters because the question tokens identify the queried attribute and the image tokens provide the visual evidence; increased mass on both groups indicates a shift away from non-evidence tokens rather than a claim that image-token attention alone explains the gain. This shift aligns with the performance differences observed across vision-centric benchmarks (Tables 1 and 2), but we treat it as diagnostic evidence rather than a direct causal proof.

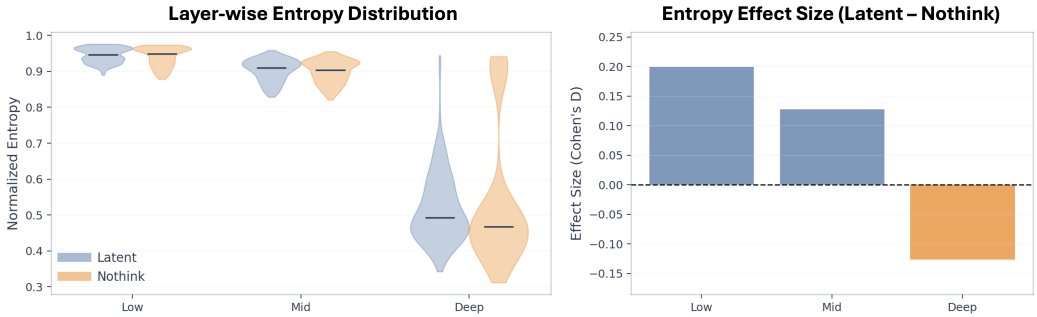

Figure 3: Layer-wise entropy of the answer-conditioned attention distribution under different RLVR prompt conditions. `LatentThink` exhibits higher entropy in lower/middle layers but lower entropy in deeper layers near the output. The right panel reports effect sizes (Cohen's d) computed as `LatentThink` minus `NoThink`; positive values indicate higher attention entropy under `LatentThink`.

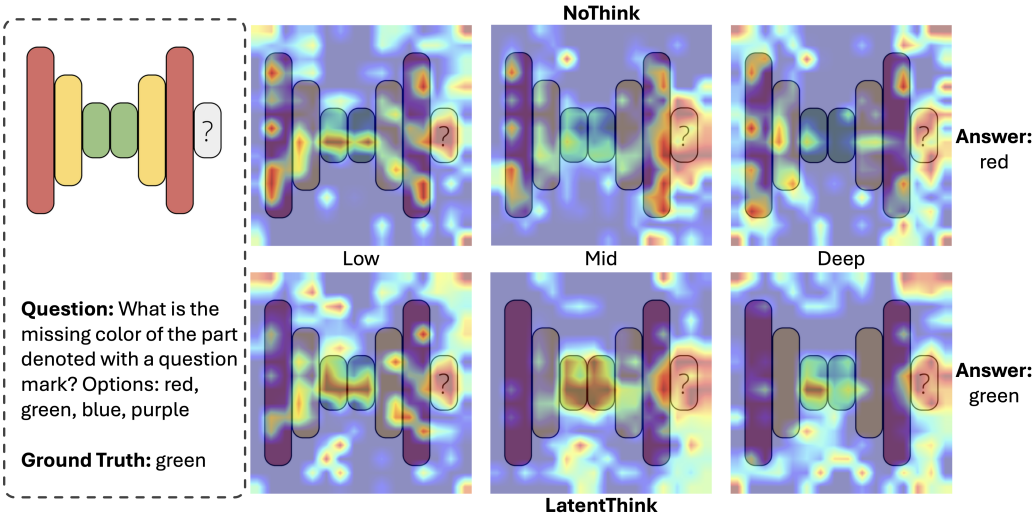

Figure 4: Qualitative attention maps across model layers for an example from PuzzleVQA. We compare `NoThink` (top row) and `LatentThink` (bottom row) at low-, mid-, and deep-layer blocks. The `LatentThink` condition allocates relatively more attention mass to the task-relevant queried region, while `NoThink` shows more diffuse attention patterns across non-target regions.

5.2 Layer-wise Entropy

As shown in Fig. 3, `LatentThink` exhibits a distinct layer-wise entropy signature compared to `NoThink`. In the lower and middle layers, `LatentThink` maintains higher entropy, with positive effect sizes (Cohen's $d \approx 0.20$ and 0.13, respectively). This suggests broader initial allocation over input signals during early-stage processing.

This trend reverses in the deeper layers near the output, where `LatentThink` has lower entropy than `NoThink` (Cohen's $d \approx -0.12$). This transition toward lower deep-layer entropy indicates a more concentrated attention allocation prior to answer emission. Such an evolution—from broad early allocation to sharper localized focus—is consistent with prompt semantics influencing internal processing patterns, while still remaining a correlational diagnostic rather than a causal intervention on attention itself.

5.3 Qualitative Attention Maps

To complement the quantitative attention and entropy analyses, we further visualize layer-wise attention maps to show how focus evolves from shallow to deep transformer blocks. Figure 4 presents qualitative attention maps for a representative PuzzleVQA example, comparing `NoThink` (top row) and `LatentThink` (bottom row) across low-, mid-, and deep-layer blocks. Under the `NoThink` strategy, attention remains relatively diffuse across layers, with noticeable mass distributed over task-irrelevant regions even in deeper blocks. The model does not progressively concentrate its focus on the missing component and relevant clues, which coincides with its incorrect final prediction ("red").

In contrast, `LatentThink` exhibits a clear layer-

Table 6: Ablation study of prompt variants under NoThink and LatentThink.

NoThink				LatentThink			
#1	#2	#3	Avg.	#1	#2	#3	Avg.
83.50	82.60	80.85	82.32	91.20	88.75	89.50	$89.82_{\uparrow 7.50}$

wise evolution. In lower layers, attention is broadly distributed over the entire scene, consistent with the higher entropy observed in early-stage processing (Figure 3). As depth increases, attention becomes increasingly concentrated on the visually informative region corresponding to the queried attribute. This sharpening of focus is most evident in the deepest layers, where attention mass is strongly localized around the relevant visual cues, coinciding with the correct final answer ("green"). These qualitative observations are consistent with the quantitative findings in Figure 2, particularly the positive mean shift ($\mu = 0.0207$) of attention mass toward task-relevant tokens under LatentThink. Moreover, the transition from broader to more localized attention aligns with the layer-wise entropy trend, where LatentThink displays higher entropy in early layers but lower entropy in deeper blocks. Together, these results indicate that prompt semantics is associated with a structured transition in attention allocation, moving from broad contextual coverage toward focused, task-aligned representations before answer generation.

5.4 Prompt Paraphrase Ablation

To verify that the observed differences arise from strategy-level semantics rather than one specific wording, we conduct an ablation study using multiple prompt variants for both NoThink and LatentThink. For each strategy, we generate three semantically equivalent but lexically diverse prompt variants using ChatGPT (Singh et al., 2025), while keeping all fine-tuning and evaluation settings identical.

As summarized in Table 6, LatentThink outperforms NoThink across all tested variants. LatentThink has an average accuracy of 89.82%, maintaining a stable margin of +7.50% over the NoThink average of 82.32%. Importantly, even the lowest-performing LatentThink variant (#2: 88.75%) exceeds the highest-performing NoThink variant (#1: 83.50%). These results do not exhaust all possible instruction-strength choices, but they reduce the likelihood that the effect comes from one lucky string. The consistent gap across variants suggests that the relevant factor is the semantic contrast between prohibiting reasoning and permitting internal reasoning under the same output constraint.

6 Discussion

What the attention shift means. Our analysis should be interpreted as evidence of prompt-induced redistribution over evidence-bearing tokens, not as proof that attention changes are the sole cause of performance gains. Because both image tokens and question tokens carry task evidence, the relevant pattern is the joint increase in Image+Question attention and the corresponding shift away from other tokens. The entropy and visualization results provide convergent diagnostics, but they remain correlational.

What LatentThink isolates. NoThink and LatentThink share the same final-answer-only output constraint. Their difference is semantic: NoThink explicitly suppresses reasoning-related behavior, while LatentThink permits reasoning to remain internal without allowing extra output text. Thus, the comparison isolates whether this semantic permission changes attention and behavior under the same verifier and optimization setup.

7 Conclusion

We study prompt semantics as a controlled intervention in vision-language RLVR. Under the same data, rewards, optimization settings, and final-answer constraints, different prompt conditions produce consistent differences in both downstream performance and attention behavior. In particular, LatentThink shifts attention toward task-relevant image and question tokens and leads to more concentrated attention in deeper layers. While these analyses are correlational rather than causal, the results consistently show that prompt semantics can influence both multimodal attention patterns and optimization outcomes under RLVR. More broadly, our findings suggest that prompts in RL-trained multimodal systems are not only formatting instructions, but also behavioral signals that affect how models allocate attention and process evidence during reasoning.

8 Limitations

Our experiments primarily use Qwen-series VLMs and vision-centric VQA settings. We do not evaluate whether the same behavior appears in text-only QA, under shuffled answer options, without answer options, or across substantially different VLM families. These settings would help distinguish model-family effects from broader RLVR behavior. In addition, our attention analyses remain correlational rather than causal. While entropy and visualization results consistently align with performance changes, they do not establish that the observed attention redistribution is itself the mechanism responsible for the gains.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv: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, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. R1-v: Reinforcing super generalization ability in vision-language models with less than $3. https://github.com/Deep-Agent/R1-V. Accessed: 2025-02-02.

Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16259–16273.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Deepanway Ghosal, Vernon Toh, Yew Ken Chia, and Soujanya Poria. 2025. Algopuzzlevqa: Diagnosing multimodal reasoning challenges of language models with algorithmic multimodal puzzles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9615–9632.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*.

Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. 2025. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, and 1 others. 2024. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. 2025. Think or not? selective reasoning via reinforcement learning for vision-language models. *arXiv preprint arXiv:2505.16854*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jingxian Xu, Mengyu Zhou, Weichang Liu, Hanbing Liu, Shi Han, and Dongmei Zhang. 2025. Twt: Thinking without tokens by habitual reasoning distillation with multi-teachers' guidance. *arXiv preprint arXiv:2503.24198*.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and Shuai Ma. 2024. Re-reading improves reasoning in large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 15549–15575.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*.