LatentDraft: Shaping Latent Thoughts into an Intermediate Modality for Interleaved Reasoning in Vision-Language Models

Anonymous ACL submission

Abstract

Vision-language models typically rely on natural-language reasoning to expand the computation budget during inference, thereby improving task performance. However, discretizing continuous intermediate thoughts into textual tokens introduces a language bottleneck, which can lead to the loss of cross-modal information and limit model performance on fine-grained visual understanding and complex reasoning tasks. In this paper, we propose **LatentDraft**, an explicit-implicit interleaved reasoning paradigm for vision-language models. LatentDraft uses natural-language reasoning as the explicit backbone, adaptively triggers latent-space reasoning at language bottlenecks, and shapes latent thoughts into an intermediate modality between vision and language. In this way, it preserves cross-modal information in continuous space while supporting a compact expansion of implicit computation. Experiments show that LatentDraft achieves an average improvement of 4.94% over the baseline across multiple benchmarks and outperforms representative explicit and implicit reasoning-enhancement methods. Further analyses demonstrate the advantages of LatentDraft in hallucination mitigation, robustness in complex reasoning, cross-task generalization, and computational efficiency, while also supporting the core design of modeling latent thoughts as an intermediate modality.

1 Introduction

Recent vision-language models (VLMs) have made substantial progress in multimodal understanding and reasoning (Liu et al., 2023). Chain-of-thought(CoT) prompting has become an important methods for improving complex reasoning by verbalizing intermediate steps (Wei et al., 2022). This paradigm has been extended to multimodal settings, where textual reasoning is expected to guide the model in using visual evidence before generating the final answer (Zhang et al., 2024).

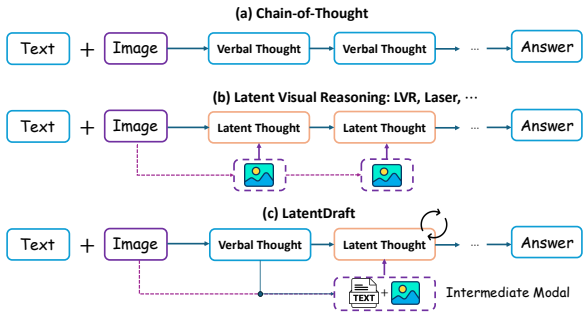


Figure 1: Unlike textual CoT and prior latent visual reasoning methods, **LatentDraft** interleaves verbal thoughts with intermediate modal latent drafts to combine interpretable language reasoning with visually grounded hidden-state computation.

However, visual reasoning is not always well suited to linguistic expression. Many intermediate conclusions, such as occlusion order, local-to-global correspondence, viewpoint change, and fine-grained spatial relations, require continuous and relational visual computation. Forcing such processes into the natural-language may create a mismatch between the content of reasoning and its medium of expression. The model may generate fluent but visually ungrounded reasoning, replacing real visual computation with linguistic priors (Long et al., 2026). Longer textual reasoning is therefore not necessarily better. Some thoughts should be computed, but not necessarily spoken. Language is suitable for expressing symbolic or high-level semantic relations (Wang et al., 2026b), while certain visually grounded computations are better preserved in hidden space, where they need not be discretized into words. Based on this motivation, we propose **LatentDraft**, a language-latent interleaved reasoning paradigm for visual reasoning. Rather than requiring all intermediate steps to be expressed as text, LatentDraft alternates between explicit natural-language reasoning and implicit hidden-state drafts. Textual segments provide a readable reasoning scaffold, while latent drafts are

inserted at key visual bottlenecks to carry continuous intermediate computations that are difficult to verbalize but necessary for subsequent reasoning.

As illustrated in Figure 1, LatentDraft differs from both textual CoT, which exposes intermediate reasoning entirely as language, and prior latent visual reasoning, which routes intermediate computation through latent states. LatentDraft preserves the interpretability of verbal thoughts while allowing hidden states to function as an *intermediate modality* between visual evidence and linguistic expression. To make such latent drafts useful for subsequent language generation, we shape them from both semantic and visual perspectives. Semantic guidance keeps latent states readable to the language pathway by aligning their LM-head projections with stage-dependent concept distributions without forcing collapse to a single token. An intermediate-modality constraint further places latent drafts between textual and visual centroids, encouraging them to preserve image-grounded evidence while remaining compatible with later textual reasoning. Compared with directly pulling latent states toward static visual representations like LVR (Li et al., 2025), these objectives enable the model to perform continuous computation in the hidden space at visual bottlenecks, followed by a return to natural language for output.

In this way, the model can reason through a mixed medium: linguistically expressible steps are spoken, visually grounded computations that are hard to express are carried by latent drafts, and the model then returns to language to generate the final answer. Experiments on five multimodal benchmarks show that LatentDraft consistently improves over the base VLM and existing latent reasoning baselines, with particularly strong gains on hallucination-sensitive and visually grounded reasoning tasks. Our contributions are threefold:

- We introduce **LatentDraft**, a language-latent interleaved reasoning paradigm that augments natural-language reasoning with hidden-state drafts at visual bottlenecks.
- We formulate latent drafts as an *intermediate modality* between visual evidence and linguistic expression, providing a geometric motivation for hidden-space reasoning in VLMs.
- We design a training objective that combines semantic guidance and an intermediate-modality constraint, enabling latent states to bridge consecutive textual reasoning segments while preserving visual grounding.

2 Related Work

2.1 Chain-of-Thought Reasoning in VLMs

CoT prompting improves LLM reasoning by eliciting intermediate natural-language steps (Wei et al., 2022), and has been extended through zero-shot prompting, self-consistency, decomposition, and rationale supervision (Kojima et al., 2022; Wang et al., 2023; Zhou et al., 2023; Zelikman et al., 2022). In VLMs, CoT has been instantiated through multimodal rationale datasets, rationale-conditioned inference, and visually grounded reasoning procedures that connect textual rationales with image regions, visual marks, or structured stages (Zellers et al., 2019; Lu et al., 2022; Zhang et al., 2023; Shao et al., 2024; Yang et al., 2023; Zhang et al., 2025a,b; Dong et al., 2025). However, longer rationales do not necessarily imply faithful reasoning: textual explanations may fail to reflect the model's actual decision process (Turpin et al., 2023), and VLM reasoning chains can remain inconsistent with visual evidence (Chen et al., 2024b). LatentDraft preserves text as an interpretable scaffold, but introduces latent drafts for visually grounded computations that are difficult to express faithfully as text.

2.2 Latent Space Reasoning

Latent reasoning studies whether intermediate computation can be carried by continuous hidden states rather than decoded as natural-language tokens. In language-only settings, implicit CoT, Coconut, and compressed reasoning representations demonstrate that computation can proceed in hidden space without verbalization (Goyal et al., 2024; Hao et al., 2025; Cheng and Van Durme, 2024; Tan et al., 2026). Recent VLM-oriented methods extend this to multimodal reasoning: Heima compresses multimodal CoT into hidden thinking tokens, LVR reconstructs visual tokens from latent states, Monet optimizes continuous visual thoughts, and Laser aligns latent states with dynamic semantic windows (Shen et al., 2025; Li et al., 2025; Wang et al., 2025a, 2026b). These methods demonstrate the promise of latent states for visual reasoning, but they typically treat latent states as either compressed CoT tokens or visual-latent embeddings. In contrast, LatentDraft formulates each latent draft as an intermediate modality between visual evidence and linguistic expression, enabling latent states to remain compatible with subsequent language generation while preserving visually grounded information.

3 Method

3.1 The LatentDraft Paradigm

Given an image I and a visual task Q, the model produces an answer $\hat{y}$. We focus on the intermediate reasoning that are essential for the final answer but difficult to faithfully verbalize, such as qualitative spatial relations, viewpoint changes, occlusion order, and correspondences between local evidence and global referents. Forcing such reasoning processes into text may replace continuous visual thoughts with fluent but under-constrained explanations, causing the model to over-rely on linguistic priors, leading to unsupported conclusions.

The central hypothesis is that *some intermediate computations in visual reasoning should be executed but need not be spoken*. We therefore propose **LatentDraft**, a language-latent interleaved reasoning paradigm that interleaves explicit textual reasoning with implicit hidden-state drafting:

$$y = u_1\,[\mathbf{H}_1]\,u_2\,[\mathbf{H}_2]\cdots u_M\,[\mathbf{H}_M]\,u_{M+1}\,\hat{y}, \quad (1)$$

where $\{u_m\}$ form a readable textual reasoning trace, and each $\mathbf{H}_m$ is a latent draft inserted at a key visual bottleneck. Text segments organise the problem, express verbalizable observations, and maintain the symbolic reasoning scaffold. Latent drafts, in contrast, preserve continuous visual computations that are difficult to discretise into tokens but must still be consumed by later reasoning. LatentDraft reallocates reasoning across two media.

Inside the m-th latent draft, hidden states are produced recursively by the model f_θ:

$$\begin{aligned} \mathbf{h}_m^{(s)} &= f_\theta\big(\mathbf{h}_m^{(s-1)} + \mathbf{e}_{\text{step}}(s)\,\big|\,\mathcal{C}_m\big), \\ s &= 1,\ldots,k_m, \quad k_m \in \{1,2,3\}, \end{aligned} \quad (2)$$

where $\mathbf{h}_m^{(0)}$ is the hidden state of the latent sentinel, $\mathcal{C}_m$ denotes the preceding KV cache, and $\mathbf{e}_{\text{step}}(s)$ distinguishes recursive steps within a draft. Crucially, the model f_θ is reused as the latent recurrent computation module rather than introducing a separate reasoner. The latent states $\mathbf{H}_m$ are not decoded into tokens. Instead, they are injected into the subsequent context as continuous hidden states, allowing u_{m+1} to attend to the implicit visual computation performed at the bottleneck.

3.2 Latent Thoughts as an Intermediate Modality

Training pursues three complementary goals: (1) the model should invoke latent drafting at genuine visual bottlenecks rather than inserting format markers mechanically; (2) each latent draft should be acted as an information bridge between u_m and u_{m+1}, ensuring subsequent text depends on the implicit computation encoded in $\mathbf{H}_m$; and (3) $\mathbf{H}_m$ should remain compatible with the language generation pathway while carrying visual evidence that text alone cannot reliably express.

We build on the semantic superposition mechanism of Laser (Wang et al., 2026b), which softly aligns the language head distribution of latent tokens with a set of concepts to preserve multiple compatible meanings in continuous space. We inherit this mechanism, but pose a further fundamental question: *where should a latent state reside in the hidden space?* Our answer is that it should collapse neither into a language token as in Laser (Wang et al., 2026b) nor into a purely visual token as in LVR (Li et al., 2025), but instead be shaped as a **intermediate modality** between visual evidence and linguistic expression. The semantic and visual components below implement this principle from two complementary directions.

Latent supervision principle. For each latent draft, we construct an ordered sequence of concept sets $\{\mathcal{S}_m^{(1)},\ldots,\mathcal{S}_m^{(k_m)}\}$. These sets follow a cognitive direction from high-level semantic reference to concrete visual evidence: earlier stages capture global anchors and spatial referents, while later stages move toward local cues and decisive visual judgments. The concept sequence serves purely as supervision for the latent states; it is never inserted into the input or output text.

Semantic Guidance: Stage-Windowed semantic supervision. At step s of draft m, the latent state $\mathbf{h}_m^{(s)}$ is projected through the language head:

$$\ell_m^{(s)} = W_U\mathbf{h}_m^{(s)}, \qquad p_{\theta,m}^{(s)} = \text{softmax}(\ell_m^{(s)}). \quad (3)$$

The distribution $p_{\theta,m}^{(s)}$ is normalized over the full vocabulary, but semantic supervision is applied only on a stage-dependent window:

$$\Omega_m^{(s)} = \left(\bigcup_{r=s}^{k_m}\mathcal{S}_m^{(r)}\right)\setminus\mathcal{X}, \quad (4)$$

where $\mathcal{X}$ removes stop words and non-informative high-frequency fragments. By construction, $\Omega_m^{(1)} \supseteq \cdots \supseteq \Omega_m^{(k_m)}$. Thus, early latent states are allowed to maintain a broader semantic superposition over possible future judgments, while later

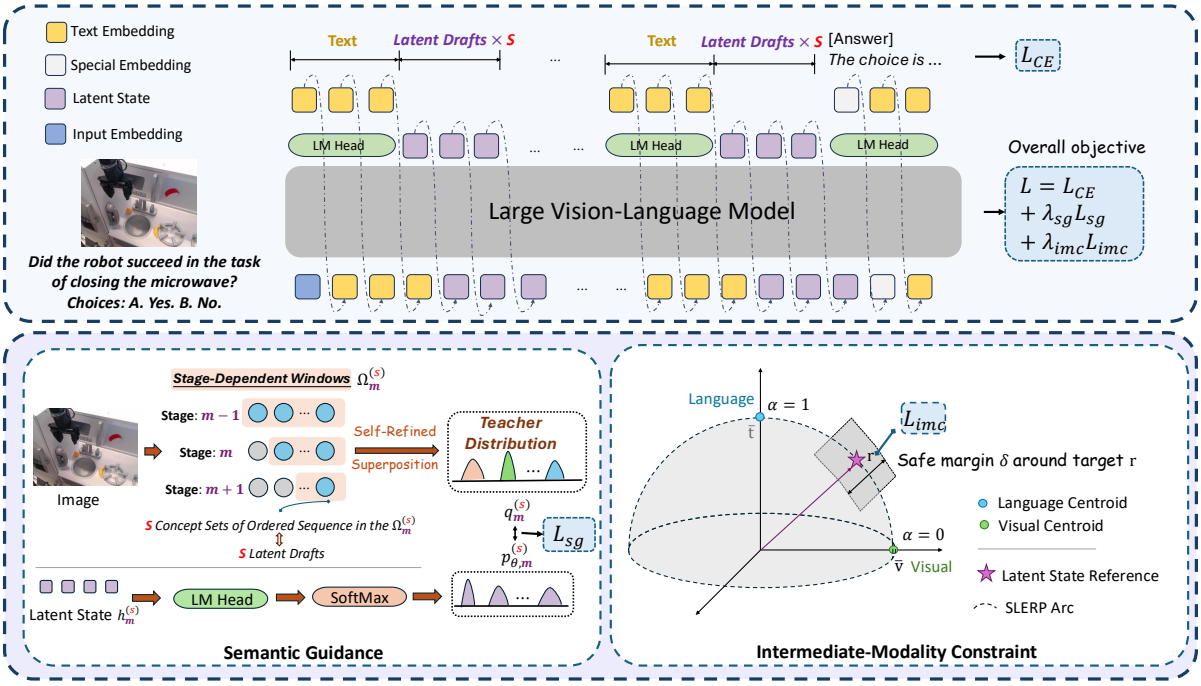


Figure 2: Architecture and training objective of LatentDraft. The model interleaves text with latent drafts during reasoning, and optimizes latent states using semantic guidance and an intermediate-modality constraint between language and visual centroids, together with the standard causal language-modeling loss.

states are gradually narrowed toward more specific image-grounded evidence.

Given the window $\Omega_m^{(s)}$, we construct the teacher distribution $q_m^{(s)}$ using the self-refined superposition mechanism (Wang et al., 2026b). Specifically, the teacher is obtained from the model's detached prediction restricted to the current window. LatentDraft constructs the stage window from the ordered stage sets inside each latent draft. This adapts semantic superposition (Wang et al., 2026b) to hidden-state drafting without requiring the stage concepts to appear in the input or output text. The full teacher construction is given in Appendix A.

The semantic loss is then

$$\mathcal{L}_{\text{sg}} = -\sum_{m=1}^{M} \sum_{s=1}^{k_m} \sum_{k \in \Omega_m^{(s)}} q_m^{(s)}(k) \log p_{\theta,m}^{(s)}(k). \quad (5)$$

This loss supervises only the stage window, while the student distribution remains normalized over the full vocabulary. As a result, the latent state remains readable through the language head, yet is not forced to collapse to a single token prediction.

Intermediate-Modality Constraint. Semantic supervision alone does not ensure that a latent state carries image-grounded information. Directly aligning $\mathbf{h}_m^{(s)}$ with visual tokens, however, may move it away from the language-compatible region required for generating u_{m+1}. We therefore construct a stage-dependent reference between language and vision. For each concept token $c_j \in \mathcal{S}_m^{(s)}$, let $\mathbf{e}(c_j)$ be its LM token embedding. The textual centroid is as in Eq. 6. Let $\{\mathbf{z}_{m,i}\}_{i=1}^{N_v}$ be the projected visual tokens of the current sample in the LM hidden space. Using $\bar{\mathbf{t}}_m^{(s)}$ as a query, we retrieve the relevant visual-token indices $\mathcal{I}_m^{(s)}$ and define the visual centroid as in Eq. 6.

$$\bar{\mathbf{t}}_m^{(s)} = \frac{1}{|\mathcal{S}_m^{(s)}|} \sum_{c_j \in \mathcal{S}_m^{(s)}} \mathbf{e}(c_j), \bar{\mathbf{v}}_m^{(s)} = \frac{1}{|\mathcal{I}_m^{(s)}|} \sum_{i \in \mathcal{I}_m^{(s)}} \mathbf{z}_{m,i}. \quad (6)$$

After normalizing the centroids as $\mathbf{t}_m^{(s)} = \bar{\mathbf{t}}_m^{(s)} / \|\bar{\mathbf{t}}_m^{(s)}\|_2$ and $\mathbf{v}_m^{(s)} = \bar{\mathbf{v}}_m^{(s)} / \|\bar{\mathbf{v}}_m^{(s)}\|_2$, we construct the intermediate reference by spherical linear interpolation:

$$\mathbf{r}_m^{(s)} = \frac{\sin((1-\tau_m^{(s)})\theta_m^{(s)})}{\sin \theta_m^{(s)}} \mathbf{v}_m^{(s)} + \frac{\sin(\tau_m^{(s)}\theta_m^{(s)})}{\sin \theta_m^{(s)}} \mathbf{t}_m^{(s)},$$
$$\theta_m^{(s)} = \arccos\left(\mathbf{v}_m^{(s)\top} \mathbf{t}_m^{(s)}\right). \quad (7)$$

The interpolation coefficient $\tau_m^{(s)}$ specifies the position of the reference between the visual and textual poles:

$$1 \geq \tau_m^{(1)} \geq \cdots \geq \tau_m^{(k_m)} \geq 0. \quad (8)$$

Model	ERQA	VSI-Bench	RealWorldQA	HallusionBench	MMStar	Overall
Zero-Shot VLMs						
Qwen2.5-VL-8B-Instruct (Bai et al., 2025b)	35.3	53.2	68.5	52.9	63.9	54.8
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	42.8	57.1	70.2	61.1	68.6	60.0
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	43.5	54.8	73.5	65.4	73.9	62.2
InternVL3.5-8B (Wang et al., 2025b)	37.0	53.9	67.5	54.5	69.3	56.4
VL-Rethinker (Wang et al., 2026a)	37.5	52.7	68.2	71.1	63.2	58.5
Vision-R1 (Huang et al., 2025)	32.5	48.2	64.7	63.8	62.7	54.4
Latent Reasoning						
Monet (Wang et al., 2025a)	34.5	51.8	68.9	56.4	60.3	54.4
LVR (Li et al., 2025)	36.8	54.5	69.3	65.2	57.9	56.7
Laser (Wang et al., 2026b)	36.0	52.9	69.5	67.7	60.3	57.3
LatentDraft (Ours)						
LatentDraft (Ours)	**47.8**	**59.2**	**73.6**	**71.8**	**72.1**	**64.9**
$\Delta\uparrow$ vs. Qwen3-VL-8B-Instruct	+5.0	+2.1	+3.4	+10.7	+3.5	+4.9
Ablation Study						
LatentDraft (w/o Latent Thought)	44.5	58.3	71.9	69.5	70.8	63.0
LatentDraft (w/o Intermediate-Modality Constraint)	44.0	57.5	72.5	71.2	72.1	63.5
LatentDraft (w/o Semantic Guidance)	43.5	59.0	70.6	68.8	69.7	62.3

Table 1: Main results comparing **LatentDraft** with baselines across four categories: Zero-Shot VLMs, Latent Reasoning methods, and our proposed LatentDraft, along with ablation studies. The best results are highlighted in **bold**. Δ denotes the absolute performance gain over Qwen3-VL-8B-Instruct (Baseline Model). Overall is computed as the average across all five benchmarks.

Larger $\tau_m^{(s)}$ places the reference closer to the textual centroid, while smaller $\tau_m^{(s)}$ moves it closer to the visual centroid. Thus, earlier latent steps are biased toward the language side to preserve compatibility with high-level semantic references, whereas later steps move toward the visual side to carry more concrete image evidence.

Finally, we encourage the latent state to prefer the intermediate reference over either endpoint:

$$\mathcal{L}_{\text{imc}} = \sum_{m=1}^{M} \sum_{s=1}^{k_m} \Big[[d(\mathbf{h}_m^{(s)}, \mathbf{r}_m^{(s)}) - d(\mathbf{h}_m^{(s)}, \mathbf{v}_m^{(s)}) + \delta]_+ + [d(\mathbf{h}_m^{(s)}, \mathbf{r}_m^{(s)}) - d(\mathbf{h}_m^{(s)}, \mathbf{t}_m^{(s)}) + \delta]_+ \Big], \tag{9}$$

where $d(\mathbf{a}, \mathbf{b}) = 1 - \cos(\mathbf{a}, \mathbf{b})$, $[\cdot]_+ = \max(0, \cdot)$, and δ is a safety margin. The hinge form produces gradients only while $\mathbf{h}_m^{(s)}$ has not entered the intermediate neighbourhood. Once the constraint is satisfied, the gradient vanishes, preventing latent states from collapsing onto a single reference point. This gives a geometric interpretation of the intermediate modality: $\mathbf{h}_m^{(s)}$ collapses neither into the language centroid nor into the visual centroid, but stably resides between them while preserving stage-level conceptual superposition.

Overall objective. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{sg}} \mathcal{L}_{\text{sg}} + \lambda_{\text{imc}} \mathcal{L}_{\text{imc}}. \tag{10}$$

Here $\mathcal{L}_{\text{CE}}$ is the cross-entropy calculated on the explicit reasoning chain and the final answer, $\mathcal{L}_{\text{sg}}$ and $\mathcal{L}_{\text{imc}}$ are the semantic-guidance loss and the intermediate-modality constraint loss applied on latent thoughts. λ_{sg} and λ_{imc} are set to 1. Together, the three terms preserve the generative textual path, keep latent drafts semantically readable, and encourage hidden-space computation to carry visual evidence that is difficult to faithfully verbalize. A more detailed geometric motivation analysis is provided in Appendix B.

4 Experiments

4.1 Main Experiments

We evaluate LatentDraft across five comprehensive benchmarks to assess its capabilities along different dimensions, with the results reported in Table 1. Experimental setups are detailed in Appendix C. LatentDraft achieves significant improvements over the baseline on all benchmarks and also outperforms existing latent reasoning methods. Its core advantages are reflected in the three aspects:

Hallucination mitigation. LatentDraft achieves the most pronounced improvement on HallusionBench, indicating that adaptively triggered latent thoughts, combined with joint visual-semantic optimization, can effectively identify and correct misalignments in visual understanding, thereby substantially reducing the model's tendency to produce

visual hallucinations.

Robustness in complex reasoning. On benchmarks such as ERQA, which emphasize spatial understanding and embodied reasoning, LatentDraft iteratively refines latent thoughts and expands computation in the latent space, enabling a more accurate understanding of scenes and entity relationships. This makes it more effective than one-shot generation or purely explicit reasoning.

Cross-task generalization. LatentDraft performs strongly across diverse benchmarks covering event reasoning, visual-spatial understanding, real-world knowledge, hallucination detection, and general multimodal capabilities. This validates the broad applicability of our method without requiring task-specific adaptation.

Overall, LatentDraft combines the strengths of explicit and implicit reasoning. It achieves broad performance gains over mainstream baselines with limited computational overhead, while preserving the logical planning, semantic progression, and interpretability of natural-language CoT. Unlike conventional text-based CoT methods, LatentDraft shifts part of the computation into adaptively triggered latent thoughts, producing compact intermediate-modality representations. This interleaved design reduces noise from long explicit reasoning traces and improves robustness on long-tail reasoning tasks.

4.2 Performances on Different Tasks

We conduct a more fine-grained analysis of LatentDraft on ERQA to examine the strengths and limitations of this paradigm.

LatentDraft achieves clear gains on Pointing, Spatial Reasoning, State Estimation, Task Reasoning, and Other. These subtasks share a strong dependence on continuous visual grounding: Pointing requires accurate target localization; Spatial Reasoning and State Estimation require understanding scene structure and object states; Task Reasoning involves tracking embodied goals; and Other includes non-standard reasoning patterns such as counterfactual reasoning. By triggering latent thoughts, LatentDraft preserves richer visual evidence and additional computational dimensions beyond discrete language tokens, leading to stronger continuous-space reasoning where natural language alone is insufficient.

In contrast, the gains on Trajectory Reasoning and Action Reasoning are more limited, as

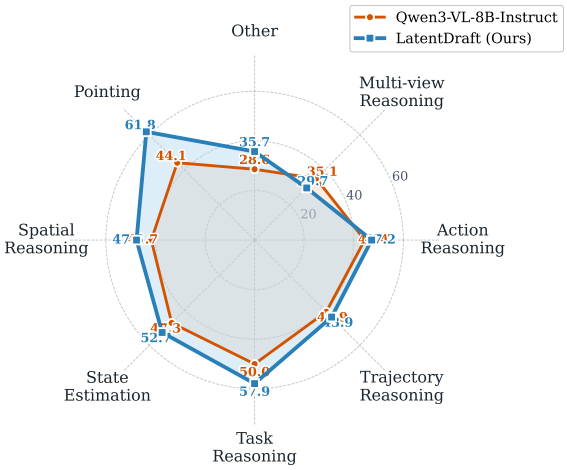


Figure 3: Performance of LatentDraft across eight subtasks of ERQA, achieving accuracy improvements over the baseline on seven of them.

these tasks require stronger spatial extrapolation and physical process modeling. Although latent thoughts retain visual information and extend computation in continuous space, they are still insufficient for accurately simulating complex physical dynamics. Multi-view Reasoning is the only subtask with a performance drop, likely because it requires reconstructing 3D structure from 2D observations, which exceeds the effective reasoning scope of the current latent thought mechanism.

In summary, LatentDraft is most effective on tasks requiring visual grounding and continuous cross-modal reasoning, while its benefits are limited on tasks that demand physical simulation or complex 3D spatial reconstruction. Results suggest that latent thoughts serve as a compact computation channel at the language bottleneck, enabling a form of adaptive reasoning that is compatible with both visual and linguistic representations.

4.3 Efficiency and Latency

We conduct a comprehensive analysis of the computational cost of LatentDraft to verify its usability. Under the same experimental settings, we measure three key metrics: FLOPs, latency, and the number of generated tokens. Figure 4 visualizes the comparison across the three configurations.

Compared with Qwen3-VL-8B-Instruct, LatentDraft introduces a 29.8% increase in FLOPs, a 67.4% increase in latency, and a 66.2% increase in the number of decoded tokens. Although LatentDraft incurs additional computation, the absolute overhead remains small. Compared with Qwen3-

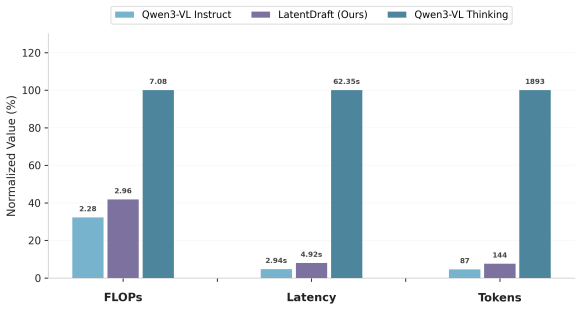


Figure 4: Efficiency comparison among Qwen3-VL-Instruct , LatentDraft, and Qwen3-VL-Thinking. LatentDraft introduces a modest computational overhead.

VL-8B-Thinking, LatentDraft demonstrates substantial efficiency advantages. It consumes only 41.8% of the FLOPs, requires only 7.9% of the runtime, and generates only 7.6% of the tokens, while achieving better accuracy.

These results distill the central idea behind LatentDraft: enabling test-time scaling through targeted and learnable implicit reasoning. Unlike Thinking models, which simply scale up visible reasoning traces, LatentDraft learns the abstract reasoning process as compact implicit representations. This allows the model to substantially improve output quality with only a small and controllable amount of additional computation. This favorable trade-off is the fundamental reason why LatentDraft is viable as a new reasoning paradigm.

4.4 Ablation Study

To verify the contribution of each component in LatentDraft, we conduct a systematic ablation study by separately removing latent thoughts during inference, the Semantic Guidance Loss, and the Intermediate-Modality Constraint during training. The results are shown in Table 1.

Latent thought is the core mechanism that enables adaptive interleaved reasoning in LatentDraft. When its triggers are suppressed, the model must rely solely on the natural-language reasoning path, resulting in a clear overall performance drop. The degradation is most pronounced in hallucination detection and complex event reasoning, where iterative correction of visual understanding is essential. This highlights the critical role of latent thought during inference.

Semantic Guidance Loss is the most critical training signal among all components. Removing this loss results in the largest overall performance drop. This is because, without the constraint of semantic decoding anchors, the latent thought mechanism loses its ability to form and evolve compositional representations of reasoning semantics.

Intermediate-Modality Constraint regularizes the manifold distribution of draft hidden states, ensuring that latent thoughts do not drift away from key visual cues during iterative refinement. Removing this loss also leads to a certain degree of performance degradation, indicating that this constraint plays an important role in enabling native visual thinking capability.

The components of LatentDraft exhibit strong synergy. Latent thought constitutes the core of the reasoning paradigm, while also serving as the carrier of the intermediate modality and the mechanism for test-time scaling. Semantic Guidance Loss and Intermediate-Modality Constraint guide the optimization of latent thoughts from two complementary perspectives, namely semantic decoding and geometric manifold regularization. Both are indispensable.

5 Modality Manifold Analysis: Latent Thought as an Intermediate Modality

To understand how LatentDraft achieves visual grounding through latent reasoning, we conduct a geometric analysis of the representation space. Specifically, we examine three modality embeddings: the vision encoder output $\mathbf{V}$, the text word embeddings $\mathbf{T}$, and the thought hidden states $\mathbf{H}$ produced by LatentDraft during latent reasoning. Our central hypothesis is that latent reasoning creates an *intermediate modality*, in which the thought hidden states $\mathbf{H}$ simultaneously encode visual and linguistic information, thereby bridging the modality gap without compromising semantic capacity.

Modality Gap Reduction. Figure 5(a) shows that training moves $\mathbf{H}$ closer to both visual and textual centroids, while the angle between $\mathbf{T}$ and $\mathbf{V}$ remains unchanged. This indicates that LatentDraft reshapes the thought hidden states without altering the base model's input representations. Figure 6 further confirms this trend: $\cos(\mathbf{H}, \mathbf{V})$ increases substantially, $\cos(\mathbf{H}, \mathbf{T})$ improves moderately, and $\cos(\mathbf{T}, \mathbf{V})$ stays stable. These results show that the thought hidden states acquire visual information while preserving and even strengthening their alignment with textual semantics.

Orthogonal Decomposition: Lossless Visual Encoding. Whether increased visual alignment

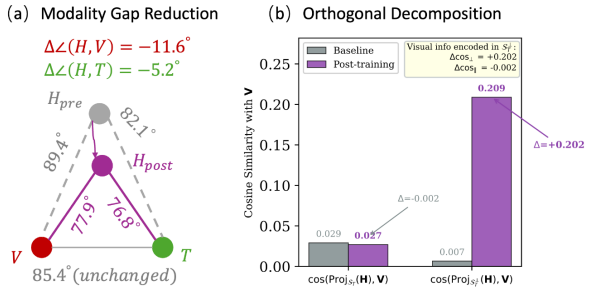


Figure 5: Modality gap and orthogonal decomposition. (a) Angular relationships among $\mathbf{V}$, $\mathbf{T}$, $\mathbf{H}$ centroids before and after training. After training, $\mathbf{H}$ moves closer to both $\mathbf{V}$ and $\mathbf{T}$, while the $\mathbf{T}$-$\mathbf{V}$ angle remains unchanged. (b) Orthogonal decomposition results: visual information is almost exclusively encoded by the orthogonal component $\mathrm{Proj}_{\mathcal{S}_T^{\perp}}(\mathbf{H})$, leaving the semantic component unaffected.

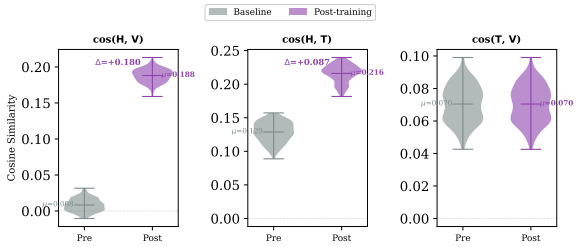


Figure 6: Per-sample cosine similarity distributions. After training, $\cos(\mathbf{H}, \mathbf{V})$ increases dramatically, $\cos(\mathbf{H}, \mathbf{T})$ shows a moderate gain, and $\cos(\mathbf{T}, \mathbf{V})$ stays constant.

comes at the expense of semantic capacity is a key question. Let $\mathcal{S}_T$ denote the subspace spanned by the top principal components of $\mathbf{T}$, and $\mathcal{S}_T^{\perp}$ its orthogonal complement. We decompose each thought hidden state as:

$$\mathbf{h} = \underbrace{\mathrm{Proj}_{\mathcal{S}_T}(\mathbf{h})}_{\text{semantic}} + \underbrace{\mathrm{Proj}_{\mathcal{S}_T^{\perp}}(\mathbf{h})}_{\text{orthogonal}}. \tag{11}$$

Figure 5(b) presents the critical finding. Before training, both components have negligible similarity with $\mathbf{V}$. After training, the orthogonal component's similarity with $\mathbf{V}$ surges dramatically, while the semantic component remains virtually unchanged. This demonstrates that virtually all visual information is encoded in directions orthogonal to the semantic subspace, leaving the text-aligned component intact. Moreover, the overwhelming majority of $\mathbf{H}$'s energy resides in $\mathcal{S}_T^{\perp}$, indicating that the high-dimensional representation space provides ample capacity for encoding visual information without interfering with the semantic subspace.

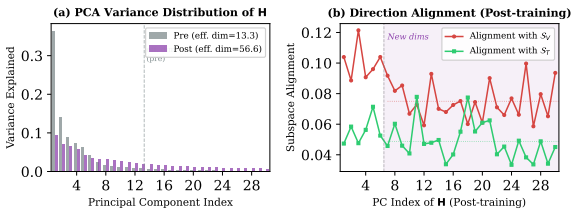


Figure 7: Dimensionality expansion and directional alignment. (a) PCA variance distribution: the effective dimensionality of $\mathbf{H}$ increases substantially after training, with variance spread across more principal components. (b) Newly activated directions predominantly align with the vision subspace $\mathcal{S}_V$.

Dimensionality Expansion. Figure 7(a) reveals a striking structural change: the effective dimensionality of $\mathbf{H}$ increases substantially after training. The pre-training $\mathbf{H}$ concentrates variance in the first few principal components, while the post-training $\mathbf{H}$ distributes variance much more uniformly across many directions. Figure 7(b) shows that these newly activated directions are predominantly aligned with the vision subspace $\mathcal{S}_V$ rather than the text subspace $\mathcal{S}_T$, confirming that the dimensionality expansion serves to accommodate visual information.

Summary. The modality manifold analysis shows that LatentDraft encodes visual information in the orthogonal complement of the semantic subspace, enabling cross-modal fusion without semantic degradation. The thought hidden states therefore form an *intermediate modality* that carries visual and linguistic information in orthogonal subspaces.

6 Conclusion

We propose a multimodal reasoning paradigm that takes natural language as the explicit reasoning backbone while adaptively invoking latent reasoning at language bottlenecks. The explicit and latent reasoning processes complement each other, forming an interleaved reasoning flow. Based on this paradigm, we introduce **LatentDraft**, which uses carefully constructed training data and tailored supervision to shape latent thoughts into an intermediate modality between vision and language. This design allows the model to retain richer cross-modal information in continuous hidden states and to use latent thoughts as compact carriers of expanded computation. Extensive experiments show that LatentDraft brings consistent improvements across multiple dimensions of VLM capability.

Limitations

Given the inherent limitations of supervised fine-tuning, LatentDraft currently still learns latent-thought triggering patterns by fitting to large-scale data. Although the triggering process is adaptive, it essentially remains a form of imitation learning and lacks genuine exploration. Future research should therefore consider incorporating Reinforcement Learning to internalize latent reasoning as an intrinsic mechanism of the model, rather than treating it merely as a form of expanded computation at inference time. In addition, due to the special nature of explicit-implicit interleaved reasoning, latent thoughts can currently support only a limited number of effective evolution steps and still struggle to carry overly fine-grained semantic evolution in the reasoning process.

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, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv e-prints*, pages arXiv–2502.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024b. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 192–210.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pages 27896–27923.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. 2026. Understanding language prior of LVLMs

by contrasting chain-of-embedding. In *The Fourteenth International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. 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.

Xuan Shen, Yizhou Wang, Yufa Zhou, Xiangxi Shi, Pu Zhao, Yanzhi Wang, and Jiuxiang Gu. 2025. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*.

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

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, and 1 others. 2025. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2026a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:30865–30891.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2025a. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. 2026b. Forest before trees: Latent superposition for efficient visual reasoning. *arXiv preprint arXiv:2601.06803*.

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.

xAI. 2024. Grok-1.5 Vision Preview. `https://x.ai/news/grok-1.5v`. Published April 12, 2024. Introduces the RealWorldQA benchmark. Accessed: 2026-05-26.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. 2018. Breaking the softmax bottleneck: A high-rank rnn language model. In *International Conference on Learning Representations*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. 2025a. Weaving context across images: Improving vision-language models through focus-centric visual chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27782–27798.

Juntian Zhang, Song Jin, Chuanqi Cheng, Yuhan Liu, Yankai Lin, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, and 1 others. 2025b. Viper: Empowering the self-evolution of visual perception abilities in vision-language model. *arXiv preprint arXiv:2510.24285*.

Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, and 1 others. 2023. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Teacher Distribution Construction

For completeness, we describe the teacher distribution used in Eq. 5. For each latent state $\mathbf{h}_m^{(s)}$, we first gather the logits on the stage window $\Omega_m^{(s)}$ and construct a detached soft target:

$$\tilde{q}_m^{(s)} = \mathrm{sg}\left[\mathrm{softmax}\left(\ell_m^{(s)}|_{\Omega_m^{(s)}}/T\right)\right]. \tag{12}$$

Similar to Laser (Wang et al., 2026b), we apply an entropy-triggered intervention when the restricted soft target is overly flat. Let $\bar{\mathcal{S}}_m^{(s)} = \mathcal{S}_m^{(s)} \setminus \mathcal{X}$. The prototype $y_m^{(s)}$ is uniform over $\bar{\mathcal{S}}_m^{(s)}$ and zero on $\Omega_m^{(s)} \setminus \bar{\mathcal{S}}_m^{(s)}$. The final teacher is

$$q_m^{(s)} = \beta_m^{(s)} y_m^{(s)} + \left(1 - \beta_m^{(s)}\right) \tilde{q}_m^{(s)}, \tag{13}$$

where

$$\begin{aligned} \beta_m^{(s)} &= \alpha_{\mathrm{mix}} \mathbb{1}\left[\bar{H}(\tilde{q}_m^{(s)}) > \eta\right], \\ \bar{H}(q) &= -\frac{\sum_{k \in \Omega_m^{(s)}} q(k) \log q(k)}{\log |\Omega_m^{(s)}|}. \end{aligned} \tag{14}$$

If $\bar{\mathcal{S}}_m^{(s)}$ is empty, we set $\beta_m^{(s)} = 0$.

B Detailed Geometric Motivation

Language readout under-saturates the hidden state. For a latent state $\mathbf{h} = \mathbf{h}_m^{(s)}$, the language head produces next-token logits through a linear readout

$$\ell(\mathbf{h}) = W_U \mathbf{h}, \qquad W_U \in \mathbb{R}^{|\mathcal{V}| \times d}. \tag{15}$$

Thus, the language distribution observes $\mathbf{h}$ only through the language-visible projection induced by W_U. A perturbation Δ changes the logits by

$$\ell(\mathbf{h} + \Delta) - \ell(\mathbf{h}) = W_U \Delta. \tag{16}$$

Consequently, directions with small language-head sensitivity, i.e., small $\|W_U \Delta\|$ relative to $\|\Delta\|$, can carry additional information while only weakly perturbing the next-token distribution.

This view is consistent with two known geometric properties of language models. First, the linear-softmax readout imposes a low-rank bottleneck on the family of output distributions (Yang et al., 2018). Second, contextual language representations are often anisotropic and concentrate in narrow conic regions (Gao et al., 2019; Ethayarajh, 2019). Together, these observations suggest that the hidden state has representational capacity that is not fully saturated by the language-generation readout alone. We refer to this slack as ***Information Under-Saturation*** (IUS).

Visual and textual information occupy separated directions. A second ingredient is the geometric separation between visual and textual representations. Prior work on multimodal representation learning observes that even semantically aligned visual and textual embeddings may occupy separated regions or cones in the shared space (Liang et al., 2022). In a VLM, visual tokens are projected into the LM hidden space before being consumed by the language model. After this projection, visual evidence need not coincide with the dominant directions used by textual representations. Instead, it can be represented along directions that are compatible with, but not identical to, the language manifold.

This separation is important for latent drafting. If visual evidence were forced to collapse directly into discrete language tokens, fine-grained spatial or relational information could be lost. If it were represented as purely visual tokens, it could become difficult for subsequent language generation to consume. The useful region for a latent draft is therefore not either endpoint, but a transition region in which visual evidence remains available while the state is still accessible to the language pathway.

Why the Intermediat-Modality works. IUS and the modality gap together provide the geometric intuition behind LatentDraft. A latent draft $\mathbf{H}_m = \{\mathbf{h}_m^{(1)}, \ldots, \mathbf{h}_m^{(k_m)}\}$ is not treated as compressed text or as a sequence of visual tokens. Instead, each latent state is shaped to preserve a language-visible semantic projection while also carrying image-grounded information in hidden-space directions that are not fully exhausted by the next-token readout.

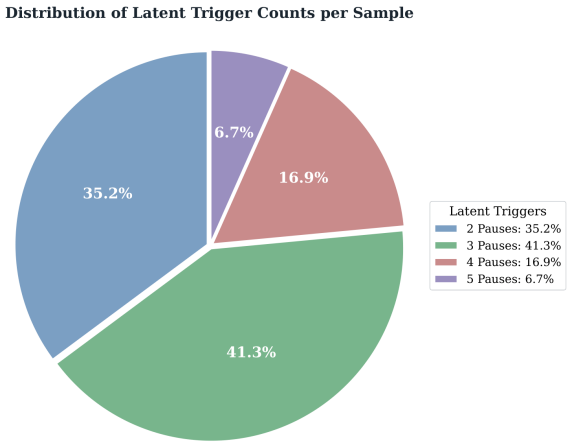


Figure 8: Percentage distribution of latent thinking pause counts per sample in the training dataset.

Under this view, the two supervision terms introduced above play complementary roles. $\mathcal{L}_{\text{sg}}$ constrains the language-visible projection of $\mathbf{h}_m^{(s)}$ through the language head, so that the latent state remains readable by the subsequent textual reasoning path. In contrast, $\mathcal{L}_{\text{imc}}$ shapes the hidden-space geometry of $\mathbf{h}_m^{(s)}$ by placing it relative to textual and visual centroids. The former controls what the latent state exposes to the vocabulary space, while the latter encourages the state to carry visual evidence that may not be faithfully expressible as a single token or short textual phrase.

This explains why the intermediate reference is preferable to directly aligning latent states with either endpoint. Aligning only with the textual centroid would preserve language compatibility but risk losing image-grounded evidence. Aligning only with the visual centroid would strengthen visual grounding but could interfere with the generation of u_{m+1}. Placing the state between the two centroids gives the model a controlled intermediate medium: the latent state can be attended to by later textual reasoning while retaining visual information that would be difficult to verbalize explicitly. The two losses are therefore designed to reduce direct conflict by acting on different aspects of the same representation: the language-visible projection and the hidden-space cross-modal geometry.

C Experimental Setup

We use `Qwen3-VL-8B-Instruct` as the base model and perform supervised fine-tuning on 50K synthetic samples constructed in the LatentDraft format. Training is conducted on 16 H20 GPUs. During evaluation, we set Temperature to 0 and use `do_sample=False` to ensure stable outputs. For each benchmark, we follow its official testing and evaluation protocol, and report the best result over three runs. The samples used in the modality analysis experiments are all drawn from ERQA and VSI-Bench.

D Data Details

For training data construction, we use the Claude-4.7-opus model to synthesize 50K training samples in the LatentDraft format, with images sourced from the VCR and VIRL-39K datasets. We further conduct a detailed statistical analysis of the distribution of latent thought trigger counts in the data, as shown in Figure 8. The dataset covers samples with 2–5 latent thought triggers, with an average of 2.95 latent reasoning steps per sample. Figure 9 shows a sample of the training data.

E Baseline Details

We compare LatentDraft with two groups of external strong baselines: general-purpose or reasoning-enhanced VLMs, and latent reasoning methods.

(1) General-purpose VLM baselines.

- **Qwen2.5-VL-8B-Instruct** (Bai et al., 2025b) is a strong open-source vision-language model with broad multimodal understanding capabilities, including visual recognition, object localization, document understanding, chart reasoning, and video understanding.
- **Qwen3-VL-8B-Instruct** (Bai et al., 2025a) is used as the main reference model in our experiments. It provides a stronger multimodal backbone and supports interleaved image, text, and video contexts.
- **Qwen3-VL-8B-Thinking** is the reasoning-oriented variant of Qwen3-VL (Bai et al., 2025a). It improves complex reasoning mainly by scaling explicit visible reasoning traces, making it a natural comparison for LatentDraft, which instead introduces compact hidden-state drafts.
- **InternVL3.5-8B** (Wang et al., 2025b) is included as another competitive open-source VLM baseline with strong general multimodal ability, reasoning capability, and inference efficiency.

(2)Reasoning-enhanced VLM baselines.
We further compare with VLMs explicitly optimized for reasoning.

- **VL-Rethinker** (Wang et al., 2026a) enhances the slow-thinking ability of VLMs through reinforcement learning and explicit self-reflection, using mechanisms such as selective sample replay and forced rethinking.
- **Vision-R1** (Huang et al., 2025) is a reasoning-oriented multimodal model trained with cold-start multimodal chain-of-thought data and reinforcement learning to encourage stronger reasoning behaviors.

These baselines evaluate whether explicit textual reflection and RL-induced reasoning are sufficient for the visual reasoning tasks considered in this work.

(3)Latent reasoning baselines.
The most direct competitors to LatentDraft are latent reasoning methods, which move part of the reasoning process from natural-language tokens to hidden or visual-latent representations.

- **Monet** (Wang et al., 2025a) enables MLLMs to reason in latent visual space by generating continuous embeddings that function as intermediate visual thoughts.
- **LVR** (Li et al., 2025) performs latent visual reasoning by training the model to generate latent states that reconstruct answer-critical visual tokens in the visual embedding space.
- **Laser** (Wang et al., 2026b) aligns latent states with dynamic semantic windows, allowing the model to preserve a superposition of plausible future semantics before committing to more specific visual judgments.

Compared with these methods, LatentDraft does not treat latent states as either compressed CoT tokens or purely visual embeddings. Instead, it formulates latent drafts as an intermediate modality between visual evidence and linguistic expression.

F Benchmark Details

We evaluate LatentDraft on five complementary multimodal benchmarks. Together, these benchmarks cover embodied reasoning, visual-spatial understanding, real-world visual grounding, hallucination-sensitive reasoning, and general multimodal capability. Unless otherwise specified, we follow the official evaluation protocol of each benchmark and report the corresponding percentage score. The overall score is computed as the unweighted average across the five benchmarks.

ERQA. ERQA, short for Embodied Reasoning Question Answering, is a multimodal benchmark for evaluating embodied reasoning capabilities in realistic physical scenarios (Team et al., 2025). It contains interleaved image-text multiple-choice questions that require models to reason about spatial relations, object states, world knowledge, and robot-relevant task progress. We include ERQA because it directly tests whether a VLM can ground its reasoning in visual evidence when answering questions about physical scenes. In our fine-grained analysis, ERQA is further divided into subtasks such as Pointing, Spatial Reasoning, State Estimation, Task Reasoning, Trajectory Reasoning, Action Reasoning, Multi-view Reasoning, and Other, allowing us to examine where latent drafting is most beneficial.

VSI-Bench. VSI-Bench is a video-based visual-spatial intelligence benchmark designed to test whether multimodal models can see, remember, and recall spaces from sequential visual observations (Yang et al., 2025). It contains more than 5,000 question-answer pairs over real indoor scenes and covers configurational, measurement-estimation, and spatiotemporal reasoning tasks. VSI-Bench is particularly relevant to LatentDraft because many of its questions require continuous spatial reasoning, scene memory, and viewpoint-aware understanding, which may be difficult to faithfully express using only discrete textual chain-of-thought.

RealWorldQA. RealWorldQA evaluates practical real-world visual understanding and spatial grounding (xAI, 2024). The benchmark consists of natural images from vehicles and other everyday scenarios, each paired with a question and an easily verifiable answer. Its questions often involve object counting, direction, relative position, scene layout, and commonsense physical interpretation. We use RealWorldQA to assess whether LatentDraft improves grounding in ordinary real-world images rather than only in synthetic or highly curated settings.

HallusionBench. HallusionBench is a diagnostic benchmark for studying entangled language

hallucination and visual illusion in large vision-language models (Guan et al., 2024). It contains human-crafted image-context reasoning questions with control structures designed to reveal whether a model is answering according to the visual evidence or relying on misleading language priors. This benchmark is central to our evaluation because LatentDraft is designed to reduce visually ungrounded reasoning by allowing difficult visual computations to be carried in hidden space before returning to natural language.

MMStar. MMStar is a vision-indispensable multimodal benchmark consisting of 1,500 human-selected samples that are designed to reduce text-only shortcuts and data leakage (Chen et al., 2024a). It evaluates six core multimodal capabilities and eighteen fine-grained axes, covering a broad range of perception and reasoning skills. We include MMStar as a general multimodal benchmark to test whether LatentDraft improves overall VLM capability beyond task-specific spatial or hallucination-sensitive settings.

G The LatentDraft Reasoning Process: A Case Study

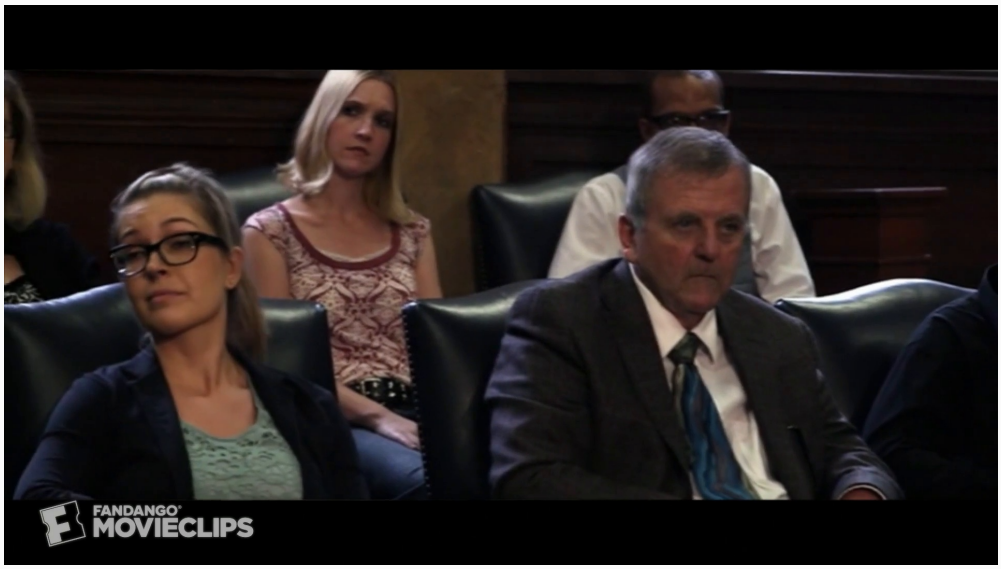
Figure 10 illustrates how LatentDraft handles a height-based visual reasoning problem. The question cannot be answered by color recognition alone: the model must first associate each colored circle with a shelf level, estimate its vertical position relative to the floor, and then compare it with the 4-foot threshold. LatentDraft allocates these operations to different reasoning media. The textual trace first states the high-level plan, while latent drafts are inserted at the key visual bottlenecks corresponding to the top, middle, and bottom shelf regions. These hidden drafts preserve continuous visual evidence such as shelf height, perspective, and floor reference, which would be difficult to faithfully verbalize as discrete tokens. The following textual segments then consume the latent drafts to make interpretable judgments: the red circle on the upper shelf and the yellow circle on the middle shelf are above the threshold, whereas the green and blue circles on the lower shelf are not. This leads to the correct answer, C.

This example highlights the central advantage of LatentDraft. Rather than producing a longer textual chain of thought, the model separates readable problem decomposition from visually grounded hidden-state computation. The language segments provide an interpretable scaffold, while the latent drafts carry the continuous spatial reasoning needed for the final decision.

A Case Study of LatentDraft Training Data

Task Type: physical

Question: Based on the heads and gazes of the people seated, where is the focal point of their attention located relative to the camera?



Reasoning with Latent Thought Triggers:

To work out where the audience is looking I first need the orientation of each visible head; let each one cast a forward gaze ray r_i. <|Latent|>

Latent Thought 1: Head Orientation and Gaze Direction

Abstract	courtroom gallery; seated audience rows; front woman head; front man head; back blonde woman head.
Concrete	back bespectacled man head; head yaw left; gaze direction left; off-axis alignment.

Those four head-rays all swing off the camera axis toward the same side of the frame, none of them facing the lens and none drifting to the opposite wall. With that bundle of rays in mind, I now ask where they intersect and at what vertical level, since a slight downward or level pitch tells me whether the target is elevated like a bench or low like a seated witness. <|Latent|>

Latent Thought 2: Sightline Convergence and Target Location

Abstract	courtroom gallery; seated audience rows; front man gaze line; front woman gaze line.
Concrete	back woman gaze line; converging sightlines; off-frame right direction; seated eye-level height.

Following those rays outward, they meet just past the right edge of the frame at roughly seated eye height, consistent with a speaker or witness positioned off-camera to that side rather than an elevated judge.

Answer: Off the camera's right side, roughly level with the seated viewers.

Figure 9: A complete LatentDraft training example. The natural-language reasoning trace contains two adaptive latent thought triggers. Each trigger is paired with abstract and concrete latent key tokens that supervise the model to preserve intermediate visual cues and refine them into task-relevant spatial evidence.

The LatentDraft Interleaved Reasoning Process: A Case Study

Question: What color circles correspond to points higher than 4 feet?

Choices:
A. Red.
B. Green, yellow, and red.
C. Yellow and red.
D. None of the above.

Please answer directly with only the letter of the correct option and nothing else.

Reasoning:
I need to identify the vertical position of each colored circle on the shelving unit and estimate its height relative to the floor. **<|latent|>** shelf top **<|/latent|>** The red circle sits on the uppermost shelf, which is clearly above the 4-foot threshold. **<|latent|>**shelf middle **<|/latent|>** The yellow circle rests on the middle shelf, and judging from the visible shelf depth and the floor reference, this shelf is also well above 4 feet. **<|latent|>** shelf bottom **<|/latent|>** The green and blue circles are on the bottom shelf area, which sits well below 4 feet, so they do not qualify. **<|latent|><|/latent|>** Since only the red and yellow circles lie higher than 4 feet, the answer is C.

Final Answer: C
Correct Answer: C

Figure 10: Case study of LatentDraft on height-based visual reasoning. The task requires the model to identify the shelf level of each colored circle and compare its height with a 4-foot threshold. LatentDraft uses natural-language text to organize the reasoning process, while inserting latent drafts at visual bottlenecks such as the top, middle, and bottom shelf regions. The subsequent textual segments consume these hidden-state drafts to conclude that the red and yellow circles are above the threshold, producing the correct answer C. The **red latent markers** are visualization labels for hidden-state drafts rather than decoded natural-language tokens.