MULTIMODAL CONTINUOUS REASONING VIA ASYMMETRIC MUTUAL VARIATIONAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are often constrained by a language-space bottleneck, forcing complex visual reasoning into discrete tokens which can lose perceptual nuance. A promising alternative is continuous latent reasoning, where the goal is to discover implicit reasoning pathways that bridge the multimodal query and the final answer. However, this introduces a severe train-inference mismatch: a training-time posterior, conditioned on the ground-truth answer, can exploit answer-dependent shortcuts. Standard variational training then forces the inference-time prior to mimic a posterior that has access to information unavailable at test time, leading to poor performance. To address this, we propose Asymmetric Mutual Variational Learning (AMVL), a framework that resolves this mismatch via a bidirectional calibration objective. A forward KL divergence trains the target-agnostic prior to match the posterior, while a novel reverse KL divergence simultaneously regularizes the posterior, preventing it from collapsing into inference-incompatible regions and mitigating this "answer leakage". We provide theoretical analysis formalizing this leakage as prior contamination and prove that our dual-KL objective reduces it. We instantiate AMVL in a latent-integrated MLLM and show that it consistently outperforms strong discrete and latent-reasoning baselines, improving the average score on the complex BLINK benchmark by +10.83 and achieving gains of up to +32.00 on individual reasoning tasks, with analyses confirming improved latent-space stability.

1 INTRODUCTION

Human reasoning is inherently multimodal. When we solve a visual puzzle or interpret a complex diagram, we think directly in perceptual, spatial, and abstract representations that language alone cannot fully capture (LeCun et al., 2022; Mahowald et al., 2024). This poses a fundamental challenge for Multimodal Large Language Models (MLLMs): if the intermediate reasoning process is constrained to the discrete token space of natural language, the model is forced to verbalize visual concepts that are intrinsically continuous and high-dimensional. The result is a systematic **language-space bottleneck**—reasoning quality is limited not by the model's representational capacity, but by the expressive constraints of the discrete language space through which all intermediate thought must pass. This bottleneck is particularly acute in vision-language tasks demanding fine-grained spatial abstraction or multi-step planning, where text-based Chain-of-Thought (CoT) can cause models to drift from the visual input, introduce hallucinations, and lose precise perceptual grounding (Xu et al., 2026; Guan et al., 2024).

These limitations have motivated a growing body of work on latent visual reasoning, where models perform intermediate steps directly in a continuous embedding space (Li et al., 2026b; Pham & Ngo, 2025; Yang et al., 2025). Recent methods, including LVR (Li et al., 2026b), Monet (Wang et al., 2026b), and Mull-Tokens (Ray et al., 2025), have shown promise by replacing discrete reasoning tokens with continuous latent states. However, these pioneering methods share a **critical and underexplored limitation**: they all rely on **explicit, hand-crafted supervision signals**—such as reconstruction objectives or alignment losses—to shape what these latent states should encode. The latent reasoning process is thus constrained to encode whatever the designer has pre-specified as important, rather than being free to discover the representations most useful for bridging the input question and the final answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We argue that a more principled alternative lies in formulating latent reasoning as a structured probabilistic inference problem (Kingma & Welling, 2013; Sohn et al., 2015). Instead of prescribing what latent states should encode, this approach allows the model to discover the intermediate representations that most naturally bridge the multimodal input and the target output. This framing immediately suggests a target-aware posterior over latent states for training and a target-agnostic prior for inference. However, applying this idea to powerful autoregressive MLLMs introduces a severe train-inference mismatch (Lin et al., 2025; Wang et al., 2026a). When the posterior has access to the reference answer, it can rely on answer-dependent shortcuts that are unavailable at test time. Under the standard evidence lower bound (ELBO) (Kingma & Welling, 2013), a forward KL term then trains the prior to imitate these posterior latents, causing the inference-time prior to inherit a latent geometry partially shaped by information leakage. As a result, the prior is poorly calibrated for test-time reasoning even though it is optimized to approximate the posterior during training.

To address this, we propose **Asymmetric Mutual Variational Learning (AMVL)**, a unified framework for multimodal continuous reasoning. AMVL directly tackles the train-inference mismatch by establishing an asymmetric mutual learning process between the prior and the posterior. It relies on two complementary regularizers: a **forward KL alignment** term that aligns the prior with the posterior-inferred latent states, and a **reverse KL regularization** term that constrains the posterior relative to the learned prior. Crucially, these are not symmetric peer-learning losses in the style of deep mutual learning; rather, they regulate two distributions with fundamentally different conditioning structures and downstream roles. This dual-KL calibration closes the latent gap from both sides, replacing hand-crafted latent supervision with a self-contained, answer-driven signal. It encourages the latent space to be both expressive during training and well-calibrated for inference. We instantiate AMVL in a latent-integrated MLLM architecture with lightweight variational heads, making it efficient and seamlessly applicable.

In summary, our contributions are as follows:

- We identify train-inference mismatch between a target-aware posterior and a target-agnostic prior as the central obstacle in variational multimodal latent reasoning, and clarify why this problem is not adequately addressed by standard one-sided ELBO training.
- We propose AMVL, an asymmetric mutual learning framework that combines forward prior alignment with reverse posterior support regularization, enabling end-to-end discovery of a continuous reasoning space without external latent supervision.
- We demonstrate through extensive experiments that AMVL consistently outperforms strong discrete and latent-reasoning baselines on challenging multimodal benchmarks, confirming the benefits of inference-compatible continuous latent reasoning.

2 Related Work

2.1 Discrete Multimodal Reasoning

Recent advancements optimize Multimodal Large Language Models (MLLMs) to explicitly *"think about images"* (Hong et al., 2025; Jiang et al., 2025; Zhang et al., 2023; Ni et al., 2025). Frameworks like Vision-R1 (Huang et al., 2026) and PAPO (Wang et al., 2026c) leverage reinforcement learning to elicit extensive natural language Chain-of-Thought (CoT). While interpretable, this paradigm suffers directly from the **language-space bottleneck**: forcing continuous, high-dimensional visual signals into a rigid discrete vocabulary inevitably loses fine-grained perceptual nuances, often causing reasoning drift and hallucinations. To mitigate this, a concurrent *"thinking with images"* paradigm (Su et al., 2025b; Sarch et al., 2025; Chen et al., 2025; Chung et al., 2025; Zhang et al., 2025) intertwines pixel-level visual features directly into intermediate reasoning steps. However, while models like PixelReasoner (Su et al., 2025a) and DeepEyes (Zheng et al., 2026) improve spatial grounding, a fundamental structural limitation remains: their overarching reasoning trajectories are still strictly bound to discrete, autoregressive text generation.

Unlike discrete paradigms, AMVL bypasses textual discretization by operating entirely in a continuous latent space. Leveraging the **inherently high information density** of continuous vectors, AMVL allows abstract spatial logic to fluidly evolve, eliminating discrete vocabulary constraints and ensuring rich perceptual details are preserved.

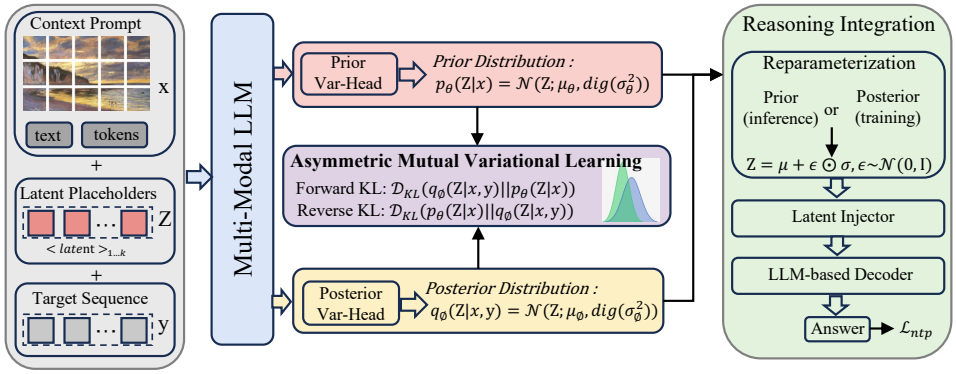


Figure 1: Overview of AMVL for multimodal continuous reasoning. Given a multimodal prompt, the model inserts k latent slots into the autoregressive sequence and uses their hidden states to parameterize a target-agnostic prior $p_\theta(Z|x)$ and a target-aware posterior $q_\phi(Z|x, y)$ over latent sequences $\mathbf{Z}$. During training, posterior samples are injected into the latent slots for decoding, while forward and reverse KL terms jointly calibrate the prior and regularize the posterior. During inference, latent features are sampled from the prior and used to guide autoregressive answer generation.

2.2 Continuous Latent Reasoning

To move beyond purely discrete reasoning trajectories, a growing body of work introduces latent tokens as internal computation slots within autoregressive models (Butt et al., 2025; Geiping et al., 2025; Hao et al., 2022; Shen et al., 2025; Wang et al., 2025a). Early explorations, such as Pause Tokens (Goyal et al., 2024), demonstrated that appending dummy tokens before generating the answer allows language models to perform implicit computation. More recently, representative continuous reasoning methods such as Mull-Tokens (Ray et al., 2025), LVR (Li et al., 2026b), Monet (Wang et al., 2026b), and Coconut (Hao et al., 2022) extend the input sequence with continuous hidden-state capacity, enabling the model to perform intermediate computation without emitting every step as discrete text. These methods suggest that continuous latent reasoning can provide additional expressiveness and computational flexibility.

While existing approaches rely on explicit supervision—reducing latent states to mere compressed proxies—we reformulate the reasoning trajectory as an unobserved stochastic variable. Through variational modeling, our continuous space is optimized solely for effective multimodal reasoning, rather than superficial signal reconstruction.

2.3 Variational Inference for Latent Reasoning

Variational inference offers a principled framework for conditional generation (Kingma & Welling, 2013; Sohn et al., 2015), though applying it to autoregressive decoders historically requires mitigations against posterior collapse (Iyyer et al., 2015; He et al., 2019; Higgins et al., 2017). Recently, this perspective has formalized LLM reasoning trajectories as latent variables (Hu et al., 2023; Zhou et al., 2025). However, whether compressing visual-semantic CoT (ReGuLaR (Wang et al., 2026a)) or steering discrete CoT, these methods bind their optimization to the **explicit alignment of step-by-step traces**. Approaches like RAVR (Lin et al., 2025) use reference-guided posteriors that construct trajectories, making them highly susceptible to **answer leakage**. The posterior exploits the answer as an **informational shortcut**, forcing the target-agnostic prior to mimic this hindsight bias and causing a severe **train-inference mismatch**.

Our approach shares a conceptual resemblance with deep mutual learning (Zhang et al., 2018), as both frameworks involve two distributions mutually regularizing each other during training. However, while traditional mutual learning employs symmetric peers collaborating over the same output space, our prior and posterior are **fundamentally asymmetric**. They possess distinct conditioning structures (target-agnostic vs. target-aware) and operational phases (inference vs. training). By jointly optimizing bidirectional KL objectives, AMVL achieves an **asymmetric mutual calibration**: the forward KL aligns the prior with posterior-inferred states, while the reverse KL restricts the posterior from drifting into inference-incompatible regions, effectively mitigating answer leakage.

3 METHOD

We present Asymmetric Mutual Variational Learning (AMVL), a unified framework for multimodal continuous reasoning that jointly learns a continuous latent space and an autoregressive decoder. As illustrated in Figure 1, AMVL frames latent reasoning as a probabilistic inference problem, allowing the model to *discover* intermediate representations that best connect a multimodal input to its target answer, rather than relying on hand-crafted supervision. Our core contribution is a bidirectional latent calibration objective designed to resolve the train-inference mismatch inherent in this setup.

3.1 A VARIATIONAL APPROACH TO LATENT REASONING

Our approach is founded on the idea that latent reasoning variables can be learned by treating them as unobserved components within a conditional generative model.

Discovering Reasoning via Latent-Variable Modeling. We introduce a sequence of continuous latent variables $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k] \in \mathbb{R}^{k \times d}$ to represent intermediate reasoning steps. The conditional log-likelihood of generating an answer $\mathbf{y}$ from a multimodal context $\mathbf{x}$ is then expressed by marginalizing over these latent variables:

$$\log p_\theta(\mathbf{y} \mid \mathbf{x}) = \log \int p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{Z})\, p_\theta(\mathbf{Z} \mid \mathbf{x})\, d\mathbf{Z}. \tag{1}$$

Here, $p_\theta(\mathbf{Z} \mid \mathbf{x})$ is a **target-agnostic prior** distribution over reasoning states used for inference, and $p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{Z})$ is the decoder that generates the answer conditioned on the latent sequence. Because the integral in Eq. 1 is intractable, we turn to variational inference (Kingma & Welling, 2013; Sohn et al., 2015). This provides a principled way to learn the latent variables by introducing a **target-aware posterior** distribution, $q_\phi(\mathbf{Z} \mid \mathbf{x}, \mathbf{y})$.

By conditioning the posterior on both the input $\mathbf{x}$ and the ground-truth answer $\mathbf{y}$, we provide the model with a powerful, self-contained training signal. The model can effectively use "hindsight" to infer the latent reasoning states $\mathbf{Z}$ that would have been most useful for bridging the gap between the question and the final answer, allowing it to discover optimal reasoning pathways organically.

The Inherent Train-Inference Mismatch. While powerful, this approach leads to the standard Evidence Lower Bound (ELBO):

$$\log p_\theta(\mathbf{y} \mid \mathbf{x}) \;\geq\; \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{Z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{Z} \mid \mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{Z} \mid \mathbf{x}))\,. \tag{2}$$

The first term trains the decoder to generate $\mathbf{y}$ conditioned on latent reasoning states, while the KL term encourages the prior to match the posterior. However, in our setting the posterior has access to the target sequence $\mathbf{y}$ during training, whereas the prior at inference depends only on $\mathbf{x}$. This asymmetry creates a critical failure mode we term **answer leakage**: rather than grounding the latent reasoning trace in the multimodal input, the posterior $q_\phi(\mathbf{Z} \mid \mathbf{x}, \mathbf{y})$ can take a shortcut by relying almost entirely on the reference answer $\mathbf{y}$ to construct $\mathbf{Z}$ (Bowman et al., 2016; Zhao et al., 2017c). Because the standard ELBO uses the forward KL term to pull the prior toward this posterior, the prior is trained to mimic these answer-dependent latent states. At inference time, however, the reference answer is unavailable, and the prior is left navigating a latent space whose structure was shaped by shortcuts it can no longer exploit. The result is a severe **train-inference mismatch**: the prior fails to produce latent states that support useful reasoning, despite having been trained to approximate the posterior. This mismatch is especially acute in strong autoregressive decoders, where the high model capacity makes answer leakage easy—the posterior can encode target information through subtle statistical dependencies that are nearly invisible to the reconstruction loss, yet inaccessible to the prior at test time.

Theoretical Analysis of Answer Leakage. To formalize this failure mode, we decompose the posterior mean into an input-grounded component and an answer-dependent shift: $\mu_\phi(\mathbf{x}, \mathbf{y}) = \mu_\phi^{(x)}(\mathbf{x}) + \delta(\mathbf{x}, \mathbf{y})$, where $\delta(\mathbf{x}, \mathbf{y})$ denotes the posterior displacement induced by the target answer $\mathbf{y}$. Under this decomposition, one can show (Lemma 1 in Appendix C.1) that minimizing the standard ELBO KL term causes the learned prior mean to absorb the average answer-dependent shift $\bar{\delta}(\mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})}[\delta(\mathbf{x}, \mathbf{y})]$ as a residual bias, producing a prior contamination of magnitude $1/2 \sum_{j=1}^{d} \left[\bar{\delta}_j(\mathbf{x})^2 / \sigma_{\theta,j}^2(\mathbf{x})\right]$. Furthermore, one can show (Lemma 2 in Appendix C.2) that forward

KL alignment alone—even with a stop-gradient on the posterior—yields the same contaminated prior optimum and exerts no direct corrective gradient on the posterior itself: the posterior's answer-dependent bias is unchanged. See Appendix C for full statements and proofs.

3.2 Asymmetric Mutual Variational Learning (AMVL)

To resolve this fundamental mismatch, we introduce **Asymmetric Mutual Variational Learning (AMVL)**, a framework that establishes a mutual learning process between the prior and posterior through a bidirectional KL regularization mechanism. While standard ELBO training uses a one-sided objective that can cause the prior to inherit the posterior's answer-leakage bias, as theoretically analyzed in Appendix C (Lemmas 1 and 2), AMVL calibrates the latent space from both directions. This dual-calibration approach is designed to make the latent space both expressive during training and well-calibrated for inference.

AMVL achieves this with two complementary losses that compose its bidirectional calibration mechanism:

- **Forward Prior Alignment ($\mathcal{L}_{\text{fwd}}$).** To improve the prior's calibration, we use a forward KL term (derived from the standard Evidence Lower Bound; see Appendix B.1) to train the prior p_θ to match the latent states inferred by the posterior q_ϕ. We use a stop-gradient ($\text{sg}[\cdot]$) on the posterior to ensure this loss updates only the prior parameters:
$$\mathcal{L}_{\text{fwd}} = D_{\text{KL}}\big(\text{sg}[q_\phi(\mathbf{Z} \mid \mathbf{x}, \mathbf{y})] \,\|\, p_\theta(\mathbf{Z} \mid \mathbf{x})\big). \tag{3}$$
This term effectively teaches the inference-time prior to approximate the latent reasoning states found to be useful during training.
- **Reverse Posterior Regularization ($\mathcal{L}_{\text{rev}}$).** Forward alignment alone is insufficient, as it does not prevent the posterior from exploiting answer leakage in the first place. We therefore introduce a reverse KL term that updates only the posterior, regularizing it against drifting into prior-incompatible regions:
$$\mathcal{L}_{\text{rev}} = D_{\text{KL}}\big(\text{sg}[p_\theta(\mathbf{Z} \mid \mathbf{x})] \,\|\, q_\phi(\mathbf{Z} \mid \mathbf{x}, \mathbf{y})\big). \tag{4}$$
This term directly penalizes posterior drift away from high-density regions of the prior. Proposition 1 formalizes this effect for the common diagonal Gaussian case.

Proposition 1 (Reverse KL suppresses incompatible posterior drift)**.** *For diagonal Gaussian latents $p_\theta(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \text{diag}(\sigma^2_\theta(\mathbf{x}))\big)$ and $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = \mathcal{N}\big(\mu_\phi(\mathbf{x}, \mathbf{y}), \text{diag}(\sigma^2_\phi(\mathbf{x}, \mathbf{y}))\big)$, the reverse KL in Eq. (4) has the closed form*

$$\mathcal{L}_{rev} = \frac{1}{2} \sum_j \left(\frac{\sigma^2_{\theta,j}}{\sigma^2_{\phi,j}} + \frac{(\mu_{\theta,j} - \mu_{\phi,j})^2}{\sigma^2_{\phi,j}} - 1 + \log \frac{\sigma^2_{\phi,j}}{\sigma^2_{\theta,j}} \right),$$

where p_θ is treated as constant by stop-gradient. Its gradient with respect to the posterior mean is

$$\frac{\partial \mathcal{L}_{rev}}{\partial \mu_{\phi,j}} = \frac{\mu_{\phi,j} - \mu_{\theta,j}}{\sigma^2_{\phi,j}}.$$

Thus, reverse KL penalizes posterior drift from prior-compatible high-density latent regions, with a mean-mismatch penalty stronger as the posterior sharpens (i.e., as $\sigma^2_{\phi,j}$ decreases).

Proof. This result is proved in Appendix C.3.

Combining $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{rev}}$ creates bidirectional calibration. While $\mathcal{L}_{\text{fwd}}$ pulls the prior toward the posterior, $\mathcal{L}_{\text{rev}}$ prevents the posterior from drifting into inference-incompatible states. Proposition 2 formalizes how this objective reduces prior contamination caused by answer leakage.

Proposition 2 (Bidirectional calibration reduces prior contamination)**.** *Under the local linear leakage model in Assumption 1, where the answer-dependent shift is written as*

$$\delta(\mathbf{x}, \mathbf{y}) = \alpha f(\mathbf{x}, \mathbf{y}),$$

and under the local linear-response stationary condition in Assumption 2, there exists a constant $c > 0$ such that the local equilibrium leakage coefficient α under AMVL satisfies

$$\alpha_{AMVL} = \frac{\alpha_{ELBO}}{1 + c\gamma/\sigma^2_{eff}},$$

where α_{ELBO} is the equilibrium leakage under one-sided ELBO training, γ is the weight of the reverse KL term, and σ^2_{eff} is an effective posterior variance. Consequently, the prior mean contamination under AMVL satisfies

$$\Delta_{AMVL}(\mathbf{x}) = \|\mu_\theta^{AMVL}(\mathbf{x}) - \mu^{(x)}(\mathbf{x})\|_2 = \alpha_{AMVL}\|\bar{f}(\mathbf{x})\|_2,$$

whereas the corresponding contamination under one-sided ELBO matching is

$$\Delta_{ELBO}(\mathbf{x}) = \alpha_{ELBO}\|\bar{f}(\mathbf{x})\|_2.$$

Therefore,

$$\Delta_{AMVL}(\mathbf{x}) < \Delta_{ELBO}(\mathbf{x})$$

whenever $\gamma > 0$, $\|\bar{f}(\mathbf{x})\|_2 > 0$, and $\alpha_{ELBO} > 0$.

Proof. This result is proved in Appendix C.4.

Proposition 2 formalizes how this dual objective reduces the prior contamination caused by answer leakage. The forward KL term then transfers this less-contaminated signal to the prior, resulting in a model that is better calibrated for inference. From a geometric perspective, these two KL directions establish a mutual mass-covering dynamic. As detailed in Appendix B.3, both KL terms are mass-covering because optimization occurs on the second argument of the divergence. The forward term forces the prior to cover the posterior, while the reverse term forces the posterior to cover the prior, discouraging it from collapsing into narrow, inference-incompatible regions. This dual calibration is the core mechanism allowing AMVL to discover an expressive yet inference-compatible latent reasoning space. Further theoretical discussion, including an information-theoretic view in Proposition 3, is available in Appendix C.

3.3 Implementation within an MLLM

We instantiate AMVL by augmenting a standard MLLM architecture with minimal, lightweight components.

Latent-Integrated Architecture. To integrate latent variables into the autoregressive process, we insert k placeholder tokens, denoted as `<latent>`, into the input sequence between the prompt $\mathbf{x}$ and target $\mathbf{y}$:

$$S = [\mathbf{x}, \texttt{<latent>}_1, \ldots, \texttt{<latent>}_k, \mathbf{y}]. \tag{5}$$

The prior is parameterized from latent placeholders conditioned only on x, whereas a separate posterior pass uses the reordered sequence $S_{\text{post}} = [x, y, \langle \texttt{latent} \rangle_{1:k}]$, enabling the latent hidden states to causally attend to the complete target y. The resulting posterior sample is then injected into the latent positions to condition the decoder for answer generation. We model both distributions as factorized diagonal Gaussians for efficient, closed-form optimization:

$$p_\theta(\mathbf{Z}|\mathbf{x}) = \prod_{i=1}^{k} \mathcal{N}\left(\mathbf{z}_i; \boldsymbol{\mu}_{\theta,i}(\mathbf{x}), \operatorname{diag}(\boldsymbol{\sigma}^2_{\theta,i}(\mathbf{x}))\right), \tag{6}$$

$$q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{k} \mathcal{N}\left(\mathbf{z}_i; \boldsymbol{\mu}_{\phi,i}(\mathbf{x}, \mathbf{y}), \operatorname{diag}(\boldsymbol{\sigma}^2_{\phi,i}(\mathbf{x}, \mathbf{y}))\right). \tag{7}$$

LLM-Native Variational Head. To seamlessly integrate this machinery into the MLLM, we design lightweight variational heads that operate on the model's hidden states. Let $H = [h_1, \ldots, h_k] \in \mathbb{R}^{k \times D}$ be the final-layer hidden states at the k latent token positions. The Gaussian parameters are computed via a projection network:

$$[\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2] = W_{\text{out}}\, \text{SwiGLU}(\text{RMSNorm}(\mathbf{H})), \tag{8}$$

where $\mu, \log \sigma^2 \in \mathbb{R}^{k \times d}$. This "LLM-native" design leverages standard components (Shazeer, 2020; Zhang & Sennrich, 2019), preserving architectural consistency. The prior and posterior use two structurally identical but separately parameterized variational heads, operating on contexts conditioned on $\mathbf{x}$ and $(\mathbf{x}, \mathbf{y})$, respectively. Unless otherwise stated, we set the latent dimension to $d = 512$.

Latent-Conditioned Decoding. We use the standard reparameterization trick (Rezende et al., 2014) to sample a latent sequence $\mathbf{Z}$:

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \boldsymbol{\sigma}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{9}$$

The sampled $\mathbf{Z} \in \mathbb{R}^{k \times d}$ is then projected back to the MLLM's hidden dimension D via a **latent injector** and used to replace the embeddings at the `<latent>` positions, conditioning the final answer generation on the continuous reasoning states.

3.4 Final Training Objective

We optimize the entire model end-to-end by minimizing a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NTP}} + \beta \mathcal{L}_{\text{fwd}} + \gamma \mathcal{L}_{\text{rev}}, \tag{10}$$

where β and γ are scalar weights, and $\mathcal{L}_{\text{NTP}}$ is the standard autoregressive next-token prediction loss, conditioned on a latent sample from the posterior:

$$\mathcal{L}_{\text{NTP}} = -\mathbb{E}_{\mathbf{Z} \sim q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})} \left[\sum_t \log p_{\text{LLM}}(y_t | \mathbf{x}, y_{<t}, \mathbf{Z}) \right]. \tag{11}$$

Closed-Form KL Computation. A key advantage of our diagonal Gaussian parameterization is that both KL terms admit a closed-form solution. For two diagonal Gaussians $\mathcal{N}_1 = \mathcal{N}(\mu_1, \text{diag}(\sigma_1^2))$ and $\mathcal{N}_2 = \mathcal{N}(\mu_2, \text{diag}(\sigma_2^2))$, the KL divergence is:

$$D_{KL}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} \sum_{j=1}^{d} \left(\frac{\sigma_{1,j}^2 + (\mu_{1,j} - \mu_{2,j})^2}{\sigma_{2,j}^2} - 1 + \log \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} \right). \tag{12}$$

This allows for efficient and stable computation of both $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{rev}}$ during training.

Asymmetric KL Scheduling. Rather than using fixed weights, we schedule β and γ to stabilize training, as detailed in Appendix E. The forward KL weight β is warmed up early to allow the prior to begin tracking the posterior. The reverse KL weight γ is introduced with a delay and kept weaker. This strategy allows the prior to become partially calibrated before the reverse regularization constrains the posterior, preventing over-regularization in early stages when the prior is still weak.

Inference. At inference time, the target $\mathbf{y}$ is unavailable. We perform deterministic mean-based inference by setting $\mathbf{Z} = \mu_\theta(\mathbf{x})$, inject the resulting latent representation into the decoder, and generate the answer autoregressively. The bidirectional calibration objective encourages the prior to produce inference-compatible latent representations that support downstream answer generation. We further evaluate robustness to stochastic prior sampling in Appendix I.

4 Experiments

4.1 Experimental Setup

Implementation details. We implement AMVL on Qwen2.5-VL-7B-Instruct (Bai et al., 2025), trained on a mixture of multimodal reasoning datasets including Visual-CoT (Shao et al., 2024), ReFocus (Fu et al., 2025), CogCoM (Qi et al., 2025), and Zebra-CoT (Li et al., 2026a). Unless otherwise specified, we use $k = 8$ latent tokens with dimension $d = 512$. Following (Li et al., 2026b; Wang et al., 2026b), the vision encoder is frozen while the language backbone and variational modules are jointly optimized. Full details are in Appendix E.

Experiment Benchmarks. We evaluate fine-grained visual perception on V^* (Wu & Xie, 2024), HRBench4K (Wang et al., 2025b), and HRBench8K (Wang et al., 2025b); complex visual reasoning on BLINK (Fu et al., 2024); and OOD robustness on VisualPuzzles (Appendix F). Broader-benchmark results are reported in Appendix J.1.

Baselines. We compare against representative SOTA MLLMs built on the same backbone (Qwen2.5-VL-7B), grouped into two paradigms: (1) **Discrete Reasoning**—text-centric approaches (*thinking about images*, e.g., Vision-R1 (Huang et al., 2026), PAPO (Wang et al., 2026c)) and visually-augmented discrete generation (*thinking with images*, e.g., PixelReasoner (Su et al., 2025a), DeepEyes (Zheng et al., 2026)); and (2) **Continuous Latent Reasoning**—LVR (Li et al., 2026b), Mull-Tokens (Ray et al., 2025), and Monet (Wang et al., 2026b). See Appendix D for details.

Table 1: Performance on V*, HRBench4K, and HRBench8K. "Avg." is the mean Overall score across benchmarks. Best and second-best open-source results are highlighted in **bold** and underlined. The last row shows absolute gains over Qwen2.5-VL-7B.

Method	V*			HRBench4K			HRBench8K			Avg.
	Overall	Attribute	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	
Proprietary Models										
GPT-4o	67.50	72.20	60.50	59.00	70.00	48.00	55.50	62.00	49.00	60.67
Open-Source Models										
Qwen2.5-VL-7B	76.44	77.39	75.00	68.00	80.25	55.75	63.75	74.25	53.25	69.40
+ SFT	81.68	83.48	78.95	68.38	78.28	58.50	61.63	70.75	52.50	70.56
+ SFT + GRPO	78.53	78.26	78.95	70.00	83.25	56.75	66.75	78.00	55.50	71.76
DeepEyes	83.25	**84.35**	81.58	71.25	83.75	58.75	65.13	77.00	53.25	73.21
LVR	81.15	80.00	82.89	70.62	83.50	57.75	64.12	77.25	51.00	71.96
PixelReasoner	81.15	80.87	81.58	72.00	84.00	**60.00**	66.12	76.50	55.75	73.09
Mull-Tokens	79.95	81.58	77.49	70.25	86.50	54.00	65.75	81.00	50.50	71.69
Monet	83.25	83.48	82.89	71.00	85.25	56.75	68.00	79.75	**56.25**	74.08
Ours-7B	**84.29**	**84.35**	**84.21**	**72.12**	**87.25**	57.00	**68.50**	**83.50**	53.50	**74.97**
Absolute gain	+7.85	+6.96	+9.21	+4.12	+7.00	+1.25	+4.75	+9.25	+0.25	+5.57

Table 2: Performance on vision-centric BLINK tasks. Obj. Loc., M-View, and Vis. Sim. denote Object Localization, Multi-view Reasoning, and Visual Similarity.

Method	IQ Test	Jigsaw	Obj. Loc.	Art Style	M-View	Spatial Relation	Vis. Sim.	Avg.
Qwen2.5-VL-7B	20.00	45.33	49.18	64.96	41.35	88.81	82.96	56.08
Vision-R1	27.33	54.67	48.36	65.81	45.86	79.72	71.85	56.23
PixelReasoner	26.00	72.00	54.10	67.52	46.62	86.71	85.93	62.70
PAPO	22.67	66.67	**56.56**	64.96	49.62	**90.21**	85.19	62.27
LVR	26.00	52.67	50.00	**70.94**	45.41	88.81	82.22	59.44
Mull-Tokens	31.33	69.33	47.54	62.42	**60.15**	84.61	80.74	62.30
Monet	29.33	42.67	52.46	62.39	50.38	79.02	85.19	57.35
Ours-7B	**32.67**	**77.33**	54.92	70.09	55.64	88.81	**88.89**	**66.91**
Absolute Gain	+12.67	+32.00	+5.74	+5.13	+14.29	+0.00	+5.93	+10.83

4.2 Main Results

Overall Performance Analysis. AMVL establishes a new SOTA among Qwen2.5-VL-7B models. On fine-grained perception benchmarks (Table 1), it achieves a 74.97 average (+5.57), driven by gains on V* (+7.85) and HRBench8K (+4.75). On complex spatial reasoning (Table 2), AMVL elevates the BLINK average by +10.83, highlighted by a +32.00 surge on the topology-heavy Jigsaw task. These gains confirm that bidirectionally regularized latent reasoning enhances both fine-grained visual search and macroscopic structural logic. The advantage extends to OOD settings (Appendix F), indicating that the learned latent space remains calibrated under distribution shift.

Compared with Discrete Multimodal Reasoning. We compare AMVL against state-of-the-art discrete reasoning models, including text-based reasoning chains (*thinking about images*, e.g., Vision-R1, PAPO) and those interleaving visual tools within discrete generation (*thinking with images*, e.g., PixelReasoner, DeepEyes). While these baselines improve upon the base model, they force high-dimensional visual concepts into a lossy text space. As shown in Tables 1 and 2, AMVL consistently outperforms these discrete paradigms, achieving superior accuracy on abstract tasks like IQ Test (32.67) and Visual Similarity in BLINK (88.89), where discrete verbalization often falls short or induces hallucinations. This empirically validates our claim that performing intermediate reasoning directly in continuous perceptual space preserves spatial and abstract representations that discrete language tokens cannot fully capture.

Compared with Continuous Latent Reasoning. We further benchmark AMVL against pioneering *latent reasoning* frameworks, including LVR, Mull-Tokens, and Monet. While these methods bypass the language bottleneck, they rely on *explicit, hand-crafted supervision signals* (e.g., predefined reconstruction targets), restricting their ability to discover optimal intermediate representations. AMVL's systematic superiority across virtually all sub-metrics provides strong evidence for our variational formulation. By framing continuous reasoning as structured probabilistic infer-

Table 3: Ablation of training objectives. "Fwd-KL" and "Rev-KL" denote the forward and reverse KL terms. The last two rows differ only in optimization order.

Method	V*	HRBench4K	HRBench8K
NTP only	81.15	70.50	67.38
Fwd-KL only	40.84	53.37	52.00
Rev-KL only	75.92	69.62	64.38
NTP + Fwd-KL	82.72	72.12	67.75
NTP + Rev-KL	82.20	71.75	67.25
NTP + Fwd-KL + Rev-KL (reverse-first)	80.63	72.38	68.25
NTP + Fwd-KL + Rev-KL (forward-first)	84.29	72.12	68.50

Table 4: Ablation study of latent configuration. We vary the number of latent tokens k and the latent dimension d while keeping other settings fixed.

Setting	Number of latent tokens k			Latent dimension d			
	4 tokens	8 tokens	16 tokens	128	256	512	768
HRBench4K	70.00	72.12	71.75	72.62	71.50	72.12	70.88
HRBench8K	66.25	68.50	67.12	66.75	66.62	68.50	67.38
V*	76.44	84.29	81.15	82.72	76.96	84.29	79.58

ence, AMVL eliminates the need for latent supervision. The consistent gains across in-distribution benchmarks demonstrate that our dual-KL regularization mitigates posterior collapse and improves the quality of the learned latent reasoning space—ensuring representations are expressive during training and stable during inference.

4.3 Ablation Study

Effect of bidirectional variational calibration. Table 3 shows the NTP baseline suffers from a train-inference mismatch. Either calibration direction partially alleviates this: forward alignment trains the prior to track posterior latents, while reverse regularization constrains posterior drift toward prior-supported regions. However, optimizing either direction alone is suboptimal—forward-only training leaves the posterior unconstrained, while reverse-only training can overly suppress useful latent variability. The full AMVL objective performs best, confirming that effective multimodal latent reasoning requires both prior alignment and posterior regularization. Direct counterfactual diagnostics of this leakage mechanism are reported in Appendix J.2.

Effect of latent configuration. Table 4 studies varying latent token count k and dimension d. Increasing k from 4 to 8 improves performance, but further increasing to 16 degrades results. Unlike discrete tokens with low information density, continuous vectors encapsulate rich logic compactly; excessive latent slots thus introduce redundancy and "information dilution," complicating variational optimization. Similarly, increasing d from 512 to 768 degrades performance, as overparameterization exacerbates the dual-KL optimization burden without adding reasoning power.

Additional analyses. Further ablations examine variational-head design, objective weighting, gradient decoupling, and inference-time sampling robustness (Appendix I). Latent-token interventions, visual grounding, reverse-KL annealing and efficiency are detailed in Appendices J.3–J.5.

5 Conclusion

In this paper, we address the language-space bottleneck and train–inference mismatch in multimodal large language models. We propose **Asymmetric Mutual Variational Learning (AMVL)**, a principled framework for continuous latent reasoning that jointly aligns the prior and regularizes the posterior. AMVL decouples latent expressiveness from answer leakage without requiring handcrafted supervision. Experiments on fine-grained perception and abstract reasoning benchmarks show that AMVL achieves state-of-the-art performance over both discrete Chain-of-Thought and prior latent-reasoning methods.

AI USE STATEMENT

Artificial intelligence tools (LLMs) were employed exclusively for copy-editing purposes, including grammar and formatting corrections. All scientific content, methodologies, and conclusions are the original work of the authors. The authors have thoroughly vetted the text and bear complete accountability for the accuracy of this paper.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the complete source code and pre-trained model weights. This release will be accompanied by detailed instructions and evaluation scripts to enable independent replication of our results by other researchers.

REFERENCES

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.

Anirudh Goyal ALIAS PARTH GOYAL, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. Z-forcing: Training stochastic recurrent networks. *Advances in neural information processing systems*, 30, 2017.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pp. 10–21, 2016.

Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft tokens, hard truths. *arXiv preprint arXiv:2509.19170*, 2025.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025.

Jiwan Chung, Junhyeok Kim, Siyeol Kim, Jaeyoung Lee, Min Soo Kim, and Youngjae Yu. Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. *arXiv e-prints*, pp. arXiv–2505, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. In *International Conference on Machine Learning*, pp. 17783–17805. PMLR, 2025.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens, 2024. URL https://arxiv.org/abs/2310.02226.

Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pp. 14375–14385, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024. *URL https://arxiv.org/abs/2412.06769*, 98, 2022.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in advances in approximate Bayesian inference, NIPS*, volume 1, 2016.

Minjie Hong, Zirun Guo, Yan Xia, Zehan Wang, Ziang Zhang, Tao Jin, and Zhou Zhao. Apo: Enhancing reasoning ability of mllms via asymmetric policy optimization. *arXiv preprint arXiv:2506.21655*, 2025.

Edward J Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. *arXiv preprint arXiv:2310.04363*, 2023.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=UZIjskfbfU.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pp. 1681–1691, 2015.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought, 2025. URL https://arxiv.org/abs/2505.16192.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=c6XIVI3TiQ.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Tianqianjin Lin, Xi Zhao, Xingyao Zhang, Rujiao Long, Yi Xu, Zhuoren Jiang, Wenbo Su, and Bo Zheng. Ravr: Reference-answer-guided variational reasoning for large language models. *arXiv preprint arXiv:2510.25206*, 2025.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

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

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540, 2024.

Tom Minka et al. Divergence measures and message passing. 2005.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning, 2025. URL https://arxiv.org/abs/2505.19702.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: A visual language model with chain-of-manipulations reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Fg0eo2AkST.

Arijit Ray, Ahmed Abdelkader, Chengzhi Mao, Bryan A. Plummer, Kate Saenko, Ranjay Krishna, Leonidas Guibas, and Wen-Sheng Chu. Mull-tokens: Modality-agnostic latent thinking, 2025. URL https://arxiv.org/abs/2512.10941.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2, 2024.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

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

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b.

Fanmeng Wang, Haotian Liu, Guojiang Zhao, Hongteng Xu, and Zhifeng Gao. Regular: Variational latent reasoning guided by rendered chain-of-thought, 2026a. URL https://arxiv.org/abs/2601.23184.

Jianwei Wang, Ziming Wu, Fuming Lai, Shaobing Lian, and Ziqian Zeng. Synadapt: Learning adaptive reasoning in large language models via synthetic continuous chain-of-thought. *arXiv preprint arXiv:2508.00574*, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *CVPR*, 2026b.

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

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru WANG, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=izbBqTL8vb.

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

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, et al. Thinking in uncertainty: Mitigating hallucinations in mlrms with latent entropy-aware decoding. *arXiv preprint arXiv:2603.13366*, 2026.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024b. URL https://arxiv.org/abs/2410.16153.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl. *arXiv e-prints*, pp. arXiv–2505, 2025.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, 2018.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017a.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017b.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 654–664, 2017c.

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

Xiangxin Zhou, Zichen Liu, Haonan Wang, Chao Du, Min Lin, Chongxuan Li, Liang Wang, and Tianyu Pang. Variational reasoning for language models. *arXiv preprint arXiv:2509.22637*, 2025.

A Main Notation Introduction

For clarity, Table 5 summarizes the main notations used throughout the paper.

Table 5: Meanings of the main notations used in the paper.

Notation	Type	Meaning
$\mathbf{x}$	sequence / context	Multimodal input context, e.g., visual tokens and text prompt.
$\mathbf{y}$	sequence	Target output sequence (answer tokens).
y_t	token	The t-th token in the target sequence.
$\mathbf{y}_{<t}$	sequence prefix	Target prefix before time step t.
$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_k]$	$\mathbb{R}^{k \times d}$	Continuous latent reasoning sequence consisting of k latent slots.
$\mathbf{z}_i$	$\mathbb{R}^d$	The continuous latent variable at the i-th latent slot.
k	integer	Number of latent slots.
d	integer	Latent dimension of each latent slot.
D	integer	Hidden dimension of the base MLLM.
$\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_k]$	$\mathbb{R}^{k \times D}$	Final-layer hidden states at the latent placeholder positions.
$\mathbf{h}_i$	$\mathbb{R}^D$	Hidden state at the i-th latent token position.
$p_\theta(\mathbf{Z}\|\mathbf{x})$	distribution	Target-agnostic prior distribution over latent reasoning states, conditioned solely on input $\mathbf{x}$.
$q_\phi(\mathbf{Z}\|\mathbf{x}, \mathbf{y})$	distribution	Target-aware variational posterior distribution over latent reasoning states, conditioned on $(\mathbf{x}, \mathbf{y})$.
θ	parameters	Parameters of the prior branch and the autoregressive decoder.
ϕ	parameters	Parameters of the variational posterior branch.
$\boldsymbol{\mu}_{\theta,i}(\mathbf{x})$	$\mathbb{R}^d$	Mean vector of the prior distribution for the i-th latent slot.
$\boldsymbol{\sigma}^2_{\theta,i}(\mathbf{x})$	$\mathbb{R}^d$	Diagonal variance vector of the prior distribution for the i-th latent slot.
$\boldsymbol{\mu}_{\phi,i}(\mathbf{x}, \mathbf{y})$	$\mathbb{R}^d$	Mean vector of the posterior distribution for the i-th latent slot.
$\boldsymbol{\sigma}^2_{\phi,i}(\mathbf{x}, \mathbf{y})$	$\mathbb{R}^d$	Diagonal variance vector of the posterior distribution for the i-th latent slot.
$\boldsymbol{\mu}$	$\mathbb{R}^{k \times d}$	Mean tensor of a given latent Gaussian distribution.
$\boldsymbol{\sigma}^2$	$\mathbb{R}^{k \times d}$	Diagonal variance tensor of a given latent Gaussian distribution.
$\log \boldsymbol{\sigma}^2$	$\mathbb{R}^{k \times d}$	Log-variance tensor predicted by the variational heads.
$\boldsymbol{\epsilon}$	$\mathbb{R}^{k \times d}$	Standard Gaussian noise used in the reparameterization trick, sampled from $\mathcal{N}(0, \mathbf{I})$.
$\mathbf{H}_Z$	$\mathbb{R}^{k \times D}$	Decoder-aligned latent features obtained by projecting $\mathbf{Z}$ through the latent injector.
$\mathrm{Injector}(\cdot)$	mapping	Linear or MLP projection from latent space $\mathbb{R}^d$ to MLLM hidden space $\mathbb{R}^D$.
$\mathrm{sg}[\cdot]$	operator	Stop-gradient operator, used to block gradient flow through a specific branch.
$\mathcal{L}_{\mathrm{fwd}}$	scalar loss	Forward KL alignment loss, $D_{\mathrm{KL}}(\mathrm{sg}[q_\phi(\mathbf{Z}\|\mathbf{x}, \mathbf{y})] \| p_\theta(\mathbf{Z}\|\mathbf{x}))$.
$\mathcal{L}_{\mathrm{rev}}$	scalar loss	Reverse KL regularization loss, $D_{\mathrm{KL}}(\mathrm{sg}[p_\theta(\mathbf{Z}\|\mathbf{x})] \| q_\phi(\mathbf{Z}\|\mathbf{x}, \mathbf{y}))$.
$\mathcal{L}_{\mathrm{NTP}}$	scalar loss	Autoregressive next-token prediction loss conditioned on latent reasoning states.
$\mathcal{L}_{\mathrm{total}}$	scalar loss	Final training objective combining $\mathcal{L}_{\mathrm{NTP}}$, $\mathcal{L}_{\mathrm{fwd}}$, and $\mathcal{L}_{\mathrm{rev}}$.
β	scalar	Scheduling weight for the forward KL loss term.
γ	scalar	Scheduling weight for the reverse KL loss term.
$D_{\mathrm{KL}}(\cdot\|\cdot)$	divergence	Kullback–Leibler divergence between two distributions.

B Variational Derivations and Theoretical Motivation

B.1 Derivation of the ELBO

We begin with the conditional log-marginal likelihood of the target sequence $\mathbf{y}$ given the multimodal context $\mathbf{x}$:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \log \int p_\theta(\mathbf{y}, \mathbf{Z}|\mathbf{x})\, d\mathbf{Z}, \tag{13}$$

where $\mathbf{Z}$ denotes the latent reasoning sequence. This quantity is the ideal training objective, since it marginalizes over all possible latent reasoning trajectories that may support the generation of the

target answer. However, the integral is generally intractable because the latent space is continuous and high-dimensional.

To obtain a tractable objective, we introduce a variational posterior $q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y})$, which approximates the target-conditioned latent distribution during training. Multiplying and dividing by $q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y})$ inside the integral

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \log \int q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y}) \frac{p_\theta(\mathbf{y}, \mathbf{Z}|\mathbf{x})}{q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y})} \, d\mathbf{Z}. \tag{14}$$

This reformulation allows us to express the marginal likelihood as an expectation with respect to the variational posterior.

Applying Jensen's inequality yields a lower bound on the log-marginal likelihood:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})} \left[\log \frac{p_\theta(\mathbf{y}, \mathbf{Z}|\mathbf{x})}{q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y})} \right]. \tag{15}$$

Using the factorization

$$p_\theta(\mathbf{y}, \mathbf{Z}|\mathbf{x}) = p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{Z}) \, p_\theta(\mathbf{Z}|\mathbf{x}), \tag{16}$$

we obtain

$$\log p_\theta(\mathbf{y}|\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{Z})] - D_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{Z}|\mathbf{x})) \,. \tag{17}$$

This is the standard evidence lower bound (ELBO) used in our formulation (Kingma & Welling, 2013).

The ELBO contains two terms with distinct roles. The reconstruction term,

$$\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{Z})],$$

encourages the decoder to generate the target answer conditioned on latent reasoning states sampled from the posterior. The KL term,

$$D_{\mathrm{KL}}(q_\phi(\mathbf{Z}|\mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{Z}|\mathbf{x})) \,,$$

encourages the target-agnostic prior to match the target-aware posterior. Together, these two terms define the standard variational objective for latent-variable conditional generation.

In our model, the conditional likelihood is parameterized autoregressively:

$$\log p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{Z}) = \sum_t \log p_\theta(y_t \mid \mathbf{x}, \mathbf{y}_{<t}, \mathbf{Z}). \tag{18}$$

Accordingly, the reconstruction term in the ELBO is instantiated as the standard next-token prediction objective conditioned on the sampled latent sequence $\mathbf{Z}$. In other words, the decoder learns to generate the answer token by token while treating the latent reasoning sequence as an additional continuous conditioning signal (Bowman et al., 2016).

This standard ELBO serves as the variational foundation of our method. However, as discussed in the main text, ELBO optimization alone is insufficient in our setting because the posterior has access to the target during training, whereas inference must rely solely on the prior (Razavi et al., 2019; ALIAS PARTH GOYAL et al., 2017). This motivates the additional reverse-KL regularization introduced in AMVL.

B.2 Evidence Upper Bound (EUBO) and Reverse KL Regularization

As discussed above, standard ELBO optimization alone is insufficient in our setting, because it only encourages the prior to match the target-aware posterior, while leaving the posterior itself weakly constrained with respect to inference-time usability (Bowman et al., 2016; Zhao et al., 2017b; He et al., 2019). This motivates introducing an additional reverse-side regularizer on the posterior branch. One theoretical motivation for such a reverse-KL term comes from the Evidence Upper Bound (EUBO) (Hinton et al., 1995; Grosse et al., 2015) view of variational inference.

While the standard ELBO relies on the forward KL divergence and provides a lower bound on the marginal likelihood, the reverse KL divergence with respect to the true target-conditioned posterior gives rise to an EUBO. Consider the reverse KL divergence from the true target-conditioned posterior $p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})$ to the variational posterior $q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})$. By non-negativity of KL divergence,

$$D_{\mathrm{KL}}(p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y}) \,\|\, q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})) \geq 0. \tag{19}$$

By the definition of conditional probability, we have

$$p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y}) = \frac{p_\theta(\mathbf{y},\mathbf{Z}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})},$$

we can expand:

$$\begin{aligned} D_{\mathrm{KL}}(p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y}) \,\|\, q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})) &= \mathbb{E}_{p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})}{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}\right] \\ &= \mathbb{E}_{p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y},\mathbf{Z}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})\, q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}\right] \\ &= \mathbb{E}_{p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y},\mathbf{Z}|\mathbf{x})}{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}\right] - \log p_\theta(\mathbf{y}|\mathbf{x}). \end{aligned} \tag{20}$$

Rearranging the terms, we obtain an upper bound on the log-marginal likelihood:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) \leq \mathbb{E}_{p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y},\mathbf{Z}|\mathbf{x})}{q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})}\right] \triangleq \mathcal{U}_{\mathrm{EUBO}}. \tag{21}$$

This establishes the exact EUBO. Minimizing $\mathcal{U}_{\mathrm{EUBO}}$ with respect to the variational parameters ϕ is mathematically equivalent to minimizing the exact reverse KL divergence. Unlike the forward KL, this reverse direction penalizes approximate distributions that assign insufficient density to regions supported by the reference distribution. This support-coverage preference (Minka et al., 2005; Bishop & Nasrabadi, 2006) is one reason why reverse-KL-style objectives are often invoked when motivating upper-bound-oriented regularization principles.

From Exact EUBO to Practical Surrogate. Directly minimizing the exact reverse KL requires taking expectations under the true posterior $p_\theta(\mathbf{Z}|\mathbf{x},\mathbf{y})$, which is analytically intractable and unavailable during end-to-end optimization.

In our setting, an ideal reference distribution for posterior regularization should reflect inference-time latent support while remaining tractable. Since the true target-conditioned posterior is not available in closed form, we instead use the learned prior $p_\theta(\mathbf{Z}|\mathbf{x})$ as a tractable reference distribution and define the practical reverse regularizer as:

$$\mathcal{L}_{\mathrm{rev}} = D_{\mathrm{KL}}(\mathrm{sg}[p_\theta(\mathbf{Z}|\mathbf{x})] \,\|\, q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y}))\,, \tag{22}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. This design yields a practical reverse-KL regularizer that faithfully inherits the support-compatibility preference of the exact EUBO perspective. Importantly, the stop-gradient ensures that the loss updates only the posterior branch.

Therefore, the $\mathcal{L}_{\mathrm{rev}}$ term used in the main text is best interpreted as an EUBO-motivated practical surrogate, rather than a direct optimization of the exact theoretical EUBO. Its primary purpose is not to optimize a formal upper bound directly, but to robustly regularize the posterior toward inference-compatible support defined by the learned prior.

B.3 Complementary Roles of Forward and Reverse KL

The two KL directions in AMVL play different and complementary roles:

$$\mathcal{L}_{\mathrm{fwd}} = D_{\mathrm{KL}}(\mathrm{sg}[q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y})] \,\|\, p_\theta(\mathbf{Z}|\mathbf{x}))\,, \tag{23}$$

$$\mathcal{L}_{\mathrm{rev}} = D_{\mathrm{KL}}(\mathrm{sg}[p_\theta(\mathbf{Z}|\mathbf{x})] \,\|\, q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y}))\,. \tag{24}$$

Because of the stop-gradient operators, the two objectives affect different parameter sets. The forward term $\mathcal{L}_{\mathrm{fwd}}$ updates the prior while keeping the posterior fixed, thereby calibrating the target-agnostic prior toward posterior-inferred latent reasoning states. In contrast, the reverse term $\mathcal{L}_{\mathrm{rev}}$

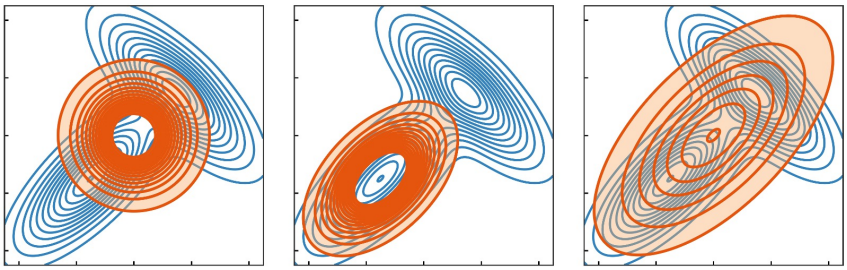


Figure 2: Geometric illustration of the two asymmetric KL objectives underlying AMVL on a bimodal latent distribution. Blue contours denote the stop-gradient reference distribution, and orange contours denote the optimized distribution. (a) Initial prior–posterior mismatch. (b) Forward KL alignment updates the prior to cover posterior-supported regions. (c) Reverse KL regularization constrains the posterior toward prior-compatible regions.

updates the posterior while keeping the prior fixed, thereby discouraging posterior solutions that are excessively sharp, overly target-specific, or poorly supported by the learned prior. It is crucial to understand the geometric nature of these KL terms. In both $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{rev}}$, the optimization is performed with respect to the *second argument* of the KL divergence (p in $\mathcal{L}_{\text{fwd}}$ and q in $\mathcal{L}_{\text{rev}}$). Minimizing $D_{\text{KL}}(P \,\|\, Q)$ with respect to Q is known to have a mass-covering effect, forcing the optimized distribution Q to spread out and cover the support of the fixed distribution P. Therefore, both KL terms in AMVL are mass-covering, not mode-seeking. The complementarity arises from the *asymmetric direction* of this mass-covering behavior:

- **Forward KL ($\mathcal{L}_{\text{fwd}}$) forces the prior p to cover the posterior q.** This encourages the prior to be expressive enough to represent all reasoning states the posterior finds useful during training.
- **Reverse KL ($\mathcal{L}_{\text{rev}}$) forces the posterior q to cover the prior p.** This regularizes the posterior, preventing it from collapsing into a narrow, answer-dependent mode that is unsupported by the prior (i.e., a region where $p(\mathbf{z})$ is low).

In essence, AMVL establishes a mutual mass-covering dynamic in which the prior and posterior are encouraged to cover each other. This differs from the common mode-seeking versus mass-covering trade-off and is the key to calibrating the latent space. The illustration in Figure 2 should therefore be interpreted as showing two asymmetric distribution-matching processes: the forward term updates the prior toward the posterior, whereas the reverse term updates the posterior toward the prior.

This separation is important in our setting. If only the forward KL is used, the prior is trained to chase the posterior, but the posterior itself is weakly constrained with respect to inference-time usability. If only the reverse term is used, posterior spread may improve, but the prior is not explicitly trained to match posterior-inferred reasoning states. Their combination therefore reduces train-inference mismatch from both directions: the prior is encouraged to approach the posterior, and the posterior is simultaneously discouraged from drifting too far away from what the prior can support at inference time.

B.4 Diagonal Gaussian Case and Closed-Form Effects

In our implementation, both the prior and posterior are factorized diagonal Gaussians (Kingma & Welling, 2013; Rezende et al., 2014) over the k latent slots:

$$p_\theta(\mathbf{Z}|\mathbf{x}) = \prod_{i=1}^{k} p_\theta(\mathbf{z}_i|\mathbf{x}), \qquad q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y}) = \prod_{i=1}^{k} q_\phi(\mathbf{z}_i|\mathbf{x},\mathbf{y}). \tag{25}$$

Therefore, the KL divergence between the full latent sequences decomposes across slots:

$$D_{\text{KL}}(q_\phi(\mathbf{Z}|\mathbf{x},\mathbf{y}) \,\|\, p_\theta(\mathbf{Z}|\mathbf{x})) = \sum_{i=1}^{k} D_{\text{KL}}(q_\phi(\mathbf{z}_i|\mathbf{x},\mathbf{y}) \,\|\, p_\theta(\mathbf{z}_i|\mathbf{x})), \tag{26}$$

and similarly for the reverse direction.

For two diagonal Gaussian distributions

$$\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \mathrm{diag}(\boldsymbol{\sigma}_1^2)), \qquad \mathcal{N}_2 = \mathcal{N}(\boldsymbol{\mu}_2, \mathrm{diag}(\boldsymbol{\sigma}_2^2)),$$

the KL divergence has the closed form

$$D_{\mathrm{KL}}(\mathcal{N}_1 \,\|\, \mathcal{N}_2) = \frac{1}{2} \sum_{j=1}^{d} \left(\frac{\sigma_{1,j}^2 + (\mu_{1,j} - \mu_{2,j})^2}{\sigma_{2,j}^2} - 1 + \log \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} \right). \tag{27}$$

This expression makes the effects of the two KL directions explicit. For the forward KL, large mean mismatch $(\mu_{1,j} - \mu_{2,j})^2$ and insufficient prior variance are penalized, which pushes the prior toward posterior-inferred latent states. For the reverse KL, the posterior is penalized when it assigns too little variance or too little mass to regions supported by the prior, thereby discouraging posterior over-concentration and improving support compatibility.

Since our implementation predicts log-variance, it is useful to rewrite the KL in terms of

$$\mathbf{L} = \log \boldsymbol{\sigma}^2.$$

Substituting $L_{1,j} = \log \sigma_{1,j}^2$ and $L_{2,j} = \log \sigma_{2,j}^2$ into the component-wise formula, we obtain:

$$D_{\mathrm{KL}}(\mathcal{N}_1 \,\|\, \mathcal{N}_2) = \frac{1}{2} \sum_{j=1}^{d} \left(\exp(L_{1,j} - L_{2,j}) + (\mu_{1,j} - \mu_{2,j})^2 \exp(-L_{2,j}) - 1 - (L_{1,j} - L_{2,j}) \right). \tag{28}$$

For the reverse direction (where $\mathcal{N}_2$ represents the posterior), when the posterior log-variance $L_{2,j}$ becomes excessively small, both the variance-ratio term and the mean-mismatch term can grow rapidly. Thus, minimizing $D_{\mathrm{KL}}(p_\theta \| q_\phi)$ explicitly penalizes posterior over-concentration relative to the learned prior.

In practice, we compute both KL terms in closed form for each latent slot and then average over latent slots and batch elements.

C Theoretical Analysis of Answer Leakage and AMVL

We provide a theoretical analysis of why bidirectional KL regularization in AMVL better mitigates answer leakage than standard one-sided ELBO training.

Setup. Our formulation builds on the standard conditional variational framework (Kingma & Welling, 2013; Sohn et al., 2015). Let $p_\theta(\mathbf{z} \mid \mathbf{x})$ denote the inference-time prior, and $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ the training-time posterior. We consider the common diagonal Gaussian parameterization

$$q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = \mathcal{N}\big(\mu_\phi(\mathbf{x}, \mathbf{y}), \mathrm{diag}(\sigma_\phi^2(\mathbf{x}, \mathbf{y}))\big),$$
$$p_\theta(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \mathrm{diag}(\sigma_\theta^2(\mathbf{x}))\big).$$

To formalize answer leakage, we decompose the posterior mean into an input-grounded component and an answer-dependent shift:

$$\mu_\phi(\mathbf{x}, \mathbf{y}) = \mu_\phi^{(x)}(\mathbf{x}) + \delta(\mathbf{x}, \mathbf{y}), \tag{29}$$

where $\delta(\mathbf{x}, \mathbf{y})$ denotes the additional posterior displacement induced by the target answer $\mathbf{y}$. When $\delta \equiv 0$, the posterior mean is fully grounded in the input $\mathbf{x}$, with no answer-dependent mean shift.

We further define the conditional mean leakage

$$\bar{\delta}(\mathbf{x}) := \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y}|\mathbf{x})}[\delta(\mathbf{x}, \mathbf{y})]. \tag{30}$$

In this appendix, we focus on *mean-level leakage*, i.e., answer-dependent posterior drift in the latent mean. This restriction is deliberate: mean drift alone already suffices to induce prior contamination under one-sided KL matching, and it admits a clean closed-form analysis of both the contaminated prior optimum and the reverse-KL corrective gradient. We leave leakage through answer-dependent posterior variance or higher-order statistics to future work.

Ideal inference-time latent distribution. Let $p^*(\mathbf{z} \mid \mathbf{x})$ denote an ideal inference-time latent distribution that depends only on $\mathbf{x}$ and supports optimal downstream prediction. In the propositions below, we will explicitly state when we assume that the mean of p^* matches the input-grounded posterior component:

$$\mu^*(\mathbf{x}) = \mu_\phi^{(x)}(\mathbf{x}). \tag{31}$$

This assumption formalizes the desideratum that the ideal test-time latent state should be input-grounded rather than answer-conditioned.

C.1 Prior Contamination Under ELBO

The KL term in the standard conditional ELBO is

$$\mathcal{L}_{\mathrm{ELBO}}^{\mathrm{KL}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\mathrm{data}}}\left[D_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\big)\right]. \tag{32}$$

Lemma 1 (Prior contamination under ELBO)**.** *Assume diagonal Gaussian latents with fixed posterior q_ϕ. Then minimizing the ELBO KL term with respect to the prior mean yields*

$$\mu_\theta^*(\mathbf{x}) = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}|\mathbf{x})}\left[\mu_\phi(\mathbf{x}, \mathbf{y})\right] = \mu_\phi^{(x)}(\mathbf{x}) + \bar{\delta}(\mathbf{x}). \tag{33}$$

Consequently, if the ideal inference-time latent mean is input-grounded as in equation 31, then one-sided ELBO matching causes the learned prior mean to inherit the average answer-dependent shift as residual bias. In particular, if $p^*(\mathbf{z} \mid \mathbf{x})$ shares the same diagonal covariance as the learned prior, then

$$D_{KL}\big(p^*(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta^*(\mathbf{z} \mid \mathbf{x})\big) = \frac{1}{2}\sum_{j=1}^{d} \frac{\bar{\delta}_j(\mathbf{x})^2}{\sigma_{\theta,j}^2(\mathbf{x})}. \tag{34}$$

More generally, equation 34 isolates the contribution due purely to mean contamination.

Proof. For fixed posterior q_ϕ, minimizing

$$\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[D_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\big)\right]$$

with respect to the Gaussian prior mean $\mu_\theta(\mathbf{x})$ is equivalent to minimizing, for each $\mathbf{x}$, the expected quadratic mean-mismatch term

$$\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[\sum_{j=1}^{d} \frac{\big(\mu_{\phi,j}(\mathbf{x}, \mathbf{y}) - \mu_{\theta,j}(\mathbf{x})\big)^2}{2\sigma_{\theta,j}^2(\mathbf{x})}\right].$$

Taking derivatives with respect to $\mu_{\theta,j}(\mathbf{x})$ and setting them to zero yields

$$\mu_{\theta,j}^*(\mathbf{x}) = \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big[\mu_{\phi,j}(\mathbf{x}, \mathbf{y})\big].$$

Using the decomposition in equation 29,

$$\mu_{\phi,j}(\mathbf{x}, \mathbf{y}) = \mu_{\phi,j}^{(x)}(\mathbf{x}) + \delta_j(\mathbf{x}, \mathbf{y}),$$

we obtain

$$\mu_{\theta,j}^*(\mathbf{x}) = \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big[\mu_{\phi,j}^{(x)}(\mathbf{x}) + \delta_j(\mathbf{x}, \mathbf{y})\big] = \mu_{\phi,j}^{(x)}(\mathbf{x}) + \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big[\delta_j(\mathbf{x}, \mathbf{y})\big] = \mu_{\phi,j}^{(x)}(\mathbf{x}) + \bar{\delta}_j(\mathbf{x}).$$

Stacking coordinates gives

$$\mu_\theta^*(\mathbf{x}) = \mu_\phi^{(x)}(\mathbf{x}) + \bar{\delta}(\mathbf{x}),$$

which proves equation 33.

Now assume that the ideal inference-time latent distribution $p^*(\mathbf{z} \mid \mathbf{x})$ has mean

$$\mu^*(\mathbf{x}) = \mu_\phi^{(x)}(\mathbf{x})$$

as in equation 31, and that $p^*(\mathbf{z} \mid \mathbf{x})$ shares the same diagonal covariance as the learned prior $p_\theta^*(\mathbf{z} \mid \mathbf{x})$, namely $\mathrm{diag}(\sigma_\theta^2(\mathbf{x}))$. Then the KL divergence between these two diagonal Gaussians is

$$D_{KL}\big(p^*(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta^*(\mathbf{z} \mid \mathbf{x})\big) = \frac{1}{2}\sum_{j=1}^{d}\left(\frac{\sigma_{\theta,j}^2(\mathbf{x})}{\sigma_{\theta,j}^2(\mathbf{x})} + \frac{\big(\mu_j^*(\mathbf{x}) - \mu_{\theta,j}^*(\mathbf{x})\big)^2}{\sigma_{\theta,j}^2(\mathbf{x})} - 1 + \log\frac{\sigma_{\theta,j}^2(\mathbf{x})}{\sigma_{\theta,j}^2(\mathbf{x})}\right).$$

Because the two distributions share the same covariance, the variance-ratio and log-determinant terms cancel, leaving

$$D_{KL}\big(p^*(\mathbf{z} \mid \mathbf{x}) \,\|\, p^*_\theta(\mathbf{z} \mid \mathbf{x})\big) = \frac{1}{2}\sum_{j=1}^{d} \frac{\big(\mu^*_j(\mathbf{x}) - \mu^*_{\theta,j}(\mathbf{x})\big)^2}{\sigma^2_{\theta,j}(\mathbf{x})}.$$

Substituting $\mu^*(\mathbf{x}) = \mu^{(x)}_\phi(\mathbf{x})$ and

$$\mu^*_\theta(\mathbf{x}) = \mu^{(x)}_\phi(\mathbf{x}) + \bar{\delta}(\mathbf{x})$$

gives

$$\mu^*_j(\mathbf{x}) - \mu^*_{\theta,j}(\mathbf{x}) = -\bar{\delta}_j(\mathbf{x}),$$

and therefore

$$D_{KL}\big(p^*(\mathbf{z} \mid \mathbf{x}) \,\|\, p^*_\theta(\mathbf{z} \mid \mathbf{x})\big) = \frac{1}{2}\sum_{j=1}^{d} \frac{\bar{\delta}_j(\mathbf{x})^2}{\sigma^2_{\theta,j}(\mathbf{x})},$$

which is exactly equation 34. Under the shared-covariance assumption, this expression isolates the contribution due purely to mean contamination.

Lemma 1 formalizes the central problem of one-sided ELBO training in our setting: answer-dependent posterior bias is not merely tolerated during training, but is absorbed by the inference-time prior (Hoffman & Johnson, 2016; Zhao et al., 2017a).

C.2 Why Forward Alignment Alone Is Insufficient

AMVL uses the forward alignment term

$$\mathcal{L}_{\text{fwd}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}}\left[D_{KL}\big(\text{sg}[q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})] \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\big)\right]. \tag{35}$$

Lemma 2 (Forward alignment alone is insufficient)**.** *For fixed posterior q_ϕ, minimizing $\mathcal{L}_{\text{fwd}}$ over θ yields the same optimal prior mean as minimizing the KL term in the standard ELBO:*

$$\mu^{*,\text{fwd}}_\theta(\mathbf{x}) = \mathbb{E}_{\mathbf{y}|\mathbf{x}}[\mu_\phi(\mathbf{x}, \mathbf{y})]. \tag{36}$$

Moreover,

$$\nabla_\phi \mathcal{L}_{\text{fwd}} = 0.$$

Therefore, in isolation, forward alignment calibrates the prior to the current posterior but exerts no direct corrective gradient on answer-dependent bias already present in the posterior.

Proof. The stop-gradient operator treats q_ϕ as constant with respect to ϕ, hence

$$\nabla_\phi \mathcal{L}_{\text{fwd}} = 0.$$

With respect to θ, the objective is identical to minimizing the same forward KL over the prior, so the stationary condition is exactly the same as in Lemma 1, yielding equation 36.

Lemma 2 clarifies the role of forward KL: it is necessary for prior calibration, but by itself it cannot eliminate answer leakage at the source.

C.3 Reverse KL Suppresses Incompatible Posterior Drift

AMVL further introduces the reverse regularizer

$$\mathcal{L}_{\text{rev}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}}\left[D_{KL}\big(\text{sg}[p_\theta(\mathbf{z} \mid \mathbf{x})] \,\|\, q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})\big)\right]. \tag{37}$$

Proposition 1 (Reverse KL suppresses incompatible posterior drift)**.** *For diagonal Gaussian latents $p_\theta(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\big(\mu_\theta(\mathbf{x}), \text{diag}(\sigma^2_\theta(\mathbf{x}))\big)$ and $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) = \mathcal{N}\big(\mu_\phi(\mathbf{x}, \mathbf{y}), \text{diag}(\sigma^2_\phi(\mathbf{x}, \mathbf{y}))\big)$, the reverse KL in Eq. (4) has the closed form*

$$\mathcal{L}_{rev} = \frac{1}{2}\sum_j \left(\frac{\sigma^2_{\theta,j}}{\sigma^2_{\phi,j}} + \frac{(\mu_{\theta,j} - \mu_{\phi,j})^2}{\sigma^2_{\phi,j}} - 1 + \log\frac{\sigma^2_{\phi,j}}{\sigma^2_{\theta,j}}\right),$$

where p_θ is treated as constant by stop-gradient. Its gradient with respect to the posterior mean is

$$\frac{\partial \mathcal{L}_{rev}}{\partial \mu_{\phi,j}} = \frac{\mu_{\phi,j} - \mu_{\theta,j}}{\sigma^2_{\phi,j}}.$$

Thus, reverse KL penalizes posterior drift from prior-compatible high-density latent regions, with a mean-mismatch penalty stronger as the posterior sharpens (i.e., as $\sigma^2_{\phi,j}$ decreases).

Proof. The closed-form expression in Proposition 1 is the standard KL divergence between two diagonal Gaussian distributions, with the prior distribution treated as constant by stop-gradient. Differentiating its quadratic mean-mismatch term with respect to $\mu_{\phi,j}$ gives

$$\frac{\partial \mathcal{L}_{\text{rev}}}{\partial \mu_{\phi,j}} = \frac{\mu_{\phi,j} - \mu_{\theta,j}}{\sigma^2_{\phi,j}}.$$

To characterize the effect of this gradient, we set it to zero:

$$\frac{\mu_{\phi,j} - \mu_{\theta,j}}{\sigma^2_{\phi,j}} = 0 \quad \Longrightarrow \quad \mu_{\phi,j} = \mu_{\theta,j}.$$

Since $\sigma^2_{\phi,j} > 0$, the objective, viewed as a function of $\mu_{\phi,j}$ with $\sigma^2_{\phi,j}$ fixed, is strictly convex:

$$\frac{\partial^2 \mathcal{L}_{\text{rev}}}{\partial \mu^2_{\phi,j}} = \frac{1}{\sigma^2_{\phi,j}} > 0.$$

Hence, this stationary point is the unique global minimizer with respect to $\mu_{\phi,j}$.

Moreover, the gradient has the same sign as the deviation $\mu_{\phi,j} - \mu_{\theta,j}$:

$$\frac{\partial \mathcal{L}_{\text{rev}}}{\partial \mu_{\phi,j}} \begin{cases} > 0, & \text{if } \mu_{\phi,j} > \mu_{\theta,j}, \\ < 0, & \text{if } \mu_{\phi,j} < \mu_{\theta,j}. \end{cases}$$

Consequently, the negative gradient direction points from $\mu_{\phi,j}$ toward $\mu_{\theta,j}$, thereby exerting a restoring effect on posterior-mean drift. The gradient magnitude is

$$\left| \frac{\partial \mathcal{L}_{\text{rev}}}{\partial \mu_{\phi,j}} \right| = \frac{|\mu_{\phi,j} - \mu_{\theta,j}|}{\sigma^2_{\phi,j}},$$

which increases as $\sigma^2_{\phi,j}$ decreases. Thus, for a fixed mean mismatch, reverse KL imposes a stronger penalty when the posterior becomes more sharply concentrated.

Substituting the decomposition in equation 29 into the posterior-mean gradient stated in Proposition 1 gives

$$\frac{\partial \mathcal{L}_{\text{rev}}}{\partial \delta_j} = \frac{\mu^{(x)}_{\phi,j}(\mathbf{x}) + \delta_j(\mathbf{x}, \mathbf{y}) - \mu_{\theta,j}(\mathbf{x})}{\sigma^2_{\phi,j}(\mathbf{x}, \mathbf{y})}. \tag{38}$$

Define the prior-centering bias as

$$b_j(\mathbf{x}) := \mu_{\theta,j}(\mathbf{x}) - \mu^{(x)}_{\phi,j}(\mathbf{x}).$$

Equation equation 38 can then be rewritten as

$$\frac{\partial \mathcal{L}_{\text{rev}}}{\partial \delta_j} = \frac{\delta_j(\mathbf{x}, \mathbf{y}) - b_j(\mathbf{x})}{\sigma^2_{\phi,j}(\mathbf{x}, \mathbf{y})}.$$

Hence, reverse KL penalizes answer-dependent posterior drift relative to the current prior center. In the low-contamination regime where $b_j(\mathbf{x}) \approx 0$—for example, after bidirectional calibration has partially aligned the prior with the input-grounded component—the gradient reduces to

$$\frac{\partial \mathcal{L}_{\text{rev}}}{\partial \delta_j} \approx \frac{\delta_j(\mathbf{x}, \mathbf{y})}{\sigma^2_{\phi,j}(\mathbf{x}, \mathbf{y})}. \tag{39}$$

Therefore, the negative gradient directly suppresses answer-dependent posterior shift, with a stronger restoring effect when the posterior is more sharply concentrated.

C.4 BIDIRECTIONAL CALIBRATION REDUCES PRIOR CONTAMINATION

We now formalize the local shrinkage effect of bidirectional calibration under a simplified linear-response model.

Recall that γ is the scalar weight controlling the strength of the reverse KL regularization term $\mathcal{L}_{\mathrm{rev}}$ in the AMVL training objective (Eq. 10). A larger γ imposes a stronger restoring force against answer-dependent posterior drift, as formalized below.

Assumption 1 (Local linear leakage model)**.** *In a local neighborhood of training, the posterior mean admits the form*

$$\mu_\phi(\mathbf{x}, \mathbf{y}) = \mu^{(x)}(\mathbf{x}) + \alpha f(\mathbf{x}, \mathbf{y}), \tag{40}$$

where $\alpha \geq 0$ denotes a scalar leakage coefficient and $f(\mathbf{x}, \mathbf{y})$ is an answer-dependent direction. Let

$$\bar{f}(\mathbf{x}) := \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}|\mathbf{x})}\big[f(\mathbf{x}, \mathbf{y})\big].$$

Assumption 1 is a local linearization of the general leakage decomposition in Eq. 29, which writes

$$\mu_\phi(\mathbf{x}, \mathbf{y}) = \mu_\phi^{(x)}(\mathbf{x}) + \delta(\mathbf{x}, \mathbf{y})$$

for a free vector-valued answer-dependent shift $\delta(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$. Here we refine that decomposition by factoring

$$\delta(\mathbf{x}, \mathbf{y}) = \alpha f(\mathbf{x}, \mathbf{y}),$$

where α captures a shared local leakage amplitude and $f(\mathbf{x}, \mathbf{y})$ captures the direction and sample-dependent shape of leakage. Freezing the local shape f and analyzing the scalar coordinate α is what makes the equilibrium analysis tractable. Under this parameterization, the conditional average leakage direction $\bar{f}(\mathbf{x})$ plays the same role as $\bar{\delta}(\mathbf{x})$ in Eq. 30, with the correspondence

$$\bar{\delta}(\mathbf{x}) \leftrightarrow \alpha\, \bar{f}(\mathbf{x}).$$

Assumption 2 (Local linear-response stationary condition)**.** *Near a local stationary point, the training dynamics along the leakage coordinate α admit a first-order linear approximation. In particular, motivated by Proposition 1, the reverse KL contributes a restoring force along the leakage coordinate that is locally proportional to*

$$\frac{\gamma\alpha}{\sigma^2_{\mathrm{eff}}},$$

where σ^2_{eff} denotes an effective posterior variance along the leakage direction. We further assume that, in the local regime of interest, the prior mean is approximately centered on the input-grounded component, i.e.,

$$\mu_\theta(\mathbf{x}) \approx \mu^{(x)}(\mathbf{x}),$$

so that the reverse-KL contribution along the leakage coordinate is first-order proportional to α. Let α_{ELBO} denote the local equilibrium leakage coefficient under one-sided ELBO training.

Here, "equilibrium" refers to the stationary point of the leakage-coordinate dynamics during training, i.e., the value of α at which the net gradient force acting on the leakage coefficient is zero. Under one-sided ELBO training, the reconstruction term may favor answer-dependent shortcuts in the posterior, while the forward KL partially resists them; α_{ELBO} is the local balance point of these effects. Under AMVL, the reverse KL introduces an additional restoring force that pushes α toward zero, thereby shifting the equilibrium to a smaller value.

Proposition 2 (Bidirectional calibration reduces prior contamination)**.** *Under the local linear leakage model in Assumption 1, where the answer-dependent shift is written as*

$$\delta(\mathbf{x}, \mathbf{y}) = \alpha f(\mathbf{x}, \mathbf{y}),$$

and under the local linear-response stationary condition in Assumption 2, there exists a constant $c > 0$ such that the local equilibrium leakage coefficient α under AMVL satisfies

$$\alpha_{AMVL} = \frac{\alpha_{ELBO}}{1 + c\gamma/\sigma^2_{eff}},$$

where α_{ELBO} is the equilibrium leakage under one-sided ELBO training, γ is the weight of the reverse KL term, and σ^2_{eff} is an effective posterior variance. Consequently, the prior mean contamination under AMVL satisfies

$$\Delta_{AMVL}(\mathbf{x}) = \|\mu_\theta^{AMVL}(\mathbf{x}) - \mu^{(x)}(\mathbf{x})\|_2 = \alpha_{AMVL}\|\bar{f}(\mathbf{x})\|_2,$$

whereas the corresponding contamination under one-sided ELBO matching is

$$\Delta_{ELBO}(\mathbf{x}) = \alpha_{ELBO}\|\bar{f}(\mathbf{x})\|_2.$$

Therefore,

$$\Delta_{AMVL}(\mathbf{x}) < \Delta_{ELBO}(\mathbf{x})$$

whenever $\gamma > 0$, $\|\bar{f}(\mathbf{x})\|_2 > 0$, and $\alpha_{ELBO} > 0$.

Proof. Under Assumption 2, the net ELBO-only gradient force along the leakage coordinate is locally linear near its stationary point. Denoting the local curvature constant by $c_1 > 0$, we write

$$G_{\mathrm{ELBO}}(\alpha) \approx c_1\big(\alpha_{\mathrm{ELBO}} - \alpha\big),$$

whose unique zero is α_{ELBO}, the local equilibrium under one-sided ELBO training.

Next, using Assumption 1, substitute

$$\mu_\phi(\mathbf{x}, \mathbf{y}) = \mu^{(x)}(\mathbf{x}) + \alpha f(\mathbf{x}, \mathbf{y})$$

into the reverse-KL mean gradient established in Proposition 1. For each latent dimension j,

$$\frac{\partial \mathcal{L}_{\mathrm{rev}}}{\partial \mu_{\phi,j}} = \frac{\mu_{\phi,j} - \mu_{\theta,j}}{\sigma^2_{\phi,j}}.$$

Under the local centering condition $\mu_\theta(\mathbf{x}) \approx \mu^{(x)}(\mathbf{x})$, the mean mismatch is, to first order, proportional to $\alpha f_j(\mathbf{x}, \mathbf{y})$. Projecting the negative reverse-KL gradient onto the leakage coordinate, averaging over latent dimensions and data samples, and absorbing the resulting local geometric and averaging factors into a constant $c_2 > 0$ gives

$$G_{\mathrm{rev}}(\alpha) = -\frac{c_2\gamma}{\sigma^2_{\mathrm{eff}}}\alpha.$$

Under AMVL, the ELBO and reverse-KL forces act simultaneously. Therefore, the local equilibrium α_{AMVL} satisfies

$$G_{\mathrm{ELBO}}(\alpha_{\mathrm{AMVL}}) + G_{\mathrm{rev}}(\alpha_{\mathrm{AMVL}}) = 0.$$

Substituting the two local force expressions yields

$$c_1\big(\alpha_{\mathrm{ELBO}} - \alpha_{\mathrm{AMVL}}\big) - \frac{c_2\gamma}{\sigma^2_{\mathrm{eff}}}\alpha_{\mathrm{AMVL}} = 0.$$

Rearranging gives

$$c_1\alpha_{\mathrm{ELBO}} = \alpha_{\mathrm{AMVL}}\left(c_1 + \frac{c_2\gamma}{\sigma^2_{\mathrm{eff}}}\right).$$

Defining $c := c_2/c_1 > 0$, we obtain

$$\alpha_{\mathrm{AMVL}} = \frac{\alpha_{\mathrm{ELBO}}}{1 + c\gamma/\sigma^2_{\mathrm{eff}}},$$

which establishes the equilibrium shrinkage relation claimed in Proposition 2.

For $\alpha_{\mathrm{ELBO}} > 0$, this equilibrium coefficient is strictly decreasing in γ, since

$$\frac{\partial \alpha_{\mathrm{AMVL}}}{\partial \gamma} = -\frac{c\,\alpha_{\mathrm{ELBO}}/\sigma^2_{\mathrm{eff}}}{\big(1 + c\gamma/\sigma^2_{\mathrm{eff}}\big)^2} < 0.$$

Under exact forward matching at the function optimum, Proposition 2 implies that the prior mean matches the conditional average of the regularized posterior mean:

$$\mu_\theta^{\mathrm{AMVL}}(\mathbf{x}) = \mathbb{E}_{\mathbf{y}|\mathbf{x}}\big[\mu_\phi(\mathbf{x}, \mathbf{y})\big].$$

Using Assumption 1, we obtain

$$\begin{aligned}\mu_\theta^{\mathrm{AMVL}}(\mathbf{x}) &= \mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[\mu^{(x)}(\mathbf{x}) + \alpha_{\mathrm{AMVL}} f(\mathbf{x},\mathbf{y})\right] \\ &= \mu^{(x)}(\mathbf{x}) + \alpha_{\mathrm{AMVL}}\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[f(\mathbf{x},\mathbf{y})\right] \\ &= \mu^{(x)}(\mathbf{x}) + \alpha_{\mathrm{AMVL}}\bar{f}(\mathbf{x}).\end{aligned}$$

Because $\alpha_{\mathrm{ELBO}} > 0$ implies $\alpha_{\mathrm{AMVL}} > 0$, it follows that

$$\Delta_{\mathrm{AMVL}}(\mathbf{x}) = \left\|\mu_\theta^{\mathrm{AMVL}}(\mathbf{x}) - \mu^{(x)}(\mathbf{x})\right\|_2 = \alpha_{\mathrm{AMVL}}\|\bar{f}(\mathbf{x})\|_2.$$

The corresponding one-sided ELBO relation follows identically:

$$\Delta_{\mathrm{ELBO}}(\mathbf{x}) = \alpha_{\mathrm{ELBO}}\|\bar{f}(\mathbf{x})\|_2.$$

Finally, if $\gamma > 0$, then

$$\frac{1}{1 + c\gamma/\sigma_{\mathrm{eff}}^2} < 1.$$

Therefore, whenever $\alpha_{\mathrm{ELBO}} > 0$ and $\|\bar{f}(\mathbf{x})\|_2 > 0$,

$$\begin{aligned}\Delta_{\mathrm{AMVL}}(\mathbf{x}) &= \frac{\alpha_{\mathrm{ELBO}}\|\bar{f}(\mathbf{x})\|_2}{1 + c\gamma/\sigma_{\mathrm{eff}}^2} \\ &< \alpha_{\mathrm{ELBO}}\|\bar{f}(\mathbf{x})\|_2 \\ &= \Delta_{\mathrm{ELBO}}(\mathbf{x}),\end{aligned}$$

which proves the strict contamination reduction claimed in Proposition 2.

Proposition 2 formalizes the key local mechanism of AMVL: reverse KL reduces the equilibrium answer-dependent posterior shift, and the forward alignment term then transfers this leakage-reduced posterior signal to the prior. Note that if $\bar{f}(\mathbf{x}) = 0$, the prior mean may exhibit no average contamination even though answer-dependent posterior variability remains; the present result concerns mean-level prior contamination induced by nonzero average leakage directions.

Corollary 1 (Monotonic improvement with stronger reverse regularization)**.** *Under the assumptions of Proposition 2,*

$$\Delta_{\mathrm{AMVL}}(\mathbf{x}) = \frac{\alpha_{\mathrm{ELBO}}\|\bar{f}(\mathbf{x})\|_2}{1 + c\gamma/\sigma_{\mathrm{eff}}^2} \tag{41}$$

is strictly decreasing in γ whenever $\|\bar{f}(\mathbf{x})\|_2 > 0$.

Proof. Differentiate with respect to γ:

$$\frac{\partial \Delta_{\mathrm{AMVL}}}{\partial \gamma} = -\frac{c\,\alpha_{\mathrm{ELBO}}\|\bar{f}(\mathbf{x})\|_2/\sigma_{\mathrm{eff}}^2}{(1 + c\gamma/\sigma_{\mathrm{eff}}^2)^2} < 0 \qquad \text{if } \|\bar{f}(\mathbf{x})\|_2 > 0.$$

Corollary 2 (Stronger benefit under posterior overconfidence)**.** *Under the same assumptions, the relative improvement factor*

$$\frac{\Delta_{\mathrm{ELBO}}(\mathbf{x})}{\Delta_{\mathrm{AMVL}}(\mathbf{x})} = 1 + \frac{c\gamma}{\sigma_{\mathrm{eff}}^2} \tag{42}$$

increases as σ_{eff}^2 decreases.

Proof. Immediate from Proposition 2.

Corollary 2 is practically important: within this local Gaussian analysis, reverse KL is most beneficial exactly when answer leakage is most dangerous, namely when the posterior becomes overly sharp and overconfident.

C.5 Information-Theoretic Interpretation

Answer leakage can also be understood through the conditional mutual information $I_q(\mathbf{Z}; \mathbf{Y} \mid \mathbf{X})$, extending standard ELBO surgery techniques (Hoffman & Johnson, 2016; Alemi et al., 2018) to the conditional setting:

$$I_q(\mathbf{Z}; \mathbf{Y} \mid \mathbf{X}) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \left[D_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \,\|\, q_\phi(\mathbf{z} \mid \mathbf{x})\big) \right], \tag{43}$$

where

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \int q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})\, p(\mathbf{y} \mid \mathbf{x})\, d\mathbf{y}.$$

Proposition 3 (Forward KL decomposition)**.** *For any posterior $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ and prior $p_\theta(\mathbf{z} \mid \mathbf{x})$,*

$$\mathbb{E}_{p(\mathbf{x},\mathbf{y})} D_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\big) = I_q(\mathbf{Z}; \mathbf{Y} \mid \mathbf{X}) + \mathbb{E}_{p(\mathbf{x})} D_{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{x})\big). \tag{44}$$

Proof. Expand the KL term:

$$D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \| p_\theta(\mathbf{z} \mid \mathbf{x})) = D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \| q_\phi(\mathbf{z} \mid \mathbf{x})) + \mathbb{E}_{q_\phi}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p_\theta(\mathbf{z} \mid \mathbf{x})}\right].$$

Taking expectation over $p(\mathbf{x}, \mathbf{y})$ yields equation 44.

A large value of $I_q(\mathbf{Z}; \mathbf{Y} \mid \mathbf{X})$ indicates that the latent variable retains substantial target-specific information even after conditioning on the input, which is precisely the signature of answer leakage. Proposition 3 shows that the one-sided KL in the ELBO mixes two effects: suppressing target dependence and fitting the prior to the aggregated posterior. In contrast, AMVL additionally constrains the reverse discrepancy $D_{KL}(p_\theta \| q_\phi)$, forcing each posterior $q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$ to remain compatible with prior-reachable high-density latent regions. This suppresses answer-specific posterior collapse onto narrow latent modes that cannot be reliably reached from $\mathbf{x}$ alone, thereby helping reduce train-inference mismatch.

Summary. The standard ELBO performs one-sided prior matching: it encourages the prior to chase a target-aware posterior, even when that posterior exploits answer leakage. AMVL decouples the two roles. The forward KL teaches the prior to approximate useful training-time latent states, while the reverse KL regularizes the posterior to remain compatible with the inference-time prior. Under the Gaussian mean-leakage analysis above, and under the local linear-response model in Assumptions 1–2, bidirectional calibration reduces the contamination transferred into the prior and therefore provides a principled mechanism for mitigating answer leakage more effectively than one-sided ELBO training.

D Baselines

We compare our method against representative state-of-the-art MLLM baselines from three families: *thinking about images*, *thinking with images*, and *latent reasoning*.

Thinking about Images. Methods in this category enhance multimodal reasoning by generating explicit chain-of-thought trajectories over visual inputs, typically through reinforcement learning or supervised reasoning traces.

- **Vision-R1 (Huang et al., 2026):** Adapts reinforcement learning to multimodal reasoning by encouraging a "think before answer" behavior. The model is optimized to produce explicit reasoning trajectories prior to final response generation, thereby improving step-by-step visual reasoning ability.
- **PAPO (Wang et al., 2026c):** Extends reinforcement-learning-based multimodal reasoning with an additional perception-oriented objective. Besides learning to generate reasoning traces, PAPO incorporates an implicit perception loss to encourage more faithful image-grounded descriptions and stronger alignment between visual understanding and reasoning.

Thinking with Images. This category augments reasoning by actively manipulating, refining, or querying visual evidence during inference, often through external tools or environment interaction.

- **PixelReasoner (Su et al., 2025a):** Introduces a tool-augmented reasoning framework in which the model iteratively edits or enhances the input image during the reasoning process. By interacting with modified visual observations rather than relying solely on the original input, PixelReasoner improves its ability to resolve fine-grained visual ambiguities and difficult perceptual details.
- **DeepEyes (Zheng et al., 2026):** Integrates external tool usage directly into a unified reinforcement learning loop, allowing the model to reason through actions such as visual grounding, web search, and code execution. Instead of treating tools as a separate post-processing module, DeepEyes makes tool invocation part of the reasoning policy itself, aiming to more closely mimic human-like visual perception and problem solving.

Latent Reasoning. This line of work explores reasoning directly in the embedding space, bypassing explicit natural-language or pixel-level intermediate decoding and instead using continuous latent representations as internal reasoning states.

- **LVR (Li et al., 2026b) (Latent Visual Reasoning):** Projects visual features into a joint semantic space and performs autoregressive reasoning by reconstructing query-relevant visual tokens, referred to as "latent visual thoughts," interleaved with text generation. This allows the model to reason over continuous visual semantics without fully materializing every intermediate step in language.
- **Mull-Tokens (Ray et al., 2025):** Introduces modality-agnostic latent tokens that act as a multimodal scratchpad. These tokens are trained with interleaved traces and relaxed supervision so that they can flexibly store intermediate visual or textual information and optimize latent trajectories toward the final answer.
- **Monet (Wang et al., 2026b):** Enables reasoning directly in latent visual space by treating continuous embeddings as intermediate visual thoughts. It combines a distillation-based supervised fine-tuning pipeline with Visual-latent Policy Optimization (VLPO), explicitly incorporating latent embeddings into policy-gradient-based optimization to improve generalization on abstract visual reasoning tasks.

E Implementation Details

Base model and processor. All experiments are built upon Qwen2.5-VL-7B-Instruct as the underlying multimodal large language model. We use the corresponding official processor and tokenizer for text-image formatting and multimodal input construction. During training, we extend the tokenizer with a set of latent-specific special tokens, including a latent start token, a latent end token, and a sequence of latent placeholder tokens used to reserve latent slots in the autoregressive input stream.

Training data. To construct a robust and diverse environment for continuous latent reasoning, we train AMVL on a comprehensive mixture of multimodal reasoning datasets. This mixture is explicitly curated to cover a wide spectrum of cognitive tasks, from fine-grained visual grounding to complex multi-step logical deduction, ensuring the learned latent space generalizes across varied reasoning paradigms. The composition includes:

- **Visual-CoT (Shao et al., 2024):** This dataset provides a foundational corpus for step-by-step multimodal reasoning. It trains the model to decompose complex visual questions into intermediate logical steps, bridging high-level semantic queries with observable visual evidence.
- **ReFocus (Fu et al., 2025):** To ensure the latent reasoning space captures precise spatial and region-level information, we include ReFocus. This dataset emphasizes grounded visual reasoning, requiring the model to maintain focus on fine-grained visual details and specific regions of interest throughout the inference process.
- **CogCoM (Qi et al., 2025):** Focusing on complex cognitive trajectories, CogCoM enhances the model's ability to perform multi-hop deductive reasoning. It provides intricate scenarios where the model must synthesize multiple pieces of visual and textual information to arrive at a valid conclusion.

- **Zebra-CoT (Li et al., 2026a):** This dataset is utilized to strengthen relational and compositional reasoning capabilities. It forces the model to track intricate relationships between multiple objects or concepts within a visual scene, discouraging reliance on superficial data correlations.

All training samples across these diverse sources are standardized into a unified chat-style multimodal format. During preprocessing, any sample lacking a valid and complete assistant response is strictly filtered out to maintain high supervision quality. For every valid sample, we programmatically insert a dedicated latent token block immediately following the assistant's prefix. Training the main AMVL model is conducted on 16 NVIDIA A100 GPUs and takes approximately 20 hours.

Latent Block Construction and Insertion. For every valid sample, we programmatically insert a dedicated latent token block immediately following the assistant's prefix. Given a predefined number of latent slots k, this block allocates k dedicated latent placeholder tokens enclosed by specific start and end markers. In our implementation, the latent token block takes the form:

$$\texttt{<abs_token>} + \texttt{latent_pad}_1, \ldots, \texttt{latent_pad}_k + \texttt{</abs_token>}.$$

These placeholders serve solely as structural anchors for latent injection and are explicitly excluded from standard language modeling supervision. The start and end markers preserve the latent span boundaries, while the interior placeholder embeddings are replaced by the inferred continuous latent features during both training and inference. This structural modification is critical: it explicitly allocates the continuous capacity required for our dual-KL optimization. By forcing the model to process this latent block before generating any discrete textual output, we ensure that the core abstract and spatial reasoning fluidly evolves within the high-density continuous space prior to textual serialization.

Variational parameterization and latent settings. Both the prior and posterior are parameterized as factorized diagonal Gaussian distributions over the latent slots. The latent dimension is set to $d = 512$ by default, and the number of latent slots is set to $k = 8$ unless otherwise specified in the ablation studies. The variational module predicts the mean and log-variance of each latent slot, and latent samples are drawn using the standard reparameterization trick. In implementation, we predict log-variance rather than variance directly for improved numerical stability.

Input construction and loss masking. Training uses teacher-forced autoregressive decoding. The language modeling loss is applied only to the answer tokens following the assistant prefix, while the latent start token, latent end token, and all latent placeholder tokens are masked out from next-token supervision. The latent-token mask is maintained separately so that the model can identify the positions where latent variables should be injected. This design ensures that the model is not trained to predict the fixed latent placeholders themselves, but instead uses them as containers for continuous latent reasoning states.

Optimization details. We train the model with AdamW using fused PyTorch implementation, bf16 mixed precision, gradient checkpointing, cosine learning-rate scheduling, and a warmup ratio of 0.05. The vision encoder is frozen throughout training, while the language backbone and variational modules are jointly optimized. We also enable TF32 matrix multiplication where available for training acceleration. The exact batch size, gradient accumulation steps, learning rate, and training epochs follow the settings reported in the main experiments.

KL scheduling. We apply different schedules to the forward and reverse KL terms for stable optimization. The forward KL weight is linearly warmed up from 0 to 1.0 over the first 2000 training steps:

$$\beta_t = \min\left(1, \frac{t}{2000}\right).$$

For the reverse KL term, we use a delayed annealing strategy. Its weight is fixed to 0 during the first 1000 steps, and then linearly increased to 0.5 over the following 2000 steps:

$$\gamma_t = \begin{cases} 0, & t < 1000, \\ 0.5 \cdot \min\left(1, \frac{t-1000}{2000}\right), & t \geq 1000. \end{cases}$$

Table 6: Performance on the out-of-distribution VisualPuzzles benchmark.

Method	Overall	Algorithmic	Analogical	Deductive	Inductive	Spatial
Qwen2.5-VL-7B	32.71	37.02	21.80	47.50	26.32	21.80
Pangea-7B	31.30	32.40	23.70	38.50	28.70	32.50
Deepeyes	32.96	37.79	27.01	41.00	26.79	27.01
LVR	27.74	28.63	23.22	36.00	28.23	24.12
LLaVA-OneVision-72B	30.80	34.70	26.50	37.00	27.30	28.70
Ours-7B	33.90	32.44	27.96	52.50	28.23	30.77

This schedule allows the prior to become partially calibrated before reverse-side posterior regularization becomes active, which improves training stability in practice.

Embedding initialization and freezing strategy. After extending the tokenizer with latent-related special tokens, we resize the model vocabulary accordingly. Newly added token embeddings are initialized using the mean of the original embedding matrix. During the initial embedding adaptation stage, gradient updates to the embedding matrix are restricted to the newly added tokens, preventing unnecessary drift in the pretrained vocabulary. In the main training stage, the language backbone is jointly optimized with the variational components, while the visual encoder remains frozen.

F Out-of-Distribution Generalization on VisualPuzzles

To further evaluate whether the learned latent reasoning space generalizes beyond the in-distribution benchmarks used in the main paper, we test our model on VisualPuzzles, an out-of-distribution multimodal reasoning benchmark designed to assess more abstract reasoning skills. VisualPuzzles contains diverse reasoning categories, including algorithmic, analogical, deductive, inductive, and spatial reasoning, and therefore provides a useful testbed for evaluating whether the learned latent representations support transferable reasoning rather than benchmark-specific pattern matching.

Table 6 reports the results. Our 7B model achieves the best overall performance (33.90), outperforming all compared baselines, including Qwen2.5-VL-7B (Bai et al., 2025), Deepeyes (Zheng et al., 2026), Pangea-7B (Yue et al., 2024b), LVR (Li et al., 2026b), and LLaVA-OneVision-72B (Li et al., 2024). Notably, our model obtains the strongest score on deductive reasoning (52.50), with additional gains on analogical and spatial reasoning compared with most competing models. These improvements suggest that the learned latent reasoning states are not merely improving in-domain answer generation, but also provide a more transferable reasoning substrate under distribution shift.

This result complements the main benchmark findings. While the in-distribution results show that AMVL improves multimodal reasoning performance on standard evaluation sets, the VisualPuzzles experiment further indicates that the learned latent space retains useful abstract structure under out-of-distribution conditions. This is consistent with our broader claim that improving train-inference compatibility in latent reasoning can lead not only to better in-domain decoding, but also to stronger reasoning generalization.

G Semantic Properties of the Latent Reasoning Space

To better understand the learned latent reasoning tokens, we conduct qualitative and quantitative probing analyses, shown in Figure 3 and Figure 4.

Visual grounding. Figure 3 visualizes token-level relevance heatmaps obtained via occlusion-based sensitivity analysis. For each latent token L_i, we mask image patches and measure the resulting representation shift, where warmer colors indicate higher sensitivity. Across different queries, the latent tokens consistently respond to task-relevant visual regions rather than diffuse global context. For example, when the query asks about the color of a motorcycle or a comb, the strongest responses localize around the queried objects and their nearby boundaries. The tokens also exhibit diverse but overlapping relevance patterns: some capture broader context, while others focus more sharply on local details. This suggests that the latent block forms a sequence of progressively refined

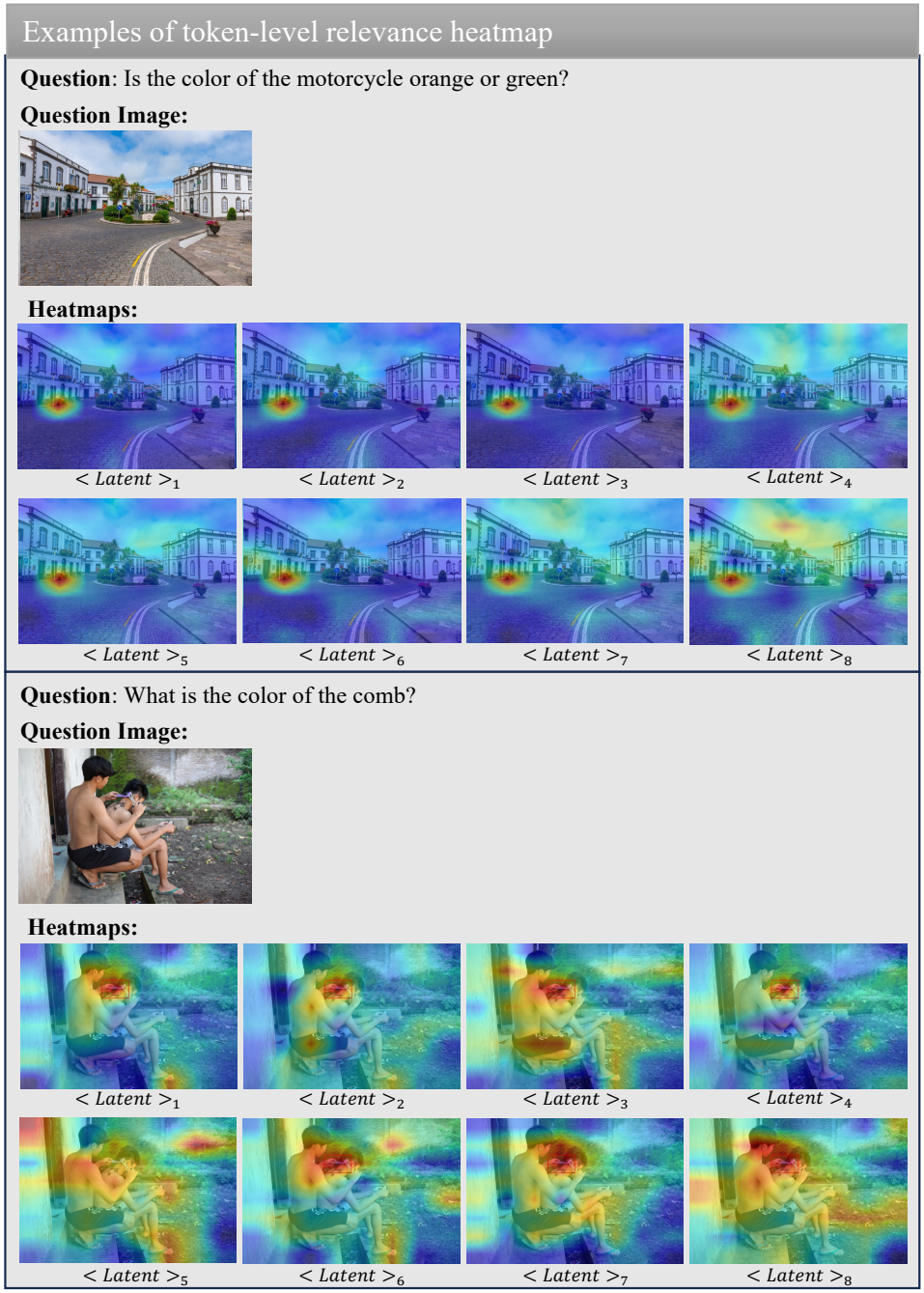


Figure 3: Token-level relevance heatmaps.

visual abstractions, rather than collapsing into redundant slots. More importantly, the strong localization to query-relevant regions is consistent with our main claim that continuous latent reasoning can preserve fine-grained perceptual grounding without forcing intermediate reasoning into discrete language tokens.

Latent geometry and sensitivity. Figure 4 provides complementary quantitative evidence. The token-wise cosine similarity matrix (left) shows a banded structure: adjacent latent tokens are more similar, while distant tokens (e.g., L_1 and L_8) are less aligned. This pattern suggests a smooth but non-collapsed latent trajectory. The perturbation analysis (right) further shows that latent represen-

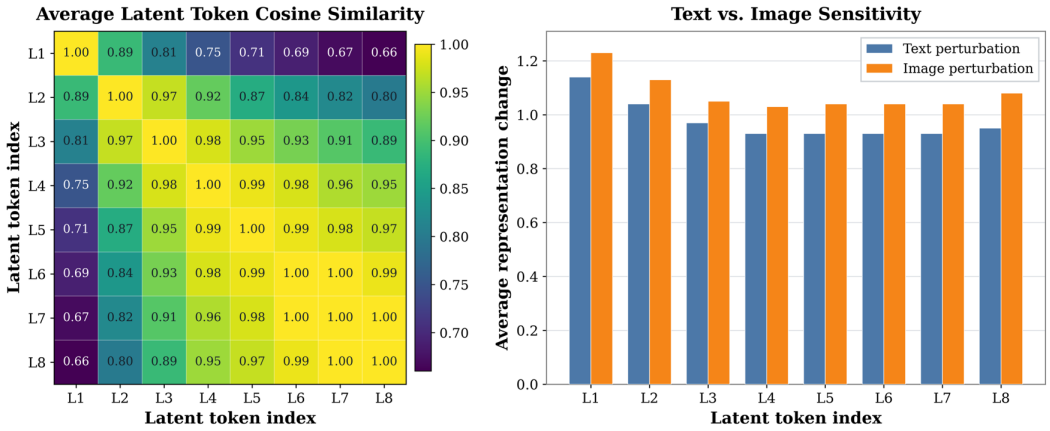


Figure 4: Quantitative analysis of latent token properties. **Left:** Average cosine similarity matrix across latent tokens. **Right:** Sensitivity of latent representations to image and text permutations.

Table 7: Latent-space spread and prior–posterior alignment statistics. Lower spread and paired L2 indicate more compact and better-aligned latent geometry, while higher cosine indicates stronger directional consistency.

Method	Prior Spread $S_p \downarrow$	Posterior Spread $S_q \downarrow$	Paired L2 $D_{\text{L2}} \downarrow$	Cosine $S_{\cos} \uparrow$	Mean Shift $D_{\text{shift}} \downarrow$
NTP	1.9243	2.2326	16.9488	-0.0190	16.6418
NTP + Rev-KL	0.4453	0.5177	3.6960	0.9594	3.6294
NTP + Fwd-KL	0.6359	0.8707	5.6528	0.8697	5.5484
AMVL	0.6786	0.8257	5.4317	0.8883	5.3240

tations are more sensitive to image permutations than to text permutations. This indicates that the learned latent space is strongly grounded in visual input, while still remaining conditioned on the textual query. Taken together, these results suggest that AMVL's latent reasoning space is strongly grounded in visual evidence, while still being shaped by the textual query. This is consistent with our broader motivation of alleviating the language-space bottleneck in multimodal reasoning.

H Latent Spread Analysis

To better understand how different training objectives shape the latent reasoning space, we analyze the dispersion of prior and posterior latent means under the four main training variants from our ablation study (Table 3): NTP, NTP + Rev-KL, NTP + Fwd-KL, and AMVL. Here, NTP denotes the baseline trained only with next-token prediction, without explicit latent regularization. NTP + Fwd-KL denotes the forward-KL alignment variant, while NTP + Rev-KL denotes the reverse-KL regularized variant. AMVL denotes the full bidirectional objective. This analysis complements the main ablation study by examining how each objective shapes the geometry of the learned latent space.

For each validation example $n \in \{1, \ldots, N\}$, we extract the prior mean $\mu_p^{(n)} \in \mathbb{R}^{k \times d}$ and posterior mean $\mu_q^{(n)} \in \mathbb{R}^{k \times d}$ from the trained model, where k is the number of latent slots and d is the latent dimension. To obtain a sample-level representation, we apply slot-wise mean pooling:

$$\bar{\mu}_p^{(n)} = \frac{1}{k} \sum_{i=1}^{k} \mu_{p,i}^{(n)}, \qquad \bar{\mu}_q^{(n)} = \frac{1}{k} \sum_{i=1}^{k} \mu_{q,i}^{(n)}.$$

For each branch (prior or posterior), we compute the global center across the validation set, denoted as c_p and c_q:

$$c_p = \frac{1}{N} \sum_{n=1}^{N} \bar{\mu}_p^{(n)}, \qquad c_q = \frac{1}{N} \sum_{n=1}^{N} \bar{\mu}_q^{(n)}.$$

We define the average latent spread for the prior (S_p) and posterior (S_q) as the mean Euclidean distance from each sample to its respective global center:

$$S_p = \frac{1}{N}\sum_{n=1}^{N}\left\|\bar{\mu}_p^{(n)} - c_p\right\|_2, \qquad S_q = \frac{1}{N}\sum_{n=1}^{N}\left\|\bar{\mu}_q^{(n)} - c_q\right\|_2.$$

Notably, these spread statistics quantify the dispersion of sample-level pooled latent means across the dataset, rather than the covariance or support width of each individual latent distribution.

Geometric Interpretation: S_p and S_q quantify the global concentration of sample-level latent representations across the dataset. Lower spread indicates that the pooled latent means are more tightly clustered around the global center, whereas higher spread suggests stronger cross-sample drift.

To comprehensively quantify the alignment between the prior and posterior branches, we formulate three paired-geometry metrics, each capturing a distinct aspect of the train-inference mismatch:

$$D_{\text{L2}} = \frac{1}{N}\sum_{n=1}^{N}\left\|\bar{\mu}_p^{(n)} - \bar{\mu}_q^{(n)}\right\|_2, \quad S_{\cos} = \frac{1}{N}\sum_{n=1}^{N}\frac{\langle\bar{\mu}_p^{(n)}, \bar{\mu}_q^{(n)}\rangle}{\|\bar{\mu}_p^{(n)}\|_2\|\bar{\mu}_q^{(n)}\|_2}, \quad D_{\text{shift}} = \|c_p - c_q\|_2.$$

- **Paired L2 Distance (D_{L2})** measures the *absolute instance-level error*. It evaluates how far apart the target-agnostic prior and the target-aware posterior are for the exact same input sample.
- **Cosine Similarity ($S_{\cos}$)** measures the *directional consistency*. Regardless of magnitude, it evaluates whether the prior and posterior point toward the same semantic region in the high-dimensional space.
- **Mean Shift (D_{shift})** measures the *systematic global bias*. It captures the macro-level translation between the entire prior distribution and the posterior distribution, indicating overall domain drift.

Table 7 summarize these statistics. Several patterns are broadly consistent with the benchmark trends reported in the main text.

NTP leads to severe latent mismatch and the weakest overall reasoning behavior. Without explicit latent regularization, NTP produces the most dispersed latent geometry for both branches ($S_p = 1.9243, S_q = 2.2326$), together with the worst prior–posterior alignment ($D_{\text{L2}} = 16.9488, S_{\cos} = -0.0190, D_{\text{shift}} = 16.6418$). This suggests that the posterior drifts freely toward target-dependent latent regions while the prior remains poorly calibrated for inference-time usage. Such severe train–inference mismatch is consistent with the weaker benchmark performance of NTP-only observed in the ablation study.

NTP + Rev-KL produces the most concentrated sample-level latent geometry, but not the strongest downstream performance. Among all variants, NTP + Rev-KL yields the smallest sample-level spread and the strongest prior–posterior alignment under the considered metrics ($S_p = 0.4453, S_q = 0.5177, D_{\text{L2}} = 3.6960, S_{\cos} = 0.9594$). This suggests that reverse-side regularization is effective at reducing cross-sample drift of latent means and improving compatibility with the learned prior. Importantly, this concentration is measured at the level of pooled latent means across validation examples, and does not imply that each individual posterior becomes sharper. In fact, reverse KL can still encourage broader support within a sample while making the sample-level latent geometry more globally concentrated. However, NTP + Rev-KL does not achieve the best downstream performance, implying that stronger geometric concentration alone is not sufficient: overly strong posterior regularization may reduce useful target-conditioned variability needed for reasoning.

NTP + Fwd-KL substantially improves over NTP, but still exhibits an imbalanced latent geometry. The forward-KL variant (NTP + Fwd-KL) resolves most of the catastrophic mismatch observed under NTP, improving both paired distance and cosine alignment by a large margin. At the same time, the prior branch remains more concentrated across examples ($S_p = 0.6359$), while the posterior branch is still more dispersed across examples ($S_q = 0.8707$), yielding a relatively large prior–posterior spread gap ($\Delta S = S_q - S_p = 0.2348$). This pattern is consistent with a one-sided

Table 8: Ablation study of variational head architectures. We compare a linear head, a standard MLP head, our LLM-native head, and a deeper variant of the same design.

Method	V*	HRBench4K	HRBench8K
Linear head	76.96	70.00	66.25
MLP head	75.92	69.88	65.50
LLM-native head	84.29	72.12	68.50
Deeper LLM-native head	75.39	70.25	65.50

calibration regime in which the prior is trained to chase the posterior, while the posterior itself remains more weakly constrained and therefore relatively target-dependent. Correspondingly, NTP + Fwd-KL improves benchmark accuracy over NTP, but remains below the full AMVL objective.

AMVL yields a more favorable trade-off between prior expressiveness and posterior regularization. Compared with NTP + Fwd-KL, AMVL slightly increases the spread of sample-level prior means ($S_p = 0.6786$) while slightly reducing the spread of sample-level posterior means ($S_q = 0.8257$), thereby significantly shrinking the prior–posterior spread gap ($\Delta S = 0.1471$). It also improves paired L2 distance, cosine similarity, and mean shift over the forward-only baseline. Importantly, although AMVL does not produce the most concentrated latent geometry overall, it achieves the strongest downstream benchmark performance. This suggests that the most effective latent reasoning space is not the one with the tightest global concentration, but the one that best balances inference-time prior expressiveness with training-time posterior regularization. This is consistent with the main design of AMVL, which combines prior alignment with posterior control to improve train-inference compatibility.

I Additional Ablation Studies

In this section, we provide extended empirical analyses to further validate the structural design, optimization stability, and inference robustness of our AMVL framework. Specifically, we investigate: (1) the impact of different **variational head architectures**, confirming the necessity and efficiency of our lightweight LLM-native design; (2) the **sensitivity to loss weights**, demonstrating that AMVL's dual-KL objective remains robust across a diverse range of hyperparameter settings; and (3) the **robustness to inference-time latent sampling** under various temperatures, which empirically verifies the smoothness and stability of the learned continuous latent space.

Effect of the variational head architecture. Table 8 compares different variational head architectures, including a linear head, a standard MLP head, our proposed LLM-native head, and a deeper variant of the same design. The proposed lightweight LLM-native head performs best across benchmarks, indicating that variational parameterization benefits from being aligned with the architectural patterns of the underlying MLLM rather than relying on a separately designed generic projection head.

Interestingly, the deeper variant performs worse than the default head. We hypothesize that, in our setting, the variational head mainly serves as a lightweight readout of latent statistics from the shared MLLM hidden states, rather than as an independent high-capacity encoder. Increasing the head depth may therefore weaken feature-space alignment with the backbone, complicate prior-posterior calibration, and make the posterior more prone to fitting target-conditioned shortcut information.

Sensitivity to loss weights. Table 9 reports a sensitivity analysis over the objective weights. Overall, AMVL remains reasonably stable across a range of coefficient settings, indicating that the gains do not depend on a narrowly tuned loss balance. At the same time, changing the relative strengths of the NTP, forward KL, and reverse KL terms affects the trade-off across benchmarks, which is consistent with their different roles in latent-space learning.

In particular, the forward KL term mainly improves prior calibration for inference, while the reverse KL term controls posterior sharpness and support compatibility. Over-emphasizing either side can hurt performance: excessively strong reconstruction pressure may weaken latent regularization, whereas overly strong reverse regularization may suppress useful target-conditioned information.

Table 9: Sensitivity analysis of loss weights. We vary the coefficients of the next-token prediction loss, forward KL alignment, and reverse KL regularization to evaluate the robustness of AMVL to objective weighting.

λ_{NTP}	β	γ	V*	HRBench4K	HRBench8K
1	0.5	1	81.68	71.62	67.75
1	1	0.5	84.29	72.12	68.50
2	1	1	80.63	73.00	68.88
2	0.5	1	81.68	72.00	69.00
2	1	0.5	81.68	70.88	67.88

Table 10: Ablation study of the stop-gradient design in our variational alignment objectives. We evaluate the effect of removing stop-gradient from the prior-alignment term, the posterior-regularization term, or both.

Method	V*	HRBench4K	HRBench8K	VisualPuzzles
w/o sg in Prior Alignment	83.25	72.12	67.62	32.53
w/o sg in Posterior Regularization	80.10	72.25	66.75	32.79
w/o sg in Both	81.68	72.50	67.75	33.05
Full	84.29	72.12	68.50	33.90

Overall, the results support the design of AMVL as a balanced objective that jointly improves prior calibration and posterior regularization.

Effect of stop-gradient design. As shown in Table 10, the full model performs best overall, demonstrating the importance of the stop-gradient design in our variational alignment objectives. Removing stop-gradient from either the prior-alignment term or the posterior-regularization term leads to clear performance drops, especially on V* and HRBench8K. The degradation is most pronounced when the prior-alignment term is allowed to update the posterior, suggesting that this term works best when the posterior serves as a fixed, answer-informed teacher for the prior. Similarly, removing stop-gradient from posterior regularization weakens the intended constraint on the posterior by allowing the prior to co-adapt.

Additionally, removing stop-gradient from both terms does not cause catastrophic failure, but still underperforms the full model across most benchmarks. This indicates that the gain of our method comes not only from the presence of bidirectional KL objectives, but more importantly from the decoupled gradient flow they enforce. Overall, the results validate our design principle that prior alignment and posterior regularization should update only their intended target distributions.

Robustness to Inference-Time Latent Sampling. To further test the stability of the learned latent reasoning space at inference time, we evaluate AMVL under stochastic prior sampling with different temperatures. In the main experiments, inference uses the prior latent mean directly. Here, we instead inject Gaussian noise scaled by the prior standard deviation and a temperature parameter τ:

$$z = \mu + \tau \cdot \epsilon \odot \sigma, \qquad \epsilon \sim \mathcal{N}(0, I), \quad \sigma = \exp\left(\tfrac{1}{2} \log \text{var}\right).$$

When $\tau = 0$, this reduces to deterministic mean-based inference. As τ increases, the sampled latent variables become increasingly stochastic. At inference time, both μ and $\log \text{var}$ are obtained from the prior branch, so the injected noise reflects uncertainty under the learned inference-time latent distribution rather than any target-aware posterior information.

We evaluate the trained model under $\tau \in \{0.0, 0.2, 0.5, 0.8, 1.0\}$ on V*, HRBench4K, HRBench8K, and VisualPuzzles. The results are reported in Table 11.

The results show that AMVL remains reasonably stable under moderate latent perturbations. On V*, performance drops from 84.29 at $\tau = 0.0$ to 83.25 at $\tau = 0.2$ and remains at the same level at $\tau = 0.5$, indicating that mild stochasticity does not substantially harm inference. HRBench4K exhibits a similarly gradual degradation, decreasing from 72.12 to 71.00 as τ increases from 0.0 to 1.0. HRBench8K is even more stable, with only minor fluctuations across temperatures. On

Table 11: Robustness to inference-time latent sampling under different temperatures. Higher is better.

Benchmark	Inference Temperature (τ)				
	$\tau = 0.0$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.8$	$\tau = 1.0$
V*	84.29	83.25	83.25	82.72	82.20
HRBench4K	72.12	71.88	71.75	71.62	71.00
HRBench8K	68.50	67.88	68.00	67.50	67.62
VisualPuzzles	33.90	34.50	33.39	33.56	32.71

VisualPuzzles, moderate stochastic sampling at $\tau = 0.2$ slightly improves the score (34.50 vs. 33.90), while higher temperatures eventually reduce performance.

Overall, these results suggest that the latent reasoning space learned by AMVL is not overly brittle to local perturbations at inference time. Moderate sampling noise only leads to small performance changes, and in some cases can even slightly improve generalization, which is consistent with the learned latent space being locally smooth rather than highly unstable. At the same time, performance gradually degrades as τ becomes large, indicating that excessive stochasticity still moves the latent state away from the most reliable inference region. Together with the latent-space analyses in Appendix H, this sampling ablation suggests that improved prior-posterior calibration in AMVL not only benefits deterministic inference, but also makes the learned latent reasoning space more robust to moderate stochastic perturbations.

J Additional Experiments and Analyses

J.1 Broader Benchmarks

General multimodal reasoning benchmarks. We evaluate Qwen2.5-VL-7B, LVR, and AMVL on MMMU (Yue et al., 2024a), MathVista (Lu et al., 2023), MathVision (Wang et al., 2024), and ScienceQA (Lu et al., 2022). All data and evaluation procedures are taken from the official benchmark repositories, and all compared methods use the same setup. MMMU is used as the representative evaluation from the MMMU benchmark family. Table 12 shows that AMVL obtains the highest score on all four benchmarks.

Table 12: Results on broader multimodal reasoning benchmarks. Best results are in **bold**.

Method	MMMU	MathVista	MathVision	ScienceQA
Qwen2.5-VL-7B	50.33	64.50	21.38	83.52
LVR	49.11	67.20	20.72	87.64
AMVL	**53.11**	**67.50**	**22.37**	**88.71**

J.2 Direct Diagnostics of Answer Leakage

We compare AMVL with a variant trained without reverse KL using three diagnostics. For each example $(\mathbf{x}, \mathbf{y}^+)$, an incorrect answer $\mathbf{y}^-$ is obtained by randomly sampling the ground-truth answer of a different question, and a counterfactual posterior latent is sampled as $\mathbf{z}^- \sim q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}^-)$. Counterfactual Latent Robust Accuracy (CLRA) is the percentage of examples for which the decoder recovers $\mathbf{y}^+$ from the original input and the misleading latent, i.e., from $p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}^-)$.

Counterfactual Latent KL measures posterior sensitivity to changing only the answer:

$$\mathcal{D}_{\mathrm{cf}} = D_{\mathrm{KL}}\big(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}^+) \,\|\, q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}^-)\big)\,. \tag{45}$$

For the Answer-Alignment Margin, we pool the posterior latent into $\mathbf{e}_z$ and average the decoder's final-layer token states for the correct and incorrect answers into $\mathbf{h}^+$ and $\mathbf{h}^-$, respectively:

$$\mathcal{M}_{\mathrm{align}} = \cos(\mathbf{e}_z, \mathbf{h}^+) - \cos(\mathbf{e}_z, \mathbf{h}^-). \tag{46}$$

Higher CLRA and lower values of the other two metrics indicate weaker dependence on the answer used to construct the posterior. Table 13 shows that reverse KL raises CLRA by 9.5 absolute points and reduces Counterfactual Latent KL and Answer-Alignment Margin by 76.6% and 60.6%, respectively.

Table 13: Direct diagnostics of answer leakage.

Metric	w/o Rev-KL	AMVL	Change
CLRA ↑	21.0%	**30.5%**	+9.5 points
Counterfactual Latent KL ↓	7.7697	**1.8176**	−76.6%
Answer-Alignment Margin ↓	2.844×10^{-5}	$\mathbf{1.120\times10^{-5}}$	−60.6%

J.3 Functional Role and Visual Grounding of Latent Tokens

Latent-token interventions. At inference time, we reduce, remove, or permute the latent tokens. Every intervention in Table 14 degrades performance relative to the full model. In particular, the drop after permutation shows that the latent slots are not interchangeable and that their order carries task-relevant information.

Table 14: Inference-time interventions on the latent sequence.

Model variant	V^*	HRBench4K	HRBench8K
Full AMVL	**84.29**	**72.12**	**68.50**
Fewer latent tokens	81.15	70.75	66.88
Removed latent tokens	80.63	70.62	65.75
Permuted latent tokens	82.72	71.75	67.88

Dependence on visual information. We additionally train a variant in which latent tokens are prevented from accessing visual information. The performance degradation in Table 15 shows that the latent variables absorb task-relevant image evidence rather than serving only as generic textual computation slots.

Table 15: Effect of preventing latent tokens from accessing visual information.

Model variant	V^*	HRBench4K	HRBench8K
Full AMVL	**84.29**	**72.12**	**68.50**
Latents w/o visual information	78.53	70.62	66.38

Spatial grounding. The occlusion-based sensitivity maps in Figure 3 provide complementary qualitative evidence. Masking different image patches and measuring the resulting latent-representation shift shows that different latent tokens respond to distinct, query-relevant regions.

J.4 Sensitivity to Reverse-KL Annealing

The default schedule delays reverse-KL regularization until the prior has become partially calibrated (Appendix E). We compare it with a variant in which reverse KL is introduced immediately without annealing. Table 16 shows that the no-annealing model remains functional but underperforms the delayed schedule on all three benchmarks.

J.5 Inference Cost

Table 17 reports inference-cost measurements under the same evaluation setup. Relative to Qwen2.5-VL-7B, AMVL adds 0.16 GB of GPU memory, increases time to first token (TTFT) by 115 ms, and changes throughput from 0.68 to 0.63 samples/s. Relative to LVR, AMVL uses 0.05 GB more memory, reduces TTFT by 218 ms, and improves throughput by 0.17 samples/s.

Table 16: Sensitivity to delayed reverse-KL annealing.

Model	V*	HRBench4K	HRBench8K
AMVL w/o annealing	82.20	71.50	67.50
AMVL with delayed annealing	**84.29**	**72.12**	**68.50**

Table 17: Inference efficiency. Lower memory and TTFT are better; higher throughput and accuracy are better.

Method	Memory (GB) ↓	TTFT (ms) ↓	Samples/s ↑	V* ↑
Qwen2.5-VL-7B	**17.58**	**1148**	**0.68**	76.44
LVR	17.69	1481	0.46	81.15
AMVL	17.74	1263	0.63	**84.29**

K Limitations and Future Work

While extended ablations (Appendix I) confirm AMVL's stability and efficacy, our current empirical validation is limited to the 7B parameter scale. A critical direction for future work is scaling AMVL to larger foundational models (e.g., 70B+) to investigate whether massive parameter counts induce the spontaneous emergence of more complex, generalized latent reasoning structures, further advancing the frontier of continuous multimodal reasoning.

L Broader Impacts

Our proposed latent thinking model aims to enhance the reasoning efficiency and interpretability of large language models. Positively, this could lower computational costs for complex reasoning tasks and make model decision-making processes more transparent to users. However, we also acknowledge potential negative societal impacts. Enhanced reasoning capabilities could potentially be misused to generate more sophisticated disinformation or automate malicious activities such as phishing.