# NEXT THOUGHTS ARE DISTRIBUTIONS: GENERATIVE AUTOREGRESSIVE REASONING IN LATENT SPACE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reasoning problems often admit multiple valid ways to proceed. Continuous reasoning promises to move computation beyond language tokens into a more compact latent space, but representing several plausible ways to think next remains difficult. We introduce Autoregressive Thought Flow (ATF), which models the next continuous thought as a multimodal distribution. A causal autoregressive model performs the reasoning computation, while a lightweight diffusion head generates a plausible next thought from the resulting condition. The sampled thought is fed back into the model, allowing continuous reasoning to unfold for a variable number of steps while preserving the pretrained backbone. Across mathematical reasoning tasks, ATF improves accuracy with compact latent traces and benefits from reinforcement learning and additional test-time thinking. Multi-sample evaluation shows broader solution coverage, indicating that its multimodal predictions capture useful diversity among reasoning paths. Our results suggest that continuous reasoning is more effective when multiple possible next thoughts remain available rather than being collapsed into a single prediction.

## 1 INTRODUCTION

The next thought is rarely unique. The same problem can support several valid lines of reasoning, each following a different sequence of intermediate computations. Chain-of-thought makes these computations easy to supervise and inspect, but also requires every step to be expressed through a discrete, sequential language interface (Wei et al., 2022). Continuous reasoning offers a different possibility: keep intermediate computation in latent space (Hao et al., 2024), where thoughts can be compact and need not be expressed in words.

This freedom creates a basic problem: what should come next? In language, a softmax defines a distribution over a fixed vocabulary. In continuous space, there is no fixed set of alternatives, and several distinct next thoughts may all be valid. Existing methods learn such latent states in different ways (Zhang et al., 2025; Kong et al., 2025; Hao et al., 2024), including by reconstructing or predicting target representations (Ray et al., 2025; Barrault et al., 2024; Tan et al., 2025; Cheng & Van Durme, 2024; Shao et al., 2025). Recent work introduces stochastic transitions or explores multiple latent trajectories (Zou et al., 2026; Zhou et al., 2026; Wu et al., 2026; You et al., 2026). Diffusion provides a natural way to model multimodal distributions over continuous outputs (Li et al., 2024; Zhou et al., 2024; Team et al., 2025; Deng et al., 2024), but diffusion-based reasoning often replaces the causal autoregressive interface with bidirectional or full-sequence denoising (Kang et al., 2025). This changes the structure of pretrained language models and makes variable-length reasoning less natural. The challenge is therefore to preserve this multimodality without giving up autoregressive reasoning.

Our key idea is to model the next thought as a multimodal distribution while keeping reasoning autoregressive. We introduce Autoregressive Thought Flow (ATF), where a causal transformer performs the reasoning computation and produces a condition for what should come next. A lightweight diffusion head (Li & He, 2026) turns this condition into a distribution over multiple plausible continuous thoughts, samples one, and feeds it back into the model. See a concrete example in Figure 1. The backbone therefore determines what is plausible, while diffusion preserves the alternative ways that the next thought can be realized. Because thoughts are generated one at a time, ATF retains the structure of pretrained causal models and is able to extend reasoning without fixing the trajectory length in advance.

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.

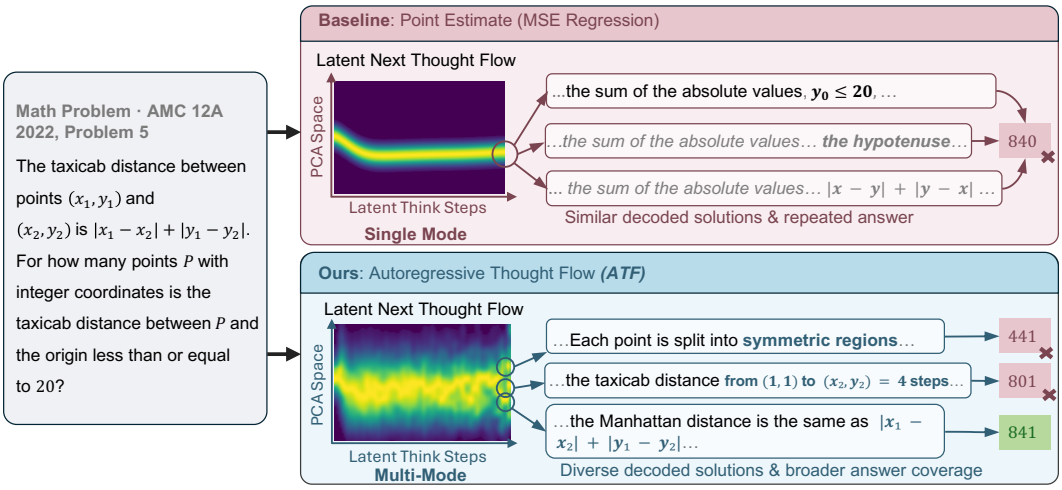



Figure 1: **Our ATF enables multiple latent continuations.** Given the same math prompt, a point-estimate continuous-thought model collapses the next-thought distribution into a single averaged trajectory, producing similar decoded continuations and an incorrect final answer. ATF instead samples next thoughts from a multi-mode latent distribution and feeds each realized thought back autoregressively, allowing distinct reasoning paths to be explored. In this example, one sampled ATF trajectory follows the correct taxicab-counting path and reaches the correct answer.

This design provides several additional benefits. Sampling from the next-thought distribution naturally exposes multiple reasoning trajectories that can be explored at test time. The diffusion formulation also inherits classifier-free guidance (Ho & Salimans, 2022) at no architectural cost, providing a built-in control over exploration at inference time. Finally, retaining the autoregressive interface makes it straightforward to apply reinforcement learning directly to the model's own unrolled latent trajectories.

Experiments and visualizations show that these capabilities translate into stronger reasoning. Across mathematical reasoning benchmarks, including the challenging OlympiadBench (He et al., 2024) and AMC 12 (Dekoninck et al., 2026), ATF improves the accuracy–reasoning-length trade-off over prior continuous-reasoning methods and short-budget language chain-of-thought (Wei et al., 2022), while using compact latent traces. Reinforcement learning further improves accuracy and leads the model to use more latent computation when useful, while generating additional thoughts at test time continues to improve performance. In multi-sample evaluation, ATF improves Pass@100 over the strongest prior baseline by 19.3 points on average across MATH-500 (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and OlympiadBench (He et al., 2024), outperforming iCoT (Deng et al., 2023), Coconut (Hao et al., 2024), CODI (Shen et al., 2025), TaH+ (Fu et al., 2025), Discrete Latent Reasoning (Su et al., 2025), Soft Thinking (Zhang et al., 2025), and the diffusion-based LaDiR (Kang et al., 2025). On OlympiadBench alone, the gain reaches 34.6 points. Even when the final answer is decoded greedily, sampling continuous thoughts substantially increases solution coverage, showing that the diversity comes from latent reasoning itself rather than text sampling. Decoded trajectories further reveal multiple plausible continuations from the same context. Our results show that preserving multimodality in next-thought prediction is useful for continuous reasoning.

## 2 Related Work

**Continuous and latent reasoning.** Chain-of-thought exposes intermediate reasoning in language (Wei et al., 2022), but long textual traces can be costly at inference time. Continuous-reasoning methods instead move this computation into latent space. Coconut feeds hidden states back into the model (Hao et al., 2024), Soft Thinking propagates soft token distributions (Zhang et al., 2025), latent-thought models introduce variational states (Kong et al., 2025), and LCM (Barrault et al., 2024; Duquenne et al., 2023), CCOT (Cheng & Van Durme, 2024), MuLTI (Ray et al., 2025), CoLaR (Tan et al., 2025), CALM (Shao et al., 2025), and MARCOS (Liu et al., 2025a) learn compact continuous representations of intermediate reasoning. These works establish continuous thoughts as a useful

reasoning interface, but many still propagate a single latent continuation or optimize toward a single target representation. Recent analyses show that such continuous reasoning can become effectively single-threaded (Wu et al., 2026), motivating methods that explicitly introduce stochasticity. Latent Thought Flow uses GFlowNets to model stochastic latent trajectories (Zou et al., 2026), LEPO uses Gumbel-Softmax sampling to explore latent reasoning policies (Zhou et al., 2026), and parallel latent reasoning injects dropout and Gaussian noise to generate multiple trajectories (You et al., 2026). ATF takes a complementary approach: it directly models the next latent thought as a conditional multimodal distribution rather than introducing stochasticity around a single predicted state.

**Generative next-state modeling.** Diffusion and flow models provide flexible tools for modeling complex continuous distributions (Song & Ermon, 2019; Ho et al., 2020; Lipman et al., 2023; Liu et al., 2023). Recent autoregressive image and multimodal models use such generative heads for next-vector prediction without discretizing continuous outputs (Li et al., 2024; Zhou et al., 2024; Team et al., 2025; Deng et al., 2024). LatentLM is especially related: it combines a causal Transformer with next-token diffusion to autoregressively generate continuous VAE latents (Sun et al., 2024). Its focus is multimodal generation and understanding, whereas ATF applies this generative interface to reasoning, where each sampled vector becomes the next intermediate thought. LaDiR also applies diffusion to latent reasoning, but denoises a reasoning trajectory at the sequence level (Kang et al., 2025). In contrast, ATF models the immediate next thought with a lightweight diffusion head while preserving autoregression across thoughts, allowing the reasoning trajectory to grow dynamically. This formulation also inherits classifier-free guidance (Ho & Salimans, 2022), providing an inference-time control over latent exploration without changing the causal backbone.

# 3 METHOD

Let $X$ be the input prompt, $Z = \{z_1, \ldots, z_T\}$ a sequence of continuous thoughts, and $Y$ the final answer. ATF generates $Z$ autoregressively before decoding $Y$. For supervision, we split each text rationale into ordered chunks and use a frozen Causal-Register VAE (CR-VAE; Section 3.4) to encode them into latent registers. With $z_{\text{stop}}$ denoting the embedding of the closing marker, each training example has the form

$$\underbrace{\text{[Prompt]}}_{X} \quad \texttt{<latent>} \quad \underbrace{\text{[Continuous thoughts]}}_{z_1,\ldots,z_T} \quad \underbrace{\texttt{</latent>}}_{z_{\text{stop}}} \quad \underbrace{\text{[Answer]}}_{Y}. \tag{1}$$

Discrete tokens use the backbone's embedding table, while continuous registers are projected into the same input-embedding space.

## 3.1 AUTOREGRESSIVE THOUGHT FLOW

ATF models each next thought as a conditional multi-mode distribution, factorizing the latent trajectory as

$$p_\theta(Z \mid X) = \prod_{t=1}^{T} p_\theta(z_t \mid z_{<t}, X). \tag{2}$$

A causal transformer reads the prompt and previously realized thoughts to produce a context state $h_t \in \mathbb{R}^d$ at the position immediately before $z_t$. A lightweight conditional flow head (Li & He, 2026) maps this state to a distribution over the next continuous thought. Sampling a thought and feeding it back preserves autoregression across steps while allowing different continuations from the same history (Figure 2).

We train the generative head using Rectified Flow (Li & He, 2026) due to its linear trajectories and effectiveness for high-dimensional vector prediction. For a target register $z_t$ and noise $\epsilon \sim \mathcal{N}(0, I)$, the interpolation at flow time $\tau$ is

$$Z_\tau = \tau z_t + (1-\tau)\epsilon, \quad \tau \in [0,1]. \tag{3}$$

The conditioned vector field $v_\psi(Z_\tau, \tau, h_t)$ predicts the noise-to-target direction through

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\tau,\epsilon}\left[\|v_\psi(Z_\tau, \tau, h_t) - (z_t - \epsilon)\|_2^2\right]. \tag{4}$$

We evaluate multiple flow times in parallel to provide lower-variance training signals to the conditioning state $h_t$.

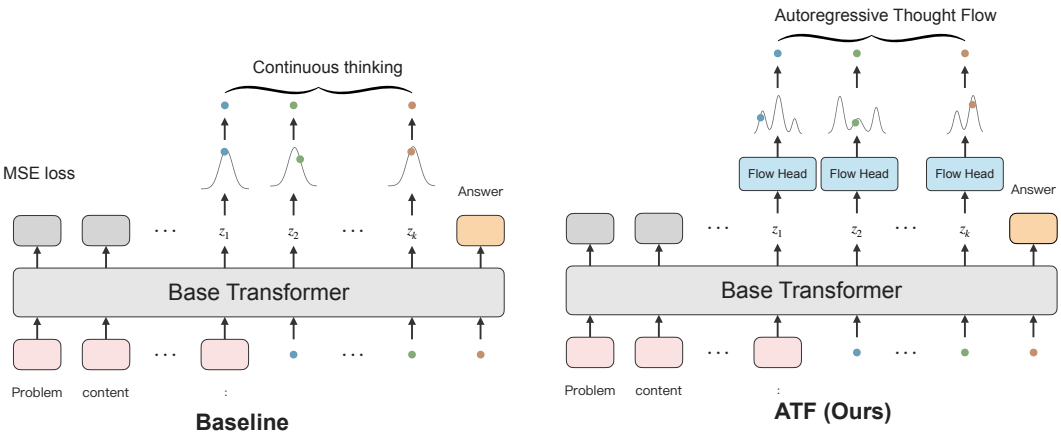



Figure 2: **ATF generates each next thought conditioned on the realized history.** A causal backbone produces the condition for a lightweight flow head, which samples a continuous thought and feeds it back into the reasoning history.

On the hybrid sequence in Eq. (1), we jointly train continuous-thought prediction and discrete-token generation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(X_{\text{discrete}}) + \lambda \mathcal{L}_{\text{flow}}(Z_{\text{latent}}). \tag{5}$$

We also supervise termination from the causal state $h_t$, so the model can decide when to end latent reasoning without reconstructing its thoughts into text. The closing marker supplies the boundary embedding $z_{\text{stop}}$ for answer generation; the rollout and stopping procedure are described in Section 3.3.

### 3.2 Reinforcement Learning for Autoregressive Thought Flow

The generative next-thought distribution provides multiple latent continuations that can be explored under an answer reward. As an optional refinement after SFT, we use group-relative optimization (Shao et al., 2024) to reinforce successful self-generated trajectories. Two levels of randomness matter: the latent reasoning path and the text answer decoded from it. Our formulation supports credit assignment at both levels while retaining the autoregressive flow model.

**Rewarding latent exploration.** For each prompt, we sample $N$ latent trajectories and decode $M$ text answers from each trajectory. Let $r_{n,m} \in \{0, 1\}$ denote answer correctness. We score a latent history by its mean answer reward and form separate advantages:

$$\begin{aligned} \bar{r}_n &= \frac{1}{M} \sum_m r_{n,m}, \qquad b = \frac{1}{N} \sum_n \bar{r}_n, \\ A_n^{\text{lat}} &= \bar{r}_n - b, \qquad A_{n,m}^{\text{text}} = r_{n,m} - \bar{r}_n. \end{aligned} \tag{6}$$

The implementation standardizes and clips these advantages. Averaging sibling answers reduces the influence of one lucky text completion on latent credit; comparing those answers within a shared history isolates the text-decoding contribution.

**Updating the generative policy.** The text-policy objective rewards correct answers conditioned on self-generated latent histories, training the backbone to better use these histories. Optionally, a latent-policy objective assigns outcome-based credit directly to the autoregressive thought flow, favoring continuations that lead to successful answers. For this latent term, we follow Flow-GRPO (Liu et al., 2025b) and inject noise at selected flow steps to record and score stochastic transitions. These transitions receive the latent advantage, while answer tokens receive the text advantage. With $\mathcal{C}$ denoting the clipped policy-update surrogate, we optimize

$$\mathcal{J}_{\text{RL}} = \lambda_z \langle \mathcal{C}(\rho^z, A^{\text{lat}}) \rangle_z + \lambda_y \langle \mathcal{C}(\rho^y, A^{\text{text}}) \rangle_y. \tag{7}$$

Here $\rho^z$ and $\rho^y$ compare current and behavior-policy transition scores and answer-token probabilities, respectively. Setting $\lambda_z = 0$ yields answer-policy refinement of the shared backbone; enabling the latent term additionally updates the flow head and its conditioning backbone. The CR-VAE remains

fixed. Transition-score definitions and normalization are given in Appendix A; multi-sample solution coverage assesses the effect on exploration.

### 3.3 Test-Time Generation and Guidance

Given a prompt, the backbone produces $h_t$ from the realized history. We draw $\epsilon \sim \mathcal{N}(0, I)$ and generate the next thought by integrating the conditioned flow from $\tau = 0$ to $\tau = 1$:

$$z_t = \epsilon + \int_0^1 v_\psi(Z_\tau, \tau, h_t) d\tau. \tag{8}$$

The sampled $z_t$ is appended to the history, updating the backbone's KV cache for the next step. Reasoning continues until the stop-margin criterion fires or the maximum latent budget is reached. The model then appends $z_{\text{stop}}$ and decodes the answer through its language modeling head. This procedure applies to both the supervised model and its RL-refined variants.

**Length and sampling budgets.** Additional inference compute can extend a trajectory or explore more trajectories. Adjusting the stopping margin permits more latent steps before answering; Figure 3 evaluates the resulting length–accuracy trade-off for ATF and ATF-RL. Independently sampling latent trajectories instead explores alternative continuations of the same prompt, even with greedy text decoding. The multi-sample coverage results in Table 2 and Appendix F assess whether these alternatives reach correct answers on more problems.

**Guidance controls exploration.** Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) adjusts the concentration of the next-thought distribution. During training, we drop $h_t$ with probability $p_{\text{uncond}}$ to learn an unconditional field alongside the conditioned one. At inference, we use

$$\tilde{v}_\tau = v_\psi(Z_\tau, \tau, \emptyset) + \gamma \left( v_\psi(Z_\tau, \tau, h_t) - v_\psi(Z_\tau, \tau, \emptyset) \right). \tag{9}$$

Larger $\gamma$ amplifies the conditional component; smaller $\gamma$ retains more influence from the unconditional field. Each guided sample enters the causal history for subsequent steps. Table 4 evaluates how this control trades single-sample accuracy against trajectory diversity and multi-sample coverage.

### 3.4 Grounding Latent Thoughts via Causal-Register VAE

To train ATF, we require ground-truth sequences of continuous latent thoughts $Z$ that correspond to valid reasoning steps. We extract these trajectories using an independent, pre-trained Causal-Register VAE (CR-VAE; Figure 5 in Appendix C). Traditional block-based VAEs encode sequence chunks independently. While sufficient for isolated representations, independent encoding is fundamentally incompatible with autoregressive sequence modeling; it duplicates information across steps and breaks the causal chain. Instead, CR-VAE models the reasoning sequence as a cumulative causal process. We insert a stochastic register $z_t$ after each reasoning block, tasked with summarizing the new reasoning information available up to that point: $q_\phi(z_t \mid X, x_{\leq t}, z_{<t}) = \mathcal{N}\left(\mu_t, \text{diag}(\sigma_t^2)\right)$. By explicitly conditioning the latent distribution on the history of previous registers $z_{<t}$, the model is forced to encode only the *residual* information introduced by the new block. The decoder then reconstructs the original sequence from this ordered stream of registers. The model is optimized using the standard evidence lower bound (ELBO): $\mathcal{L}_{\text{CR-VAE}} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}}$.

Once trained, the CR-VAE is frozen. It serves as an offline tokenizer for reasoning, converting raw text demonstrations into ordered, uncertainty-aware latent trajectories that serve as the target data $Z$ for training ATF. The CR-VAE is not used to choose final answers during benchmark evaluation; final correctness is measured from the text produced by the autoregressive decoder after latent thinking terminates. The frozen decoder instead provides an interpretability lens: we can decode any generated continuous thought $\hat{z}_t$ back into natural language to understand the internal reasoning process.

## 4 Experiments

**Baselines.** We compare discrete CoT (Wei et al., 2022), matched MSE regression, and ATF under the same Qwen3 backbones (Table 1). CoT uses unrestricted or capped intermediate text budgets; the zero-thought ATF ablation tests the contribution of latent thinking. MSE shares CR-VAE targets, adaptive stopping, and the final text decoder with ATF, replacing only the flow head with deterministic regression to isolate generative next-thought modeling. We directly reuse baseline scores reported in LaDiR (Kang et al., 2025) and CoLaR (Tan et al., 2025), using their respective LLaMA-3.1-8B and

Table 1: **ATF improves compact reasoning over discrete and regression baselines.** Accuracy (%) and generation lengths on Qwen3 backbones, with saved answers rescored using a common grader. Budget limits intermediate text tokens for discrete CoT and latent steps for continuous methods. $N_{\text{think}}$ and $N_{\text{ans}}$ report the corresponding realized thinking length and final-answer tokens. ATF-RL applies answer-policy refinement. Bold and underline mark best and second-best accuracy within each model size, excluding unbounded discrete CoT.

| Method | Budget | GSM8K | | | MATH-500 | | | OlympiadBench | | | AMC 12 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Acc.** | $N_{\text{think}}$ | $N_{\text{ans}}$ | **Acc.** | $N_{\text{think}}$ | $N_{\text{ans}}$ | **Acc.** | $N_{\text{think}}$ | $N_{\text{ans}}$ | **Acc.** | $N_{\text{think}}$ | $N_{\text{ans}}$ |
| ***Qwen3-1.7B*** | | | | | | | | | | | | | |
| *Discrete CoT* | $\infty$ | 79.15 | 1252 | 242 | 64.40 | 2582 | 324 | 27.37 | 4132 | 166 | 33.73 | 4127 | 209 |
| *Discrete CoT* | $\leq 1000$ | 67.17 | 599 | 535 | 60.60 | 879 | 627 | 28.06 | 986 | 469 | 26.51 | 984 | 711 |
| *Discrete CoT* | $\leq 500$ | 76.65 | 438 | 337 | 63.20 | 494 | 556 | 22.20 | 501 | 472 | 25.30 | 501 | 646 |
| *MSE regression* | $\leq 32$ | **82.41** | 4.0 | 300 | 62.40 | 4.1 | 550 | 27.88 | 4.6 | 604 | <u>37.35</u> | 4.6 | 669 |
| *ATF* (Ours) | 0 | 28.05 | 0 | 698 | 30.00 | 0 | 733 | 14.63 | 0 | 747 | 15.66 | 0 | 740 |
| *ATF* (Ours) | $\leq 32$ | 79.76 | 4.2 | 299 | <u>66.80</u> | 4.9 | 549 | <u>30.46</u> | 6.5 | 607 | **42.17** | 6.4 | 699 |
| *ATF-RL* (Ours) | $\leq 32$ | <u>81.73</u> | 4.0 | 289 | **69.00** | 4.9 | 493 | **32.19** | 6.1 | 631 | **42.17** | 5.7 | 654 |
| ***Qwen3-4B*** | | | | | | | | | | | | | |
| *Discrete CoT* | $\infty$ | 91.21 | 804 | 267 | 79.40 | 2020 | 417 | 44.58 | 3721 | 243 | 56.63 | 3469 | 397 |
| *Discrete CoT* | $\leq 1000$ | 47.84 | 551 | 721 | 49.20 | 865 | 391 | 20.14 | 988 | 101 | 25.30 | 983 | 140 |
| *Discrete CoT* | $\leq 500$ | 72.18 | 430 | 264 | 36.00 | 492 | 125 | 6.88 | 501 | 20 | 13.25 | 501 | 30 |
| *MSE regression* | $\leq 32$ | 90.98 | 4.0 | 295 | 75.40 | 15.4 | 546 | 39.59 | 8.6 | 580 | 44.58 | 4.0 | 714 |
| *ATF* (Ours) | 0 | 48.07 | 0 | 504 | 25.60 | 0 | 685 | 15.49 | 0 | 662 | 19.28 | 0 | 629 |
| *ATF* (Ours) | $\leq 32$ | <u>91.28</u> | 4.1 | 291 | <u>78.40</u> | 6.8 | 533 | <u>42.51</u> | 8.1 | 601 | <u>56.63</u> | 8.0 | 728 |
| *ATF-RL* (Ours) | $\leq 32$ | **92.04** | 5.3 | 283 | **78.60** | 12.9 | 493 | **44.41** | 21.4 | 622 | **57.83** | 20.8 | 666 |

LLaMA-3.2-1B-Instruct backbones and benchmark suites. These comparisons retain source-specific training procedures, including CoLaR's GSM8K-Aug/MATH recipe; Appendix B details objectives, score sources, and protocols.

**Datasets.** Qwen3 training mixes OpenMathReasoning (Moshkov et al., 2025) OpenR1Math (Hugging Face, 2025), and DeepScaleR (Luo et al., 2025). Table 1 evaluates GSM8K (Cobbe et al., 2021) (1,319 questions), MATH-500 (Hendrycks et al., 2021) (500), the canonical single-answer OlympiadBench (He et al., 2024) split (581), and AMC 12 (Dekoninck et al., 2026) (83), spanning grade-school and competition mathematics. We additionally use AIME 24/25 to study guidance. LLaMA training uses DART-MATH-derived data for the 8B model and GSM8K-Aug/MATH for the 1B model; evaluation splits are detailed in Appendix B.

**Evaluation metrics.** We report final-answer accuracy, mean thinking length $N_{\text{think}}$ (text tokens for CoT; latent steps for continuous methods), and mean answer length $N_{\text{ans}}$ in text tokens. Realized lengths differ from budgets and do not directly measure compute cost. Pass@1 and Pass@100 assess single-sample success and solution coverage with ATF; Pass@$K$ measures the fraction of problems with at least one correct answer among $K$ samples. Trajectory diversity measures mean pairwise cosine distance. For Table 1, we rescore saved answers with a common Math-Verify v0.9.0 grader, retaining original generation settings; accuracy and lengths use the same outputs.

**Implementation details.** We use `Qwen3-1.7B-Base` and `Qwen3-4B-Base` (Yang et al., 2025), filtering training examples whose thinking trace exceeds 4096 tokens. CR-VAE uses chunk size 512 and a compression curriculum from 2:1 to 128:1, ending with 4 registers per chunk; it is trained for one epoch at learning rate $3 \times 10^{-5}$. We then freeze CR-VAE and train ATF for one epoch on its latent trajectories. Unless otherwise stated, continuous methods use adaptive stopping with at most 32 latent steps. Qwen3 ATF-RL rows use answer-policy refinement, separately from the joint latent/text RL model in the LLaMA-3.1-8B comparison.

### 4.1 Main Results

**ATF improves the compact-reasoning frontier.** Table 1 shows that ATF improves over MSE on the harder 1.7B settings and all 4B settings while using roughly 4–8 latent steps. Additional Pass@16 and self-consistency results are reported in Appendix F.

Table 2: **Multimodal next-thought prediction unlocks strong test-time scaling.** ATF is competitive with prior methods at Pass@1, but the gap widens sharply as more latent trajectories are sampled. At Pass@100, ATF outperforms every baseline, reaching 83.2 on MATH-500, 97.6 on GSM8K, and 52.4 on OlympiadBench—a 34.6 point gain over the strongest prior result on OlympiadBench. The result suggests that ATF does not merely improve a single reasoning trajectory; **its multimodal next-thought distribution reaches diverse solution modes that become increasingly valuable under test-time sampling.** With ATF (Greedy), the final answer is decoded deterministically, isolating continuous thinking as the source of this diversity. All methods use LLaMA-3.1-8B; baseline scores are reported by LaDiR (Kang et al., 2025). Bold marks the best result for each metric and benchmark.

| Family | Method | MATH-500 | | GSM8K | | OlympiadBench | |
|---|---|---|---|---|---|---|---|
| | | Pass@1 | Pass@100 | Pass@1 | Pass@100 | Pass@1 | Pass@100 |
| *Hidden states* | *iCoT* (Deng et al., 2023) | 35.2 | 37.9 | 61.8 | 63.9 | 4.3 | 7.1 |
| | *Coconut* (Hao et al., 2024) | 37.3 | 39.3 | 68.3 | 74.3 | 5.8 | 6.3 |
| | *CODI* (Shen et al., 2025) | 38.5 | 45.1 | 76.3 | 81.7 | 7.6 | 14.8 |
| | *TaH+* (Fu et al., 2025) | 46.1 | 49.4 | **85.9** | 89.7 | 12.2 | 15.2 |
| *Token-based* | *Disc. Latent* (Su et al., 2025) | 43.2 | 47.3 | 83.9 | 88.6 | **13.3** | 17.8 |
| | *Soft Think* (Zhang et al., 2025) | 44.3 | 46.7 | 83.7 | 86.6 | 10.4 | 13.1 |
| *Block diffusion* | *LaDiR w/o Stage 2* | 30.7 | 35.8 | 57.8 | 62.6 | 5.9 | 10.5 |
| | *LaDiR* (Kang et al., 2025) | **46.2** | 63.7 | 84.8 | 93.7 | 12.9 | 15.3 |
| *Next-thought flow* | ***ATF*** (Greedy) | 43.2 | 71.6 | 84.9 | 93.1 | 13.0 | 38.9 |
| | ***ATF*** | 42.2 | **83.2** | 84.6 | **97.6** | 12.7 | **52.4** |

Table 3: **Comparison with CoLaR.** CoLaR (Tan et al., 2025) and ATF use LLaMA-3.2-1B-Instruct under the GSM8K-Aug/MATH recipe; mean $\pm$ standard deviation over five runs. We directly reuse CoLaR's number and setting, which is the cloest related work that use a Gaussian distribution (single mode distribution) to model the next prediction. For the ATF here we use exactly CoLaR's setting for fair comparison. Bold marks the better mean within each dataset.

| Distribution | Method | GSM8K | | MATH | |
|---|---|---|---|---|---|
| | | Acc. (%) ↑ | Latent steps ↓ | Acc. (%) ↑ | Latent steps ↓ |
| *Single Gaussian* | *CoLaR* | $26.8 \pm 0.17$ | $\mathbf{5.57} \pm 0.02$ | $7.08 \pm 0.07$ | $16.1 \pm 0.14$ |
| *Flow-based* | ***ATF*** | $\mathbf{30.17} \pm 1.26$ | $6.43 \pm 0.01$ | $\mathbf{8.59} \pm 0.05$ | $\mathbf{7.17} \pm 0.03$ |

**Broader solution coverage through latent exploration.** Table 2 compares LLaMA-3.1-8B methods using baseline scores reported in LaDiR (Kang et al., 2025); ATF uses DART-MATH-derived data and RL refinement. With $T_{\text{text}} = 0.7$, ATF achieves the highest Pass@100 among the listed methods on all three benchmarks despite comparable or lower Pass@1 than LaDiR. On MATH-500, ATF reaches 42.2% Pass@1 versus LaDiR's 46.2%, but 83.2% Pass@100 versus 63.7%. On OlympiadBench, nearly equal Pass@1 (12.7% versus 12.9%) accompanies substantially higher Pass@100 (52.4% versus 15.3%). With greedy answer decoding, ATF already exceeds LaDiR's Pass@100 on both benchmarks (71.6% and 38.9%); GSM8K coverage is similar (93.1% versus 93.7%). In this greedy-text setting, variation arises from latent initial noise, showing that sampling latent trajectories alone can produce useful alternative continuations. These results support useful latent exploration, complementing the matched MSE comparison. Cross-paper protocol alignment and additional temperature results appear in Appendix B and Table 7.

**Comparison with a Gaussian latent baseline.** CoLaR (Tan et al., 2025) already models a stochastic next latent with a Gaussian head, making it a stronger comparison than deterministic regression alone. Under the shared LLaMA-3.2-1B-Instruct backbone, GSM8K-Aug/MATH training data, and evaluation protocol in Table 3, ATF improves accuracy by 3.37 percentage points on GSM8K and 1.51 on MATH. On MATH, it also uses 7.17 rather than 16.1 latent steps on average; on GSM8K, it uses slightly more steps (6.43 versus 5.57). This comparison extends the evidence for generative continuous reasoning beyond the deterministic MSE baseline.

**Additional continuous thoughts improve reasoning within a compact budget.** Figure 3 varies the stopping margin on MATH-500. For this sweep, stopping is checked every four latent steps: the stop token must be the top prediction with a top-1/top-2 logit gap above $m$, subject to a cap of 32 steps. Over the displayed range, ATF improves from $76.6\%$ to $77.4\%$ as its mean latent length increases from $4.01$ to $7.86$ steps; ATF-RL follows a similar trend. At $m = 8$, MSE uses $15.42$ steps and reaches $75.4\%$, whereas ATF attains higher accuracy with about half as many steps. This comparison supports the utility of the learned continuous thoughts beyond the default stopping point.

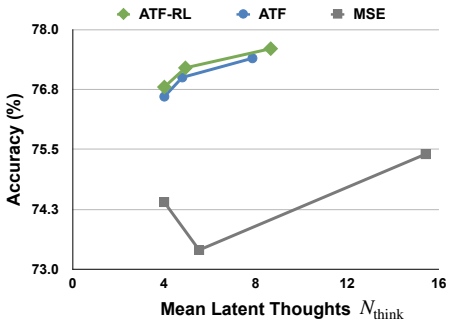



Figure 3: **ATF benefits from more latent steps.** MATH-500 accuracy at 4B for stopping margins $m \in \{0, 4, 8\}$, with $\gamma = 1$ and greedy answer decoding. The horizontal axis reports realized mean latent length. ATF achieves a stronger accuracy–latent-length trade-off than the matched regression baseline in this sweep.

**CFG controls exploration within the learned distribution.** Table 4 examines this coverage–accuracy trade-off by varying CFG for Qwen3-4B-Base ATF on AIME (Dekoninck et al., 2026). Diversity uses normalized latent trajectories to measure directional spread rather than latent magnitude. As $\gamma$ increases, Pass@1 improves initially while Pass@16 and Diversity fall.

Table 4: **CFG provides a built-in control over latent mode exploration.** For the 4B ATF on AIME, the guidance scale $\gamma$ trades off single-sample accuracy against sampled solution coverage. Larger $\gamma$ concentrates generation around high-confidence continuations, favoring exploitation, whereas smaller $\gamma$ preserves more diverse latent modes for exploration. Bold marks the best result for each metric.

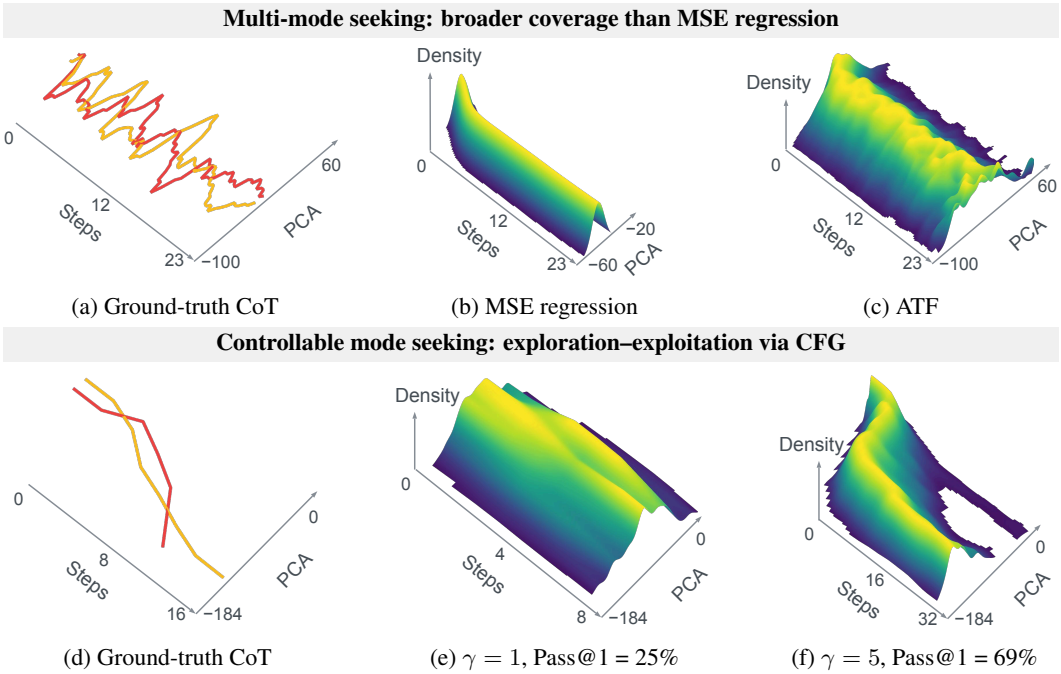
| $\gamma$ | AIME 24 | | | AIME 25 | | |
|---|---|---|---|---|---|---|
| | **Pass@1** | **Pass@16** | **Diversity** | **Pass@1** | **Pass@16** | **Diversity** |
| 1.0 | 8.54 | **30.00** | **0.7503** | 6.25 | **23.33** | **0.7518** |
| 2.5 | 9.38 | 26.67 | 0.6357 | **7.08** | 20.00 | 0.6407 |
| 5.0 | **9.58** | 26.67 | 0.4147 | 6.88 | 20.00 | 0.4499 |

**Generative next thoughts preserve multiple modes.** Figure 4(a–c) illustrates how the learning objective shapes next-thought geometry: MSE regression concentrates near a single continuation, while ATF retains multiple modes. Table 5 gives the decoded continuations. These qualitative examples complement the sampled solution-coverage results.

Table 5: **Decoded latent thoughts reveal regression collapse versus generative next-thought diversity.** MSE regression produces similar premise-restatement fragments, while ATF realizes distinct continuations from the same autoregressive state. Blue highlights shared premise restatements; orange highlights contrasting continuations.

| **Method** | **Trace** | **Decoded excerpt** |
|---|---|---|
| MSE regression | A | "Okay, so I need to find all natural numbers $N$ such that **$N(N-101)$ is a perfect square** ... First, I remember that **$N(N-101)$ needs to be a square** ..." |
| MSE regression | B | "If $N$ is a natural number, then **$N(N-101)$ must equal the square** ... So, **$N(N-101)$ must be a perfect square** ..." |
| ATF | A | "**Expressed in the form $N(N-101) = x^2$** ... $N = x^2$ is a perfect square ..." |
| ATF | B | "**$N$ is the perimeter of a square** ... identify which numbers from $1 \times 1$ to 69 are squares." |

**Guidance concentrates decoded continuations.** Figure 4(d–f) illustrates the shift toward exploitation. In Table 6, low-guidance samples follow different row constructions, while both high-guidance samples reproduce the same initial row sequence.




(a) Ground-truth CoT (b) MSE regression (c) ATF

(d) Ground-truth CoT (e) $\gamma = 1$, Pass@1 = 25% (f) $\gamma = 5$, Pass@1 = 69%

Figure 4: PCA feature visualizations of **ATF's latent reasoning modes.** We sample 500 times to estimate the distribution for our diffusion model, and plot their density map. Compared with MSE regression, ATF preserves more diverse next-thought continuations (top). Our diffusion head design allows us to apply the classifier-free guidance then adjusts the concentration of this learned distribution, trading broader exploration for more focused exploitation (bottom). The rows illustrate different questions; Pass@1 refers to the bottom-row example. Tables 5 and 6 give decoded excerpts.

Table 6: **Stronger CFG concentrates decoded continuations around the reference reasoning pattern.** Low guidance produces more varied, sometimes off-mode continuations, while high guidance favors similar row-pattern reconstructions. Orange marks contrasting low-guidance fragments; blue marks the repeated high-guidance row sequence.

| CFG $\gamma$ | Trace | Decoded excerpt |
|---|---|---|
| 1 | A | "Okay, so the problem is about finding the units digit ... If $n = 7$, then **row 7 is (1, 5, 6, 5, 1)** ..." |
| 1 | B | "Okay, let's see ... triangular array with 15th row ... additional entries are constructed by **reflecting a row with TicTacToe** ..." |
| 5 | A | "Alright, let's try to figure out this problem ... The first rows are **1, 1 1, 1 3 1, 1 5 5 1, 1 7 11 7 1** ..." |
| 5 | B | "Alright, let's try to figure out this problem ... The rows are **1, 1 1, 1 3 1, 1 5 5 1, 1 7 11 7 1** ..." |

**Inference cost of continuous thoughts.** In a separate Qwen3-4B profiling study on one RTX PRO 6000 (batch size 1), ATF uses one backbone step and 24 lightweight flow-head steps per thought. Length-based estimates reduce MATH-500 end-to-end time from 31.89 s for text CoT to 7.19 s for ATF (4.44×). Direct timing measurements agreed within 3.5% of the estimates. Appendix E reports all four benchmarks and a matched per-thought compute comparison.

## 5 Conclusion

We introduced ATF, which treats the next continuous thought as a multimodal distribution rather than a single prediction. By combining a causal reasoning backbone with a lightweight generative head, ATF preserves autoregressive reasoning while allowing multiple plausible continuations to remain possible. Across challenging reasoning tasks, this design improves accuracy, supports test-time scaling, and yields substantially greater solution coverage through diverse latent trajectories. Our results suggest that modeling uncertainty over what to think next is a key ingredient for effective continuous reasoning.

### AI USE STATEMENT

Generative AI assistance was used to aid and polish writing, revise the manuscript's text, organization, and presentation, and support literature retrieval and discovery. LLM assistance was also used for code generation and experiment execution. It was also used to draft the reinforcement-learning method description from the authors' slides and implementation, adapt the manuscript's LaTeX submission format, and transcribe and format previously reported experimental results as tables. The scientific ideas, experimental design, and interpretation of the results originated with the authors.

## REFERENCES

Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. 2026. URL https://arxiv.org/abs/2605.00674.

Chaorui Deng, Deyao Zhu, Kunchang Li, Shi Guang, and Haoqi Fan. Causal diffusion transformers for generative modeling. *arXiv preprint arXiv:2412.12095*, 2024.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023. URL https://arxiv.org/abs/2311.01460.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*, 2023.

Tianyu Fu, Yichen You, Zekai Chen, Guohao Dai, Huazhong Yang, and Yu Wang. Think-at-hard: Selective latent iterations to improve reasoning language models. *arXiv preprint arXiv:2511.08577*, 2025. URL https://arxiv.org/abs/2511.08577.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Haoqiang Kang, Yizhe Zhang, Nikki Lijing Kuang, Nicklas Majamaki, Navdeep Jaitly, Yi-An Ma, and Lianhui Qin. Ladir: Latent diffusion enhances llms for text reasoning. *arXiv preprint arXiv:2510.04573*, 2025. URL https://arxiv.org/abs/2510.04573v6.

Deqian Kong, Minglu Zhao, Dehong Xu, Bo Pang, Shu Wang, Edouardo Honig, Zhangzhang Si, Chuan Li, Jianwen Xie, Sirui Xie, et al. Latent thought models with variational bayes inference-time computation. *arXiv preprint arXiv:2502.01567*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Jiayu Liu, Zhenya Huang, Anya Sims, Enhong Chen, Yee Whye Teh, and Ning Miao. Marcos: Deep thinking by markov chain of continuous thoughts. 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. *arXiv preprint arXiv:2505.05470*, 2025b. URL https://arxiv.org/abs/2505.05470.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

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. *arXiv preprint arXiv:2512.10941*, 2025.

Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. Continuous autoregressive language models. *arXiv preprint arXiv:2510.27688*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025. URL https://arxiv.org/abs/2502.21074.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. Token assorted: Mixing latent and text tokens for improved language model reasoning. *arXiv preprint arXiv:2502.03275*, 2025. URL https://arxiv.org/abs/2502.03275.

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

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *arXiv preprint arXiv:2505.16552*, 2025.

NextStep Team, Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, et al. Nextstep-1: Toward autoregressive image generation with continuous tokens at scale. *arXiv preprint arXiv:2508.10711*, 2025.

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

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, et al. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking. In *International Conference on Learning Representations*, volume 2026, pp. 4533–4548, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Runyang You, Yongqi Li, Meng Liu, Wenjie Wang, Liqiang Nie, and Wenjie Li. Parallel test-time scaling for latent reasoning models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44703–44717, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2069. URL https://aclanthology.org/2026.acl-long.2069/.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Yuyan Zhou, Jiarui Yu, Hande Dong, Zhezheng Hao, Hong Wang, Jianqing Zhang, and Qiang Lin. Lepo: Latent reasoning policy optimization for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14416–14427, 2026.

Xiandong Zou, Jing Huang, Jianshu Li, and Pan Zhou. Latent thought flow: Efficient latent reasoning in large language models. *arXiv preprint arXiv:2606.16222*, 2026.

## A Implementation Details for RL Refinement

The optional RL extension in Section 3.2 refines the existing multimodal next-thought model using outcome rewards. This appendix specifies the transition scores and optimization details behind the compact main-text objective. The CR-VAE remains fixed, and $\theta$ collects the trainable backbone and flow-head parameters.

For $N, M > 1$, the centered advantages in the main text are standardized by the sample standard deviation across latent-mean rewards and within-latent answer rewards, respectively, with a small stabilizer and symmetric clipping. When $N = 1$, only the within-trajectory text advantage is used. A flat $M = 1$ variant uses the ordinary group-relative advantage across complete rollouts for both terms.

**Scoring stochastic flow actions.** A flow-generated thought has no categorical next-token probability. Instead of evaluating its intractable marginal density, we record stochastic denoising transitions within each autoregressive thought. Let $u_{n,t,k}$ denote the intermediate flow state for thought $t$ and denoising step $k$. At selected steps $k \in \mathcal{W}$, noise injection gives a Gaussian transition

$$\begin{aligned} u_{n,t,k+1} &= \mu_\theta(u_{n,t,k}, \tau_k, h_{n,t}) + \sigma_k \xi_{n,t,k}, \quad \xi_{n,t,k} \sim \mathcal{N}(0, I), \\ q_\theta(u_{n,t,k+1} \mid u_{n,t,k}, h_{n,t}) &= \mathcal{N}(\mu_\theta(u_{n,t,k}, \tau_k, h_{n,t}), \sigma_k^2 I). \end{aligned} \tag{10}$$

The mean is obtained from the flow vector field using the chosen stochastic transition rule; $\sigma_k$ is fixed by the sampler. Remaining denoising steps use deterministic integration. Once a thought is completed, it is appended to the causal history before generating the next thought. During optimization, recorded states are held fixed and rescored using the current backbone condition and flow head; we do not backpropagate through the rollout sampler.

For latent dimension $d$, a dimension-averaged Gaussian log-density, up to terms independent of the policy parameters, is

$$\ell^{\mathrm{Gauss}}_{n,t,k}(\theta) = -\frac{\|u_{n,t,k+1} - \mu_\theta(u_{n,t,k}, \tau_k, h_{n,t})\|_2^2}{2d\sigma_k^2}. \tag{11}$$

Our implementation supports this variance-scaled score and a raw squared-error score $s_{n,t,k} = -\|u_{n,t,k+1} - \mu_\theta\|_2^2/d$. The latter is a policy-score surrogate: omitting $1/(2\sigma_k^2)$ changes the relative weighting of denoising steps, so it is not an exact log-density. Neither score is the marginal log-probability of the completed thought.

**Policy update.** For a stored action with current score $s$ and fixed behavior-policy score $s_{\mathrm{old}}$, define

$$\begin{aligned} \rho &= \exp(s - s_{\mathrm{old}}), \\ \mathcal{C}(\rho, A) &= \min(\rho A, \mathrm{clip}(\rho, 1-\varepsilon, 1+\varepsilon)A)\,. \end{aligned} \tag{12}$$

For text actions, $s$ is the answer-token log-probability at the training temperature. For latent actions it is the selected transition score above; with raw squared error, $\rho$ is a surrogate score ratio. With the Gaussian score, dimension averaging makes it a dimension-normalized likelihood ratio. We maximize

$$\mathcal{J}_{\mathrm{RL}} = \lambda_z \big\langle \mathcal{C}(\rho^z, A^{\mathrm{lat}}) \big\rangle_z + \lambda_y \big\langle \mathcal{C}(\rho^y, A^{\mathrm{text}}) \big\rangle_y\,, \tag{13}$$

where the two averages mask padding and normalize over valid stochastic transitions and answer tokens, respectively. Separate clipping thresholds may be used for the two score scales. Without clipping, the corresponding update is the advantage-weighted score-function objective. The optional stop loss is an advantage-weighted binary surrogate; stopping in the rollout remains a deterministic margin decision and is not assigned an importance ratio.

## B Baseline Definitions and Comparison Protocols

**Matched regression baseline.** To isolate the effect of generative next-thought modeling, we compare against a baseline with the same backbone, CR-VAE targets, hybrid sequence format, and stopping rule. The only change is the latent prediction head: instead of a flow model, the baseline predicts a deterministic vector $\hat{z}_t = g_\omega(h_t)$ and minimizes

$$\mathcal{L}_{\mathrm{MSE}} = \mathbb{E}_t \left[ \|g_\omega(h_t) - z_t\|_2^2 \right]. \tag{14}$$

At inference, $\hat{z}_t$ is fed back as the next continuous thought. This baseline is therefore a point-estimate version of ATF rather than a separately engineered system.

**LLaMA-3.1-8B comparison.** The non-ATF entries in Table 2 are reported in Table 1 of LaDiR (Kang et al., 2025), including both LaDiR training stages. MATH-500 denotes the 500-question subset reported as MATH in LaDiR. Baselines retain their reported training procedures; this is a cross-paper comparison, not a matched ablation of the prediction head or RL objective.

Table 7 reports both Pass@1 and Pass@100 for the methods in the main-text comparison, including an additional text-temperature setting. ATF has the highest Pass@100 in this roster, while the best Pass@1 scores belong to LaDiR on MATH-500, TaH+ on GSM8K, and Discrete Latent on OlympiadBench.

ATF uses the LLaMA-3.1-8B backbone, DART-MATH-derived training data, and joint latent/text RL. The main-table greedy and text-temperature-0.7 settings use 100 independently sampled latent trajectories per question, with one answer per trajectory. Evaluation uses ODE latent sampling and DART grading. OlympiadBench uses the 674-question split with multi-answer cases retained. The temperature-1 row provides an additional evaluation setting. LaDiR reports answer sampling temperature 0.7.

Table 7: **Complete single-sample and multi-sample comparison.** All methods use the LLaMA-3.1-8B backbone; ATF is trained on DART-MATH-derived data and RL-refined. Each cell reports Pass@1 / Pass@100 (%). Baseline scores are reported by LaDiR. Bold marks the best value for each metric.

| Method | MATH-500 | GSM8K | OlympiadBench |
|---|---|---|---|
| iCoT | 35.2 / 37.9 | 61.8 / 63.9 | 4.3 / 7.1 |
| Coconut | 37.3 / 39.3 | 68.3 / 74.3 | 5.8 / 6.3 |
| CODI | 38.5 / 45.1 | 76.3 / 81.7 | 7.6 / 14.8 |
| Discrete Latent | 43.2 / 47.3 | 83.9 / 88.6 | **13.3** / 17.8 |
| Soft Think | 44.3 / 46.7 | 83.7 / 86.6 | 10.4 / 13.1 |
| TaH+ | 46.1 / 49.4 | **85.9** / 89.7 | 12.2 / 15.2 |
| LaDiR w/o Stage 2 | 30.7 / 35.8 | 57.8 / 62.6 | 5.9 / 10.5 |
| LaDiR | **46.2** / 63.7 | 84.8 / 93.7 | 12.9 / 15.3 |
| ATF ($T_{\text{text}} = 0$) | 43.2 / 71.6 | 84.9 / 93.1 | 13.0 / 38.9 |
| ATF ($T_{\text{text}} = 0.7$) | 42.2 / 83.2 | 84.6 / 97.6 | 12.7 / 52.4 |
| ATF ($T_{\text{text}} = 1$) | 41.4 / **84.8** | 84.5 / **98.3** | 12.6 / **52.5** |

**LLaMA-3.2-1B-Instruct comparison.** Table 3 compares ATF and CoLaR using a shared LLaMA-3.2-1B-Instruct backbone, GSM8K-Aug/MATH training data, and evaluation protocol. We report mean and standard deviation over five runs. The benchmark name MATH follows that experiment's reporting and is not equated with the MATH-500 split above. Scores from the two tables are therefore compared only within their respective experimental settings.

## C CR-VAE Architecture

Figure 5 shows the CR-VAE used to obtain continuous thought targets. As described in Section 3.4, causal registers summarize the reasoning prefix into stochastic latent variables. After reconstruction training, the encoder is frozen and used offline to produce the latent trajectories for ATF supervision; the decoder supports qualitative inspection of generated thoughts.

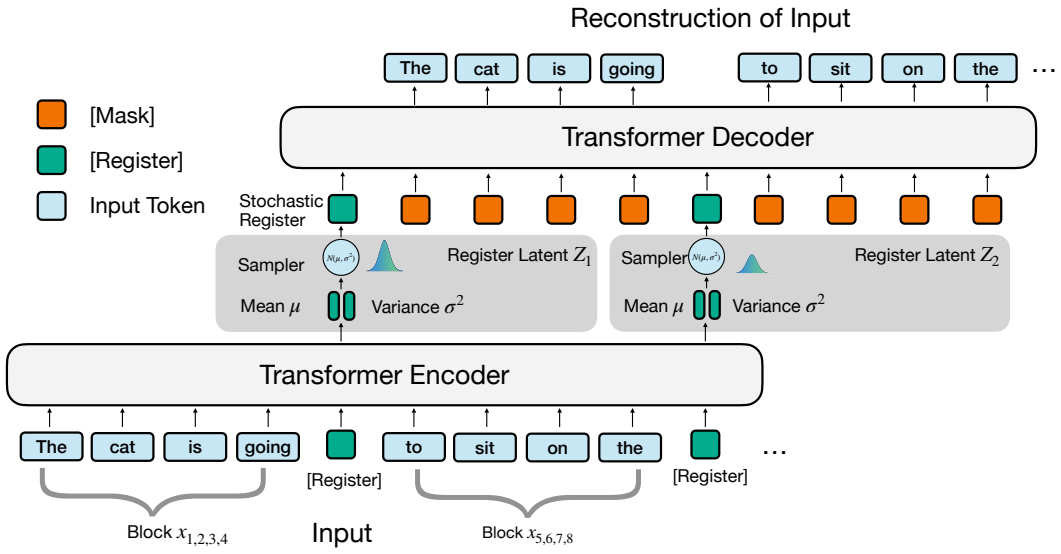



Figure 5: **Causal-Register VAE architecture.** Register tokens inserted after reasoning blocks provide the encoder states used to parameterize Gaussian posteriors. Sampled latent registers condition the transformer decoder, which reconstructs the input text using mask placeholders. Causal encoding preserves the ordering of the reasoning prefix; the resulting latent registers serve as continuous thought targets for ATF.

# D Ablation Studies

## D.1 CR-VAE Compression

**How compact should the latent thought space be?** On Qwen3-1.7B-Base, we ablate the CR-VAE compression ratio to test whether ATF requires token-faithful reconstruction, or only a compact latent space that preserves the reasoning information needed for downstream answers. Table 8 shows a clear separation between reconstruction fidelity and reasoning utility: stronger compression substantially lowers autoregressive token reconstruction, yet downstream accuracy remains stable or improves. This suggests that discrete chain-of-thought contains substantial token-level redundancy for reasoning; ATF can therefore reason efficiently in a compact latent space, avoiding the accumulation of errors that arises when long reasoning traces must be reproduced token by token.

Table 8: **More compact CR-VAE latents can improve downstream reasoning despite lower token reconstruction at 1.7B.** Using the Qwen3-1.7B-Base backbone, AR reconstruction accuracy measures in-distribution reconstruction of discrete thoughts, while GSM8K and OlympiadBench report CFG = 1 reasoning accuracy. Higher VAE compression discards token-level surface form, yet downstream accuracy remains strong, indicating that compact latents preserve the reasoning structure needed by ATF.

| VAE Comp. Ratio | AR Recon Acc. | GSM8K | Olympiad |
|---|---|---|---|
| 16:1 | **99.1** | 76.80 | 20.31 |
| 64:1 | 75.2 | 78.09 | **22.03** |
| 128:1 | 25.7 | **80.52** | 21.17 |

# E Inference Efficiency

**End-to-end cost.** We profile reasoning and answer decoding separately on one NVIDIA RTX PRO 6000 at batch size 1. Table 9 combines profiled operator costs with actual generated lengths under each method's evaluation configuration; direct timing measurements agreed within $3.5\%$ of these estimates. This is a separate profiling study, rather than a remeasurement of the evaluation outputs selected for Table 1. ATF uses 24 flow-head denoising steps per latent thought. Answer decoding accounts for most of its estimated runtime, so shorter latent traces do not by themselves determine end-to-end speed.

Table 9: **End-to-end latency estimates for Qwen3-4B.** Reasoning, answer, and total times on one RTX PRO 6000, batch size 1. ATF uses 24 flow-head denoising steps per thought.

| Benchmark | Method | Think (s) | Answer (s) | Total (s) | Speedup |
|---|---|---|---|---|---|
| GSM8K | Text CoT | 10.53 | 3.49 | 14.02 | $1.00\times$ |
| | ATF | 0.22 | 3.80 | 4.02 | $3.49\times$ |
| MATH-500 | Text CoT | 26.43 | 5.46 | 31.89 | $1.00\times$ |
| | ATF | 0.22 | 6.97 | 7.19 | $4.44\times$ |
| OlympiadBench | Text CoT | 48.69 | 3.18 | 51.86 | $1.00\times$ |
| | ATF | 0.26 | 7.89 | 8.15 | $6.40\times$ |
| AMC 12 | Text CoT | 45.60 | 5.21 | 50.81 | $1.00\times$ |
| | ATF | 0.41 | 9.56 | 9.97 | $5.09\times$ |

**Cost of generating one thought.** Table 10 isolates per-thought generation under the same LLaMA-3.1-8B backbone, $5.5$:1 compression, hardware, and batch size 1. ATF runs the backbone once per thought and performs its 24 denoising steps in the flow head, whereas LaDiR runs the backbone at each denoising step. This comparison measures the cost of the next-thought interface, independently of the number of thoughts or answer tokens generated for a benchmark.

Table 10: **Per-thought compute under matched conditions.** LLaMA-3.1-8B backbone, 5.5:1 compression, identical hardware, and batch size 1. The step counts denote flow-head evaluations for ATF and full-backbone evaluations for LaDiR.

| Method | GFLOPs / thought | Latency / thought (ms) |
|---|---|---|
| ATF | **25.2** | **35.8** |
| LaDiR, 10 steps | 139.6 | 41.4 |
| LaDiR, 50 steps | 697.9 | 207.0 |

## F Additional Sampling Analysis

**ATF provides broader sampled answer coverage than MSE regression.** On both MATH-500 and AMC 12, Table 11 shows higher Pass@16 for ATF. On MATH-500, ATF also improves self-consistency accuracy (73.00 versus 69.00), showing that broader exploration can increase both solution coverage and majority-vote success. On AMC 12, MSE has higher self-consistency despite lower Pass@16, illustrating that answer concentration and solution coverage need not improve together. The consistent coverage gain supports the usefulness of sampled latent alternatives. Figure 4(a–c) and Table 5 provide complementary qualitative illustrations of different continuations.

Table 11: **Multi-sample rollouts show broader answer coverage for ATF at 4B.** Using the Qwen3-4B-Base backbone, we evaluate $K = 16$ adaptive-stopping rollouts at temperature 0.1 on **MATH-500** and **AMC 12**. Pass@16 measures whether any sampled answer is correct, while self-consistency accuracy uses majority voting across the same samples.

| Method | MATH-500 | | AMC 12 | |
|---|---|---|---|---|
| | Pass@16 | Self-consistency Acc. | Pass@16 | Self-consistency Acc. |
| *MSE regression* | 77.60 | 69.00 | 72.29 | **62.65** |
| *ATF* **(Ours)** | **79.80** | **73.00** | **74.70** | 57.83 |