Latent Imagination Thinking: Beyond recursive models for reasoning

Karim Farid[*,1] **Jelena Bratulić**[1] **Sudhanshu Mittal**[1]

Cordelia Schmid[2] **Thomas Brox**[1]
[1]University of Freiburg [2]Inria, École Normale Supérieure, CNRS

Abstract

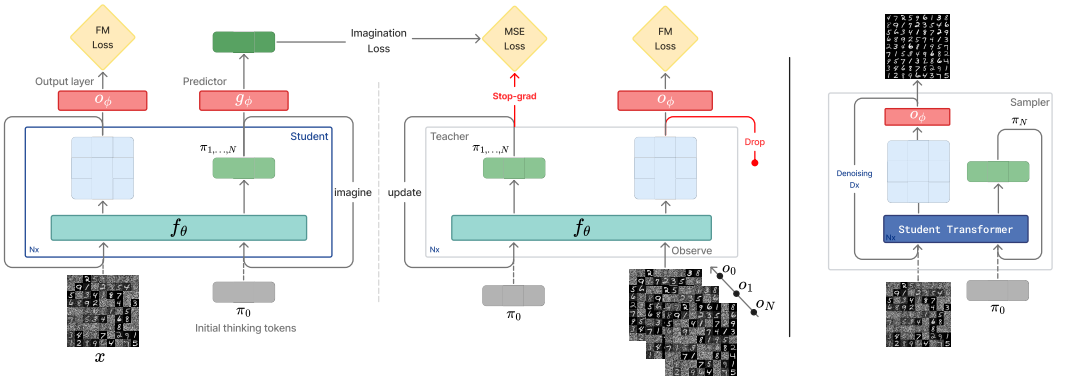
Reasoning capabilities of generative AI have recently improved by making models *think* through recursion: LLMs re-consume their own tokens (Chain-of-Thought), and diffusion models iteratively refine pixels or reconstruction-trained latents. While practical, this common design reduces reasoning to an *observational* space and conflates two roles: *latent reasoning* (discovering a task-appropriate internal language and maintaining a belief over solutions) and *modeling the observational space*. We introduce **Latent Imagination Thinking (LIT)**, a teacher–student learning paradigm that treats tokens and pixels as partial observations rather than the language of thought. To provide learning guidance for latent states, a posterior model (teacher) refines its belief using additional task-relevant observations, and a prior model (student) is trained to imagine these refinements via an imagination loss with stop-gradient targets. This turns recurrence into a latent belief update rather than repeated prediction in the observation space. We evaluate LIT on hard Sudoku puzzles in language and visual (MNIST) spaces. Increasing the number of thinking steps improves reasoning under a fixed compute budget more reliably than state-of-the-art recursive baselines acting in observation space. LIT closes the vision–language gap: our visual model reaches performance on par with the second-best state-of-the-art language model, solves the visual baseline (∼100%) while producing diverse solutions, and improves over the visual state-of-the-art (51%). Adding our imagination inductive bias to the best language model improves accuracy by 14.8%.

1 Introduction

The reasoning capabilities of generative AI have significantly improved over the past couple of years across multiple modalities, especially language and vision. To unlock reasoning and realistic generation, most current approaches stimulate models to *think* by recursively consuming their own outputs—whether through Chain-of-Thought (CoT) in LLMs or iterative denoising in diffusion models (either in pixel space or in a latent grid learned via an observation-reconstruction objective). While these approaches look distinct, they share a common principle: in both modalities, reasoning and learning are reduced to an *observational* space. This design choice is practical—it makes the reasoning process interpretable (read or visualized) and directly optimizable (with known targets)—but it inherently conflates two distinct roles: *latent reasoning*, i.e., discovering a task-appropriate internal language and maintaining a belief over solutions, and *modeling the observational space*.

Learning in observation space is often suboptimal. If there exists an optimal language for solving a reasoning task, then using another language is suboptimal. By definition, modeling the observation space (e.g., the grammatical and stylistic form of natural language) instead of the underlying latent structure of the problem incurs additional computation—a kind of tax for working in a suboptimal representation. Richens & Everitt (2024) argued that

[*]Correspondence to: faridk@cs.uni-freiburg.de.



(a) Prior Model (student) (b) Posterior Model (teacher) (c) Outer Sampling Loop

Figure 1: **Latent Imagination Thinking Model (LIT).** LIT comprises two recurrent models with shared weights, producing an intermediate latent grid and a latent belief state π_i by unrolling the model f_θ for N *thinking steps*. The key asymmetry lies in the conditioning: (a) the prior (student) model predicts a latent belief trajectory given the input x and initial π_0, whereas (b) the posterior (teacher) model, further infers a posterior belief given a set of task-relevant observations $\mathcal{O} = \{o_i\}_{i=1}^N$. (c) For inference, we discard the teacher and use only the student. We implemented the framework in a standard denoising process, where the student model with N *thinking steps* implements one out of D denoising steps, i.e., there is an outer loop of denoising iterations, each running the recursive model.

robust world models are essentially causal; without the underlying causal representations, the model cannot generalize. Many tasks have a latent representation structure that does not naturally reside in token sequences or pixel grids: implicit rules, physical intuition, abstract relations, and compositional regularities (e.g., gravity, zero, and physical interactions) are only *indirectly* accessible through observations. Some of them cannot be expressed well in language either.

Scaling observational data can mask this issue—we partially attribute the success of current LLMs to the sheer scale and abstraction of human language as a representational shortcut for the reasoning language—but it can also *bottleneck* the need for true representation learning and reasoning. We argue that this bottleneck is responsible for the limited reasoning performance of LLMs (Glazer et al., 2025; Chollet et al., 2025), and video reasoning models (Wiedemer et al., 2025), and more broadly, the deterioration of generality in late layers of LLMs, diffusion models, and their collapse to low rank updates in ViT architectures (Skean et al., 2024; Yu et al., 2025; Jacobs et al., 2026; Alain & Bengio, 2016; Papyan et al., 2020). This effect is consistent with the common information bottleneck objective: an internal representation for a given task should maximize mutual information with the target while minimizing mutual information with the input, thereby discarding extraneous details. In this work, we treat pixel and language spaces as observation spaces, and instead target inference of a *latent language* as the emergent space on which reasoning resides.

For reasoning and algorithmic learning, the field has increasingly turned toward Chain-of-Thought (CoT) and recursive models Geiping et al. (2025). Theoretical foundations suggest that recurrence or CoT allows transformer models to become Turing-complete, hence capable of unbounded computation (Dehghani et al., 2019; Wei et al., 2022; Merrill & Sabharwal, 2024). CoT attempts to harness this power through " verbalizing" the algorithm into a sequence of intermediate language tokens. Latent CoT variants like iCOT (Deng et al., 2024) and CoConut (Hao et al., 2025) attempt to remove the language space bottleneck by feeding in the intermediate hidden states as inputs instead of language tokens, which excels in reasoning tasks that require substantial search during planning. For visual models, researchers employed diffusion models (recursive denoising from Gaussian noise to data) to think in frames (Arnab et al., 2025) or to solve a Sudoku puzzle recursively in a cell-by-cell fashion (Wewer et al., 2025) using the learned uncertainty, yet all still reside in the observation space (frames, pixels, patches).

Approaches like the Hierarchical Reasoning Model (HRM) (Wang et al., 2025) and Tiny Recursive Model (TRM) (Jolicoeur-Martineau, 2025) demonstrate that iteratively improving a latent representation using a smaller network can emulate the depth of larger models and compete with large models on reasoning tasks of ARC-AGI, maze navigation, and Sudoku. Geiping et al. (2025) explicitly frames this as scaling test-time compute via *vertical* latent-space recurrence in language models, contrasting it with more expensive *horizontal* scaling through token-level reasoning traces, and reports gains on mathematical and code reasoning. However, these recursive models typically provide no explicit learning guidance for their intermediate latent states beyond the observational end-task objective, and their internal state computation is therefore free to degenerate into observation refinement rather than learning the underlying latents and belief update for solving the task.

However, recurrence alone does not imply algorithmic learning or latent inference. A recurrent model can simply refine a single observational guess without relying on the true latent state or revising a latent belief state in a structured way. Long-horizon goals, planning under partial observability, and problems requiring diverse viable solutions demand more than depth: they demand latent belief update. Reasoning is not merely "more steps"; it is an iterative process that integrates evidence, manages uncertainty, and resolves competing hypotheses not only for the solution but for the underlying latent language.

Can we train models to reason by latent imagination, instead of thinking in tokens or pixels?

In this paper, we treat modalities (language and images) as partial observations, rather than as the language of thought. We introduce a latent belief state (a "thinking state") that is updated recurrently and can help decode the observation, but is not limited to it. This decouples reasoning from decoding: the model is free to run through iterations in its own abstract representational space to find a solution without repeatedly projecting its representation into language tokens or pixels.

However, how can the model be fostered to learn such abstract reasoning representation? As seen in Fig. 1, a posterior model (teacher) updates its thinking tokens $\{\pi_1, \ldots, \pi_N\}$ using observations $\{o_1, \ldots, o_N\}$ drawn from different augmentations (noising and masking) of the solution (task). At step n, earlier observations at steps $< n$ are dropped, forcing abstraction, and forming an abstract belief state that approximates the task posterior latent states. To teach the model iterative imagination, a prior model (student) is trained not only to predict the final observational answer, but also to track the teacher's abstract belief state via an imagination loss with stop-gradient targets. The result is a recurrent model that behaves like amortized message passing: each step updates the belief latent state in response to evidence, rather than merely producing another partial observational output.

If our model performs latent-state imagination rather than observation-space reasoning, then increasing the number of thinking steps should (i) monotonically increase the probability of solving within a fixed budget in comparison to observation-space recursion and (ii) maintain coverage of multiple valid hypotheses without premature collapse. Our results consistently show these trends across different models and modalities.

In summary, our contributions are:

- A student–teacher scheme that decouples latent reasoning from observation-space decoding (tokens/pixels).
- The proposed imagination loss improves test-time scaling with more thinking steps versus non-imaginative recursive baselines.
- A diffusion instantiation of the approach achieving state-of-the-art results on the SRM bench for solving Sudokus.

2 Related Work

World models and latent imagination. Model-based RL has long advocated learning compact latent dynamics for *imagination* and planning, starting with early latent world-model agents (Ha & Schmidhuber, 2018). The Dreamer family learns a latent dynamics

model and uses imagined rollouts for reinforcement learning (RL), with successive versions improving stability and scalability (Hafner et al., 2020; 2021; 2024). Recent Dreamer variants moved toward richer observation models, abandoning the latent-space approach in favor of more expressive (diffusion-based) decoders (Hafner et al., 2025). Beyond recurrent dynamics, alternatives such as masked latent Transformers (Burchi & Timofte, 2025). This extension is challenged as their encoder model drops objective-relevant conditioning on the current observation x_t in $q_\phi(z_t \mid h_t, x_t)$ due to their auto-regressive *horizontal* token prediction over time. Related to our imagination loss, consistency objectives for latent dynamics (e.g., latent overshooting) regularize multi-step latent predictions to improve long-horizon imagination, but was shown to be ineffective with recurrent state space models (RSSM) in the predecessor version of Dreamer, PlaNet (Hafner et al., 2019). In contrast, LIT *vertically* predicts/encodes observations over depth, not new-tokens through recurrent belief update: it trains intermediate latent states to implement belief updates that support reasoning.

Amortized inference, partial observability, and belief update. More broadly, iterative inference can be viewed as message passing/belief propagation, in which each new observation provides evidence and the belief state is updated accordingly (sum-product, variational message passing). LIT implements this principle with a teacher–student scheme in which a posterior teacher refines belief upon observing new information, while a prior student learns to track these refinements via an imagination loss, turning recurrence into a structured belief update. Another fundamental challenge in the Dreamer work (Hafner et al., 2020) identified by (Bayer et al., 2021) is restricting the posteriors instead of fully-conditioned posteriors. LIT avoids this problem by learning a better approximation of the full posterior.

Self-supervised latent objectives and representation learning. Self-supervised learning has shown that strong representations can be learned without reconstructing observations. DINO-style self-distillation aligns student and teacher representations across views, avoiding pixel-level losses (Caron et al., 2021; Zhou et al., 2021; Oquab et al., 2024). JEPA-style methods generalize this by predicting missing content in representation space rather than in observation space (Assran et al., 2023), and video variants extend this principle to temporal prediction in latent space (Assran et al., 2025; Bardes et al., 2023). These methods learn abstract features but typically perform *single-shot* latent prediction rather than iterative belief refinement, and they do not explicitly train intermediate latent trajectories to perform inference. LIT inherits the "predict in latent space" philosophy but adds an explicit multi-step imagination latent inference objective, training the latent state to evolve as a belief latent state in a belief update-style over the latents and the solution.

3 Learning through latent Imagination

In this section, we introduce our method LIT for learning latent imagination. Fig. 1 depicts our training paradigm for **LIT** as a coupled *prior–posterior*. Both models share the same recurrent backbone and weights: a Transformer block (with L layers) is unrolled for N *thinking steps*, producing an intermediate latent grid/state and a latent *belief / thinking* state.

```
def training_step(x, y_true, N):
    t = sample_t(); e = randn_like(x)
    z = t * x + (1 - t) * e; v = (x - z) / (1 - t)
    o1, .. oN = get_observations(y_true)
    tt = tt_init; z_post, tt_post = z, tt_init

    loss = 0
    for n in range(N):
        z, tt = fθ(z, t, None, tt);
        v_pred = oϕ(z, t, tt)
        z_post, tt_post = fθ(z_post, t, o[n], tt_post);
        v_post = oϕ(z_post, t, tt_post)
        loss += mse(v, v_pred) + mse(v, v_post) + mse(gϕ
              (tt), tt_post.detach())

    return loss
```

Figure 2: Training step.

```
def sampling_step(z, t, t_next, N):
    tt = tt_init
    for n in range(N):
        z, tt = fθ(z, t, None, tt)
        v_pred = oϕ(z, t, tt)
        z = z + (t_next - t) * v_pred
    return z
```

Figure 3: Sampling (Denoising) step (Euler).

An *output layer* decodes the final belief into the observation language (tokens/pixels) and is trained with the task loss (FM loss in the figure). The key asymmetry is the conditioning: the *prior* (student) conditions on the standard input x and an initialization π_0, while the *posterior* (teacher) is given additional observations $\{o_0, \ldots, o_N\}$ derived from task supervision (e.g., progressively revealing solution evidence). The posterior, therefore, forms a more informed belief state. LIT trains the prior not only to solve the task, but also to *track the posterior's latent belief update* via an auxiliary imagination loss: a lightweight predictor maps the prior's intermediate belief to a target provided by the posterior, with gradients stopped through the posterior (stop-grad), making the teacher signal stable.

Training the prior only. Training the prior only resembles generative models with recurrent "register" tokens, spanning LLMs, diffusion models, and recent recursive variants. In the *prior-only* setting (Fig. 1a), a recurrent model is trained to solve the task using only the standard observation x and initial register tokens π_0. Unrolling the Transformer block for N thinking steps yields a sequence of latent thinking states $\{\pi_n^{\mathrm{pr}}\}_{n=1}^N$ and predictions $\{y_n^{\mathrm{pr}}\}_{n=1}^N$, where $\hat{y}_n^{\mathrm{pr}}$ is decoded from hidden states h_n^{pr} via the output layer. The model is optimized with the observational task loss.

Training the posterior only. In the *posterior-only* setting (Fig. 1b), the same recurrent backbone is trained, but conditioned on additional observations $\{o_0, \ldots, o_N\}$ available during training relevant to the task. Concretely, the posterior produces a sequence $\{\pi_n^{\mathrm{po}}\}_{n=1}^N$ and predictions $\{y_n^{\mathrm{po}}\}_{n=1}^N$, where each step can incorporate additional evidence (and drop hidden states computed with earlier observations), yielding a refined belief state closer to a task posterior. The model is still trained with the task loss through the output layer (FM loss) to abstract the observations, but crucially, the posterior's intermediate computation is informed by richer evidence. This yields a stronger latent belief trajectory for $\pi_{1:N}^{\mathrm{po}}$, yet by itself cannot be used for inference as it is trained on seeing new observations, and it does not teach the prior how to imagine these refined beliefs when only x is available at test time.

Training LIT (prior + posterior + imagination loss). LIT couples the two models during training. The posterior (teacher) uses $\{o_i\}$ to produce a refined latent trajectory $\{\pi_n^{\mathrm{po}}\}_{n=1}^N$, which we treat as targets for the prior (student). The prior runs on x and π_0 to imagine $\{\pi_n^{\mathrm{pr}}\}_{n=1}^N$ and $\{y_n^{\mathrm{pr}}\}_{n=1}^N$; a linear predictor $g(\cdot)$ maps π_n^{pr} into the teacher space, and we apply an imagination loss between $g(\pi_n^{\mathrm{pr}})$ and π_n^{po} with stop-gradient through the teacher. The overall effect is that recurrence is no longer "more steps to sharpen an output," but a learned *latent belief update* procedure: each thinking step is trained to move π_n^{pr} toward a posterior-consistent refinement. We show the LIT loss function in Equation 1. At test time, we discard the teacher and scale inference *vertically* by increasing the number of thinking steps N, improving reasoning by refining the latent belief state while decoding only when needed. In addition to the LIT diffusion version shown in the main Fig. 1, training algorithm Fig. 2, sampling algorithm Fig. 3, we show the TRM-LIT version in Fig. 4.

```
# Deep Supervision
for x_input, y_true in train_dataloader:
   y, z, tt = y_init, z_init, tt_init
   for step in range(N_supervision):
      x = input_embedding(x_input)
      y_star = get_observations(y_true) #Masking, Augmentation
      (y, z, tt), y_hat, q_hat = deep_recursion(x, y, z, tt)
      (y_post, z_post, tt_post), y_hat_post, q_hat_post = deep_recursion(x, y_post, z_post, tt_post, y_star=y_star)
      loss = softmax_cross_entropy(y_hat, y_true)
      loss += softmax_cross_entropy(y_hat_post, y_true)
      loss += binary_cross_entropy(q_hat, (y_hat == y_true))
      loss += mse_loss(tt_post.detach(), tt)
      loss.backward()
      opt.step()
      opt.zero_grad()
      if q_hat > 0: # early-stopping
         break
```

Figure 4: Pseudocode of Tiny Recursion Models (TRM) with LIT.

$$\mathcal{L}_{\text{LIT}} = \sum_{n=1}^{N} \mathcal{L}_{\text{FM}}\big(\hat{y}_n^{\text{pr}}, y\big) + \lambda_{\text{post}} \sum_{n=1}^{N} \mathcal{L}_{\text{FM}}\big(\hat{y}_n^{\text{po}}, y\big) + \lambda_{\text{im}} \sum_{n=1}^{N} \big\| g\big(\pi_n^{\text{pr}}\big) - \text{sg}[\pi_n^{\text{po}}] \big\|_2^2 . \quad (1)$$

4 Experimental Setup

We focus our experimental setup on the game of Sudoku—a 9x9 logic puzzle with the objective of assigning digits from 1 to 9 to empty cells while respecting constraints that each digit appears only once in each row, column, and smaller 3x3 block. A solution is considered valid if all the constraints are satisfied. Depending on the difficulty of the starting Sudoku grid, the given grid can have one or more valid solutions, while the difficulty is defined by the number of seen digits in the initial grid and the number of backtracking steps.

Due to its structured constraint satisfaction and rules, Sudoku has become an interesting test-bed to many recent reasoning models Wang et al. (2025); Wewer et al. (2025); Jolicoeur-Martineau (2025). We use the two standard benchmarks from recent work: the MNIST Sudoku from SRM Benchmark (Wewer et al., 2025) and Sudoku Extreme (Wang et al., 2025). For both benchmarks, we report accuracy, the percentage of valid grids, and the L1 distance, which measures the number of violated constraints.

SRM MNIST Sudoku. SRM is a visual reasoning benchmark proposed by Wewer et al. (2025). We focus solely on the MNIST Sudoku task, in which the grids are represented with MNIST (Deng, 2012) digit images. Following prior work, we assess the validity of the solved grid by applying a pre-trained MNIST digit classifier to each cell and checking whether all Sudoku constraints are satisfied. To limit the digit classification errors, we use only 1000 MNIST training examples per class. The dataset defines three difficulty levels based on the number of masked cells in the starting grid: easy, medium, and hard. We use only the hard subset of the dataset, where $[55, 81]$ cells in the starting grid are masked.

Sudoku Extreme. Sudoku Extreme is a challenging puzzle dataset introduced in Wang et al. (2025) that requires advanced deductive steps or backtracking. The difficulty of the grid is measured by the number of backtracks required for search Wang et al. (2025). We use the smaller version of the proposed dataset, which comprises 1000 training and 423K test puzzles with unique solutions. Given the small size of the training data, we follow prior setups and apply strong augmentations during training without violating Sudoku constraints. For the visual version of the dataset, we combine the grids with the MNIST digits.

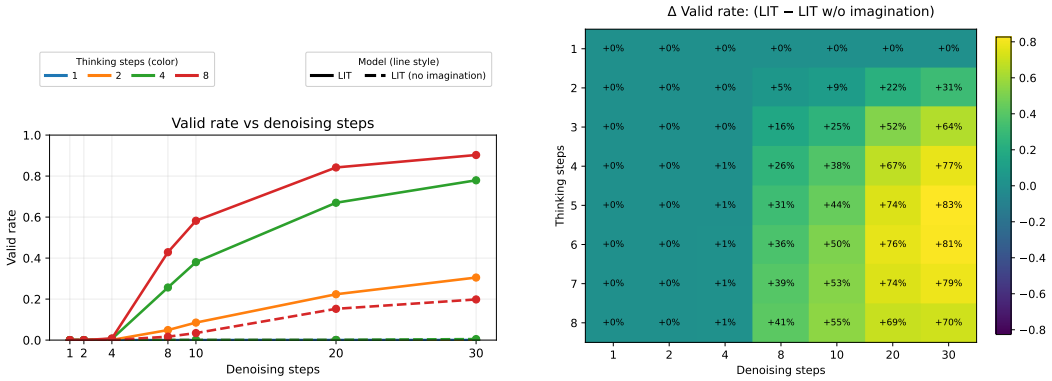
Baselines. We compare LIT, against recent reasoning models, including SRM Wewer et al. (2025), HRM Wang et al. (2025), and TRM Jolicoeur-Martineau (2025). We compare with SRM on the hard MNIST Sudoku dataset, training on the full training set and evaluating on the official test data. We compare with HRM and TRM on Sudoku Extreme with unique solutions only on the smaller 1k subset. We evaluate HRM and TRM on the language-based version only, whereas LIT is evaluated on both symbolic and visual variants.

5 Is recursion enough for reasoning?

In this section, we address the central claim of our work, leveraging the setup in Section 4.

Thinking steps vs. Denoising steps. Fig. 5a isolates two orthogonal axes of compute: *denoising steps* (the observation-space refinement budget) and *thinking steps* (the number of latent state updates). Increasing denoising steps acts as a threshold on validity, after which gains are amplified when the model is allowed to perform additional latent thinking steps: for the same denoising budget, deeper latent iterations yield consistently higher valid rates.

The delta map in Fig. 5a(a) details the effect of the two axes: imagination-driven belief updates provide the largest improvements at higher denoising budgets, suggesting that recursion in the observation space benefits from a learned latent inference process rather than acting as a standalone mechanism for reasoning.



(a) Absolute Valid Rate

(b) Delta Valid Rate

Figure 5: **Analysis of the different roles of denoising and thinking steps on the SRM MNIST Sudoku.** We plot model performance as a function of denoising steps for various thinking budgets (color intensity). (a) Solved rate; (b) Relative valid rate improvement. While increasing thinking steps consistently improves logical consistency and task success. The denoising steps serve as a hard gating, where, after 4 denoising steps, performance scales with the thinking steps.

Since denoising diffusion models inherently support diverse sampling, we investigate the effect of allocating different compute budgets to the thinking steps and k sampling attempts (parallel solutions). Figure 7 shows that deeper thinking (fewer samples with more thinking steps) outperforms shallow sampling (more samples with fewer thinking steps each).

Is recursion alone sufficient? LIT with vs. without imagination. Table 1 answers the core question directly by ablating the imagination objective. A purely recursive model without imagination ($\lambda_{\text{im}} = 0$) achieves only 19.83 final accuracy (27.44 solved-by-step), whereas LIT reaches 90.28 final (99.96 solved-by-step), while simultaneously reducing the L_1 error from 7.6229 to 0.2922. Fig. 5 explains that this gap is not an artifact of arbitrary scaling in either axis. Both variants can be run with the same total recursive compute budget (denoising x thinking) steps, yet after a certain threshold of denoising steps, the one with more thinking is significantly better. The imagination loss provides *learning guidance for intermediate latent states*, turning recurrence into a latent belief update instead of repeated observation-space prediction.

Table 1: **Main Results and Ablation of Imagination Loss on SRM MNIST Sudoku.** Comparison of LIT against diffusion baselines. Accuracy is reported as **Final (Solved-by-step)** in percent, where Solved-by-step is the cumulative solved-by-step rate (solved at any step/budget $t \leq K$). The variant with $\lambda_{img} = 0$ removes the imagination loss. Diffusion Model and SRM results are taken from Wewer et al. (2025).

Model / Approach	Sampling Strategy	Accuracy (%) ↑	L1 Error ↓
Diffusion Model	Parallel	8.00 (—)	14.120
SRM	Parallel	1.00 (—)	19.156
SRM	Predicted Order	51.60 (—)	3.2120
LIT ($\lambda_{img} = 0$)	**Parallel**	19.83 (27.44)	7.6229
LIT	**Parallel**	**90.28 (99.96)**	**0.2922**

Scaling test-time compute: puzzle difficulty and diversity (solution-level and visual). Unlike models that uniformly apply fixed compute regardless of difficulty, effective reasoning systems should allocate compute according to problem complexity, whether finding a single hard solution or exploring multiple viable solutions. As seen in Fig.9, easy problems of ranking=0, where success depends on rule understanding, the model achieves 90% accuracy with minimal compute (N=3) and reaches 90% at (N=7). However, for harder instances,

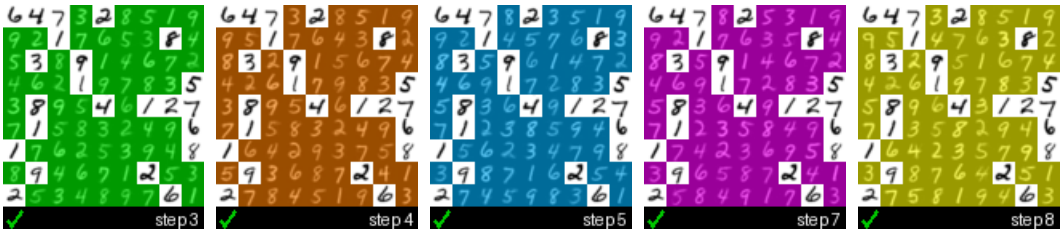


Figure 6: **Diversity examples across solutions / steps on SRM MNIST Sudoku**. The puzzle is shown in the white background/black numbers cells. The step number shows the thinking step at which each answer was found. Different colors indicate different solutions for the same puzzle at different steps.

we observe that additional thinking steps improve reasoning capability, enabling the model to progressively solve more challenging puzzles.

Beyond accuracy, Fig. 6 illustrates diverse valid solutions produced across thinking steps. Quantitatively, LIT generates *on average 5 solutions per puzzle* on Sudoku SRM and reaches the *first valid solution by the 3rd thinking step on average*. These results align with the intended behavior of latent imagination: different computation budgets correspond to solving harder problems or providing multiple viable solutions without collapse prematurely to a single hypothesis. In addition for each instance of the grid cell, e.g. the number 2, one can see visual/observational diversity.

LIT as an intervention on TRM yields a new state-of-the-art. We observe the same phenomenon in a non-diffusion recursive model. Fig. 8 shows that TRM-LIT improves steadily with training compute and overtakes TRM without imagination. On Sudoku Extreme (Table 2), TRM-LIT achieves 91.9% versus 77.4% for TRM and 55.0% for HRM, establishing a new state-of-the-art among recursive baselines. This supports the central claim: recursion is a powerful *compute* paradigm, but without an explicit latent inference objective, it can plateau as observational refinement.

Closing the visual–language gap. The modality gap: for a true latent reasoning model, pixel and language space reasoning models must infer the same hidden rules and factors. Visual LIT reaches 54.5% on Sudoku Extreme (Table 2), essentially matching the second-best language recursive baseline (HRM at 55.0%), despite operating on MNIST observations. Moreover, our per-rating analysis indicates that for puzzles whose difficulty is dominated by rule adherence (as opposed to long search), the model achieves near-perfect success (around 0.99), consistent with learning the latent constraints rather than overfitting to observation

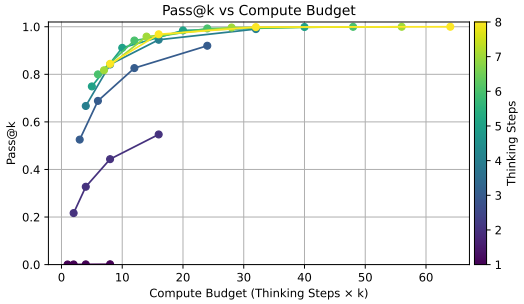


Figure 7: Pass@k on the SRM MNIST Sudoku. For a fixed compute budget, the thinking steps beat the k different random initializations for the diffusion model.

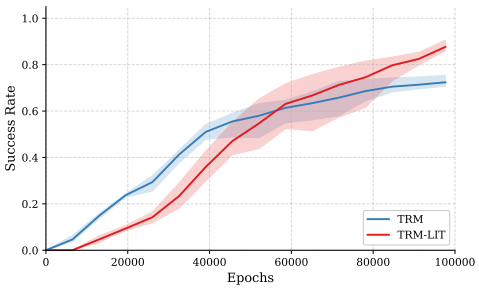


Figure 8: Performance on Sudoku Extreme with training compute of TRM-LIT vs TRM w/o imagination over three seeds.

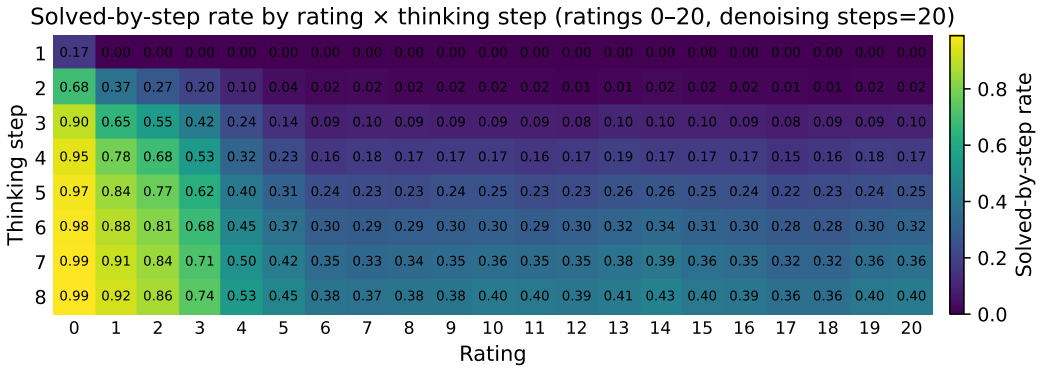


Figure 9: Solved-by-step rate by rating × thinking step of the visual-LIT on the Sudoku Extreme dataset (ratings 0–20, denoising steps=20). Rating quantifies the backtracking steps required by the tdoku solver (Dillon) to solve the puzzle (higher is harder).

artifacts. This is the regime we target: to discover a task-appropriate *language of reasoning* shared across modalities.

Table 2: **Sudoku Extreme Performance** Comparison of our Latent Imagination Thinking (LIT) framework against large-scale pretrained models and recursive-only baselines. CoT, HRM, and Direct Pred results are taken from the Jolicoeur-Martineau (2025) work. TRM is reproduced, and the best out of 3 seeds is taken.

Category	Method	# Params	Sudoku (%) ↑
CoT	Deepseek R1	671B	0.0
	Claude 3.7 8K	?	0.0
	o3-mini-high	?	0.0
Recursive, small-sample training	Direct pred	27M	0.0
	HRM	27M	55.0
	TRM	7M	77.4
Recursive, small-sample training + Latent Imagination	**Visual LIT**	81M	54.5
	TRM-LIT	7M	**91.9**

6 Conclusion

Our work aims to answer a fundamental question for generative reasoners: *is recursion in tokens or pixels sufficient, or do models need a learned latent language for belief update?* We find that recursion alone is not sufficient. We introduce **Latent Imagination Thinking (LIT)**, a teacher–student paradigm that decouples reasoning from decoding by training recurrent latent states to approximate latent posterior updates. Across diffusion and recursive transformer instantiations on Sudoku, LIT demonstrates: (i) better test-time compute scaling compared to observation-space baselines, (ii) maintenance of diverse solution hypotheses without premature collapse, (iii) effective allocation of computational resources proportional to problem difficulty, and (iv) closing the visual-language gap. These results suggest that treating modalities as partial observations—rather than as the language of thought itself—enables more principled reasoning through learned latent abstractions. LIT's formulation is general: any task where relevant observations can be provided fits within the representation learning framework. For video world models, observations can be the intermediate frames before reaching a goal; for counterfactual reasoning, observing interventions would teach the model to simulate it at test time. By relaxing the observation learning tax, LIT points toward reasoning models that invest compute in discovering the task's underlying rules and latents rather than reproducing observational details.

Acknowledgments

This research was funded by the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS" (19A23014R), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 499552394–SFB 1597, and grant number 539134284, through EFRE (FEIH_2698644) and the state of Baden-Württemberg. The compute used in this project was also funded by the German Research Foundation (DFG) under grants 417962828 and 539134284. Karim Farid and Jelena Bratulić are part of the European Lab for Learning and Intelligent Systems (ELLIS) PhD program. The authors acknowledge support from the state of BadenWürttemberg through bwHPC. We thank Arian Mousakhan, Silvio Galesso, and Johannes Dienert for valuable discussions and feedback during the development of this work.

The Use of Large Language Models (LLMs)

We used OpenAI (2025b) as a coding assistant during implementation and GPT-5 (OpenAI, 2025a) to polish the writing. All core contributions and the initial draft were done by the authors.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal Chain of Thought: Long-Video Understanding by Thinking in Frames, July 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL `https://arxiv.org/abs/2506.09985`.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Learning Representations*, 2021.

Maxime Burchi and Radu Timofte. Accurate and Efficient World Modeling with Masked Latent Transformers. *International Conference on Machine Learning*, 2025. doi: 10.48550/arxiv.2507.04075.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step, 2024. URL `https://arxiv.org/abs/2405.14838`.

T. Dillon. tdoku. URL `https://github.com/t-dillon/tdoku`. Accessed 2026-02-01.

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, February 2025.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025. URL `https://arxiv.org/abs/2411.04872`.

David Ha and Jürgen Schmidhuber. World Models. March 2018. doi: 10.5281/zenodo.1207631.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi 0002. Dream to Control: Learning Behaviors by Latent Imagination. *ICLR*, 2020. doi: 10.48550/arXiv.1912.01603.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi 0002, and Jimmy Ba. Mastering Atari with Discrete World Models. *ICLR*, 2021. doi: 10.48550/arXiv.2010.02193.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models, April 2024.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training Agents Inside of Scalable World Models, September 2025.

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, November 2025.

Mozes Jacobs, Thomas Fel, Richard Hakim, Alessandra Brondetta, Demba Ba, and T. Andy Keller. Block Recurrent Dynamics in Vision Transformers. In *The Fourteenth International Conference on Learning Representations*, October 2026.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

William Merrill and Ashish Sabharwal. The Expressive Power of Transformers with Chain of Thought. *ICLR*, 2024. doi: 10.48550/arXiv.2310.07923.

OpenAI. Introducing GPT-5. `https://openai.com/index/introducing-gpt-5`, 2025a. Blog post. Accessed: September 22, 2025.

OpenAI. Introducing codex. OpenAI, May 2025b. URL https://openai.com/index/introducing-codex/. Accessed 2026-02-02. Published 2025-05-16 (updated 2025-06-03).

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pOoKI3ouv1.

Oscar Skean, Md Rifat Arefin, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*, 2025.

Jason Wei, Xuezhi Wang 0002, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia 0002, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*, 2022. doi: 10.48550/arXiv.2201.11903.

Christopher Wewer, Bartlomiej Pogodzinski, Bernt Schiele, and Jan Eric Lenssen. Spatial reasoning with denoising models. In *International Conference on Machine Learning (ICML)*, 2025.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.