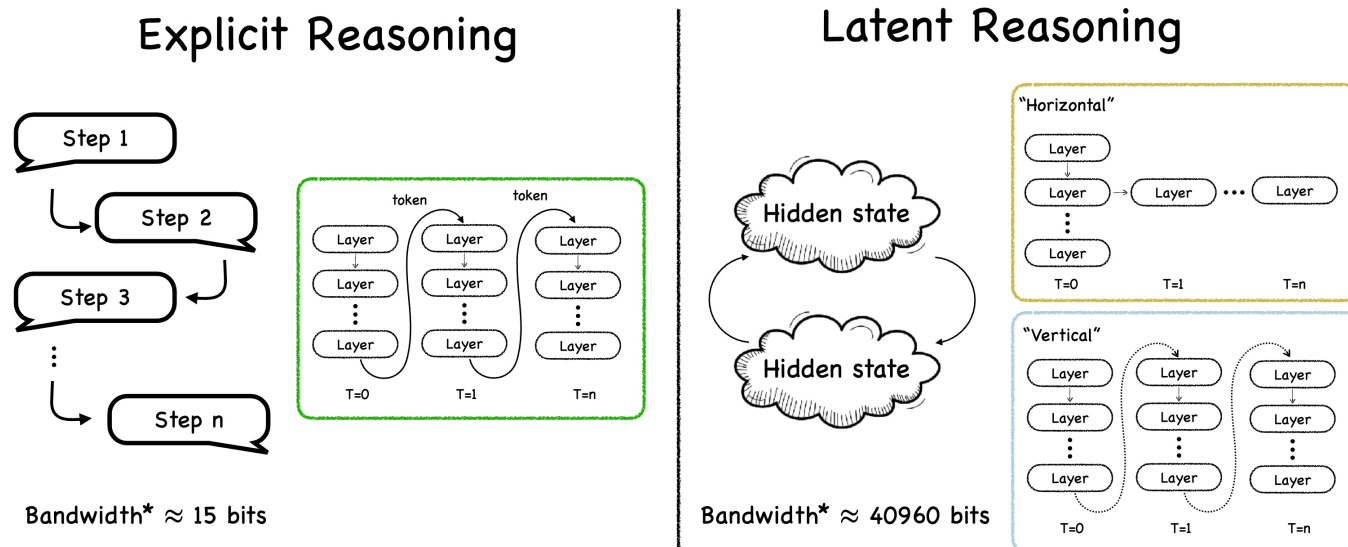


	 Reasoning	 Planning	 Modeling	 Perception	 Memory	 Collaboration	 Embodiment
Architecture	 Heima  Huginn  Looped Trans  PonderLM-2  Ouro ...	 LCR-SER  PILOT  ATP-Latent  CoLT ...	 CoCoMix  PHD-Tran.  DLCM  Dreamer  LoopFormer ...	 AURORA  AlignVLM  LaViT  VL-JEPA  HIVE ...	 IMM  G-MemLLM  CoMEM  CoMEM-Agent  PolarMem ...	 MAS4TS  KVCA  Wormhole ...	 UniVLA  ThinkAct  OccVLA  LCDrive ...
Representation	 HCoT  COCONUT  SoftCoT  Soft Thinking  SIM-CoT ...	 CTRLS  iCLP  LaSER ...	 LFJ  LatentGuard  LatentBreak  UniCog ...	 CGC-VTD  3DThinker  CoVT  LRP  LatentLens ...	 KAVA  SALS  CLaRa  DeltaKV ...	 C2C  ThoughtComm  Interlat  LatentMas  Primitives ...	 LatBot  VITA  Motus  LangForce  ColaVLA ...
Computation	 CCoT  Assorted  CODI  CoLaR  COT2 ...	 LARES  FR-Ponder  Latent  I2B-LPO  PLaT ...	 PRE  SpiralFormer  PonderLM-3  AdaPonderLM ...	 LVR  LS  DMLR  LIVR  Laser ...	 MemGen  VisMem  FlashMem ...	 HYP1/2  L2-VMAS  LatentMem ...	 LatentVLA  RD-VLA  Future-VLA  LaST-VLA ...
Optimization	 LaTRO  BoLT  SoftCoT++  LatentTTS  Latent-SFT ...	 LatentSeek  HRPO  LTO  LTPO  SofT-GRPO ...	 LF-Steering  LatentPrompt  GANPO  ConceptLM  MaD-Mix ...	 VTI  Mirage  LaCoT  VLPO  CogSense ...	 LatentEvolve  NextMem ...		 LAPA  ATE  SRPO  GLaD  CLAP ...

The past year has witnessed a surge of research centered on the latent space.

Figure from [1] Xinlei Yu et al. “The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook”

Background: The Bottleneck



Information Preservation

- Discrete Tokens: ~ 15 bits each...
- Hidden States (2560-d FP16): $\sim 40,960$ bits each.
- A 2.7×10^3 -fold bandwidth gap [1]

Communication Efficiency

- Decoding long CoT messages token-by-token.
- Generating tokens just for grammatical coherence.
- Forcing rich thoughts into a single linear sequence.

Figure from [1] Rui-Jie Zhu, et al. "A Survey on Latent Reasoning"

The Bottleneck

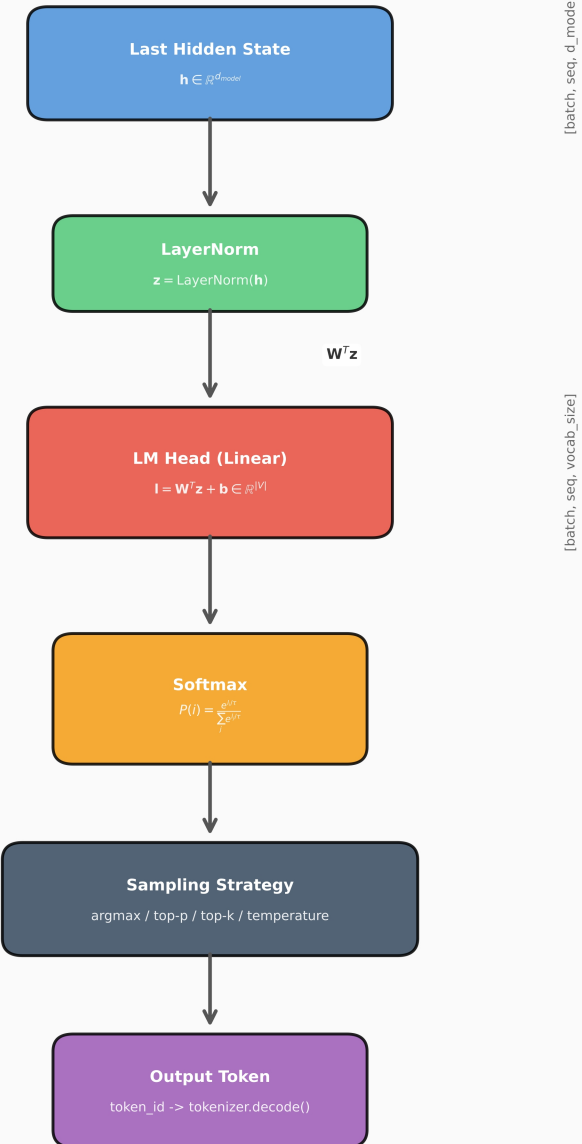
A standard process from a last hidden state to a token.

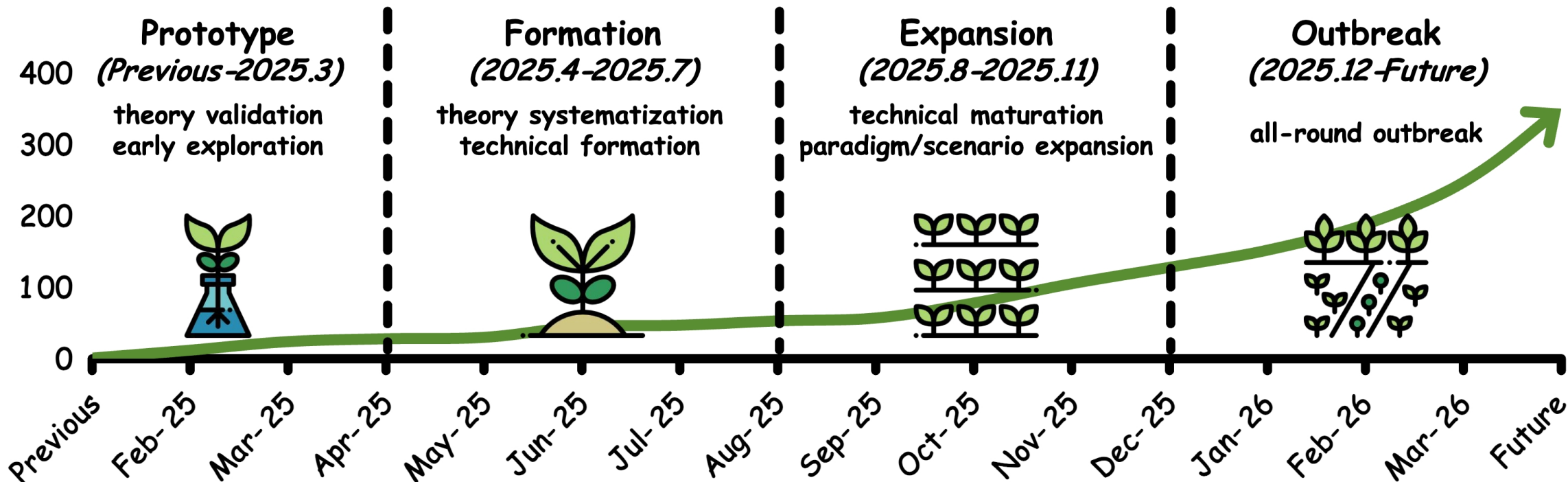
From $[1, d_model]$ $\rightarrow$ $[1]$.

A rich latent state holds parallel plans (e.g., "check desk" or "check drawer"). Sampling forces one path, permanently erasing alternative hypotheses.

LLM Decoding Flow: Last Hidden State to Token

Final layer output of Transformer





From the prototype like COCONUT where we first dared to let models think in silence to theoretical formalization that gave us the mathematical vocabulary to understand why it works; from there, it expanded across frontiers like multi-agent "mind-reading," implicit memory architectures, embodied intelligence, and cross-modal reasoning, now ushering in a future of unprecedented promise, where latent thinking becomes the native language of intelligence itself.

Figure from [1] Xinlei Yu et al. "The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook"

Zhuoyun Du, et al. "Enabling Agents to Communicate Entirely in Latent Space". ACL 2026.

Enabling Agents to Communicate Entirely in Latent Space

ACL 2026, San Diego, CA, USA

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao,
Xiaoyong Zhu, YuCheng, Bo Zheng, Wei Chen, Haochao Ying

duzy@zju.edu.cn

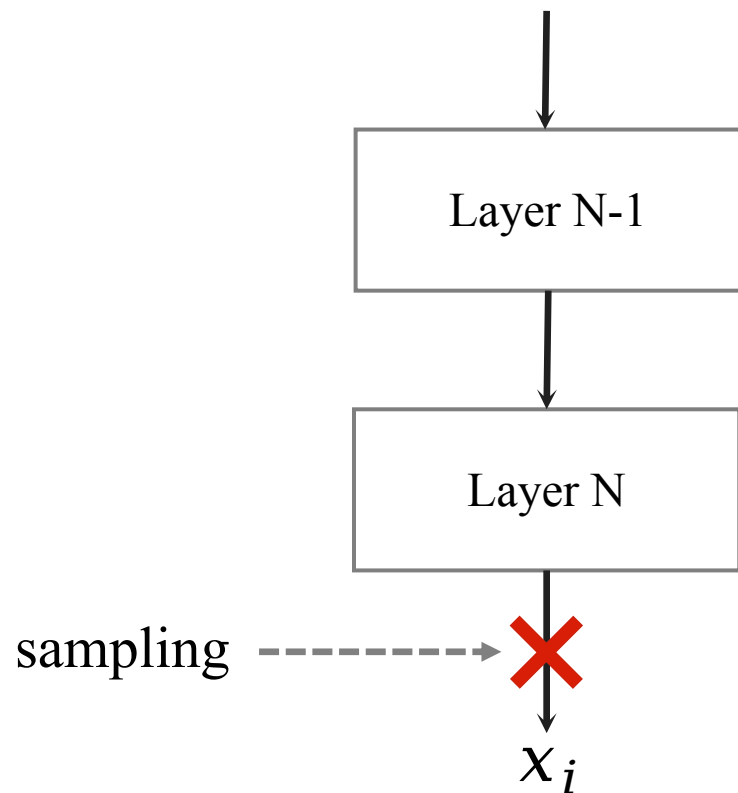
Zhejiang University, Alibaba group

Can multi-agent systems only communicate through natural language?

"We do not have organs of communication. Our brains can display our thoughts to the outside world, thereby achieving communication."

—Cixin Liu, The Dark Forest

Background

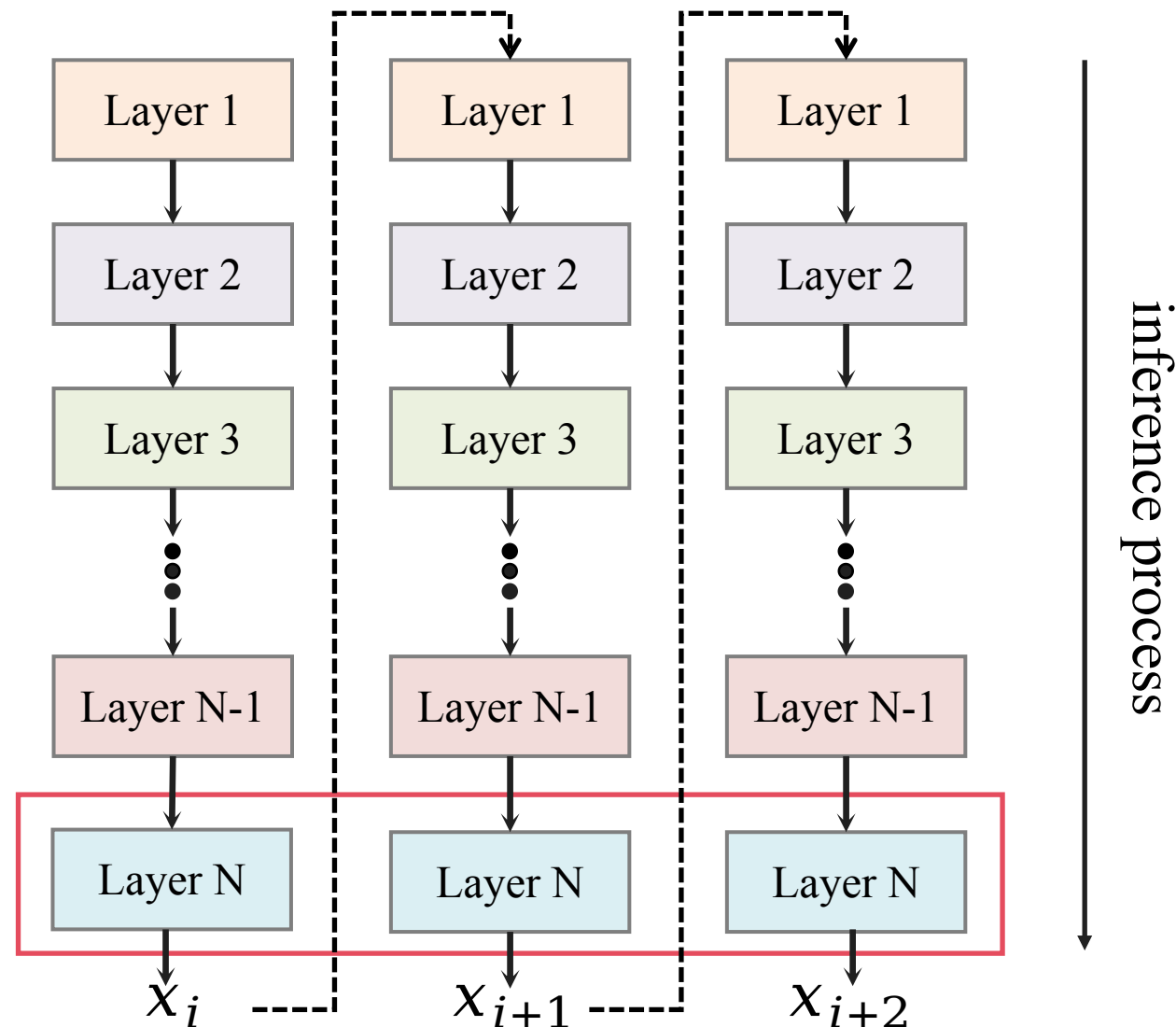


~~Use Natural Language~~
Bypassing token sampling to
preserve information
parallelism and diversity.

Motivation

- Humans often struggle to express complex thoughts efficiently through words.
- While human “telepathy” is sci-fi, open-source LLMs explicitly expose their rich internal states.

The last-layer hidden states:
naturally **preserve task-critical information**
derived from the entire inference process.

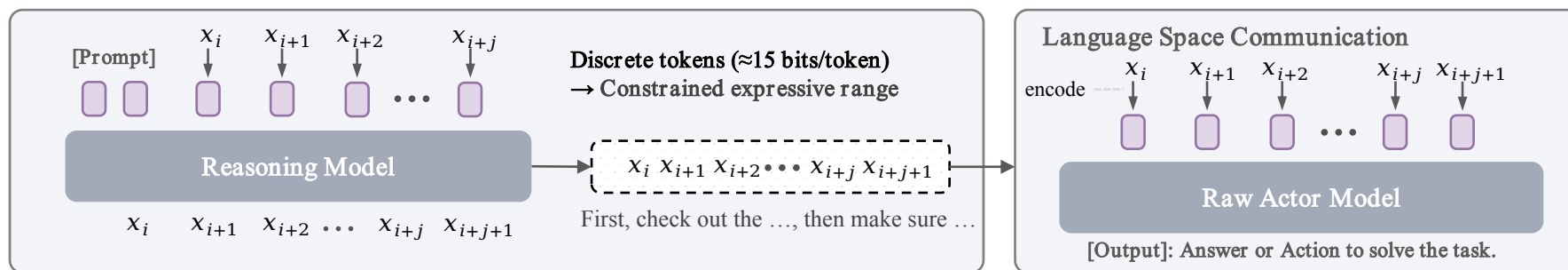


Stage 1: Enable latent space communication

Stage 2: Compress message for efficiency

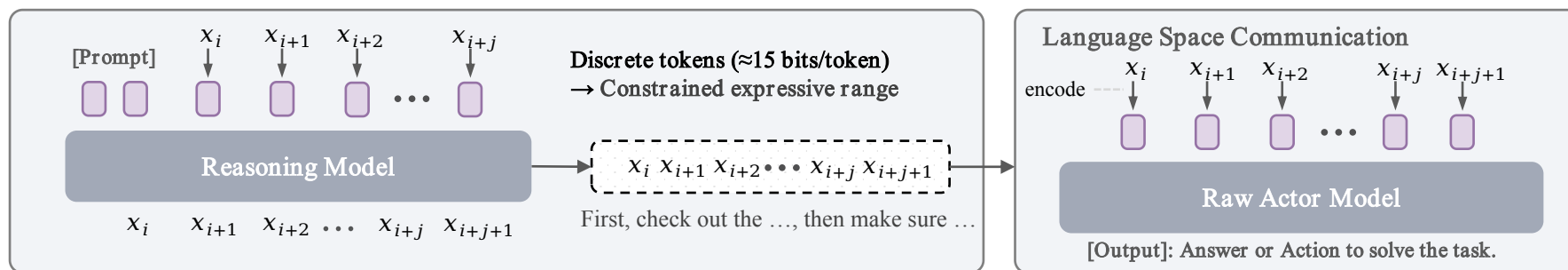
Latent space communication (Stage 1)

Baseline: Transmitting discrete tokens decoded by the language head.

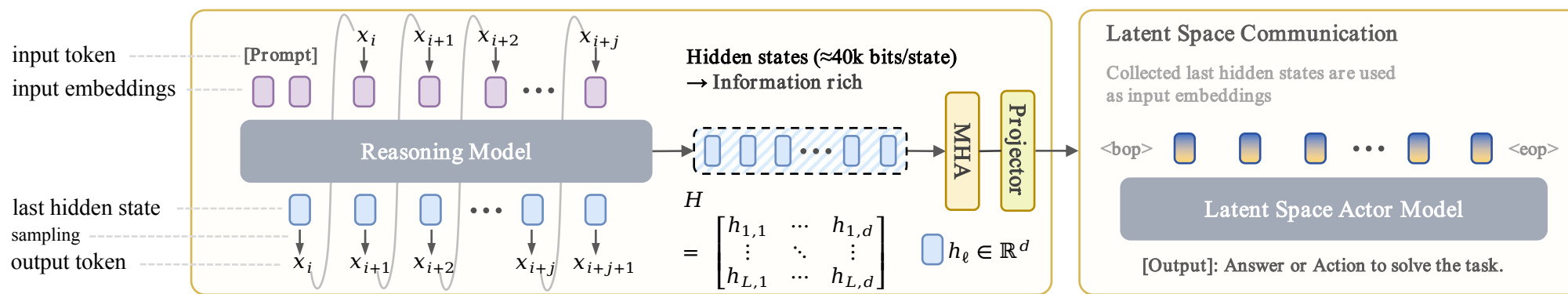


Latent space communication

Baseline: Transmitting discrete tokens decoded by the language head.

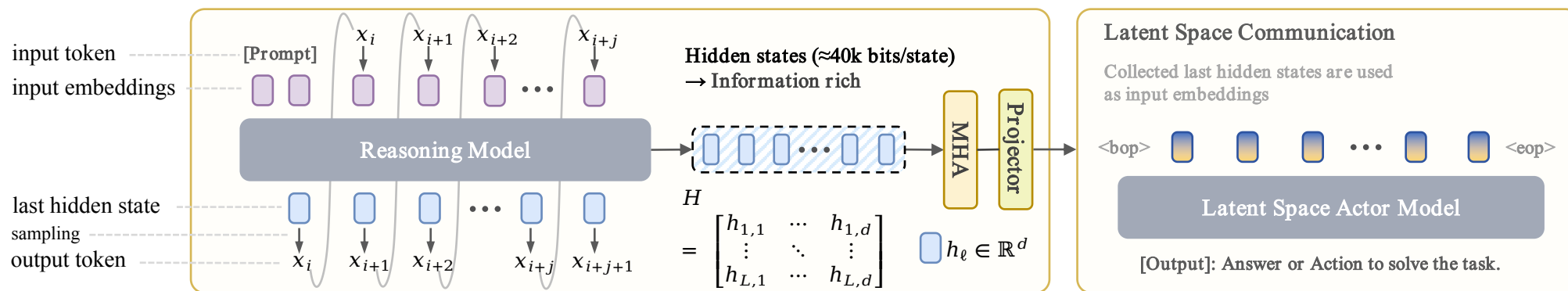


Interlat: Directly transmitting continuous last-layer hidden states as input embeddings.



Latent space communication

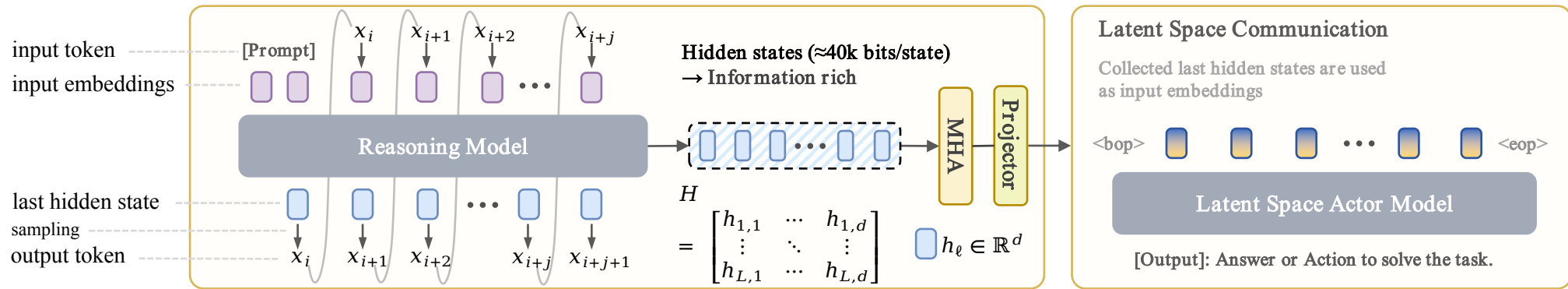
Interlat: Directly transmitting continuous last-layer hidden states as input embeddings.



- A lightweight, trainable module (Multi-Head Attention + Projection) for rescaling and interpretation.
- Bypasses the limits of natural language.
- Requires no parameter sharing or memory coupling.

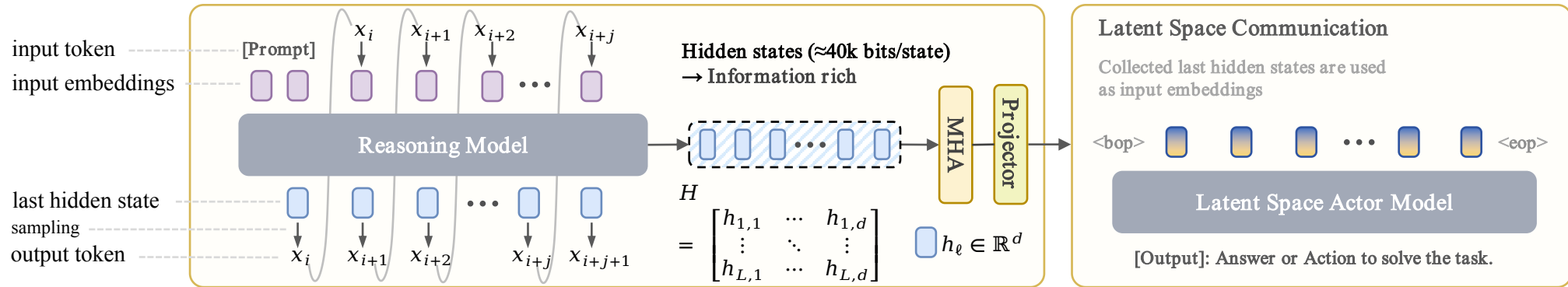
Latent space communication

Interlat: Directly transmitting continuous last-layer hidden states as input embeddings.



- L_{task} (Cross-Entropy): Ensures the actor produces accurate and coherent actions for the task.
- L_{sep} (JS Divergence): Forces the actor to genuinely utilize task-relevant latent information.
- L_{plan} (Plan Regulation): Regularizes latent predictions against text plans to prevent objective exploitation.

Latent space communication



- L_{sep} (JS Divergence): $-\frac{1}{|S|} \sum_{t \in S} \text{JS}(p_{\theta}(\cdot | C_t, H), p_{\theta}(\cdot | C_t, \tilde{H}))$.
- L_{plan} (Plan Regulation): $\frac{\beta}{|S|} \sum_{t \in S} \text{KL}(p_{\theta}(\cdot | H) \| p_{plan}(\cdot | P)) + \frac{\alpha}{|S|} \sum_{t \in S} (1 - \cos(\mathbf{z}_{\theta}(H), \mathbf{z}_{plan}(P)))$.
- $H / \tilde{H}$: Matched vs. Mismatched latents.
- $p_{\theta}(\cdot)$: Actor model's output distribution.
- P : Natural language plan.
- $\mathbf{Z}$: Logits for geometric alignment.

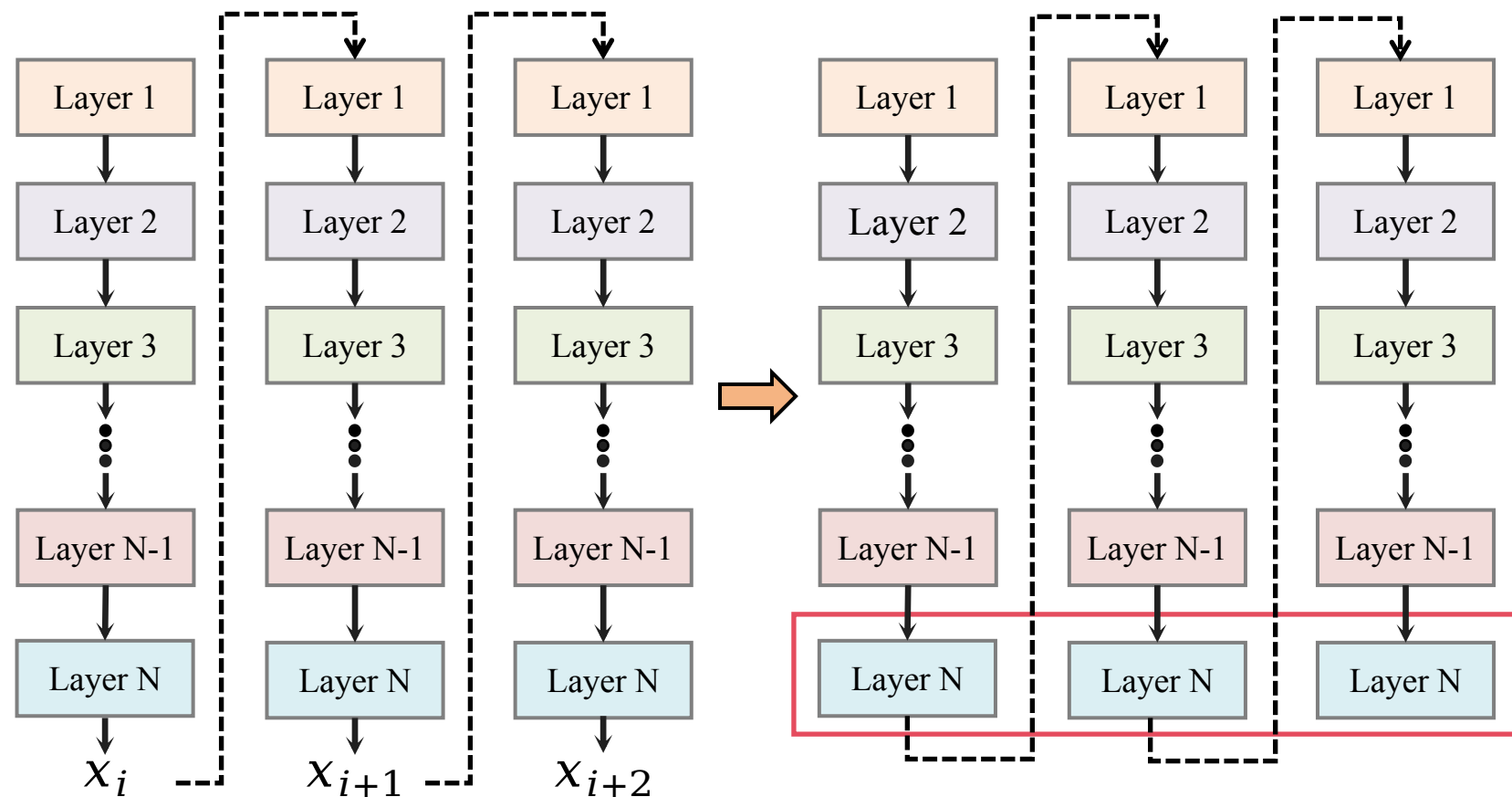
Compression process (Stage 2)

- Transmitting full-length latent space communication still requires full generation steps, offering no temporal speedup over text.
- Unlike natural-language tokens, continuous latent states are over-parameterized, making much of their temporal structure redundant.
- Our goal is therefore to learn an information-preserving bottleneck to aggressively compress latents while retaining task utility.

~~Generate full length~~
Retain core information while
maximizing inference
efficiency.

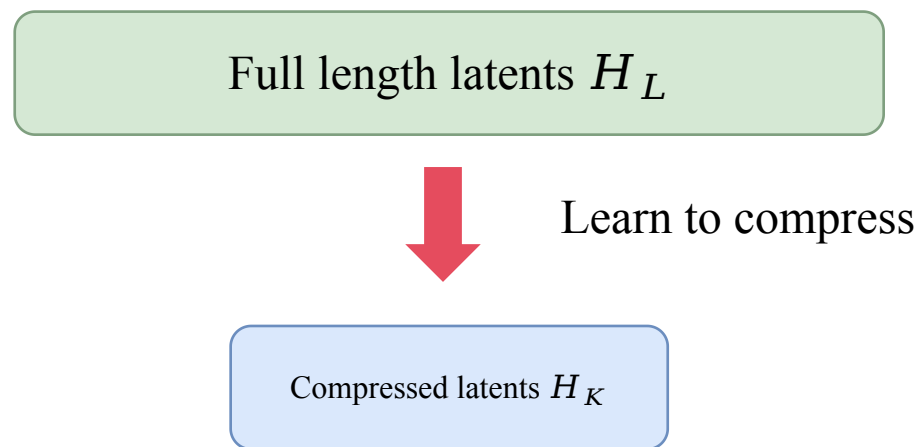
Compression via Latent-Space Reasoning

- Bypass the language head.
- Latent Autoregression
 - Feed the last hidden state directly back as the next input embedding
 - No discrete tokens; just latent states.



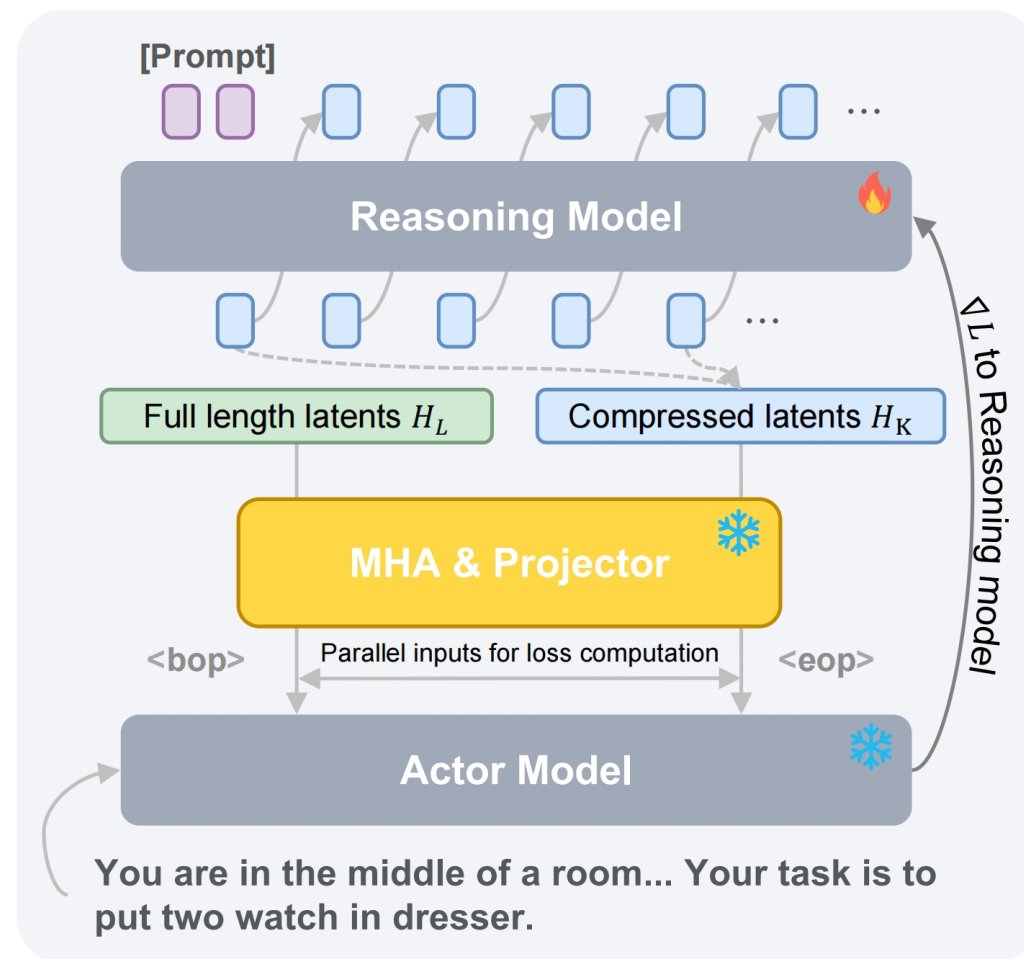
Compression design

- Find a way to let model forward less while maintain information.
- Train a separate reasoning model to generate compact messages directly in the latent space.



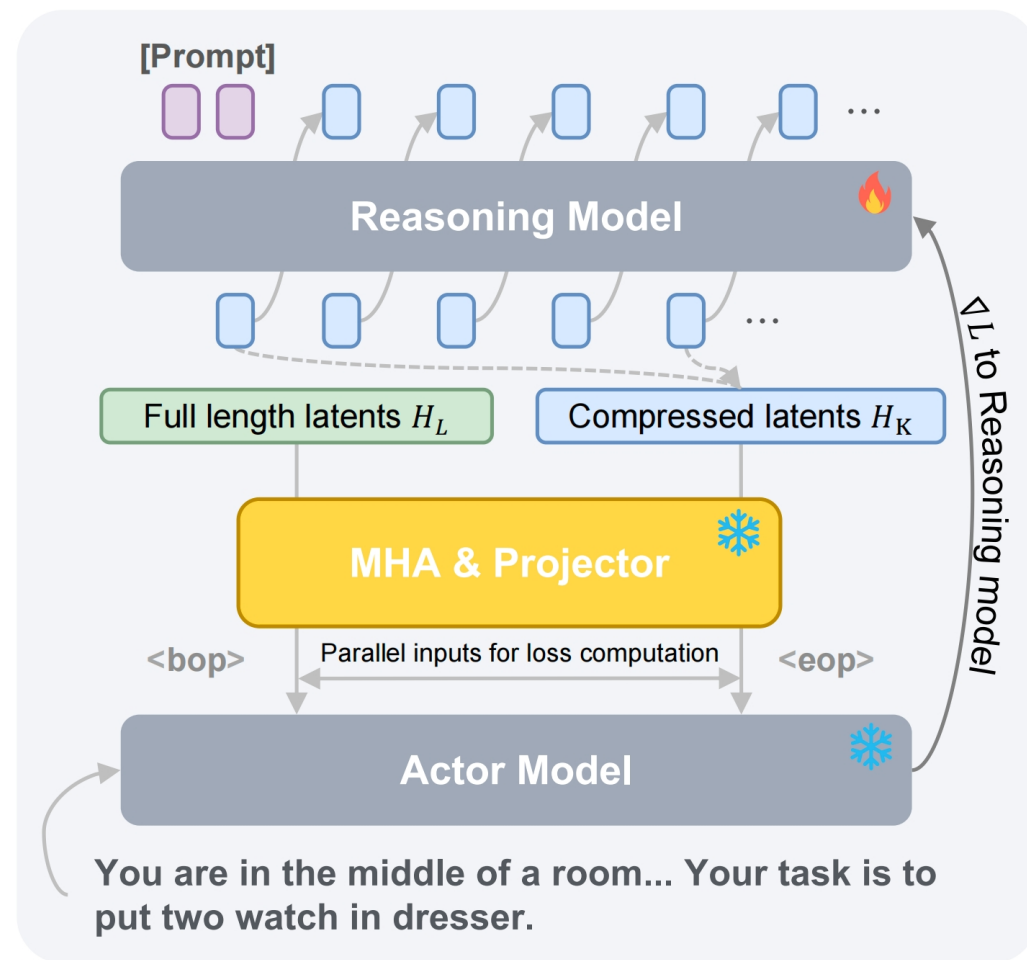
Compression design

- L_{task} (Task Utility): Ensures the compressed H_K still solves the task successfully.
- L_{pref} (Behavioral Agreement): Aligns the actor's behavior under compressed vs. full-length latents.
- L_{geom} (Geometric Alignment): Preserves the global semantic orientation of the original thought.



Compression design

- $L_{\text{pref}} : \mathcal{L}_{\text{pref}} = \sum_{t \in S} w_t KL(p_t^{(D)} || p_t^{(A)})$
 - Weights: $w_t \propto \max(\mathcal{H}(p_t^{(B)}) - \mathcal{H}(p_t^{(D)}), 0)$
 - Intuition: Focuses only on positions where latents meaningfully reduce predictive uncertainty.
- $L_{\text{geom}} : \mathcal{L}_{\text{geom}} = 1 - \cos(\bar{z}^{(A)}, \bar{z}^{(D)})$
 - $\bar{z}^{(A)}$ and $\bar{z}^{(D)}$ are the step-averaged directions of actor-side latent features.



Experiments Setup

Benchmarks & Tasks

- Alfworld: Multi-step embodied interactive planning (episodes capped at 20 steps), train dataset from [1].
- MATH: Complex symbolic reasoning.

Evaluated Models

- Actor Agents (Receivers): Qwen2.5-7B-Base, Qwen2.5-0.5B-Base, and LLaMA3.1-8B-Base.
- Reasoning Agents (Senders): Instruction-tuned variants for full CoT/latents generation; Base models for compression training.

Training Configurations

- Optimization: AdamW optimizer with a 3% linear warmup.
- Learning Rates: 1×10^{-5} for Actor tuning , 5×10^{-5} for Compression training.
- Infrastructure: bfloat16, FlashAttention-2, and DeepSpeed (ZeRO).

[1] Yifan Song, et al. Trial and error: Exploration-based trajectory optimization for llm agents. ACL 2024

Zhuoyun Du, et al. “Enabling Agents to Communicate Entirely in Latent Space”. ACL 2026.

Alfworld experiments

Method	Qwen2.5-7B-Base				Qwen2.5-0.5B-Base				LLaMA3.1-8B			
	Seen	Steps	UnSeen	Steps	Seen	Steps	UnSeen	Steps	Seen	Steps	UnSeen	Steps
Interlat												
Ours	70.48	9.41/12.54	65.42	9.86/13.37	61.19	10.55/14.22	57.46	9.38/13.90	70.71	8.02/12.58	70.90	8.21/12.96
Text	64.29	8.76/12.77	62.44	9.79/13.63	54.52	9.50/14.28	47.26	9.70/15.13	62.86	7.91/12.94	60.82	8.14/13.21
No-Comm	62.14	10.19/13.90	62.19	10.23/13.92	50.48	8.23/14.06	44.03	9.10/15.20	63.57	8.35/12.59	58.40	9.47/13.85
Baselines												
CoT (full)	67.14	8.15/12.04	64.93	9.02/12.87	57.86	8.30/13.23	50.75	8.94/14.39	69.35	7.62/12.32	70.82	7.88/12.47
No-CoT	65.71	8.23/12.27	62.69	9.15/13.20	57.14	8.96/13.69	50.25	9.80/14.87	67.18	7.85/12.61	70.34	8.02/12.88
Variants												
CrossTask	61.43	8.42/12.89	61.94	9.51/13.50	53.57	9.40/14.32	47.01	10.06/15.33	65.00	8.05/12.24	63.43	9.86/13.57
Noised												
CovNoise-0.5×	64.29	8.54/12.63	60.95	8.71/13.12	53.33	8.80/14.03	46.77	9.64/15.16	64.29	8.10/12.50	65.68	9.47/12.77
CovNoise-1.0×	63.81	8.66/12.76	63.68	8.72/12.82	53.10	8.96/14.14	44.53	9.68/15.40	58.57	7.80/12.80	64.93	9.66/13.28
WhiteNoise	61.90	8.65/12.97	61.19	9.32/13.46	57.38	8.00/13.11	57.21	9.18/13.81	61.43	8.01/12.64	64.93	9.52/13.19
CovGauss-0 μ	60.00	8.79/13.27	61.94	9.59/13.55	13.81	11.25/18.79	13.18	12.93/19.07	57.86	8.04/13.08	66.42	9.51/13.03
CovGauss- μ	65.71	8.58/12.50	64.93	8.63/12.62	44.52	9.21/15.20	34.33	10.19/16.63	60.71	7.69/12.53	64.93	8.85/12.76
RandomRot	57.86	8.43/13.31	63.68	9.37/13.23	59.05	8.24/13.06	51.99	9.12/14.34	57.86	7.67/12.86	63.44	9.04/13.25
Cross family												
Qwen2LLaMA	70.95	8.47/12.01	71.39	9.21/13.05	–	–	–	–	–	–	–	–

- Outperforms both language-based communication and single-agent CoT baselines.
- Promotes longer, yet highly successful trajectories, indicating structured exploratory behavior.

Alfworld experiments

Method	Qwen2.5-7B-Base				Qwen2.5-0.5B-Base				LLaMA3.1-8B			
	Seen	Steps	UnSeen	Steps	Seen	Steps	UnSeen	Steps	Seen	Steps	UnSeen	Steps
Interlat												
Ours	70.48	9.41/12.54	65.42	9.86/13.37	61.19	10.55/14.22	57.46	9.38/13.90	70.71	8.02/12.58	70.90	8.21/12.96
Text	64.29	8.76/12.77	62.44	9.79/13.63	54.52	9.50/14.28	47.26	9.70/15.13	62.86	7.91/12.94	60.82	8.14/13.21
No-Comm	62.14	10.19/13.90	62.19	10.23/13.92	50.48	8.23/14.06	44.03	9.10/15.20	63.57	8.35/12.59	58.40	9.47/13.85
Baselines												
CoT (full)	<u>67.14</u>	8.15/12.04	<u>64.93</u>	9.02/12.87	<u>57.86</u>	8.30/13.23	50.75	8.94/14.39	<u>69.35</u>	7.62/12.32	<u>70.82</u>	7.88/12.47
No-CoT	<u>65.71</u>	8.23/12.27	<u>62.69</u>	9.15/13.20	<u>57.14</u>	8.96/13.69	50.25	9.80/14.87	<u>67.18</u>	7.85/12.61	<u>70.34</u>	8.02/12.88
Variants												
CrossTask	61.43	8.42/12.89	61.94	9.51/13.50	53.57	9.40/14.32	47.01	10.06/15.33	65.00	8.05/12.24	63.43	9.86/13.57
Noised												
CovNoise-0.5×	64.29	8.54/12.63	60.95	8.71/13.12	53.33	8.80/14.03	46.77	9.64/15.16	64.29	8.10/12.50	65.68	9.47/12.77
CovNoise-1.0×	63.81	8.66/12.76	63.68	8.72/12.82	53.10	8.96/14.14	44.53	9.68/15.40	58.57	7.80/12.80	64.93	9.66/13.28
WhiteNoise	61.90	8.65/12.97	61.19	9.32/13.46	57.38	8.00/13.11	<u>57.21</u>	9.18/13.81	61.43	8.01/12.64	64.93	9.52/13.19
CovGauss-0 μ	60.00	8.79/13.27	61.94	9.59/13.55	13.81	11.25/18.79	<u>13.18</u>	12.93/19.07	57.86	8.04/13.08	66.42	9.51/13.03
CovGauss- μ	65.71	8.58/12.50	64.93	8.63/12.62	44.52	9.21/15.20	34.33	10.19/16.63	60.71	7.69/12.53	64.93	8.85/12.76
RandomRot	<u>57.86</u>	8.43/13.31	<u>63.68</u>	9.37/13.23	59.05	8.24/13.06	51.99	9.12/14.34	57.86	7.67/12.86	63.44	9.04/13.25
Cross family												
Qwen2LLaMA	70.95	8.47/12.01	71.39	9.21/13.05	–	–	–	–	–	–	–	–

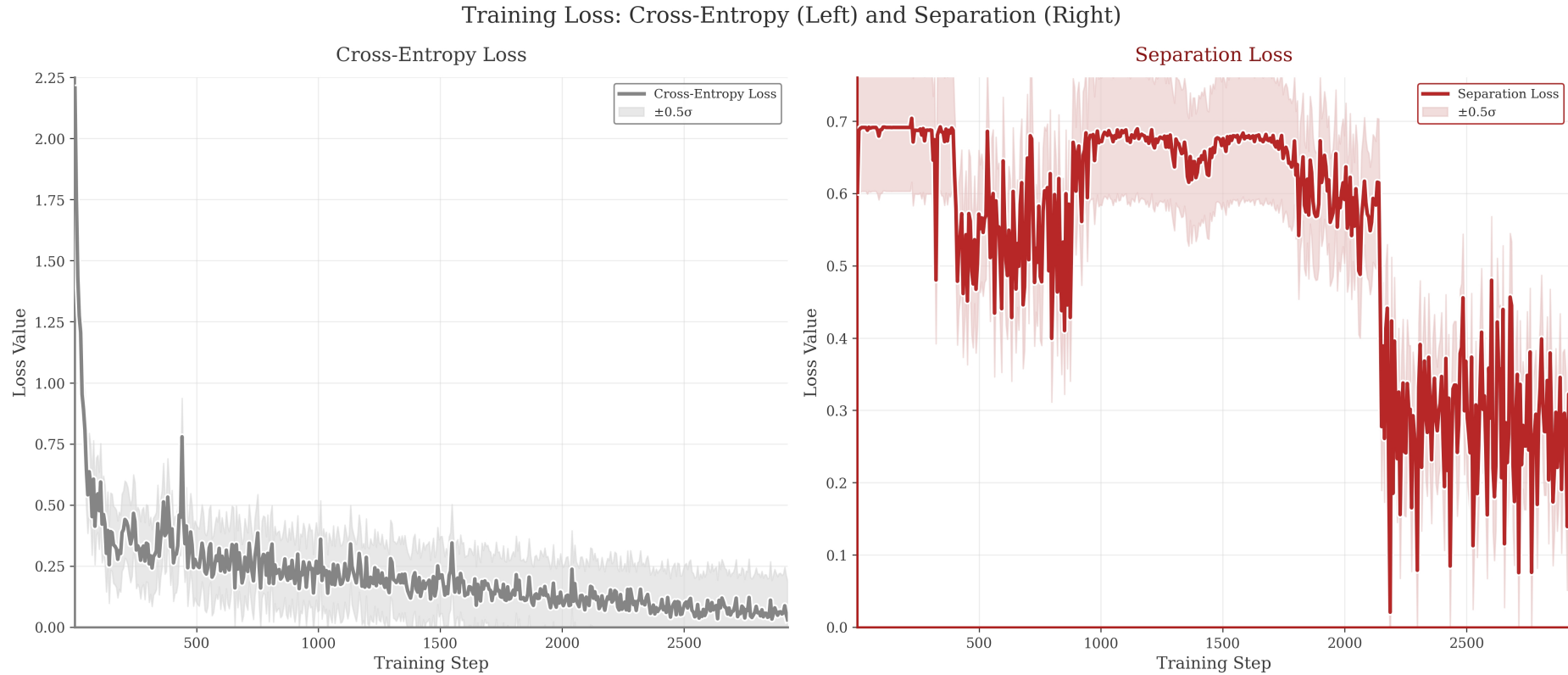
- Performance drops significantly under structured perturbations, proving reliance on semantic geometry rather than superficial moments.
- Exceptional cross-model communication ability

MATH experiments

Method	Overall	Level-3	Level-4	Level-5
Ours	<u>36.88</u>	40.08	<u>27.45</u>	15.80
Text	34.35	37.60	26.30	14.20
No-Comm	33.27	36.40	26.20	13.10
CoT (full)	38.35	45.65	31.19	15.05
No-CoT	36.25	<u>40.10</u>	26.80	14.80

- Surpasses the strong CoT baseline on the most challenging Level-5 tasks.
- Linearization regularizes simple tasks but causes "premature collapse" on complex ones.

‘Aha’ moment in training process

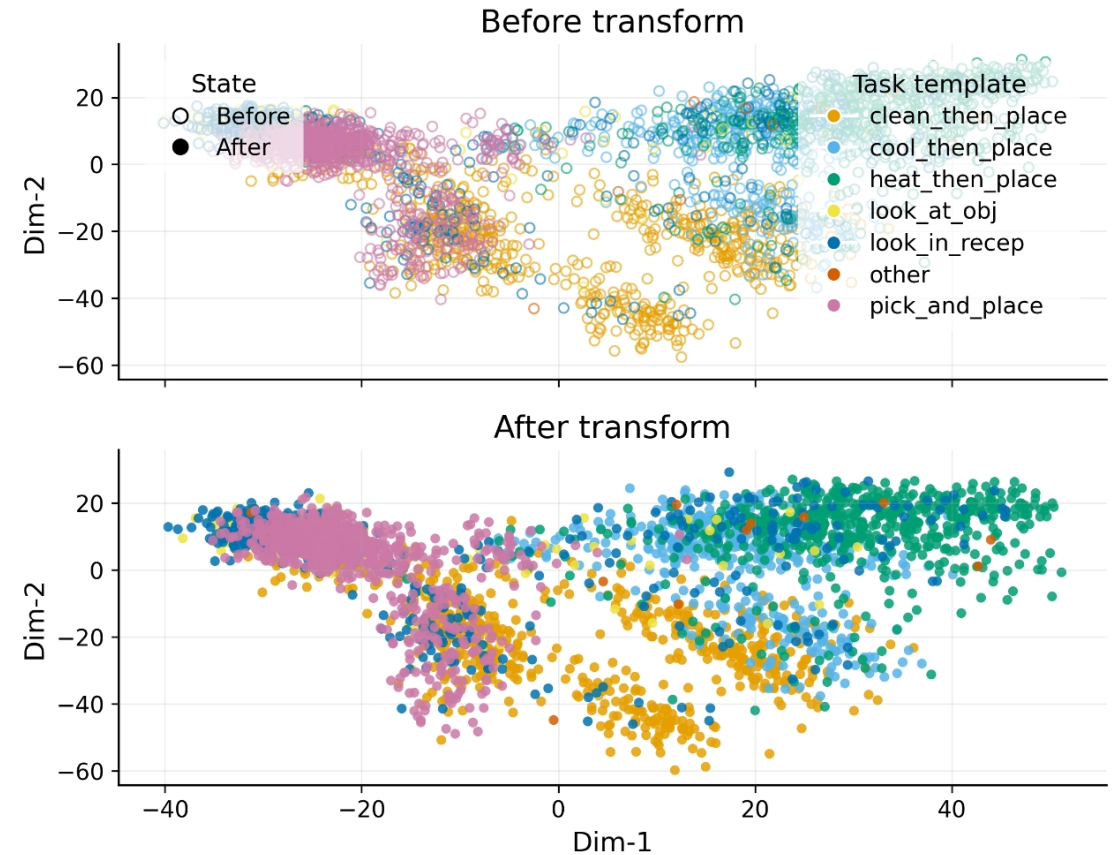


- Loss hovers near $\ln 2 \approx 0.69$ indicating no distinction between matched and mismatched latents.
- The Sharp Drop: A sudden plunge marks the model's exact "aha" moment.

Semantics in latent space communication

Structured Semantics:

- Task-Aware Clustering: Raw latent states naturally form distinguishable semantic structures based on task templates.
- Geometric Preservation: The communication adapter successfully retains and aligns these core semantic geometries.
- Not just noise—genuine semantic understanding.



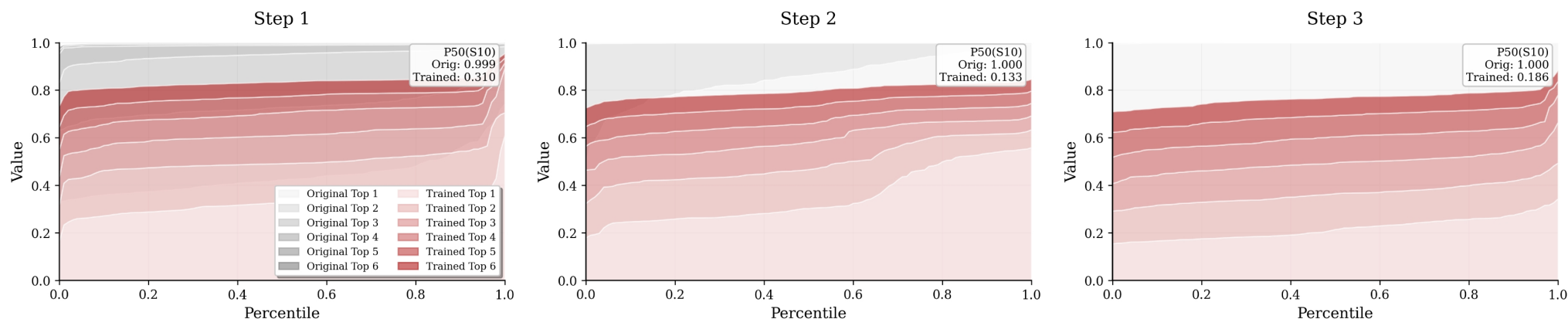
Compression experiments

Ratio	Seen	Unseen	Time
Untrained (An instruction-tuned Model)			
Full	<u>70.48</u> ± 1.01	<u>65.42</u> ± 0.87	9.19s
90%	68.57 ± 1.63	67.16 ± 1.97	-
80%	68.10 ± 1.83	61.69 ± 1.43	-
70%	67.14 ± 1.82	63.43 ± 2.24	-
60%	66.43 ± 1.63	59.20 ± 3.69	-
50%	72.14 ± 1.48	61.19 ± 2.84	-
40%	66.90 ± 2.31	59.95 ± 2.64	-
30%	65.95 ± 2.12	62.19 ± 1.58	-
20%	67.86 ± 3.23	61.44 ± 1.58	-
10%	67.86 ± 2.12	62.44 ± 2.64	-
5%	64.29 ± 1.12	60.95 ± 1.35	-
0%	62.14 ± 2.01	62.14 ± 2.32	-

Ratio	Seen	Unseen	Time
Untrained (An instruction-tuned Model)			
128L	64.55 ± 2.26	60.25 ± 2.06	3.55s
64L	66.23 ± 1.95	61.53 ± 4.32	1.83s
32L	63.57 ± 2.01	60.18 ± 3.58	1.03s
16L	64.29 ± 1.34	60.00 ± 3.01	0.62s
8L	64.00 ± 2.18	57.46 ± 2.69	0.39s
Trained			
128L	68.10 ± 1.93	62.94 ± 2.03	2.25s
64L	<u>67.14</u> ± 1.56	<u>61.94</u> ± 2.13	1.16s
32L	66.90 ± 1.46	61.94 ± 2.56	0.60s
16L	66.43 ± 2.05	61.69 ± 2.56	0.33s
8L	66.43 ± 1.22	60.45 ± 2.23	0.20s

- Untrained truncation degrades significantly under aggressive compression ratios.
- Trained models maintain competitive performance even at extreme lengths.
- Up to $24 \times$ Speedup

Why compression is effective.



- Avoids Probability Collapse: Untrained models quickly collapse to a single Top-1 trajectory.
- Maintains Parallel Hypotheses: Trained latents show stable gaps between Top-k bands across steps.
- Broader Distribution: Lower $P_{50}(S_{10})$ scores confirm the compressed latents successfully keep multiple reasoning options open.

Ablation study

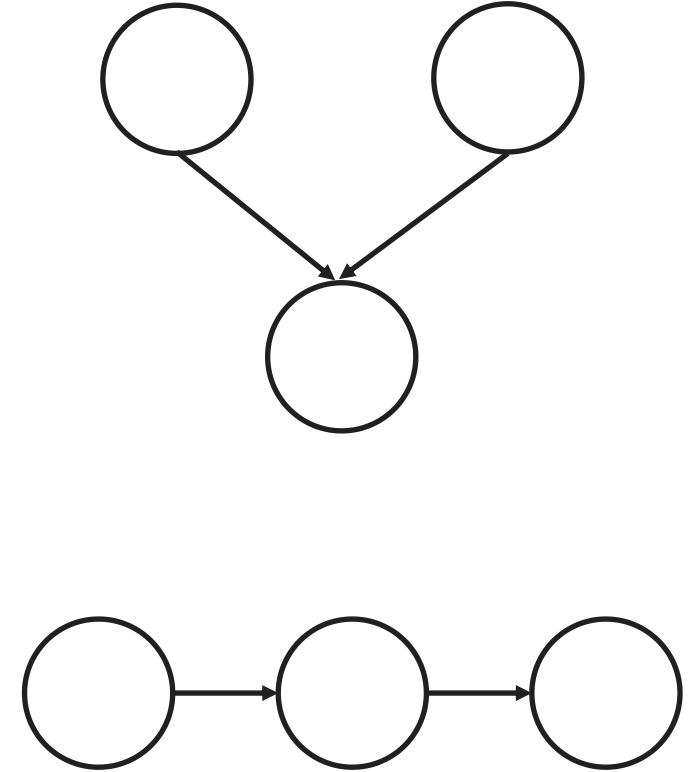
Actor	Seen	Unseen
Ours Full	70.48 ± 1.01	65.42 ± 0.87
w/o curri	33.10 ± 2.97	20.65 ± 2.15
w/o $\mathcal{L}_{\text{sep}}$	<u>58.81</u> ± 1.41	<u>60.70</u> ± 5.50
w/o $\mathcal{L}_{\text{align}}$	56.90 ± 1.41	53.98 ± 3.35
w/o adapter	4.05 ± 1.70	4.48 ± 1.31

Reasoning	Seen	Unseen
Ours Full	68.10 ± 1.93	<u>62.94</u> ± 2.03
w/o $\mathcal{L}_{\text{task}}$	65.71 ± 1.43	63.18 ± 3.47
w/o $\mathcal{L}_{\text{pref}}$	64.76 ± 2.97	60.20 ± 3.13
w/o $\mathcal{L}_{\text{geom}}$	64.05 ± 3.55	59.45 ± 3.01

- Removing the communication adapter completely cripples the actor's ability to solve tasks.
- Without curriculum learning, the actor fails to effectively learn how to interpret latent communications.
- The geometry alignment loss is the most vital component for the reasoning model to compress latents successfully.

Scale Interlat

- Experimental Setup:
 - Benchmark & Model: ALFWorld using Qwen2.5-0.5B-Base.
 - Communication: All non-output agents generate fixed-length (N=32) messages entirely in latent space.
- Sequential Chain (Left):
 - Strategist $\rightarrow$ Compiler $\rightarrow$ Actor: Passes information sequentially to progressively refine a high-level goal into actionable latent representations.
- Parallel Tree (Right):
 - [Explorer + Critic] $\rightarrow$ Actor: Parallel generation of diverse reasoning paths and verification constraints.
 - Fusion Mechanism: Dynamic merging via a projection-augmented adapter.



Scale Interlat

Communication	Topology (N)	Seen (%)	Unseen (%)	Structural Description
<i>Text (CoT)</i>	text-chain-2	57.86	50.75	Sequential: 1 Sender $\rightarrow$ 1 Actor
	text-chain-3	58.34	51.89	Sequential: Sender $\rightarrow$ Sender $\rightarrow$ Actor
	text-tree-3	61.19	56.75	Parallel: 2 Senders $\rightarrow$ Actor (Merged by agent)
<i>Latent (Ours)</i>	chain-2 (Base)	61.19	57.46	Sequential: 1 Sender $\rightarrow$ 1 Actor
	chain-3	62.14	59.44	Sequential: Sender $\rightarrow$ Sender $\rightarrow$ Actor
	tree-3	62.85	60.77	Parallel: 2 Senders $\rightarrow$ Actor (Adapter merged)

- Latent vs. Text: Latent topologies consistently outperform their natural language CoT counterparts.
- Scaling Up: 3-agent configurations improve upon the 2-agent Interlat baseline.
- Best Performer: The parallel tree topology with adapter merging achieves the highest success rates (62.85% Seen, 60.77% Unseen).

Conclusions

- Interlat:
 - Bypasses natural language constraints
 - Promotes effective exploration and higher success rates
 - Supports aggressive compression (up to $24 \times$ speedup)
- Future Directions
 - Scale to larger agent teams and develop tools for better interpretability and controllability

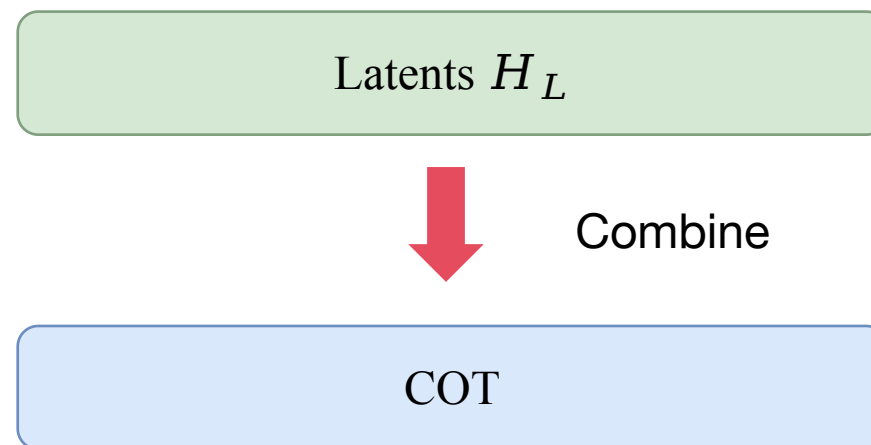
What's more?

Implications for E-Commerce Scenarios

Communication:

Hybrid Transmission

To preserve interpretability, agents can transmit natural language alongside hidden states, fusing both as joint inputs. Furthermore, we can train soft-prompts to serve as input prefixes, significantly enhancing the quality of agent generation.



What's more?

Memory: Dynamic Latent Memory

By aggregating and utilizing hidden states, we can maintain a dynamic agent memory (MemGen). This acts as a high-density, continuous representation prefix that is directly injected into the input stream.

Reasoning:

Latent-to-Text Efficiency

Customer service agents can execute short-step reasoning entirely within the latent space, and subsequently decode back to the natural language space for the final response, drastically reducing inference latency.

Reading Recommendations

Surveys:

- A Survey on Latent Reasoning
- [The Latent Space: Foundation, Evolution, Mechanism, Ability, and Outlook](#)

Latent space reasoning:

- Training large language models to reason in a continuous latent space
- Codi: Compressing chain-of-thought into continuous space via self-distillation
- Parallel Latent Reasoning for Sequential Recommendation
- Think silently, think fast: Dynamic latent compression of llm reasoning chains
- Token assorted: Mixing latent and text tokens for improved language model reasoning

Latent space communication:

- Thought communication in multiagent collaboration
- Cache-to-cache: Direct semantic communication between large language models
- Latent Collaboration in Multi-Agent Systems

There are also papers on vision in the latent space.

Resources

- Codes:
 - Code is available at: <https://github.com/XiaoDu-flying/Interlat>
- About me:
 - Homepage: <https://xiaodu-flying.github.io/>

Thank you!