Diffusion Language Models

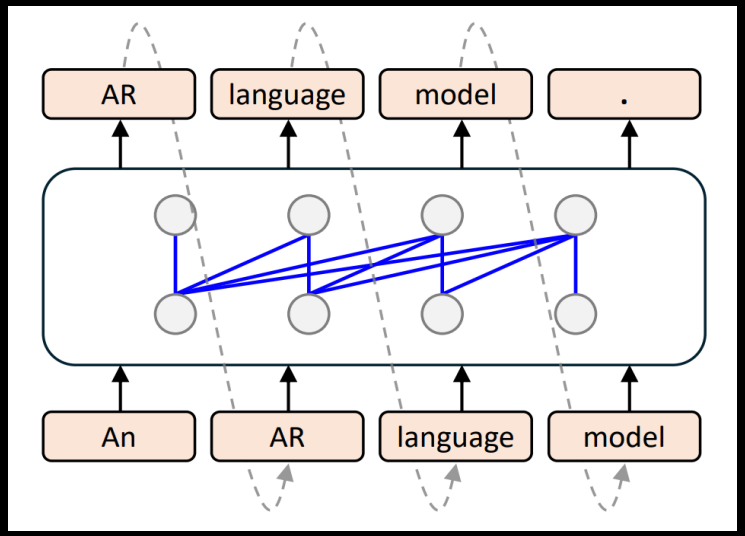
— Top 10 Open Challenges Steering the Future of Diffusion Language Model and Its Variants

https://noah-dllm.github.io/

Yunhe Wang

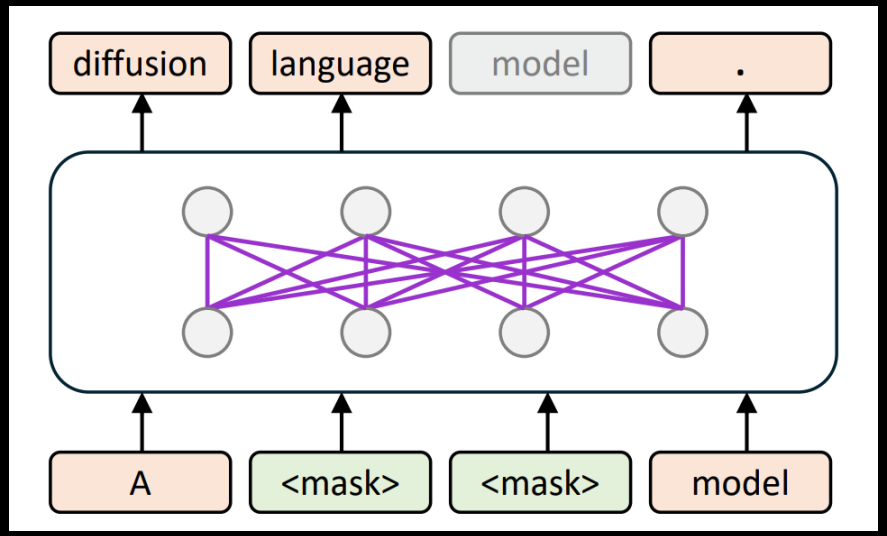
Huawei Noah's Ark Lab

What and Why Diffusion Models



Autoregressive Language Models

V.S.



Diffusion Language Models

Main technical concepts on diffusion models:

1. **Bidirectional attention:** process entire context simultaneously, enabling better global planning/coherence.
2. **Parallel decoding and efficiency:** generates many complete reasoning trajectories in parallel.
3. **Potential to unify the architecture:** can be applied to unify multimodal understanding and generation.

Figure from Zhang, Siyue et al., 2025

Next-Block Diffusion: A Principled Adaptation Path for Diffusion LLMs

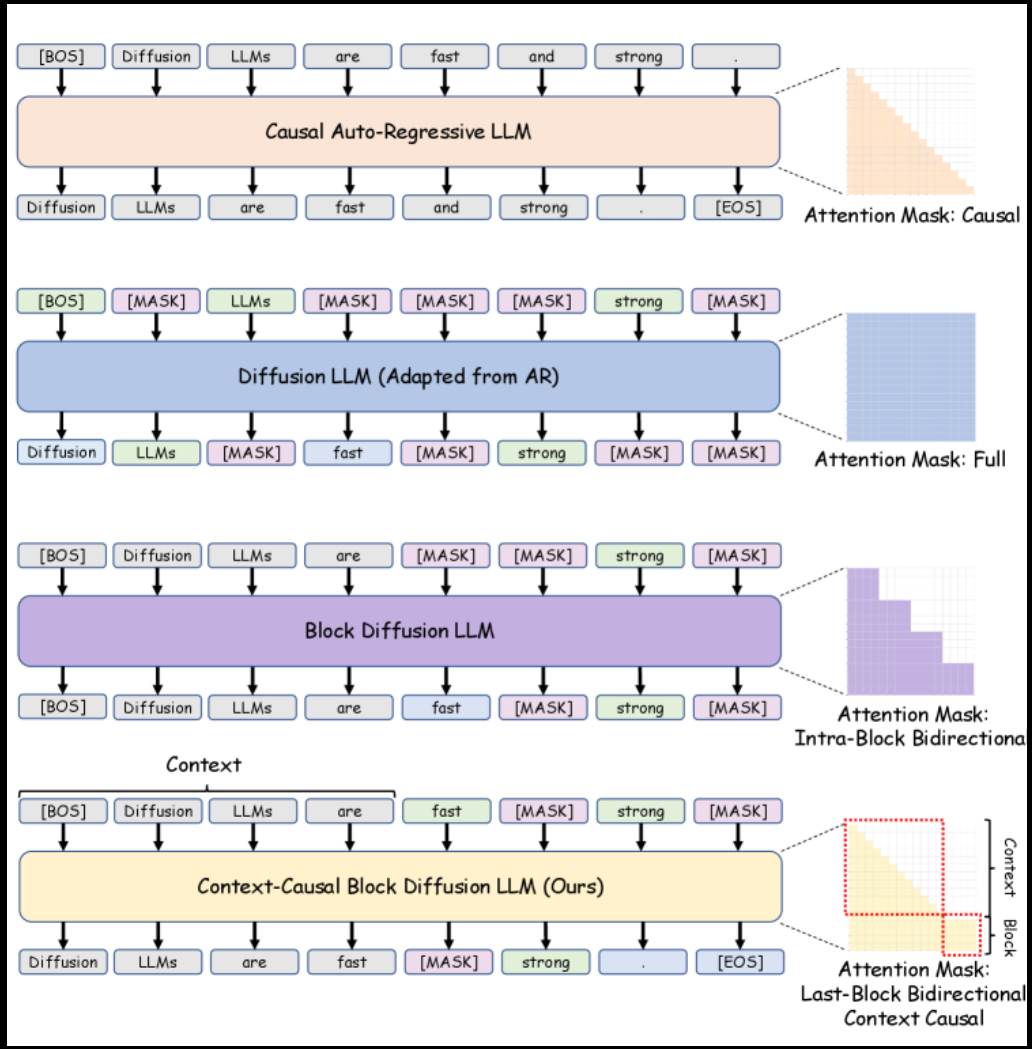
Our Goal: State-of-the-Art DiffLLM

Full-Sequence Diffusion :
➤ severe loss fluctuations

Block-Diffusion:
➤ Smoother loss, stabler training

Blockwise Causal Context :
➤ Harder to adapt from AR

AR Causal Context :
➤ Easier to adapt from AR

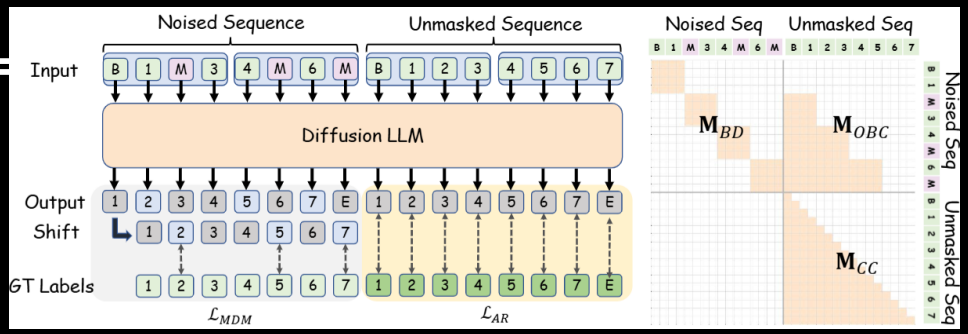


Next-Block Diffusion: A Principled Adaptation Path for Diffusion LLMs

Efficient Adaptation
Key: AR Models == Block-Diff with *blocksize=*

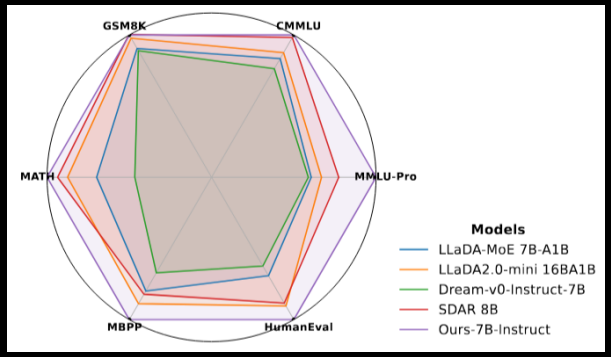
1.AR Loss regularization.

2.Gradual Block Growth.



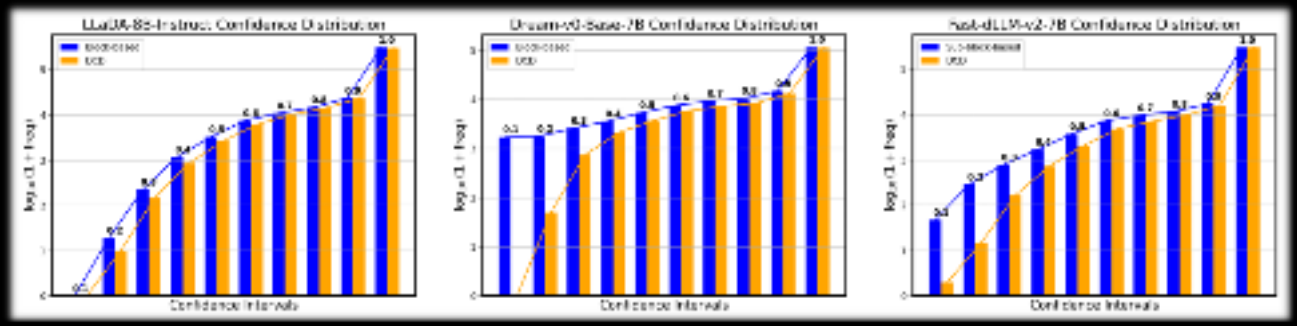
SOTA performance on math, coding, general benchmarks

1. GSM8K 91%; MATH 84%
2. MBPP 87%; HumanEval 89%



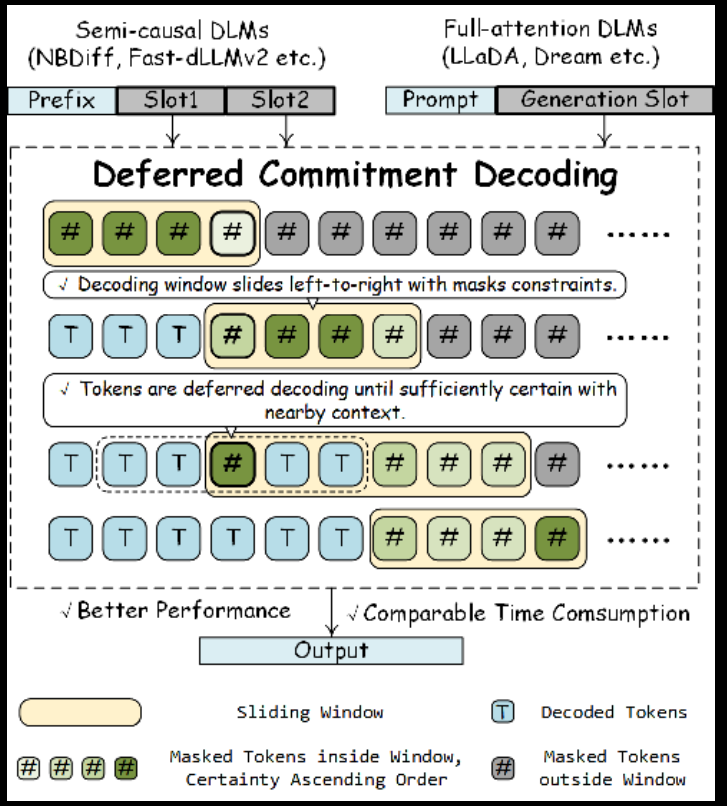
Deferred Commitment Decoding: Beyond Fixed Blocks in Diffusion LLMs

Problem with Block-Diffusion:
BICT => Block-Induced Context Truncation



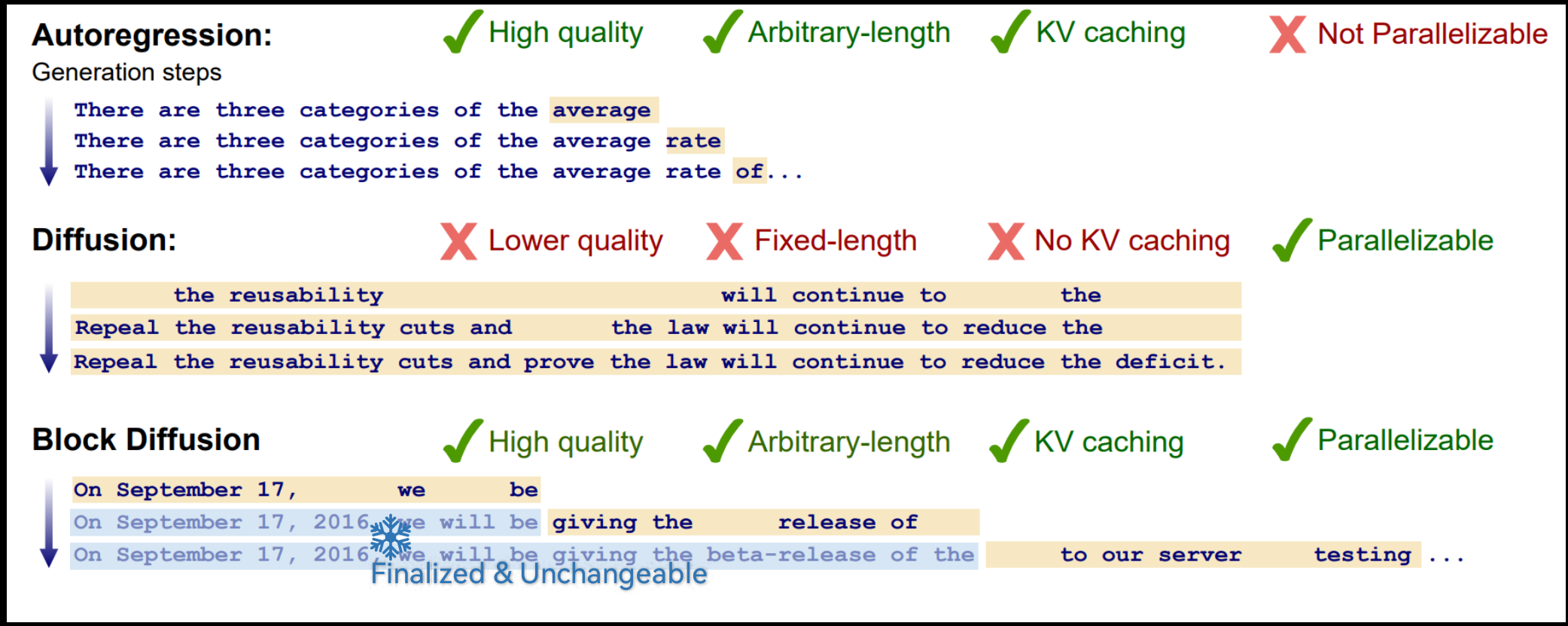
DCD improves DiffLLM confidence!

Our Solution:
DCD => Deferred Commitment Decoding



Diffusion in Diffusion: Breaking the Autoregressive Bottleneck in Block Diffusion Models

Myopia: Limited Lookahead.
Irrevocability: No mechanism to correct early errors.



Current models can generate, but they cannot correct.

Figure from Arriola et al., 2025

[1] https://arxiv.org/abs/2601.13599

Diffusion in Diffusion: Breaking the Autoregressive Bottleneck in Block Diffusion Models

Structual Block Diffusion:

✓ High-quality ✓ Global conetxt ✓ KV caching ✓ Controllable

Stage 1

```
On                  1977,
On Jan. 17, 1977, Montréal held      Show of      XV           , also
On Jan. 17, 1977, Montréal held the Show of the XV Olympiad, also known as The 1976 Summer Olympics
```

On Jan. 17, 1977, Montréal held the Show of the XV Olympiad, also known as The 1976 Summer Olympics ✗

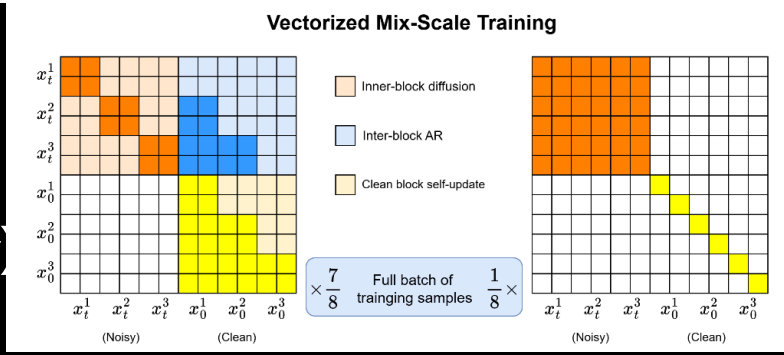
Confidence-Based Remasking

Stage 2

```
On                  , Montréal held the       of         Olympiad,                    1976 Summer Olympics
On July 17          , Montréal held the Games of         Olympiad,       known     The 1976 Summer Olympics
On July 17, 1976, Montréal held the Games of the XXI Olympiad, also known as The 1976 Summer Olympics
```

On July 17, 1976, Montréal held the Games of the XXI Olympiad, also known as The 1976 Summer Olympics ✓

Step 1: Drafting (Small Blocks)
Step 2: Revising (Global Blocks)
Key: Mixed-Scale Training (For single model capability)



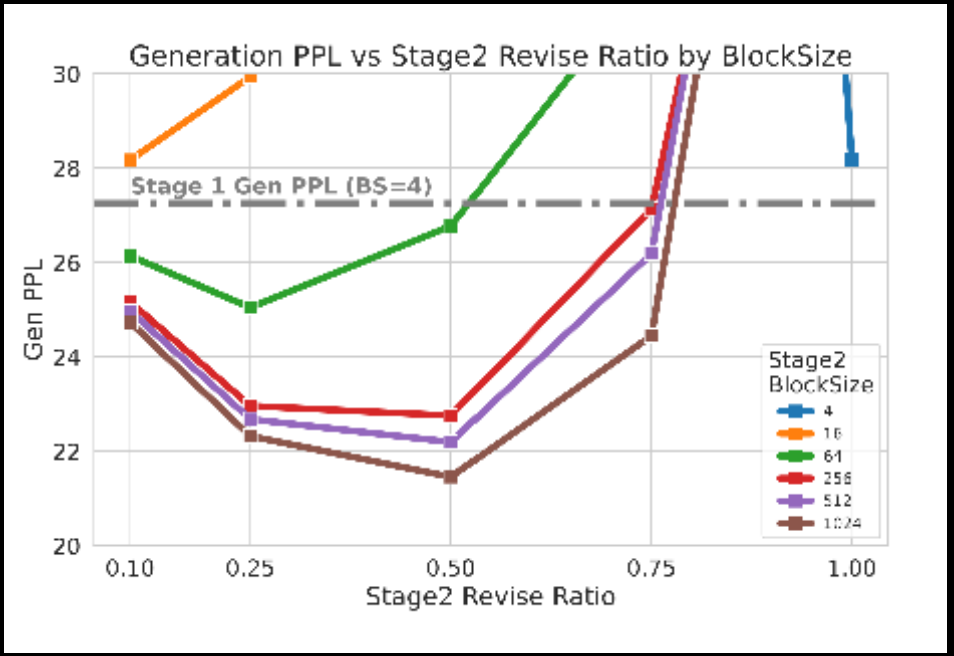
Diffusion in Diffusion: Breaking the Autoregressive Bottleneck in Block Diffusion Models

21.9 PPL (New SOTA)

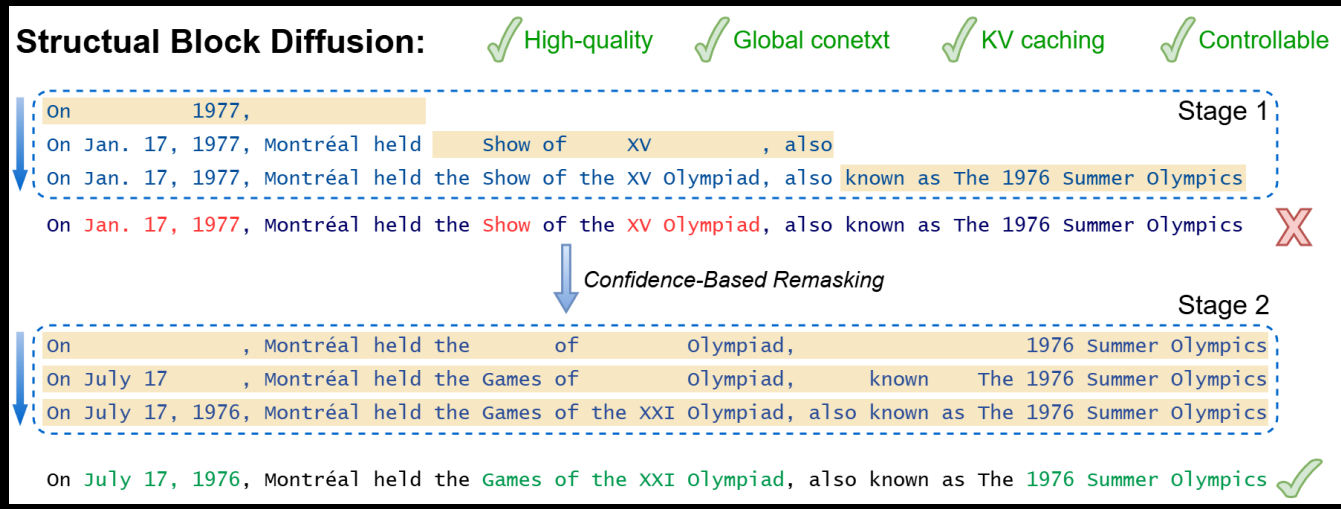
26% Data (Training Budget)

Model	$L = 1024$ Gen. PPL	$L = 1024$ NFEs	$L = 2048$ Gen. PPL	$L = 2048$ NFEs
AR	***14.1***	1K	***13.2***	2K
SEDD	52.0	1K	41.3	2K
MDLM	46.8	1K	35.3	2K
Prior Block Diffusion Work				
SSD-LM ($L' = 25$)	37.2	40K	35.3	80K
BD3-LM ($L' = 16$)	33.4	1K	31.5	2K
BD3-LM ($L' = 8$)	30.4	1K	28.2	2K
BD3-LM ($L' = 4$)	25.7	1K	23.6	2K
	25.0	1.5K	22.8	3K
Structural Block Diffusion (Using 26% Tuning Data)				
Ours (Stage 1 only)	27.4	1.0K	25.1	2.0K
Ours (Full 2-Stage)	24.6	1.1K	22.5	2.2K
	22.6	1.2K	21.2	2.5K
	21.9	1.5K	**20.6**	3.0K

Bigger Revise Window = Better



Diffusion in Diffusion: Breaking the Autoregressive Bottleneck in Block Diffusion Models



Model	$L = 1024$ Gen. PPL	$L = 1024$ NFEs	$L = 2048$ Gen. PPL	$L = 2048$ NFEs
AR	***14.1***	1K	***13.2***	2K
SEDD	52.0	1K	41.3	2K
MDLM	46.8	1K	35.3	2K
Prior Block Diffusion Work				
SSD-LM ($L' = 25$)	37.2	40K	35.3	80K
BD3-LM ($L' = 16$)	33.4	1K	31.5	2K
BD3-LM ($L' = 8$)	30.4	1K	28.2	2K
BD3-LM ($L' = 4$)	25.7	1K	23.6	2K
	25.0	1.5K	22.8	3K
Structural Block Diffusion (Using 26% Tuning Data)				
Ours (Stage 1 only)	27.4	1.0K	25.1	2.0K
Ours (Full 2-Stage)	24.6	1.1K	22.5	2.2K
	22.6	1.2K	21.2	2.5K
	21.9	1.5K	**20.6**	3.0K

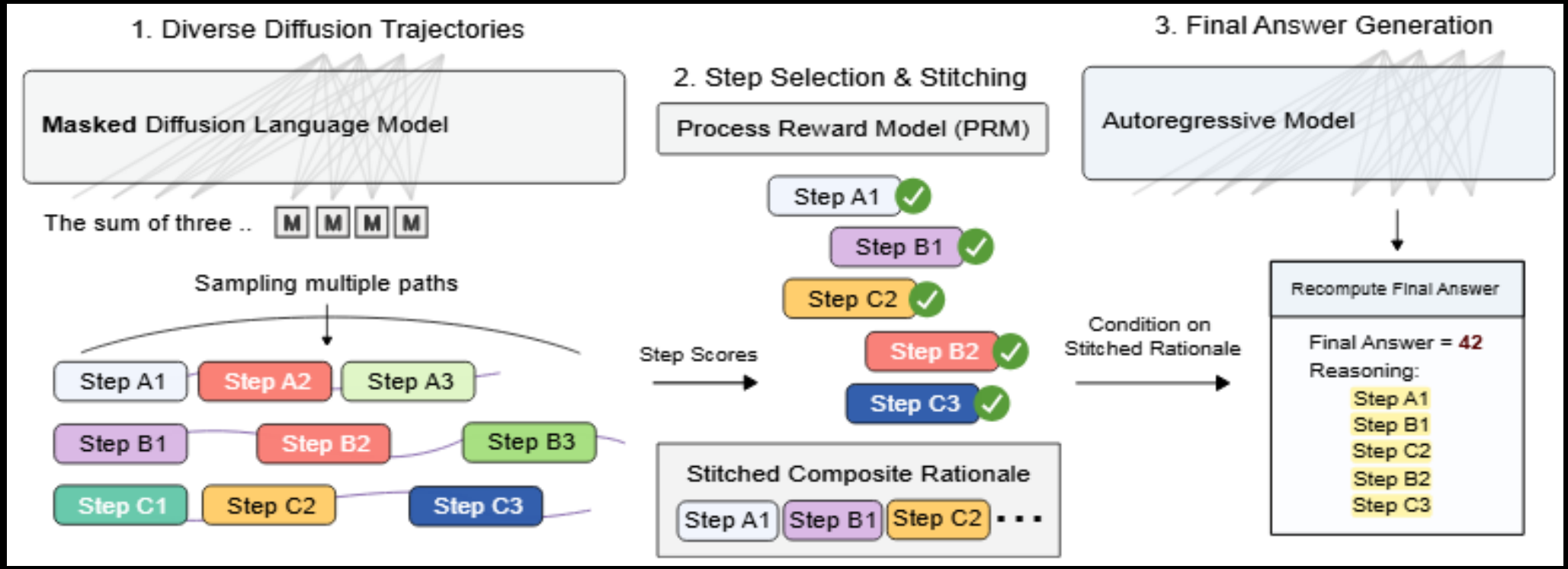
Summary:

- ✓ New Paradigm
- ✓ Effectiveness and Efficiency
- ✓ Controllable Revise Strengths

Next:

- ☐ Scaling UP
- ☐ Adaptive Recursive Refinement
- ☐ Towards “System 2” for Diffusion:

Diffusion and AR Combination: Stitching Noisy Diffusion Thoughts for Better Reasoning



Sample → Score → Stitch → Refine

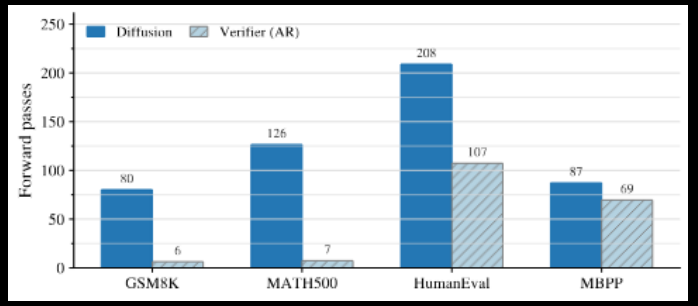
Diffusion and AR Combination: Stitching Noisy Diffusion Thoughts for Better Reasoning

		Coding				Math		Avg
Model Arch	**Size**	**HumanEval**	**HumanEval+**	**MBPP**	**MBPP+**	**GSM8k**	**MATH500**	
Qwen3	4B	57.32	50.61	67.00	80.69	77.48	-	-
Qwen3	8B	64.63	56.71	69.40	83.07	81.80	-	-
LLaDA	8B	32.32	27.44	40.80	51.85	70.96	36.2	43.3
Dream	7B	54.88	49.39	56.80	74.60	77.18	-	-
Block Diff	4B†	56.10	51.22	54.60	69.84	82.87	-	-
TiDAR (Trust AR)	8B‡	55.49	52.44	65.40	79.63	79.83	-	-
TiDAR (Trust Diff)	8B‡	57.93	55.49	65.40	80.95	80.44	-	-
Ours	8B	**70.37**	**64.02**	**74.61**	**81.75**	**90.30**	**55.0**	72.7

Model	**GSM8K**	**MATH500**	**Avg**
Qwen-Math-Instruct	9(256)40)	7(256)504)	
Phi-4	92.4 (295)	70.8 (791)	
Baselines			
LLaDA-1.5	84.0 (101)	38.0 (163)	
Dream	82.6 (512)	48.2 (512)	
LLaDA 2.0	88.0 (256)	44.6 (256)	
LLaDA 2.0	88.2 (1024)	65.6 (1024)	
LLaDA 2.0	90.1 (2048)	73.2 (2048)	
with Stitching			
LLaDA	90.3 (86)	55.0 (133)	
LLaDA-1.5	90.8 (83)	53.8 (130)	

5x Efficiency, Higher Accuracy.

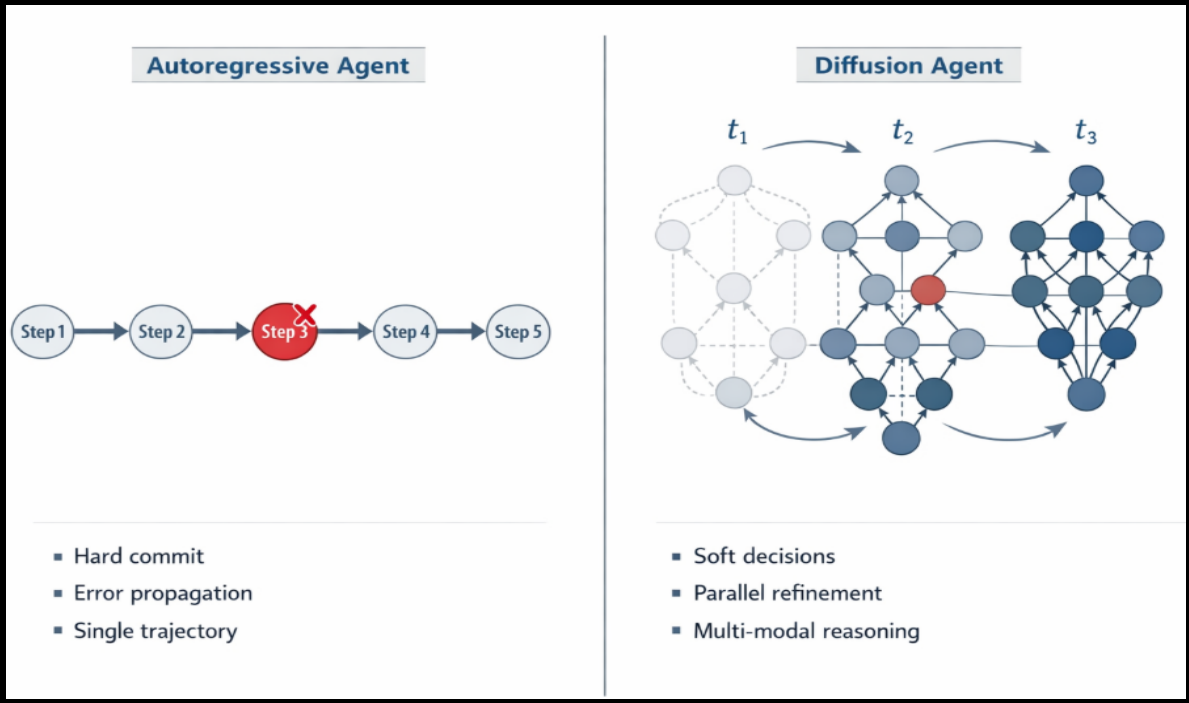
	GSM8K (0-shot)		MATH500 (0-shot)		Countdown (0-shot)	
Model / Seq Len	256	512	256	512	256	512
LLaDA	76.7	78.2	32.4	36.2	19.5	16.0
+ SFT	78.8	81.1	32.6	34.8	14.5	23.8
d1-LLaDA	81.1	82.1	38.6	40.2	32.0	42.2
Ours	**90.2**	**90.3**	**47.0**	**55.0**	**37.5**	**45.7**



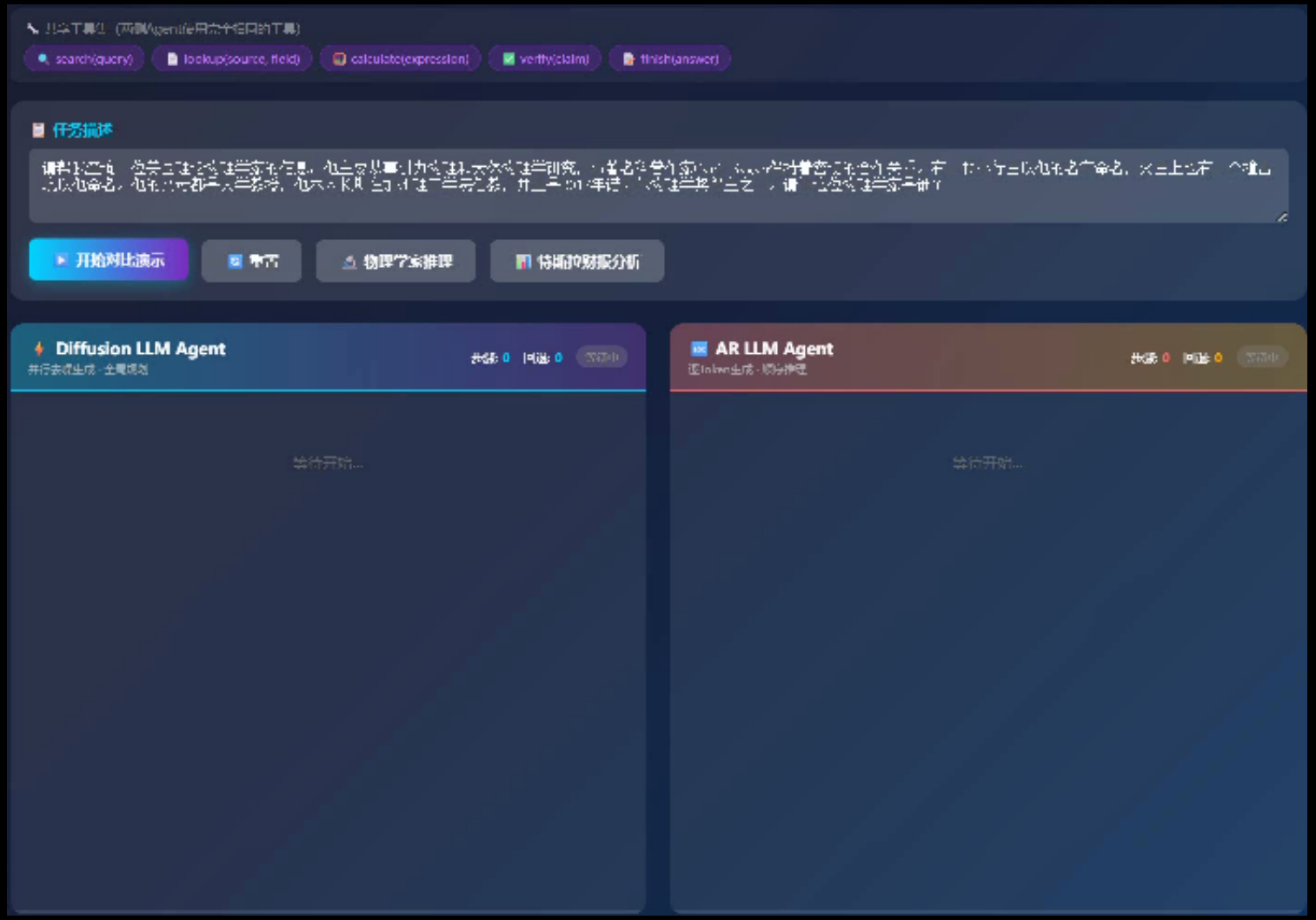
Diffusion Model for Deep Research Agent: Motivation

AR Limit: Linear & Irreversible.

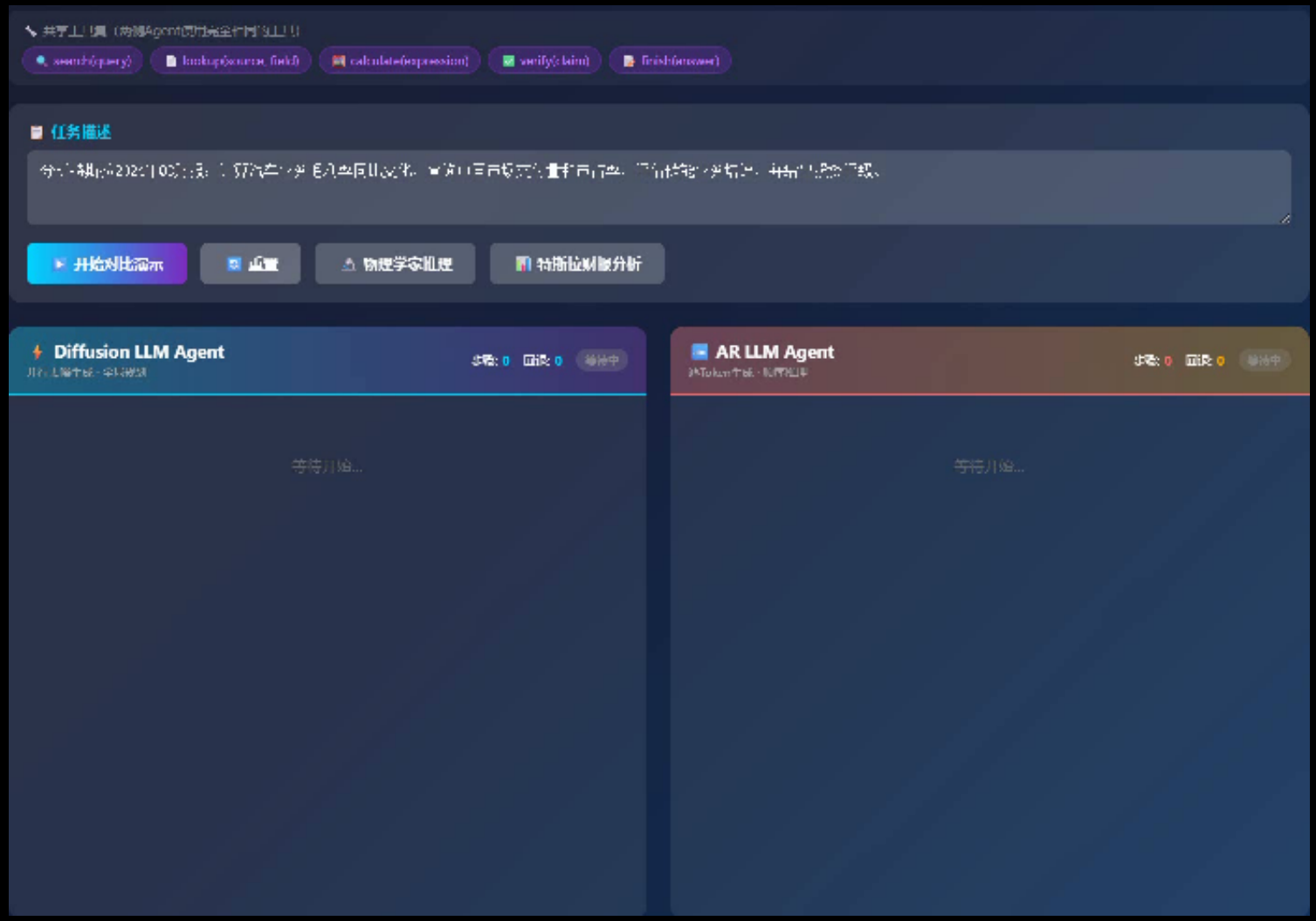
Diffusion Advantage: Global & Revisable.



Diffusion Model for Deep Research: Demo on Constraint Searching



Diffusion Model for Deep Research: Demo on Report Generation



Top 10 Open Challenges Steering the Future of Diffusion Language Models and Its Variants

Top 10 Open Challenges Steering the Future of Diffusion Language Models and Its Variants

Yunhe Wang[1], Kai Han[1], Huiling Zhen[1], Yuchuan Tian[2], Yufei Cui[1], Hanting Chen[1], Yongbing Huang[1], Dacheng Tao[3]

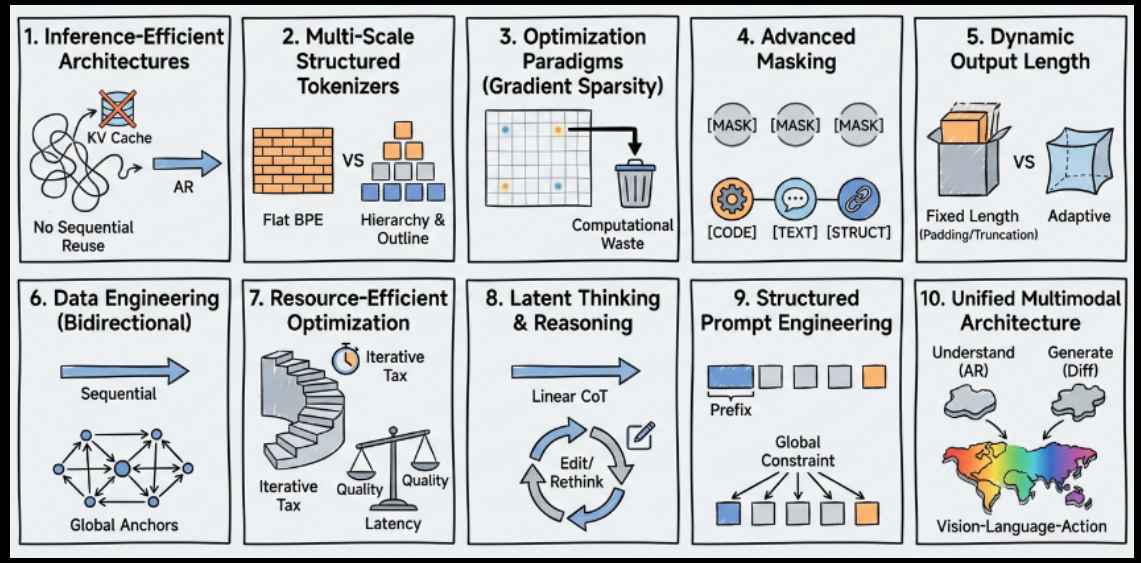
[1]Huawei Noah's Ark Lab, Beijing, China.
[2]Peking University, Beijing, China.
[3]College of Computing & Data Science, Nanyang Technological University, Singapore.

Contributing authors: yunhe.wang@huawei.com;

Abstract

The paradigm of Large Language Models (LLMs) is currently defined by auto-regressive (AR) architectures, which generate text through a sequential "brick-by-brick" process. Despite their success, AR models are inherently constrained by a causal bottleneck that limits global structural foresight and iterative refinement. Diffusion Language Models (DLMs) offer a transformative alternative, conceptualizing text generation as a holistic, bidirectional denoising process akin to a sculptor refining a masterpiece. However, the potential of DLMs remains largely untapped as they are frequently confined within AR-legacy infrastructures and optimization frameworks. In this Perspective, we identify ten fundamental challenges ranging from architectural inertia and gradient sparsity to the limitations of linear reasoning that prevent DLMs from reaching their "GPT-4 moment." We propose a strategic roadmap organized into four pillars: foundational infrastructure, algorithmic optimization, cognitive reasoning, and unified multimodal intelligence. By shifting toward a "diffusion-native" ecosystem characterized by pyramid tokenization, active remasking, and latent thinking, we can move beyond the constraints of the causal horizon. We argue that this transition is essential for developing next-generation AI capable of complex structural reasoning, dynamic self-correction, and seamless multimodal integration.

Keywords: Large Language Models, Diffusion Models, Transformers



Arxiv: https://arxiv.org/abs/2601.14041

Website: https://noah-dllm.github.io/

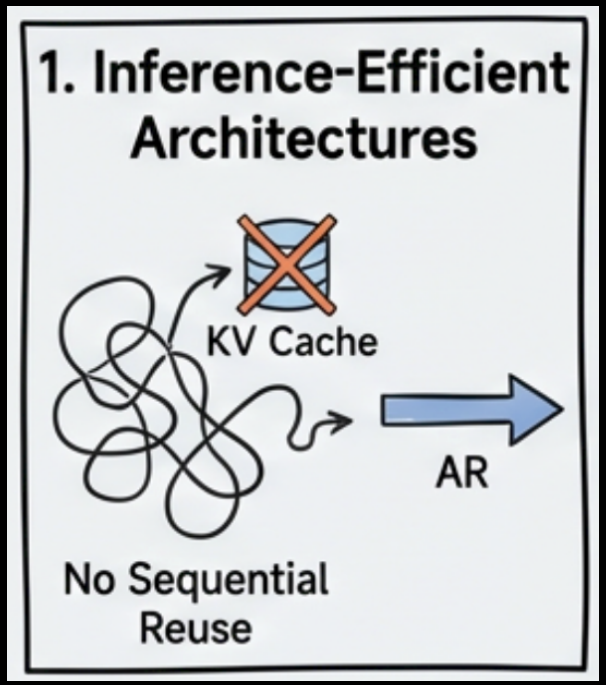
Challenge 1: Efficient Attention and Model Architectures for KV Cache

Problem:

- Inefficient Architecture Inheritance: Current diffusion models inefficiently borrow Autoregressive (AR) attention frameworks—specifically QKV computations and KV caching—which limits inference efficiency.
- Failure of KV Cache Reuse: Unlike AR's next-token prediction, the randomness of mask positions in diffusion prevents valid KV cache reuse. This structural mismatch significantly hinders high inference speeds and wider model adoption.

Potential idea:

- Enforced Left-to-Right Denoising: By enforcing a strict left-to-right denoising order for mask tokens, the system can successfully enable AR-style KV cache reuse and facilitate attention strategies natively designed for diffusion.



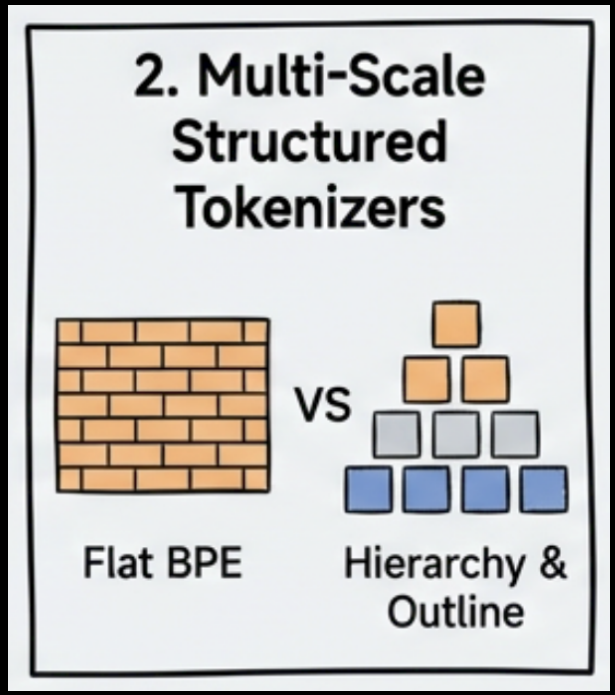
Challenge 2: Specific Tokenizer for Diffusion Language Models

Problem:

- Uniform Granularity: Current AR encoders rely on statistical splitting algorithms that produce tokens of a single, uniform scale.
- Lack of Structure: Existing tokenizers fail to mimic the multi-scale nature of human cognition, which naturally moves from global outlines to local details.
- Incompatibility with Diffusion: The standard AR paradigm is unsuitable for diffusion models, which require a more hierarchical approach to handle different levels of content generation.

Potential Idea:

- Vocabulary Pyramid: Implement a structured tokenizer with overlapping scales, where different tokens manage paragraph-level connections versus detailed local edits.
- Global-to-Local Paradigm: Adopt a multi-scale processing approach—inspired by image generation—that adjusts training and inference to leverage this hierarchical vocabulary.



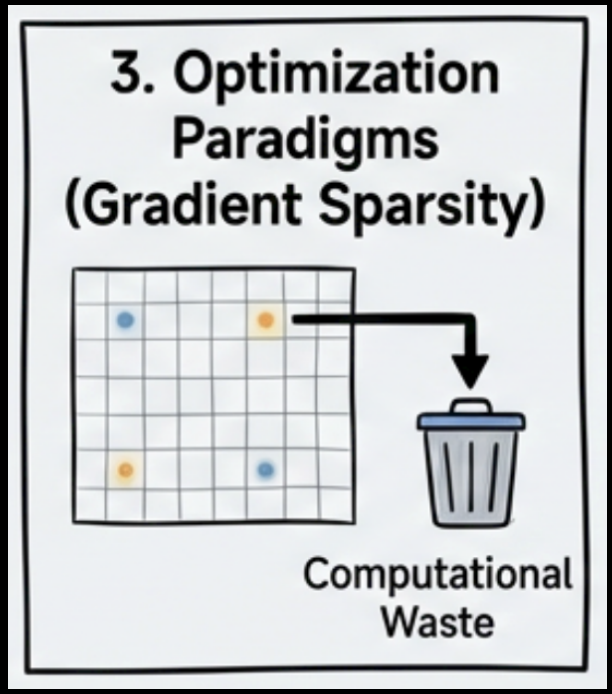
Challenge 3: Novel Training and Optimization Paradigms

Problem:

- Computational Inefficiency: Gradient calculation is highly inefficient; even with a single masked token in a long sequence (e.g., 128k tokens), a full forward and backward pass is required, leading to high costs and lower accuracy compared to AR models.
- Training Stage Inconsistency: There is a significant mismatch between pre-training (random masks) and SFT (full-answer masking), which creates difficulties for downstream Reinforcement Learning.
- Masking Ratio Complexity: Determining the optimal masking ratio and managing its dynamic adjustments across different training stages remains a critical, unresolved challenge.

Potential Idea:

- Enhanced Stage Fusion: Improve training outcomes by introducing more fusion schemes during the Pre-training and SFT stages to bridge the transition between them more effectively.



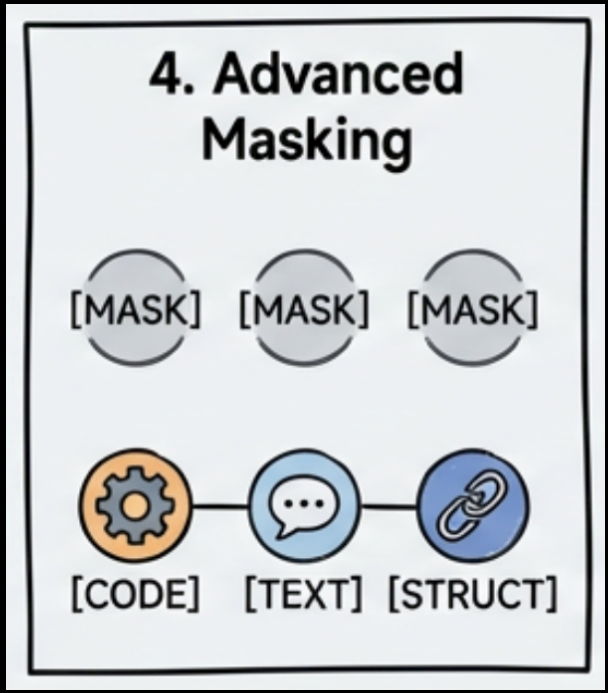
Challenge 4: Masking Strategies in Diffusion Language Models

Problem:

- Limited Token Diversity: Mainstream diffusion models rely on a single mask tokenizer, treating all positions identically and restricting functional versatility.
- Uniform Processing Probability: The current paradigm lacks granularity, processing all image or data areas with equal probability regardless of their complexity.
- Lack of Structural Linkage: Masked positions operate independently without a structured mechanism or spatial connectivity between them.

Potential idea:

- Enhanced Masking Architecture: Transition from a single token to multiple, interacting mask tokens integrated with prior assumptions to improve efficiency.
- Structured Mechanisms: Implement advanced, structured masking methods specifically designed to handle complex scenarios like code generation.



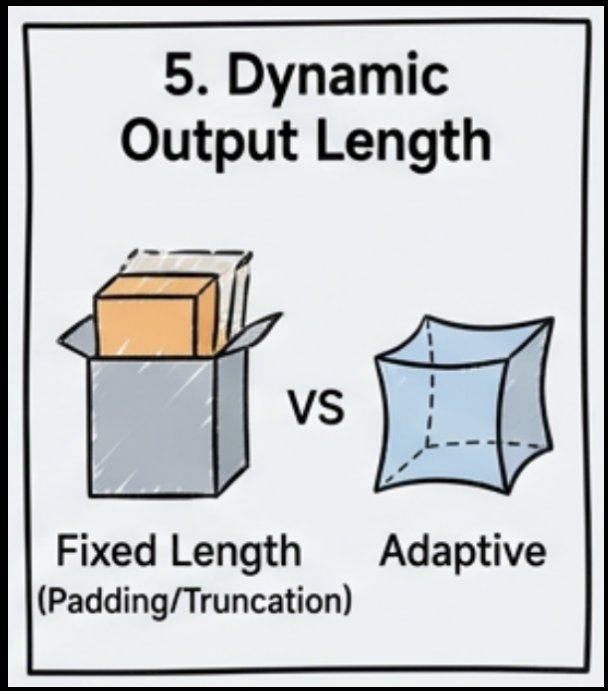
Challenge 5: Dynamic Output Length for Better Inference

Problem:

- Fixed Length Constraint: Diffusion models typically require a pre-defined sequence length and an EOS token, making it difficult to generate content that naturally falls outside those specific limits.
- Training Inefficiency: While training with variable lengths offers some adaptability, it remains inefficient for handling extreme edge cases or specific comparative logic.
- Static Inference: Models struggle to adaptively determine the optimal output length based on the specific context of a given question.

Potential Idea:

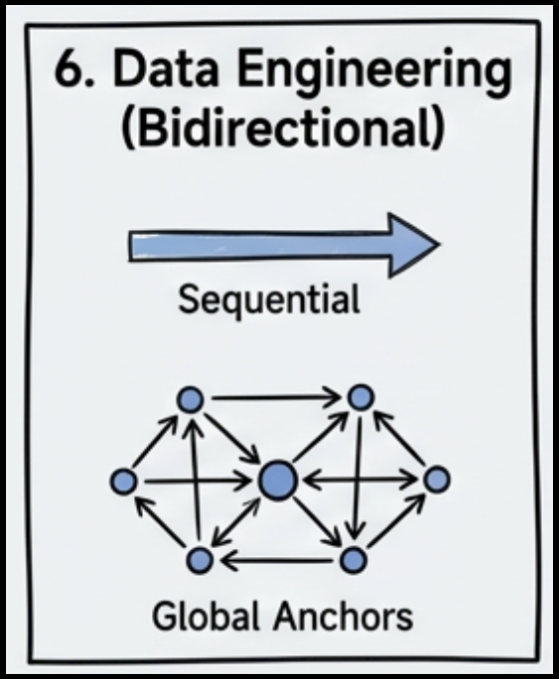
- EOS Position Prediction: Integrate EOS position forecasting during training to allow the model to perceive the required length during inference and minimize redundant computation.
- Parameter Reuse: Employ parameter reuse techniques to enable effective extrapolation when the required output exceeds the initial length.



Challenge 6: Data Engineering for Diffusion Models

Problem:

- Dependency on AR Paradigms: Current diffusion models largely reuse data and techniques optimized for Auto-Regressive (AR) models, which limits their potential for reasoning and structural knowledge.
- Optimization Mismatch: There is a significant technical challenge in tailoring data to align specifically with the random mask-token characteristics inherent to diffusion training.



Potential idea:

- Targeted Data Annotation: Improve learning efficiency and performance by incorporating mask position information and structured annotations into pre-training, SFT, and RL data.

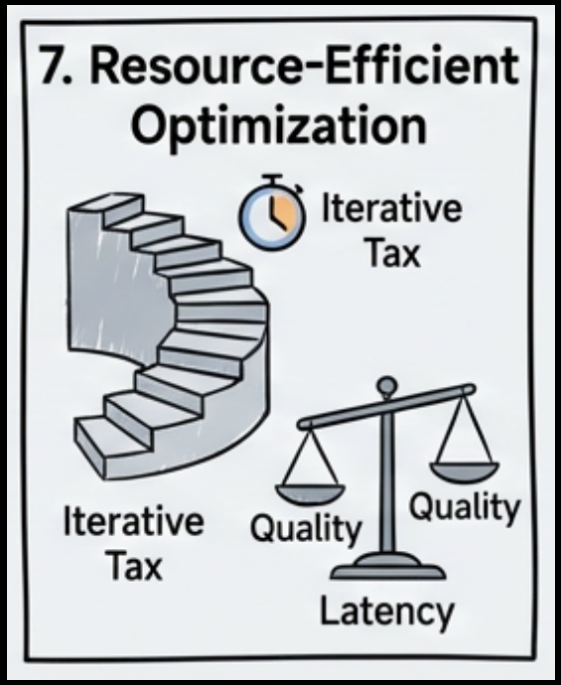
Challenge 7: Resource-Efficient Model Optimization and Compression

Problem:

- Inference Efficiency Gap: Global diffusion inference is currently less efficient than Autoregressive (AR) models as batch sizes increase.
- Denoising Overhead: The inherent complexity of the denoising process limits performance across all diffusion formats.
- Benchmark Performance: Current benchmarks indicate that AR models still maintain a clear competitive advantage over diffusion.

Potential idea:

- Optimization Techniques: Prioritize multi-step distillation, speculative inference, and low-bit post-quantization to reduce overhead.
- Hybrid Integration: Combine Diffusion for long-sequence single-batch tasks with AR for large-batch information integration.



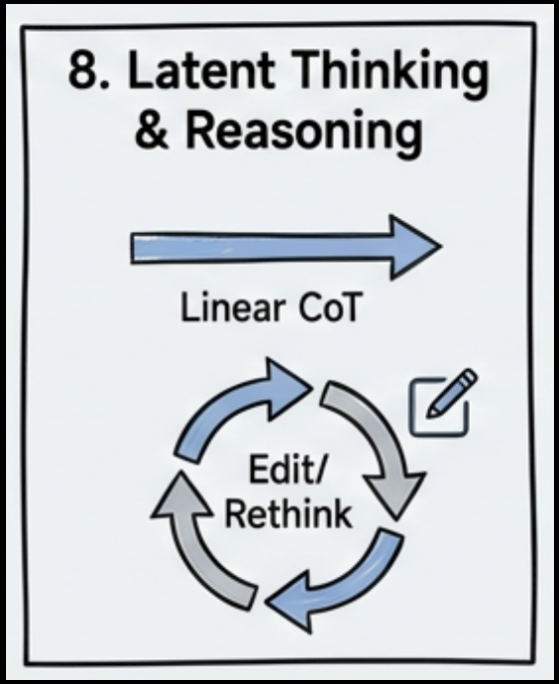
Challenge 8: Reasoning and Latent Thinking

Problem:

- Fixed-Length Inef ficiency: Standard dif fusion SFT generates answers via iterative denoising within a f ixed length, which limits the flexibility of the reasoning process.
- Suboptimal CoT Alignment: Traditional, sequential Chain-of-Thought (CoT) is poorly suited for diffusion models, failing to fully leverage their unique architectural potential.
- Resource Underutilization: Signif icant gaps exist in current data and optimization strategies, hindering the realization of "slow thinking" in these models.

Potential idea:

- Dif fusion-Specif ic CoT: Develop a new CoT framework that utilizes dif fusion-specif ic traits, such as generating outlines before supplementing details for code or agent tasks.
- Latent Editing & Re-masking: Implement "re-masking" during inference to reset low-conf idence tokens, enabling multi-granular thinking through active editing during generation.



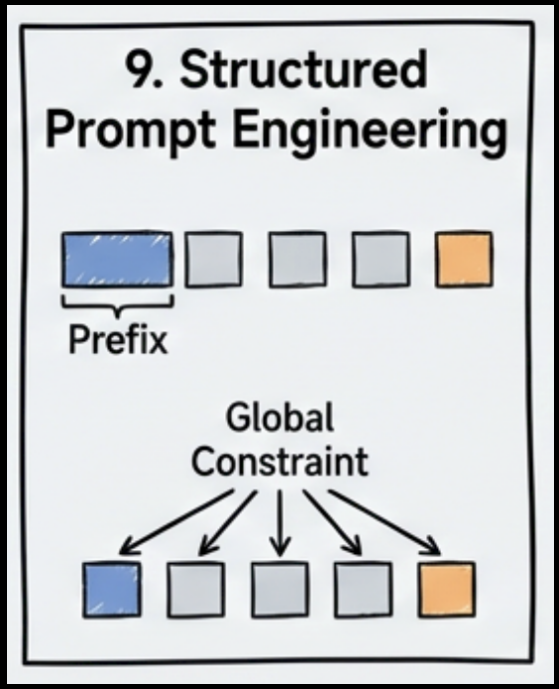
Challenge 9: Structured Prompt Engineering and Memory

Problem:

- Incompatibility of Paradigms: Diffusion models and Autoregressive (AR) models use fundamentally different masking and decoding mechanisms; diffusion can look both forward and backward, making traditional AR prompt formats suboptimal.
- Decoding Inefficiency: There is a lack of high-efficiency prompting methods that can trigger full-process decoding using minimal global key tokens, which limits performance in complex scenarios like code generation and autonomous agents.

Potential idea:

- Cloze-style Prompting: Replace standard "Q&A style" prompts with "cloze-style" formats and prompt self-evolution to better align with the diffusion mode and provide essential steering information.



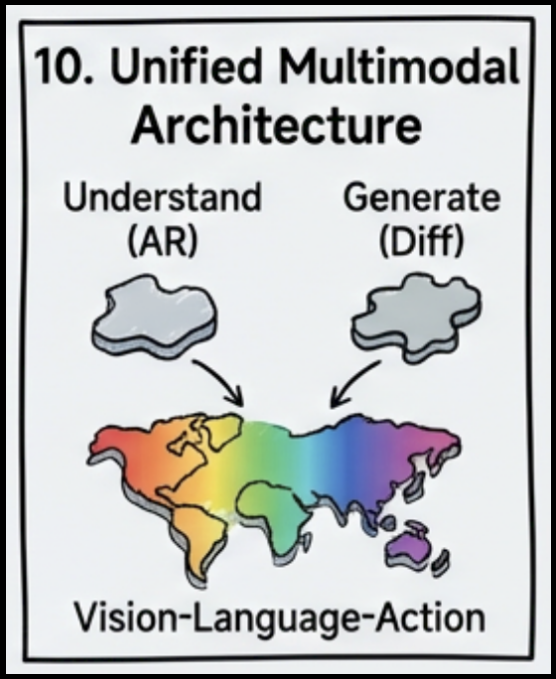
Challenge 10: Uniform Architectures through Diffusion

Problem:

- Architectural Fragmentation: Current multimodal systems rely on disparate paradigms, specifically AR models for understanding / language and Diffusion schemes for generation or action.
- Integration Barriers: Scaling toward a "from scratch" unified architecture is hindered by the difficulty of aligning different modalities, balancing data ratios, and reconciling conflicting optimization objectives.
- Execution Discontinuity: In models like VLA, there is a structural split where Vision and Language use AR, while Action requires continuous Diffusion for smooth output, preventing a truly singular training paradigm.

Potential idea:

- Discrete Diffusion Fusion: Explore using Discrete Diffusion Models as a unified structure to fuse Vision, Language, and Action into a single training paradigm and architecture.



WeChat Group

Whatsapp Group

Thank You!

Recruiting: Noahlab1@huawei.com