From AR to Diffusion: Efficiently Adapting Large Language Models with Strictly Causal and Elastic Horizons

Xiangyu Ma[1]*, **Teng Xiao**[2]*, **Zuchao Li**[1]†, **Lefei Zhang**[2]

[1]School of Artificial Intelligence, Wuhan University, Wuhan, China,
[2]School of Computer Science, Wuhan University, Wuhan, China
{maxiangyu, xiaoxiao, zcli-charlie, zhanglefei}@whu.edu.cn

Abstract

Diffusion models promise efficient parallel text generation but rely on bidirectional attention, creating a structural mismatch with pre-trained Autoregressive (AR) models. This incompatibility precludes reusing robust AR priors, necessitating prohibitive pre-training from scratch. To bridge this gap, we propose FLUID, a framework that efficiently adapts AR backbones to the diffusion paradigm. By enforcing Strictly Causal Alignment, FLUID enables seamless initialization from standard GPT-style checkpoints, circumventing the need for massive pre-training. Furthermore, we introduce Elastic Horizons, an entropy-driven mechanism that dynamically modulates denoising strides based on local information density rather than fixed schedules. Experiments demonstrate that FLUID achieves state-of-the-art performance while reducing training costs by orders of magnitude, effectively reconciling established AR foundations with efficient parallel generation. Our code is available at https://huggingface.co/MYTH-Lab/FLUID.

1 Introduction

Autoregressive (AR) language models, trained via the next-token prediction paradigm, constitute the cornerstone of modern natural language processing. By conditioning each token solely on its preceding context, AR models ensure rigorous logical consistency and training stability, powering breakthroughs in reasoning-intensive tasks such as code generation and mathematics (Delétang et al., 2024). However, this strictly sequential formulation imposes a bottleneck: inference latency grows linearly with sequence length (Luohe et al., 2025; Tang et al., 2025; Zhao et al., 2025). As model scales and context windows expand, this serial decoding cost increasingly dominates deployment budgets, creating an urgent demand for parallel generation paradigms (Yang et al., 2025; Chen et al., 2025b).

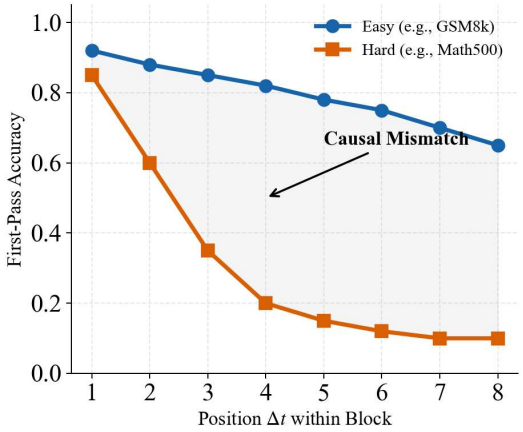


Figure 1: Causal mismatch in fixed-size block diffusion. First-pass accuracy decays significantly faster for high-entropy data (e.g., complex reasoning steps in MATH500) than low-entropy text (e.g., GSM8k) as lookahead increases.

To transcend this limitation, Discrete Diffusion Models (DLMs) have emerged as a promising alternative, offering the capability to generate multiple tokens in parallel through iterative denoising (Han et al., 2023a). Approaches such as LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025a) demonstrate that diffusion can model global context effectively. However, standard diffusion models typically rely on bidirectional attention mechanisms. While effective for global coherence, this architecture creates a critical structural mismatch with the pre-trained priors of ubiquitous AR models (e.g., Llama, GPT). This incompatibility precludes the efficient reuse of existing checkpoints, necessitating computationally prohibitive pre-training from scratch.

Recent "Block Diffusion" strategies (Arriola et al., 2025; Cheng et al., 2025) attempt to bridge

*These authors contributed equally to this work.

†Corresponding Author. This work was supported by the National Natural Science Foundation of China (No. 62306216), the Technology Innovation Program of Hubei Province (No. 2024BAB043).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20914–20927

this gap by adopting a "semi-autoregressive" hybrid design: sequences are partitioned into fixed blocks, inheriting AR's serial dependency between blocks while applying bidirectional diffusion within them. While this mitigates some efficiency bottlenecks, we identify a critical limitation in the fixed-size nature of these strategies.

Despite these advancements, we identify a critical limitation in fixed-size block diffusion: the *Entropy-Horizon Dilemma*. As illustrated in Figure 1, rigid generation windows fundamentally misalign with the dynamic entropy of natural language. Text exhibits varying information density, oscillating significantly between deterministic functional phrases and high-uncertainty reasoning steps. In high-entropy regions (e.g., MATH500), large blocks cause a rapid decay of the "causal horizon," necessitating aggressive error correction that negates parallelization gains. Conversely, in low-entropy regions (e.g., GSM8K), conservative blocks fail to fully exploit parallel decoding potential. This static approach decouples generation cadence from the intrinsic semantic rhythm, limiting both inference speed and quality.

To resolve these structural and dynamic mismatches, we propose FLUID (Flexible Unidirectional Inference Diffusion), a novel framework designed to efficiently adapt AR models into parallel diffusers. Unlike traditional diffusion models that enforce bidirectional dependencies, FLUID introduces Strictly Causal Alignment by imposing a unidirectional attention mask during the diffusion process. This design realigns the generative mechanism with the inductive biases of pre-trained AR backbones, enabling seamless initialization and efficient fine-tuning without the need for massive pre-training. Furthermore, we replace fixed block boundaries with Elastic Horizon Modeling, a mechanism driven by real-time entropy estimation. This allows FLUID to dynamically modulate the generation window—accelerating through high-confidence segments while allocating dense computation to complex reasoning steps—thereby achieving a seamless balance between inference speed and generation quality.

Comprehensive Experiments across general understanding, mathematical reasoning, and code generation demonstrate that FLUID matches the performance of top-tier autoregressive models while significantly outperforming existing bidirectional diffusion baselines. By enforcing a unidirectional attention mask to realign with the inductive biases of pre-trained backbones, our framework validates strictly causal diffusion as a paradigm that superiorly balances training efficiency, inference latency, and logical consistency. This alignment—enabling seamless initialization from standard GPT-style checkpoints to reduce training costs by orders of magnitude—theoretically facilitates native KV cache support, establishing a decisive efficiency advantage over non-linear bidirectional diffusers. Leveraging an entropy-driven Elastic Horizon mechanism, FLUID dynamically modulates its generation window, accelerating through predictable segments while ensuring high-fidelity reasoning in complex transitions to effectively mitigate the semantic fracture inherent in fixed-size strategies. Ultimately, our work establishes a scalable foundation for transforming established AR foundations into efficient parallel diffusers without compromising structural integrity or logical coherence.

2 Related Work

2.1 Discrete Diffusion for Language Modeling

DLMs offer a non-autoregressive alternative to traditional modeling by iteratively denoising discrete tokens (Austin et al., 2021a; Li et al., 2022). Utilizing masking or absorbing states, DLMs enable parallel generation and mitigate exposure bias (Zhou et al., 2024; Zeng et al., 2025). While initially computationally demanding (Gulrajani and Hashimoto, 2023), recent Masked Diffusion Models have bridged this efficiency gap. Lou et al. (2024) matched GPT-2 performance, and subsequent scaling to the billion-parameter level—exemplified by LLaDA (Nie et al., 2025) and Mercury Coder (Khanna et al., 2025)—has achieved parity with strong AR baselines like LLaMA3 (Team, 2024), establishing DLMs as a scalable foundation for language generation.

2.2 Adapting Autoregressive Models to Diffusion

To circumvent prohibitive pre-training costs, recent research adapts pre-trained AR backbones to diffusion, leveraging robust priors to reframe sequential generation as parallel denoising. DiffuLLaMA (Gong et al., 2025) pioneered this by relaxing causal masking and applying parameter-efficient fine-tuning. Subsequent hybrid models like SDAR (Cheng et al., 2025) integrate AR consistency with diffusion refinement to ensure scal-

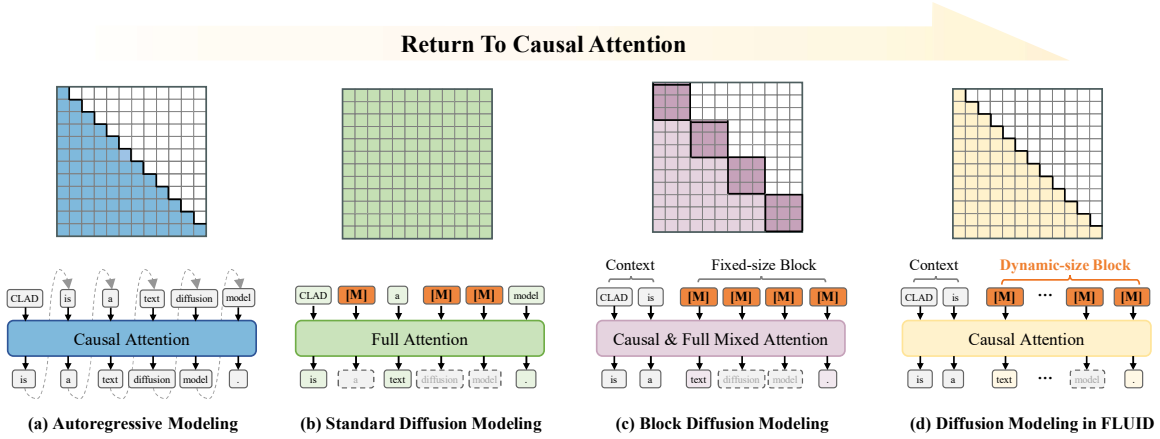


Figure 2: From Autoregressive to FLUID: Evolution of four language model generation paradigms. The diagram highlights the pivotal transition to FLUID (d), which integrates Unidirectional Attention to maintain strictly causal alignment with AR priors, and Dynamic Mask Blocks to adaptively modulate generation horizons based on sequence entropy, resolving the structural and dynamic mismatches inherent in prior paradigms.

able quality. Meanwhile, efficiency-focused methods such as Fast-DLLM (Wu et al., 2025b) and Fast-DLLM-v2 (Wu et al., 2025a) mitigate inference bottlenecks via training-free acceleration. Collectively, these works validate the computational efficiency of transforming AR foundations into robust diffusion decoders.

2.3 Block-Wise Adaptation and Hybrid Approaches

While capturing global context, full-sequence diffusion incurs high computational overhead and misaligns with the causal inductive biases of AR pre-training. Although recent caching mechanisms (Liu et al., 2025b; Ma et al., 2025; Wu et al., 2025b) attempt to alleviate these inefficiencies, they do not fundamentally alter the generation paradigm. To enable scalable generation, Semi-Autoregressive (or Block-Wise) approaches (Han et al., 2023b; Arriola et al., 2025) partition sequences into segments, denoising the current block given committed priors. While balancing quality and speed, these methods typically rely on rigid boundaries and bidirectional attention, restricting dynamic adaptability.

3 Preliminaries

Consider a sequence of tokens $\mathbf{x} = [x_1, \ldots, x_L]$ from a vocabulary $\mathcal{V}$. Let $\mathcal{D}$ denote the data distribution. We aim to learn a model p_θ that approximates $\mathcal{D}$.

3.1 Autoregressive Modeling

Standard AR models decompose the joint probability distribution into a product of conditional probabilities via the chain rule:

$$p_{\text{AR}}(\mathbf{x}) = \prod_{i=1}^{L} p_\theta(x_i \mid \mathbf{x}_{<i}), \tag{1}$$

where $\mathbf{x}_{<i}$ denotes the history tokens preceding position i. Crucially, this factorization enforces a *strictly causal dependency* structure.

3.2 Discrete Diffusion Modeling

To enable non-autoregressive generation, DLMs reframe sequence generation as a corruption-and-denoising process.

Forward Process. The forward process $q(\mathbf{x}_t \mid \mathbf{x}_0)$ progressively corrupts the clean data $\mathbf{x}_0$ over continuous time $t \in [0, 1]$. Following the absorbing state mechanism (Austin et al., 2021a), tokens are independently masked based on a noise schedule α_t. The transition probability for a token x_i is:

$$\begin{aligned} q(x_{t,i} \mid x_{0,i}) = \alpha_t \mathbb{I}(x_{t,i} = x_{0,i}) \\ + (1 - \alpha_t)\mathbb{I}(x_{t,i} = \texttt{[M]}), \end{aligned} \tag{2}$$

where [M] denotes the mask token. As $t \to 1$, the sequence converges to a fully masked state.

Reverse Process and Optimization. To reverse this corruption of generative process, following Ye et al. (2025b), we optimize a weighted cross-entropy objective as a tractable surrogate for the variational lower bound (ELBO):

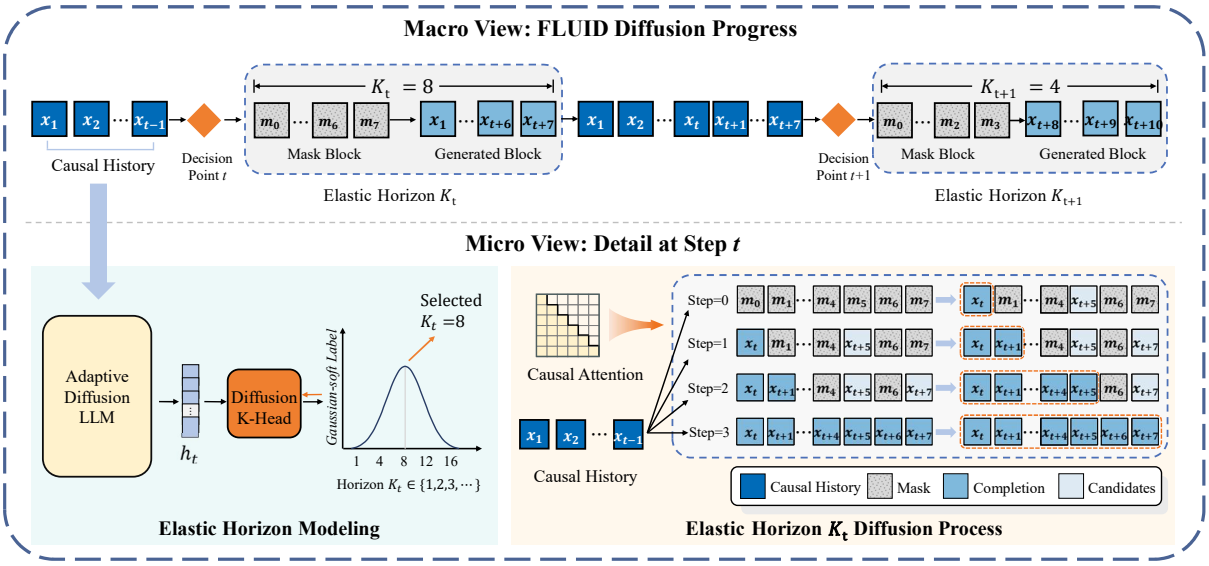


Figure 3: The Schematic Framework of FLUID. The Macro View (Top) depicts the Adaptive Diffusion Progression, where the model dynamically adjusts the lookahead horizon K_t. The Micro View (Bottom) details the mechanism at step t: the Elastic Horizon Modeling derives the optimal window size from hidden states to guide the Elastic Horizon Diffusion Process, which iteratively refines masked tokens under strict causal constraints.

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t}\left[w(t)\sum_{i=1}^{L}\mathbb{1}_{[x_{t,i}=\texttt{[M]}]}\log p_\theta(x_{0,i} \mid \mathbf{x}_t)\right], \tag{3}$$

where $w(t)$ is a time-dependent weighting term.

Typically, the denoiser p_θ uses bidirectional attention to model global context $\mathbf{x}_t$ (Ye et al., 2025b).

3.3 Block-Wise Semi-Autoregressive Strategies

As shown in Figure 2, Block Diffusion strategies (Arriola et al., 2025) bridge AR and diffusion by partitioning sequences into blocks $\mathbf{x} = [\mathbf{x}^1, \ldots, \mathbf{x}^B]$. This approach factorizes the likelihood autoregressively across blocks while employing diffusion:

$$p_\theta(\mathbf{x}) = \prod_{b=1}^{B} p_\theta(\mathbf{x}^b \mid \mathbf{x}^{<b}). \tag{4}$$

In conventional settings, blocks have a fixed size L_{block}, and intra-block generation retains bidirectional visibility.

4 FLUID

4.1 Strictly Causal Diffusion Backbone

Building upon the discrete diffusion framework (§3.2; Figure 3), FLUID enforces a *strictly causal* dependency structure. Departing from standard bidirectional implementations that model global noisy context (Ye et al., 2025b; Nie et al., 2024), we impose a structural constraint to preserve the autoregressive inductive bias of pre-trained LLMs.

Formally, we inject a lower-triangular attention mask $\mathbf{M}$ into the Transformer. For a query at position i and key at position j, the attention scores are modulated as:

$$\text{Attention}(i,j) = \begin{cases} \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d_k}} & \text{if } j \le i, \\ -\infty & \text{otherwise.} \end{cases} \tag{5}$$

This constraint restricts the conditional probability of restoring token x_i to depend solely on the history $\mathbf{x}_{t,<i}$, effectively pruning all connections to acausal future positions.

4.2 Elastic Horizon Modeling

Standard fixed-size blocks neglect the variable entropy of natural language, fracturing semantic units in high-uncertainty regions while wasting cycles in deterministic ones. We resolve this *Entropy-Horizon Dilemma* via *Elastic Horizons*, dynamically modulating the generation stride K_t based on local confidence.

4.2.1 Probabilistic Horizon Estimator

We append a lightweight *Diffusion K-Head* to the frozen backbone to predict the optimal stride. Unlike scalar regression, we model the horizon as

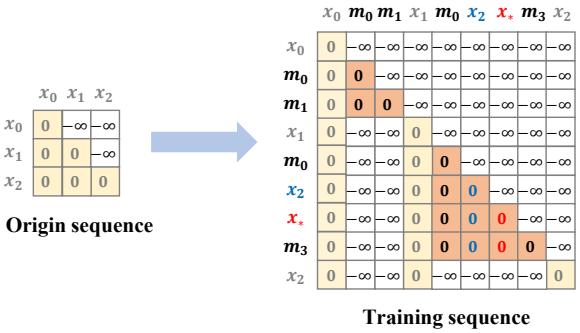


Figure 4: Illustration of Training Objective. The original sequence is transformed via dynamic expansion. Block x_1 ($k = 4$) exhibits a mixed state of stochastically restored ground-truth (x_2) and injected noise (x_*), governed by a strictly causal attention matrix.

a probability distribution to capture the inherent ambiguity of semantic boundaries. Formally, the K-Head maps the final hidden state h_t to a categorical distribution over $k \in \{1, \dots, K_{\max}\}$:

$$\begin{aligned} \mathbf{z}_t &= \text{MLP}(h_t), \\ P_\phi(k \mid h_t) &= \text{Softmax}(\mathbf{z}_t)_k. \end{aligned} \tag{6}$$

This probabilistic formulation enables the model to express uncertainty, favoring conservative steps when boundaries are ambiguous.

4.2.2 Competence Boundary Supervision

In the absence of explicit labels, we frame horizon prediction as learning a *Competence Boundary*. We define the oracle horizon K_t^* as the maximum span of high-confidence generation, derived from the probed future loss sequence $\mathcal{L}$:

$$K_t^* = \max \left\{ k \mid \frac{1}{k} \sum_{j=1}^{k} \mathcal{L}_{t+j} < \tau \right\}. \tag{7}$$

To reflect the ordinal nature of this boundary, we supervise the K-Head using a Gaussian soft target $\mathcal{Q}$ centered at K_t^*:

$$\mathcal{Q}(k) \propto \exp\left(-\frac{(k - K_t^*)^2}{2\sigma^2} \right). \tag{8}$$

Minimizing the KL divergence $D_{\text{KL}}(\mathcal{Q} \parallel P_\phi)$ compels the estimator to align its planning horizon with the backbone's intrinsic generative capability.

4.3 Training

Training follows a two-stage curriculum: causal denoising, then horizon calibration.

4.3.1 Stage I: Joint Causal Backbone Training

Stage I fine-tunes the backbone θ with the K-Head frozen, employing a hybrid objective that synergizes autoregressive generation with masked denoising to preserve linguistic priors.

As depicted in Figure 4, for each sequence $\mathbf{x}$, we sample a horizon $K \sim \mathcal{U}[1, K_{\max}]$ to partition the input into history $\mathbf{x}_{\text{obs}}$ and targets $\mathbf{x}_{\text{mask}}$. To enhance robustness, we apply *Stochastic Restoration* by injecting 10% noise into the masks (§4.1).

The optimization minimizes a hybrid loss function:

$$\begin{aligned} \mathcal{L}_{\text{Stage1}} = & \underbrace{- \sum_{x_i \in \mathbf{x}_{\text{obs}}} \log p_\theta(x_i \mid \mathbf{x}_{<i})}_{\mathcal{L}_{\text{AR}}} \\ & + \underbrace{\mathbb{E}_t \left[- \sum_{x_j \in \mathbf{x}_{\text{mask}}} w_t \log p_\theta(x_j \mid \mathbf{x}_t) \right]}_{\mathcal{L}_{\text{Diff}}}. \end{aligned} \tag{9}$$

The $\mathcal{L}_{\text{AR}}$ term maintains the semantic stability of the prefix, while $\mathcal{L}_{\text{Diff}}$ drives the parallel denoising of the K-token span under strict causal constraints.

4.3.2 Stage II: Probabilistic Horizon Training

In Stage II, we freeze the backbone to exclusively optimize the K-Head parameters ϕ, framing horizon estimation as a distribution matching task.

We probe the frozen backbone by inserting the maximum mask span ($K_{\max}$) to elicit its intrinsic uncertainty boundaries. Based on the competence boundary K^* derived in Eq. 7, we construct a Gaussian soft target $\mathcal{Q}$.

The K-Head is trained to align its predicted horizon distribution P_ϕ with the target $\mathcal{Q}$ by minimizing the Kullback-Leibler divergence:

$$\mathcal{L}_{\text{Stage2}} = D_{\text{KL}}\left(\mathcal{Q} \parallel P_\phi(\cdot \mid h_t)\right). \tag{10}$$

This objective calibrates FLUID to learn a smooth, ordinal representation of its capabilities, inducing conservative strides in high-entropy contexts.

4.4 Dynamic Causal Inference

FLUID reframes inference as a dynamic causal diffusion process, orchestrating an interplay between entropy-aware planning and parallel denoising (Algorithm 1).

At step t, the K-Head projects the causal hidden state h_t to determine the horizon $K_t = \text{argmax}_k P_\phi(k|h_t)$. This dynamically modulates the window—expanding in deterministic

Algorithm 1 Dynamic Causal Diffusion Inference

Require: Model p_θ, K-Head f_ϕ, Prompt x, Threshold γ, Limit L

while length$(x) < L$ and not EOS **do**
  $h \leftarrow \text{Encoder}(x)$; $K \leftarrow \text{argmax} f_\phi(h)$ {Plan Horizon}
  $Y \leftarrow [M]^K$; $step \leftarrow 0$
  while $step < K$ **do**
    $P \leftarrow p_\theta([x; Y])$ {Parallel Denoising Sweep}
    $Y[step] \leftarrow \text{argmax}(P[step])$ {Force Update}
    $\Delta \leftarrow 0$; $chain\_valid \leftarrow$ True
    for $j \leftarrow step + 1$ **to** $K - 1$ **do**
      if $\max(P[j]) > \gamma$ **then**
        $Y[j] \leftarrow \text{argmax}(P[j])$ {Confidence Gating}
        if $chain\_valid$ **then**
          $\Delta \leftarrow \Delta + 1$
        end if
      else
        $chain\_valid \leftarrow$ False
      end if
    end for
    $x \leftarrow [x; Y[step : step + \Delta]]$ {Commit Valid Chain}
    $step \leftarrow step + 1 + \Delta$ {Adaptive Stride}
  end while
end while
return x

phases and contracting for high-entropy transitions—before initiating parallel denoising on a mask of span K_t.

To ensure generation quality, we apply a *confidence gating* mechanism: while the immediate next token is unconditionally accepted, future predictions at relative positions $j \in \{1, \ldots, K_t\}$ are accepted only if their confidence exceeds a threshold γ. The decoding cursor then advances by $1+\Delta$, where Δ denotes the length of the continuous chain of confident denoisings immediately following the current position, implicitly truncating the predicted horizon K_t if an early low-confidence token interrupts the sequence. This adaptive stride mechanism allows FLUID to "sprint" through coherent segments (e.g., code boilerplate) and seamlessly revert to fine-grained autoregression when entropy spikes, ensuring strictly causal correctness without the rigidity of fixed-block decoding.

5 Experiments

5.1 Experimental Setup

We initialize FLUID with the openPangu-Embedded-7B (Chen et al., 2025a) checkpoint to ensure fair comparison with baselines of equivalent scale, while capitalizing on its robust linguistic priors.

Adaptation follows a two-stage curriculum: Stage I fine-tunes the backbone for 32,000 iterations (detail in Appendix B); Stage II freezes the backbone to exclusively train the Diffusion K-Head for 2,000 steps, targeting the oracle horizon K_t^*. For calibration, we employ a threshold $\tau = 2.8$ (Eq. 7); a comprehensive sensitivity analysis is provided in the Ablation section.

Training Data. Our adaptation corpus is constructed by distilling responses from openPangu-Embedded-7B over several public instruction-tuning datasets, including **Infinity-Instruct-7M** (Li et al., 2025), **deepctrl-sft-data** (zh) (DeepCtrl, 2024), **moss-003-sft** (no-tools) (Sun et al., 2024), and **UltraChat** (Ding et al., 2023). This distilled mixture covers diverse task distributions, including large-scale instruction-following data, Chinese multi-turn conversational data, open-domain dialogue, and broad high-quality instructional conversations, thereby providing broad supervision for adapting the autoregressive backbone to our strictly causal diffusion objective.

Implementation Details. Input sequences are set to a length of 1024, with a global batch size of 80. For parameter efficiency, we apply Rank-16 LoRA to the backbone, while the K-Head (a two-layer MLP) is the only newly introduced dense module trained from scratch. Following (Arriola et al., 2025), we set $K_{max} = 16$ for our dynamic analysis. Optimization utilizes the AdamW optimizer with a learning rate of 2×10^{-4}, a 0.1 warmup ratio, and a cosine decay schedule.

Benchmarks. We conduct comprehensive evaluations across three core domains. **General Knowledge and Instruction Following:** We employ MMLU (Hendrycks et al., 2020) for broad semantic understanding and IFEVAL (Zhou et al., 2023) to assess the model's ability to adhere to objective constraints and formatting prompts. **Mathematical Reasoning:** Performance is measured on GSM8K (Cobbe et al., 2021) for multi-step arithmetic and the more rigorous MATH500 (Lightman et al., 2023) for competition-level problem solving. **Code Generation:** We utilize HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021b) to evaluate the functional correctness and structural integrity of synthesized code.

5.2 Main Results

Results and Analysis. Table 1 compares FLUID-7B with leading AR and diffusion baselines. In rea-

Model	Type	Tokens	General		Math & Science		Code	
			MMLU	IFEVAL	GSM8K	MATH500	HEval	MBPP
LLaMA-3-8B†	AR	15T	68.4 (5)	49.7 (0)	78.3 (4)	27.4 (0)	59.8 (0)	57.6 (4)
Qwen-2.5-7B†	AR	18T	76.6 (5)	74.7 (0)	91.6 (0)	72.8 (0)	84.8 (0)	79.2 (4)
Gemma2 9B†	AR	8T	–	–	76.7 (0)	–	68.9 (0)	74.9 (0)
Deepseek 7B¶	AR	2T	49.4 (0)	–	63.0 (0)	–	48.2 (-)	35.2 (-)
LLaDA-8B	Diff	2.0T	65.5 (5)	59.9 (0)	78.6 (4)	36.2 (0)	47.6 (0)	34.2 (4)
LLaDA-1.5-8B	Diff	2.0T	66.0 (5)	58.2 (0)	83.3 (4)	42.6 (0)	52.4 (0)	42.8 (4)
Dream-7B	Diff	0.6T	67.0 (5)	**62.5 (0)**	81.0 (4)	39.2 (4)	55.5 (0)	**58.8 (4)**
FLUID-7B (Ours)	**Diff**	**2.7B**	**67.8 (5)**	57.7 (0)	**91.9 (4)**	**61.8 (4)**	**60.4 (0)**	53.6 (4)

Table 1: Main Comparison on Standard Benchmarks. We compare FLUID against state-of-the-art Diffusion and AR baselines under Instruct settings. "Tokens" denotes the amount of data used for pre-training (for AR) or adaptation (for Diffusion). Note that FLUID achieves comparable performance to strong baselines using **orders of magnitude less training data** (2.7B vs. Trillions). Results indicated by † and ¶ are sourced from Chu et al. (2024), Yang et al. (2024), and Bi et al. (2024). Best results in diffusion methods are **bolded**, and our results are highlighted in .

soning, FLUID-7B achieves significant advantages, scoring 91.9 on GSM8K and 61.8 on MATH500, surpassing diffusion models like Dream-7B and LLaDA-8B by 10.9 and 13.3 points on GSM8K, respectively, and matching top AR models like Qwen-2.5-7B (91.6). These results validate our strictly causal diffusion framework. Unlike bidirectional models, which disrupt deductive chains by conditioning on noisy future contexts, FLUID's causal masking preserves logical consistency.

FLUID-7B further excels in code generation (60.4 on HumanEval) and instruction following (57.7 on IFEval), outperforming LLaMA-3-8B-Instruct and substantially leading LLaDA-8B. This superiority is driven by the Elastic Horizon Modeling. Whereas fixed-block strategies often cause "semantic fracture" by truncating syntactic units, our entropy-driven mechanism dynamically modulates the generation window. By expanding for predictable syntax and contracting for high-entropy logic, FLUID ensures structural integrity while bridging the gap between diffusion-based planning and AR-style stability.

5.3 Semantic Quality Evaluation

Beyond surface-level metrics, we assess generation quality using Skywork-Reward-V2 (Liu et al., 2025a), a preference model designed to evaluate intrinsic helpfulness and logical coherence. As shown in Figure 5, FLUID achieves the highest reward scores across all five domains, validating its superior alignment with human intent.

Notably, FLUID significantly outperforms bidirectional diffusion baselines (e.g., LLaDA, Dream),

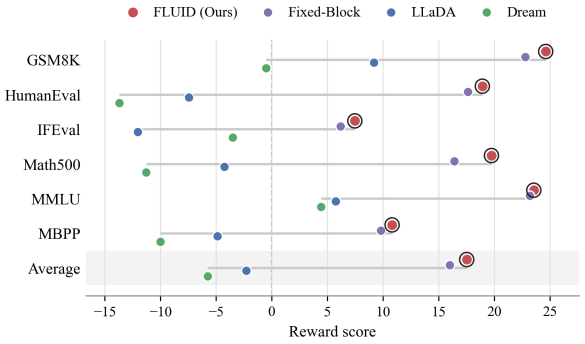


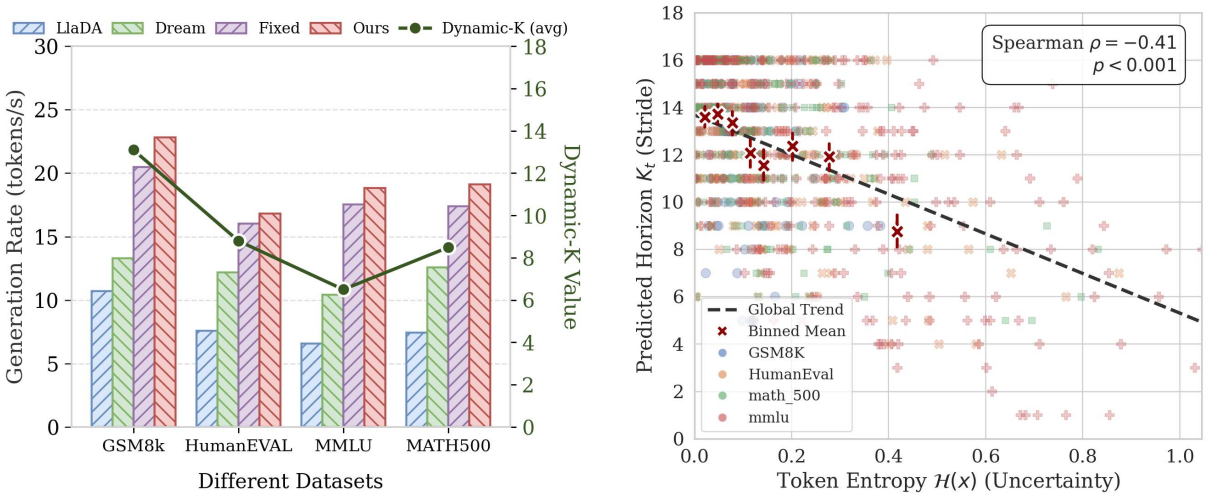
Figure 5: Semantic Quality Evaluation via Skywork-Reward-V2.

particularly in reasoning-intensive tasks like GSM8K and MATH500. This confirms that strictly causal attention preserves deductive chains often disrupted by the noisy, acausal contexts inherent to standard diffusion models. Furthermore, FLUID maintains a consistent lead over Fixed-Block strategies. By replacing rigid boundaries with entropy-driven Elastic Horizons, our model effectively mitigates "semantic fracture"—ensuring structural integrity in code generation and continuity in complex reasoning without incurring the waiting costs of conservative blocking.

5.4 Ablation Study

In this section, we analyze the effectiveness of the proposed adaptation methodologies.

Impact of Causal Attention. We evaluate the necessity of strictly causal masking. As shown in Table 2, the bidirectional attention baseline significantly underperforms, lagging behind FLUID by 9.4% on GSM8K. This confirms that conditioning



(a) Inference Throughput & Dynamic Horizon **(b) Entropy-Horizon Correlation**

Figure 6: Efficiency and Accuracy Comparison. (a) All models are evaluated without KV cache or FlashAttention to ensure a fair architectural comparison. (b) Verification of K-Head Accuracy.

on noisy, acausal future contexts disrupts the pre-trained autoregressive logic chain, leading to incoherent reasoning. By enforcing strict causal masking, FLUID realigns the diffusion process with the backbone's AR priors, effectively restoring the reasoning capabilities essential for complex tasks.

Benefit of Elastic Horizon. We further compare FLUID against the Fixed Block Size ($k = 16$) baseline. While fixed blocks improve stability over standard diffusion, they still suffer from *semantic fracture*, the arbitrary truncation of coherent logic units—resulting in a 5.5% deficit on HumanEval. FLUID's adaptive horizon addresses this by dynamically expanding generation windows based on local entropy. This ensures that atomic syntactic units (e.g., function definitions) remain intact, yielding superior performance in structure-sensitive tasks like code generation.

	Components		Benchmarks		
Method	**Causal**	**Elastic**	**GSM8K**	**MATH500**	**HEval**
Baseline	✗	✗	82.0	51.2	42.2
Baseline	✗	✓	82.5	53.6	42.8
Baseline	✓	✗	90.6	59.2	54.9
FLUID (Ours)	✓	✓	**91.9**	**61.8**	**60.4**

Table 2: Ablation study on adaptation strategies. Impact of strictly causal masking (*Causal*) and entropy-aware dynamic horizons (*Elastic*) relative to a bidirectional fixed-block baseline.

Inference Latency. As shown in Figure 6(a), FLUID establishes a distinct efficiency advantage. Compared to standard diffusion baselines (e.g., LLaDA, Dream) that require intensive iterative refinement, FLUID achieves approximately 2× speedup.

Crucially, FLUID outperforms the aggressive Fixed-Block ($K = 16$) baseline even with a smaller average stride, highlighting the efficiency of our Elastic Horizon. On challenging tasks like MMLU, forcing large strides incurs significant "waiting costs" due to high entropy. Conversely, FLUID autonomously contracts its horizon (avg $K = 6.5$) to prioritize robust updates. By mitigating error correction overhead, FLUID achieves higher throughput (18.82 vs. 17.52 tokens/s), demonstrating that optimal latency stems from *semantic synchronization* rather than rigid block maximization.

Noise Ratio	**GSM8K**	**MATH500**	**HEval**
0%	91.0	60.8	59.8
5%	91.3	61.1	59.8
10% (Ours)	**91.9**	**61.8**	**60.4**
15%	91.1	61.5	60.0

Table 3: Ablation on the stochastic restoration noise ratio in Stage I training. Impact of varying proportions of injected noise within the masked span.

Impact of Stochastic Restoration Ratio. In Stage I, FLUID adopts stochastic restoration by injecting a small amount of noise into the masked span to improve robustness during causal denoising. To study its effect, we vary the noise ratio from 0% to 15% while keeping all other settings fixed. As shown in Table 3, a moderate ratio gives the best trade-off, with 10% achieving the strongest overall

performance on GSM8K, MATH500, and HEval.

Without stochastic restoration (0%), the model is trained on easier denoising patterns and is less robust to imperfect intermediate states, leading to lower performance across all benchmarks. Introducing mild noise (5%–10%) improves stability and better prepares the model for noisy decoding trajectories. In contrast, excessive noise (15%) harms performance by distorting the target signal and making optimization under strict causal constraints more difficult, especially on MATH500 and HEval. Overall, these results show that a moderate restoration ratio is most effective, balancing robustness and target fidelity.

Threshold τ	GSM8K	MATH500	HEval
2.6	90.9	60.4	59.8
2.7	91.1	61.5	59.9
2.8 (Ours)	**91.9**	**61.8**	**60.4**
2.9	90.5	61.5	59.5
3.0	90.4	60.2	59.3

Table 4: Ablation study on the competence boundary τ.

Impact of the Competence Boundary τ. The competence boundary τ governs the trade-off between decoding aggressiveness and causal reliability by controlling how far FLUID can safely expand its dynamic horizon. As shown in Table 4, $\tau = 2.8$ achieves the best overall results across GSM8K, MATH500, and HEval. When τ is set to smaller values (e.g., 2.6 or 2.7), FLUID tends to adopt a more conservative horizon. While this improves decoding stability, it also limits the benefit of parallel generation and leads to slightly weaker overall performance. In contrast, larger values (e.g., 2.9 or 3.0) encourage more aggressive horizon expansion, but also increase error correction costs once the model overestimates its confidence, especially on reasoning-intensive tasks such as MATH500.

Overall, the results exhibit a clear bell-shaped trend centered at $\tau = 2.8$. This suggests that an appropriate competence boundary is crucial for balancing decoding efficiency and causal consistency, and confirms that $\tau = 2.8$ provides the most effective operating point for FLUID.

5.5 Verification of K-Head Accuracy

To verify that the Elastic Horizon reflects the model's internal confidence, we analyze the correlation between predicted horizon K_t and information density. Figure 6(b) reveals a significant negative correlation between uncertainty and stride length.

Task-specific strides in Figure 6(a) corroborate this adaptivity. On GSM8K, where the model demonstrates high competence (91.9% accuracy), the K-Head confidently expands the stride (average $K = 13.1$) to maximize speed. Conversely, on MMLU, where decision ambiguity is higher, the horizon naturally contracts (average $K = 6.5$). This confirms that the K-Head functions as a *semantic gear stick*—sprinting through high-confidence reasoning chains while downshifting to a cautious, fine-grained pace during challenging transitions.

6 Conclusion

In this work, we introduced FLUID, a framework that reimagines text diffusion by breaking free from the rigidity of fixed-step generation. By enforcing a strictly causal attention mechanism, FLUID bridges the structural gap between diffusion paradigms and pre-trained autoregressive priors, enabling efficient adaptation without the prohibitive costs of pre-training from scratch. Our experiments demonstrate that FLUID acts much like its namesake—dynamically modulating its generation horizon via Elastic Horizon Modeling. It flows rapidly through predictable sequences and contracts cautiously during high-entropy transitions, thereby resolving the trade-off between inference latency and reasoning integrity. Ultimately, FLUID establishes a new paradigm for efficient generative modeling, proving that aligning causal consistency with dynamic flexibility is the key to unlocking the full potential of non-autoregressive text synthesis.

Limitations

FLUID is designed to efficiently adapt pre-trained autoregressive models into a diffusion framework. Consequently, its performance is inherently bounded by the capabilities of the source model. If the base autoregressive model suffers from hallucinations or reasoning deficits, FLUID is likely to inherit these behaviors. Furthermore, our current experiments focus primarily on adapting general-purpose LLMs (e.g., OpenPangu). The efficacy of FLUID on highly specialized domains (e.g., biomedical or legal texts) or different architectures (e.g., MoE models) has yet to be extensively verified, warranting further investigation.

References

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021a. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021b. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, and 67 others. 2024. Deepseek LLM: scaling open-source language models with longtermism. *CoRR*, abs/2401.02954.

Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, and 1 others. 2025a. Pangu embedded: An efficient dual-system llm reasoner with metacognition. *arXiv preprint arXiv:2505.22375*.

Jinglin Chen, Qiwei Li, Zuchao Li, Baoyuan Qi, Liu Guoming, Haojun Ai, Hai Zhao, and Ping Wang. 2025b. Faster in-context learning for LLMs via n-gram trie speculative decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18040–18051, Suzhou, China. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and 1 others. 2025. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

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. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

DeepCtrl. 2024. deepctrl-sft-data. https://www.modelscope.cn/datasets/deepctrl/deepctrl-sft-data.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ishaan Gulrajani and Tatsunori B. Hashimoto. 2023. Likelihood-based diffusion language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023a. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023b. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo,

Yanis Miraoui, Akash Palrecha, Stefano Ermon, and 1 others. 2025. Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*, 1.

Jijie Li, Li Du, Hanyu Zhao, Bo wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. 2025. Infinity instruct: Scaling instruction selection and synthesis to enhance language models. *Preprint*, arXiv:2506.11116.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. 2025b. dllm-cache: Accelerating diffusion large language models with adaptive caching. *CoRR*, abs/2506.06295.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion language modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning (ICML)*.

Shi Luohe, Zuchao Li, Lefei Zhang, Baoyuan Qi, Liu Guoming, and Hai Zhao. 2025. KV-latent: Dimensional-level KV cache reduction with frequency-aware rotary positional embedding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1550, Vienna, Austria. Association for Computational Linguistics.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. dkv-cache: The cache for diffusion language models. *CoRR*, abs/2505.15781.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2024. Scaling up masked diffusion models on text. *arXiv preprint arXiv:2410.18514*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, and 1 others. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*, 21(5):888–905.

Zicong Tang, Shi Luohe, Zuchao Li, Baoyuan Qi, Liu Guoming, Lefei Zhang, and Ping Wang. 2025. SpindleKV: A novel KV cache reduction method balancing both shallow and deep layers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28428–28442, Vienna, Austria. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2025a. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025b. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Haoqi Yang, Yao Yao, Zuchao Li, Baoyuan Qi, Liu Guoming, and Hai Zhao. 2025. XQuant: Achieving ultra-low bit KV cache quantization with cross-layer compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9785–9800, Suzhou, China. Association for Computational Linguistics.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025a. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025b. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Yiming Zeng, Jinghan Cao, Zexin Li, Yiming Chen, Tao Ren, Dawei Xiang, Xidong Wu, Shangqian Gao, and Tingting Yu. 2025. Treediff: Ast-guided code generation with diffusion llms. *arXiv preprint arXiv:2508.01473*.

Yi Zhao, Zuchao Li, and Hai Zhao. 2025. IAM: Efficient inference through attention mapping between different-scale LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 19522–19533, Vienna, Austria. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1438–1451.

A Empirical Analysis of Decoding Trajectories in Bidirectional Diffusion

To further substantiate the structural and dynamic mismatches discussed in our Introduction (§1), we analyze the decoding trajectories of LLaDA-8B, a representative bidirectional masked diffusion model. While bidirectional attention offers theoretical global flexibility, our empirical findings reveal a "behavioral degeneration" during inference. The model's pursuit of high-confidence sampling paths unintentionally mimics the causal constraints of AR models, without achieving their computational efficiencies.

A.1 Sequential Degeneration in Single-Step Decoding

When we constrain the decoding process to a single-token-per-step regime-effectively a diffusion-based greedy search—the model exhibits a nearly linear, left-to-right progression, as shown in Figure 7(a). This *sequential degeneration* highlights a critical paradox: in natural language, the conditional probability $P(x_i|x_{<i})$ is typically far more deterministic than acausal lookahead predictions. Consequently, a confidence-driven remasking strategy naturally gravitates toward a causal path, rendering the bidirectional attention over future mask positions an expensive redundancy. This observation provides the empirical bedrock for our Strictly Causal Alignment in FLUID, as it aligns the model's architectural capacity with its actualized generative behavior, thereby enabling the seamless integration of KV Cache optimizations that are otherwise unattainable in bidirectional frameworks.

A.2 Non-linear Convergence and the Causal Mismatch

As the generation stride increases, the trajectory shifts toward a bi-terminal convergence pattern, as illustrated in Figure 7(b). The model tends to commit to both the sequence prefix and suffix tokens early in the denoising process, subsequently "filling in" the intermediate tokens. While this "pyramid" path appears to leverage global context, it introduces what we define as the *Causal Mismatch*.

Specifically, pre-sampling future tokens (e.g., sequence endings) without a coherent chain of intermediate reasoning risks *semantic fracture* (§4.4)—a state where the generated "ends" are logically irreconcilable with the high-entropy reasoning steps that follow. Moreover, such non-contiguous filling precludes any spatial locality in memory access, fundamentally breaking standard caching mechanisms. This observed "filling-from-ends" behavior validates the necessity of our Elastic Horizon Modeling, which replaces this rigid, potentially incoherent global lookahead with a confidence-driven, dynamically modulated window that ensures each parallel step is supported by its causal history.

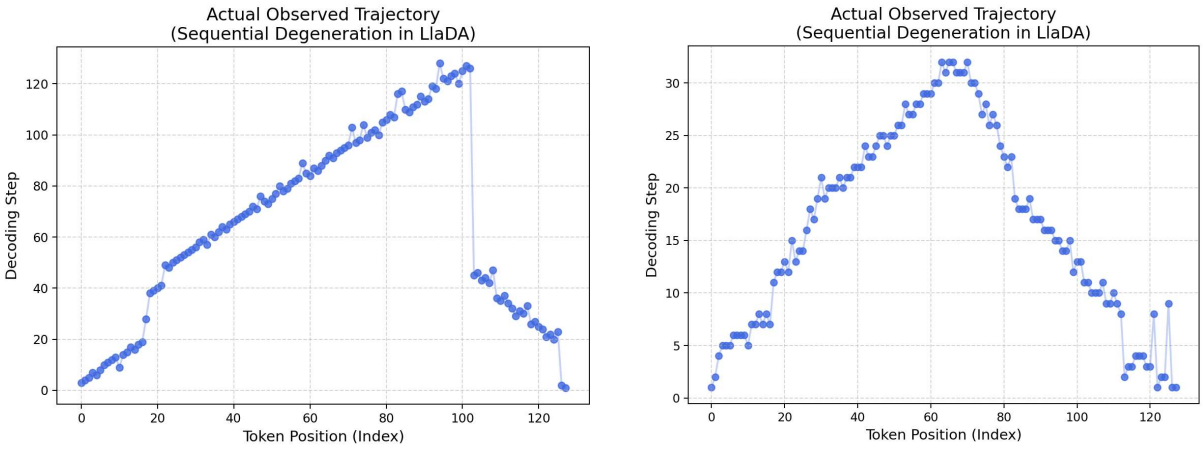
A.3 Synthesis: From Redundancy to Fluidity

Synthesizing these observations, it becomes evident that bidirectional diffusion models for text generation are often caught in an "architectural limbo." They carry the $O(N^2)$ computational burden of acausal attention, but either degenerate into sequential paths for quality or adopt non-linear paths that jeopardize logical consistency. By reframing the diffusion process within a strictly causal framework and modulating the decoding horizon elastically, FLUID resolves these contradictions, transforming the latent causal bias of LLMs from a hidden degeneration into an explicit, exploitable advantage for efficient parallel generation.

B Training Dynamics and Convergence

We monitor FLUID's training stability to assess the efficiency of our adaptation curriculum. Figure 8 shows the training loss (weighted cross-entropy) during the first stage of joint causal backbone training (§4.3).

The plot reveals a rapid initial descent within the first 1,000 iterations, indicating effective realignment of the pre-trained openPangu-Embedded-7B priors with the causal diffusion objective. After around 10,000 steps, the loss stabilizes, reflecting



(a) single-token per-step decoding (b) four-token per-step decoding

Figure 7: Decoding Trajectories of Bidirectional Diffusion. (a) Incremental decoding ($K = 1$) triggers an unintended sequential path, revealing a structural redundancy in bidirectional attention. (b) Block-wise decoding ($K = 4$) exhibits bi-terminal convergence, where non-linear filling from sequence boundaries leads to causal mismatch. These behaviors substantiate FLUID's design of strictly causal alignment and adaptive horizon modulation.

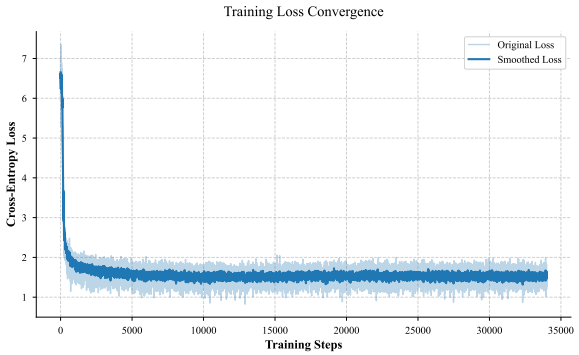


Figure 8: Training loss convergence of FLUID during Stage I. The rapid stabilization post-initialization highlights the efficiency of adapting pre-trained AR models to the diffusion paradigm via strictly causal alignment.

mastery of the unidirectional denoising task under triangular attention constraints.

The convergence remains stable throughout the 32,000 iterations of Stage I, with the smoothed loss maintaining a plateau. This confirms that our hybrid loss function (§4.3) preserves linguistic competencies while adapting to parallel generation. The subsequent 2,000 steps of Stage II, focused on calibrating the Diffusion K-Head, converge quickly due to the frozen backbone's established confidence boundaries.

C Case Study

To better understand how FLUID adapts its decoding horizon during reasoning, we visualize its decoding trajectory on an arithmetic example in Figure 9. The figure shows that FLUID does not decode with a fixed generation block. Instead, its horizon changes dynamically with local uncertainty, as reflected by the token-level entropy within the current window.

A clear low-entropy expansion and high-entropy contraction pattern can be observed. In Block 1, FLUID selects a large horizon ($K = 15$) and commits a long contiguous span, since the corresponding tokens lie in a relatively stable low-entropy region and follow a predictable local reasoning template. As decoding approaches a semantic transition region, however, the entropy rises sharply around tokens such as "Yeah" and "seems", indicating increased uncertainty. FLUID therefore shrinks its horizon to $K = 4$ in the following steps and switches to more conservative updates. This behavior shows that FLUID accelerates only in confident regions, while automatically slowing down near uncertain decision points.

This example highlights the key advantage of Elastic Horizons under strictly causal decoding. Natural language reasoning is heterogeneous: some spans are highly predictable, whereas others correspond to semantic transitions that require more cautious generation. Fixed-block decoding cannot adapt to such variation, and may therefore either over-commit across difficult regions or sacrifice efficiency in easy ones. In contrast, FLUID dynamically adjusts its horizon according to local uncertainty, accelerating through stable segments while preserving causal reliability near high-entropy regions. As a result, it achieves a better balance

Figure 9: Visualization of FLUID's decoding trajectory, token entropy, and adaptive horizon on an arithmetic reasoning prompt.

between decoding efficiency and generation quality.

Stage	Trained Module	Hardware	Est. Time
Stage I	LoRA Adapter	4×64GB GPUs	∼80 hours
Stage II	K-Head	4×64GB GPUs	negligible
Total	–	4×64GB GPUs	∼320 GPU-hours

Table 5: Computational overhead of FLUID adaptation. The total cost is approximately 320 GPU-hours, with Stage I accounting for nearly all of the training time.

D Training Cost Analysis

A practical advantage of FLUID is its low adaptation cost compared with training diffusion language models from scratch. Since FLUID reuses a pretrained autoregressive backbone and only introduces lightweight adaptation modules, the overall training overhead remains modest. In our setting, the full adaptation requires approximately 320 GPU-hours in total, corresponding to about 80 wall-clock hours on 4×64 GB GPUs.

As summarized in Table 5, Stage I accounts for nearly all of the training cost. In this stage, we adapt the backbone using Rank-16 LoRA for 32K iterations over approximately 2.62B tokens. By contrast, Stage II trains only a lightweight two-layer K-Head while keeping the backbone frozen, thereby introducing negligible additional overhead.

Importantly, FLUID follows the standard recipe of parameter-efficient fine-tuning rather than expensive full-model pretraining. Most backbone parameters remain frozen, and the training footprint is comparable to typical PEFT-based autoregressive fine-tuning. Therefore, FLUID can be readily scaled with standard multi-GPU training frameworks, without requiring the heavy compute budget typically associated with training diffusion language models from scratch.