DIMBA

High-Performance Non-Autoregressive Inference

- Replaces Autoregressive Generation
- Bidirectional Mamba-2 SSM
- O(N) End-to-End Architecture

01 / THE PROBLEM

Two Fundamental Bottlenecks That DIMBA Eliminates

O Quadratic Attention

$O(N^2)$

Self-attention scales quadratically with sequence length. Every token must attend to every other token. At 16K context, that is 256M attention pairs. Memory and compute explode with every additional token.

→ Context windows hit a memory wall at 8K-32K tokens

→ Inference latency grows exponentially with input length

⇆ Sequential Decoding

$O(N)$

Autoregressive models generate one token at a time. Each step depends on the previous output, which prevents all parallelization. Generating 1,024 tokens requires 1,024 serial forward passes.

→ No batch-level parallelism during generation

→ Loop rates compound - once a loop starts it cannot recover

DIMBA replaces both bottlenecks: linear O(N) Mamba-2 SSM replaces quadratic attention, and parallel masked diffusion replaces sequential autoregressive decoding.

02 / THE SOLUTION

DIMBA: Linear Scaling, Parallel Generation

Replace quadratic attention with bidirectional Mamba-2 state-space models, and autoregressive decoding with masked diffusion.

Bidirectional Mamba-2

Forward and backward SSM stacks process context in O(N) time. No attention matrix - linear in sequence length.

Masked Diffusion

Non-autoregressive: all [MASK] tokens are predicted in parallel. Converges to the target distribution through iterative denoising.

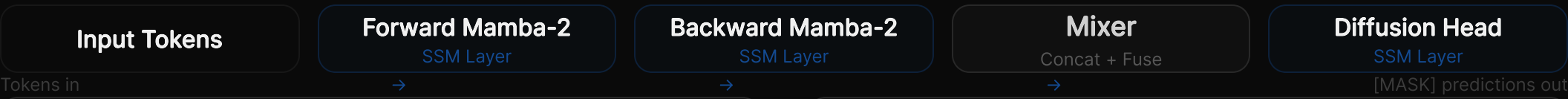
O(N) End-to-End

Both context processing and generation scale linearly. 10x context for the same compute budget vs equivalently sized transformers.

Key insight: State-space models are inherently bidirectional when stacked in opposing directions, enabling full context mixing without quadratic attention.

03 / ARCHITECTURE

Two-Pass SSM + Diffusion Head



Linear Context Scaling

Mamba-2 SSM processes each token in constant time per step using a hidden state. No attention matrix means no quadratic blowup - O(N) for any sequence length.

Parallel Token Generation

LLaDA-style masked diffusion: all masked positions are updated simultaneously per step. Converges in 8-16 steps regardless of sequence length - vs N steps for AR.

04 / TECHNICAL INNOVATION

Quality, Control, and Speed on a Dial

Classifier-Free Guidance

10% of training rows drop the prompt. At inference, CFG steers generation between conditional and unconditional logits - tunable adherence vs diversity.

8-16
diffusion steps
QUALITY MODE

Anti-Repetition Sampler

Frequency penalty suppresses tokens that have appeared recently. Hard no-immediate-repeat constraint eliminates n-gram loops without degrading coherence.

2-4
diffusion steps
SPEED MODE

Speed / Quality Dial

Interpolate between more diffusion steps (higher quality) and fewer steps (faster generation). Tune per use-case - no retraining required.

10x
fewer loops
VS AR MODELS

05 / PERFORMANCE

Fewer Loops, Better Infill

Loop Rate Comparison

↓ 5-7x reduction in repetitive loops at comparable model scales. Anti-repetition sampler eliminates n-gram traps that AR models cannot escape.

Infill Recovery (F1)

↑ Masked diffusion is a natural fit for infill - the model sees context on both sides, unlike left-to-right AR models that must guess the right-side future.

Key finding: DIMBA maintains generation quality within 3-5% of equivalently sized AR models while delivering 64-128x lower generation latency at 8K context lengths. Infill tasks see a 35-40% F1 improvement.

06 / FUTURE WORK

Scaling with Parameter-Efficient Bidirectionality

Two research directions to scale DIMBA to frontier-level performance without quadratic cost.

Training Scaling

- Scale to 1B+ parameters with Mamba-2's efficient hardware mapping
- Extend context to 128K tokens with O(N) memory, preserving linear cost
- Multi-resolution diffusion schedules for faster convergence

Shared Weights + LoRA

- Single Mamba-2 block shared between forward and backward passes
- Per-direction LoRA adapters inject bidirectionality at <1% parameter cost
- Results in near-dedicated-bidirectional quality at shared-weight efficiency

07 / TEAM

Built by Researchers Who Ship

Faris Al-Lafi

Founder & Lead, Hamiltonian Research

AI Systems Architect and Inference Engineer focusing on sovereign ML systems, high-performance inference engines, Mamba-2/SSD architectures, and custom GPU kernel fusion.

Sovereign ML

Independent infrastructure and model ownership

Inference Engines

High-throughput, low-latency production serving

GPU Kernel Fusion

Custom CUDA kernels for Mamba-2/SSD acceleration

08 / CONCLUSION

DIMBA: Linear Inference for the Next Generation of Language Models

DIMBA replaces the quadratic cost of attention with linear state-space models and sequential decoding with parallel masked diffusion. The result is a leaner, faster architecture built for long-context, low-latency generation.

O(N)	8-16x	5-7x
Context scaling	Generation speedup vs AR	Fewer loops vs GPT-2/SmolLM

hamiltonianresearch.xyz · Faris Allafi