take | 2 | add | [MASK]

Encoder representation

Large Encoder processes clean tokens

Small Decoder denoises tokens while attending to encoder representations

Decoding steps produce new tokens without invoking encoder

What | is | 2 | + | 5

Prompt (+ accumulated samples)

[MASK] | 2 | add | [MASK]

Noisy input block

✓ **Faster generation** compared to decoder-only AR and diffusion

✓ **Extends Pareto frontier** for quality-throughput trade-off

✓ **More efficient training** than decoder-only block diffusion