Block Diffusion: Interpolating between Autoregressive and Diffusion Language Models

ICLR 2025 Oral

Marianne Arriola, PhD student @ Cornell Tech


Marianne Arriola


Aaron Gokaslan


Justin T. Chiu


Zhihan Yang


Zhixuan Qi

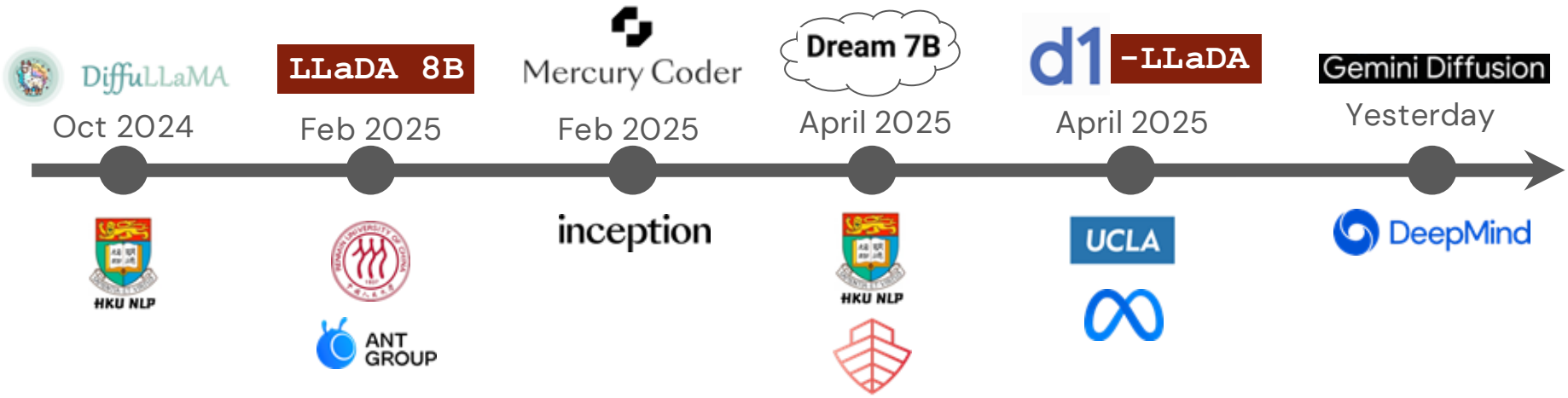

Jiaqi Han


Subham Sahoo


Volodymyr Kuleshov

Anatoly

Diffusion LLMs are here!

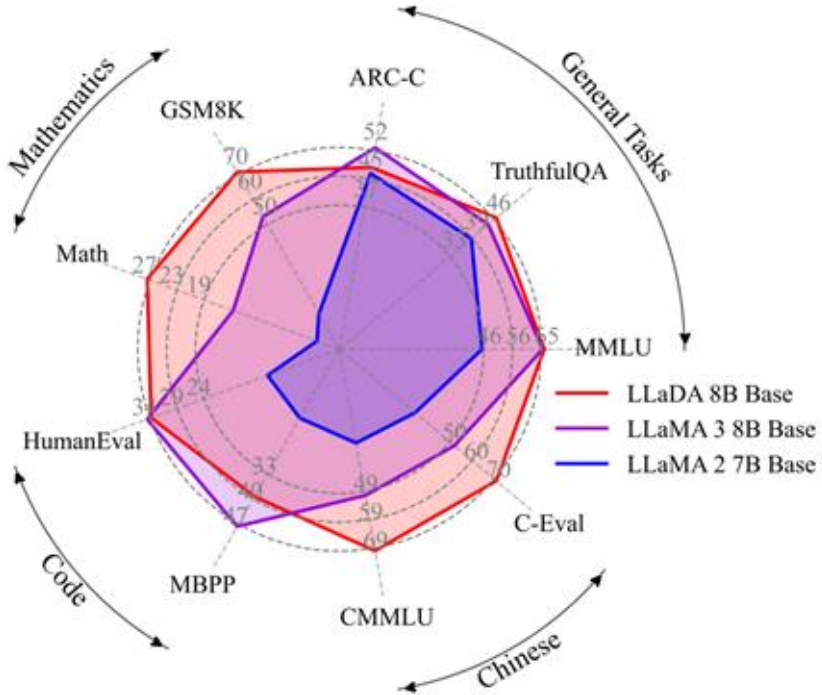


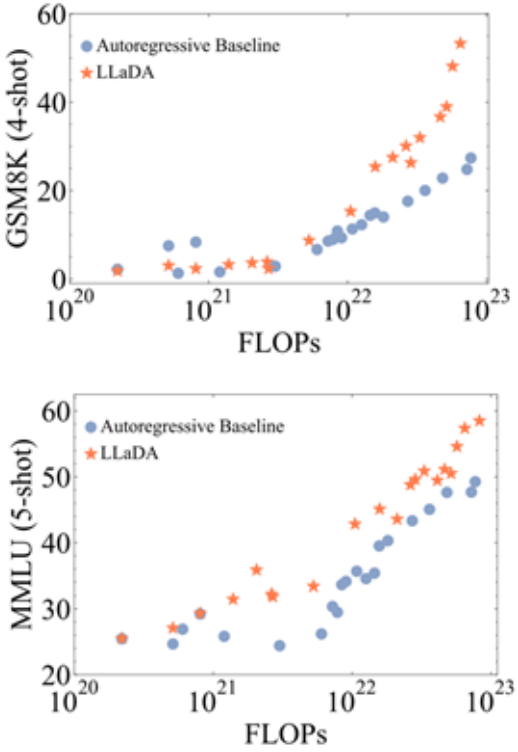
Diffusion LLMs generate text ≤**10x faster** than traditional LLMs



MercuryCoder video credits: Inception Labs.

Diffusion is competitive on **benchmarks** and shows strong **scalability**





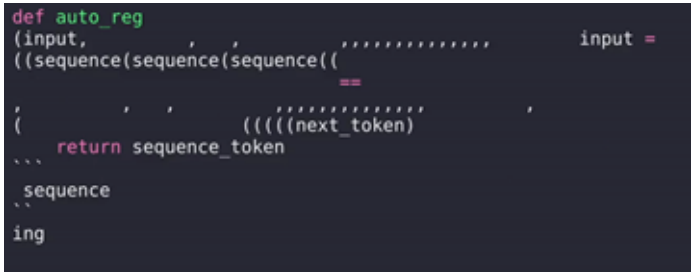
Nie et al. "Large Diffusion Language Models." arXiV preprint 2025.

Diffusion-generated text is **more controllable**

Better Guidance

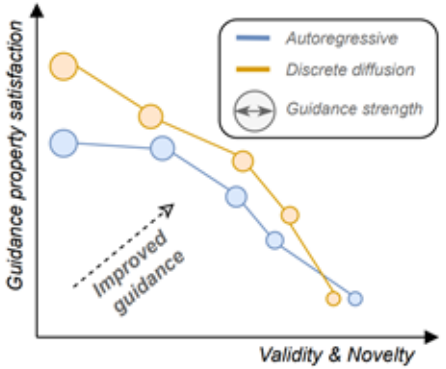
❌ Autoregressive models make **“local”** predictions

✅ Diffusion models make **“global”** refinements



MercuryCoder video credits: Inception Labs.

Molecule design



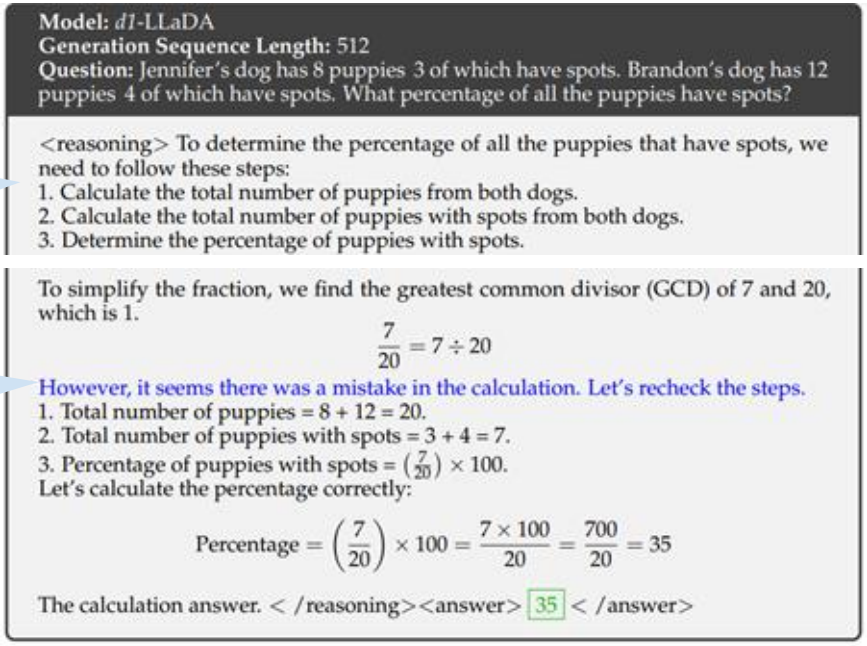
Schiff et al. “Simple Guidance Mechanisms for Discrete Diffusion Models.” ICLR 2025.

Diffusion-generated text is **more controllable**

RL for reasoning

Planning

Self-correction



Zhao et al. "d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning." Blog post 2025.

Diffusion still faces practical limitations

Autoregression

There

- ❌ Sequential
- ❌ Causal context only
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching

Diffusion

Hirsch
account

- ✅ Parallel
- ✅ Global context
- ❌ Lower quality
- ❌ Fixed-length
- ❌ No KV caching

Can we make diffusion LMs that are

- ➢ **practical** like traditional AR LMs
- ➢ while preserving their **core advantages**?

Can we combine strength from both paradigms?

Autoregression

There are three possible categories of the typical rate of return for companies that share...

- ❌ Sequential
- ❌ Causal context only
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching

Block Diffusion *(ours)*

On

- ✅ Parallel
- ✅ Semi-global context
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching

Diffusion

Hirsch will need to take account data that's to be released by tomorrow.

- ✅ Parallel
- ✅ Global context
- ❌ Lower quality
- ❌ Fixed-length
- ❌ No KV caching

Simple and Effective

Masked Diffusion Language Models

NeurIPS 2024

On

Block Diffusion:

Interpolating between AR and Diffusion LMs

ICLR 2025 Oral

Future Directions

Simple and Effective

Masked Diffusion Language Models

NeurIPS 2024

Block Diffusion:

Interpolating between AR and Diffusion LMs

ICLR 2025 Oral

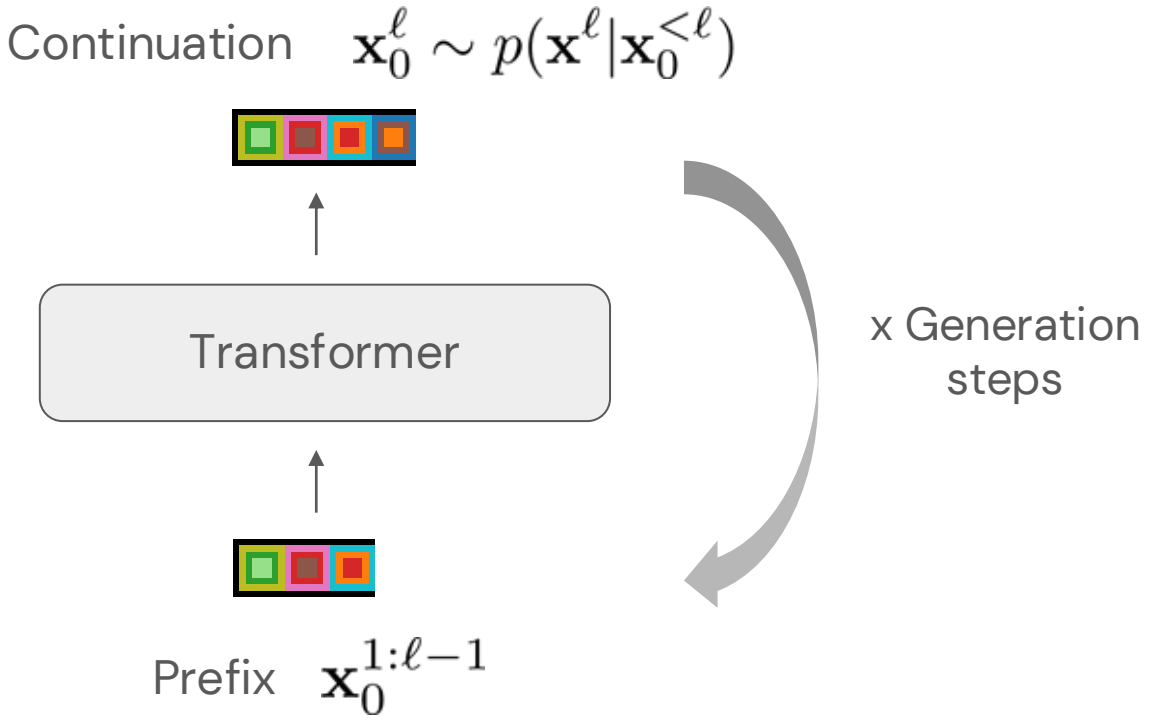
Future Directions

Goal: Approximate the distribution of language

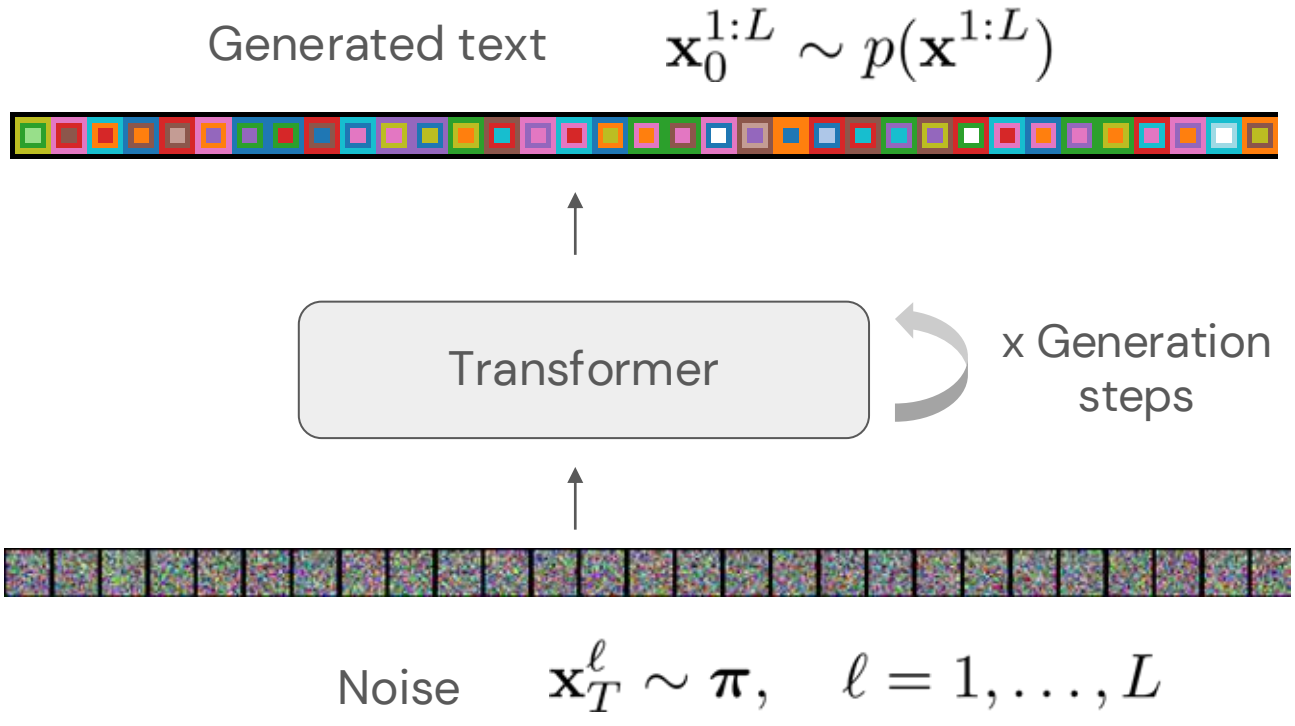
Generated text $\mathbf{x}_0^{1:L} \sim p(\mathbf{x}^{1:L})$



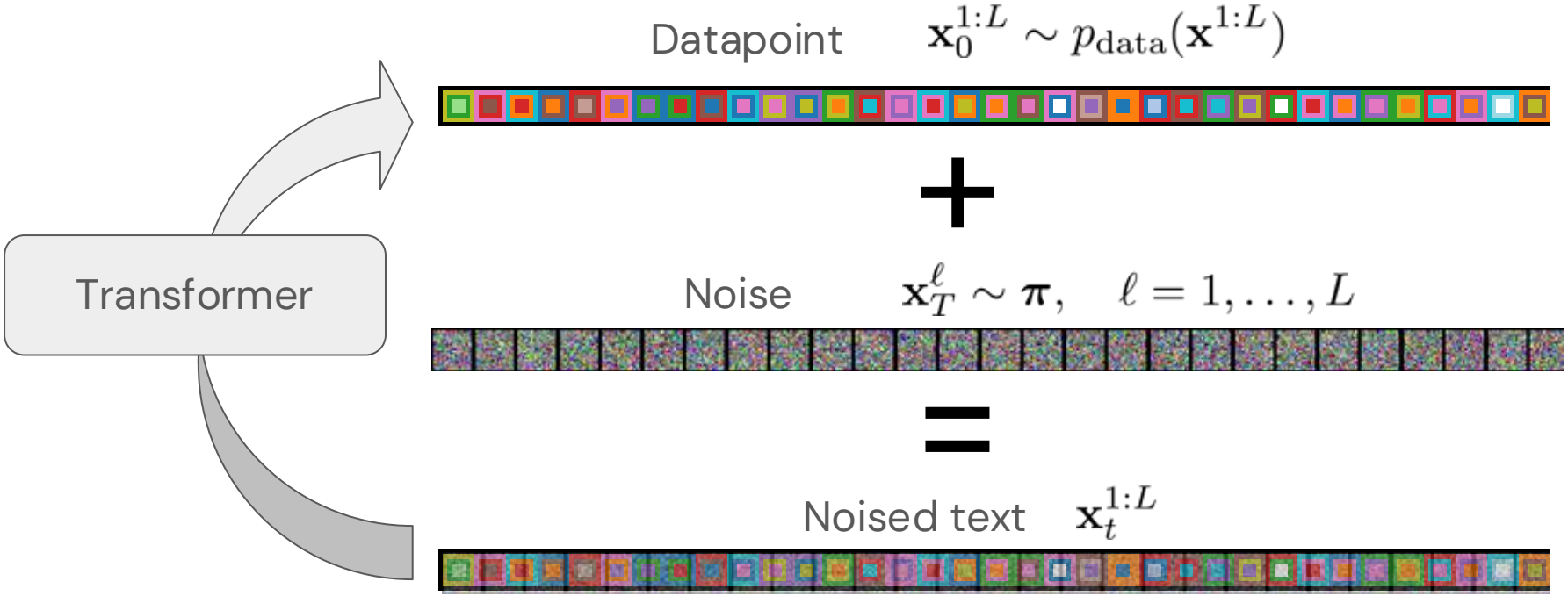
Goal: Approximate the distribution of language



Diffusion LMs generate text from noise



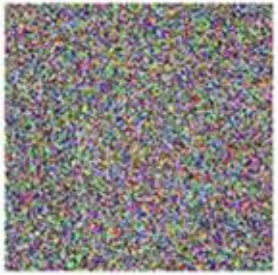
Training a model to undo noise



How do we add noise to text?



+



=



The quick brown fox jumps over the lazy dog

+



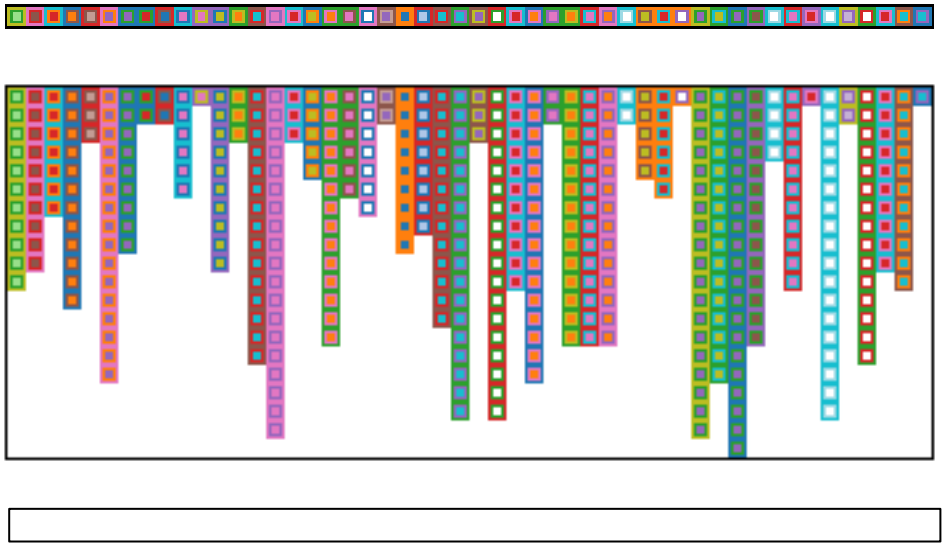
=

?

Inspired from Weinberger, K. Future of LLMs workshop.

Training a model to unmask text

Datapoint $\mathbf{x}_0^{1:L} \sim p_{\text{data}}(\mathbf{x}^{1:L})$

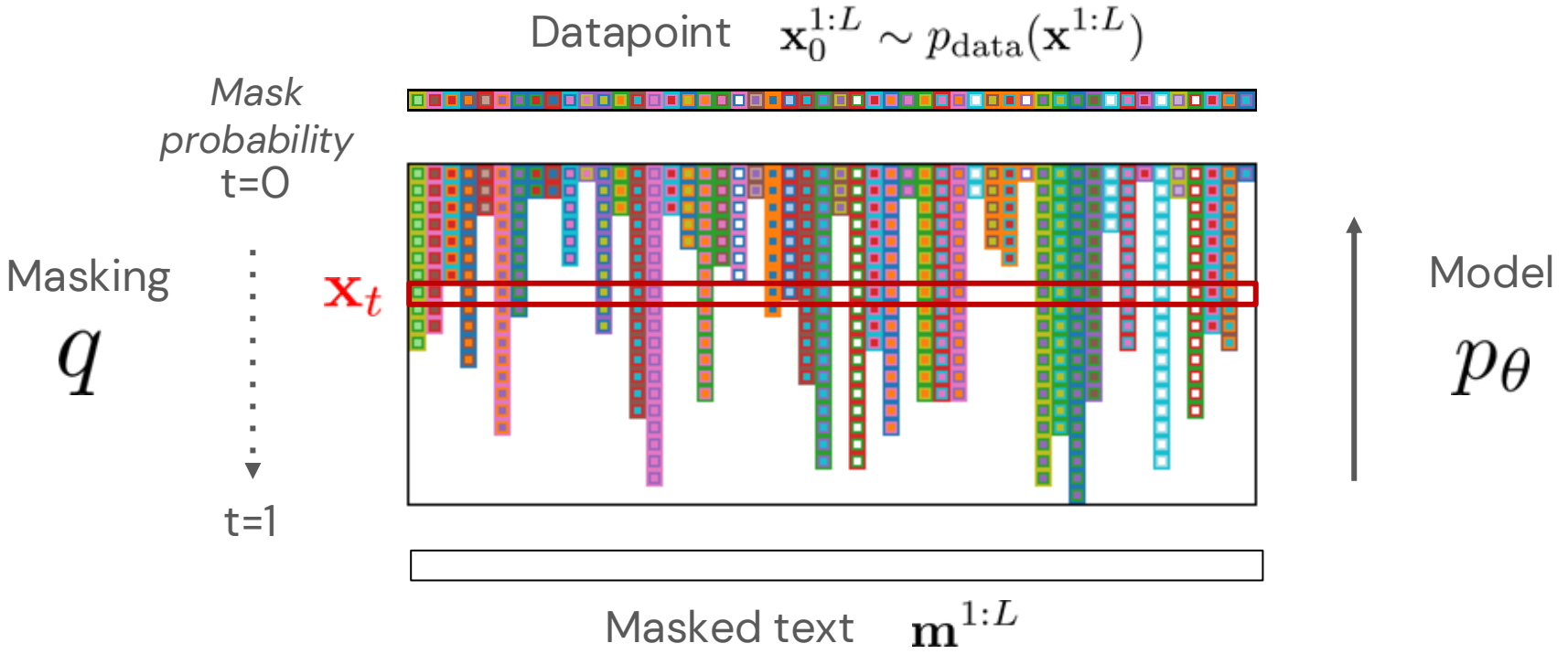


Model p_θ

Masked text $\mathbf{m}^{1:L}$

Austin et al., "Structured Denoising Diffusion Models in Discrete State-Spaces." NeurIPS 2021

Training a model to unmask text



Austin et al., "Structured Denoising Diffusion Models in Discrete State-Spaces." NeurIPS 2021

Training a model to unmask text



Austin et al., "Structured Denoising Diffusion Models in Discrete State-Spaces." NeurIPS 2021

Simple and Effective Masked Diffusion Language Models

NeurIPS 2024

Hirsch account


Subham Sahoo


Marianne Arriola


Yair Schiff


Aaron Gokaslan


Edgar Marroquin


Justin T. Chiu


Alexander Rush


Volodymyr Kuleshov

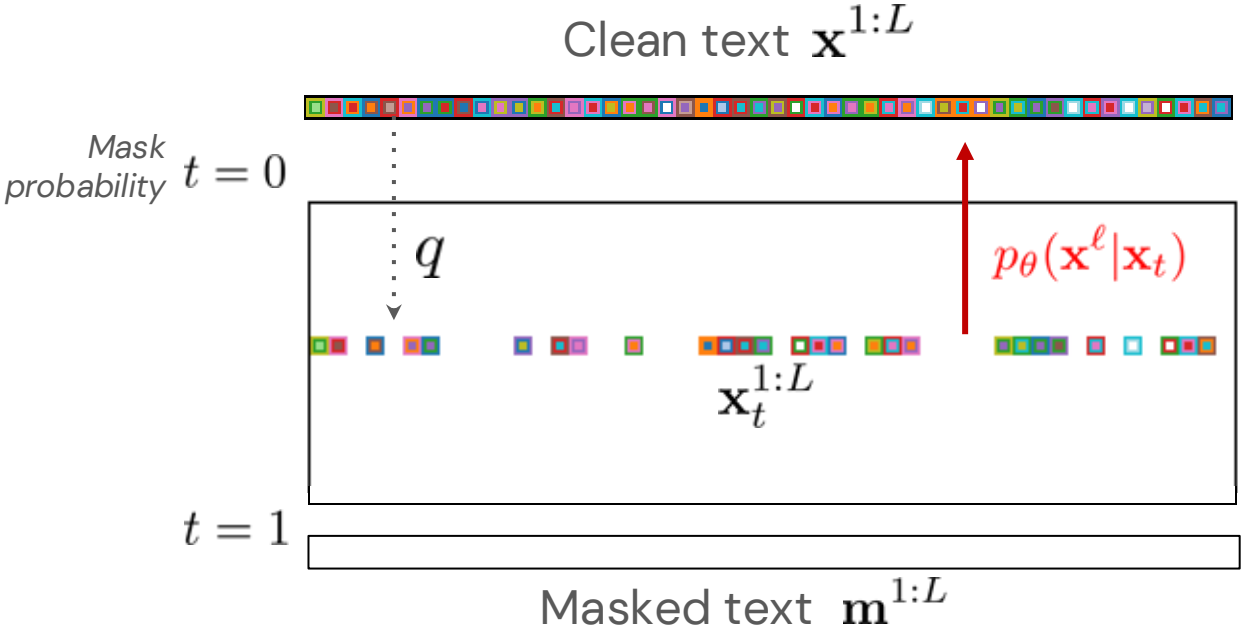




Slides inspired from Rush, A. "Simple Diffusion Language Models." YouTube.

A simple likelihood objective

$\mathbf{x}^\ell$ one-hot vector over the vocabulary



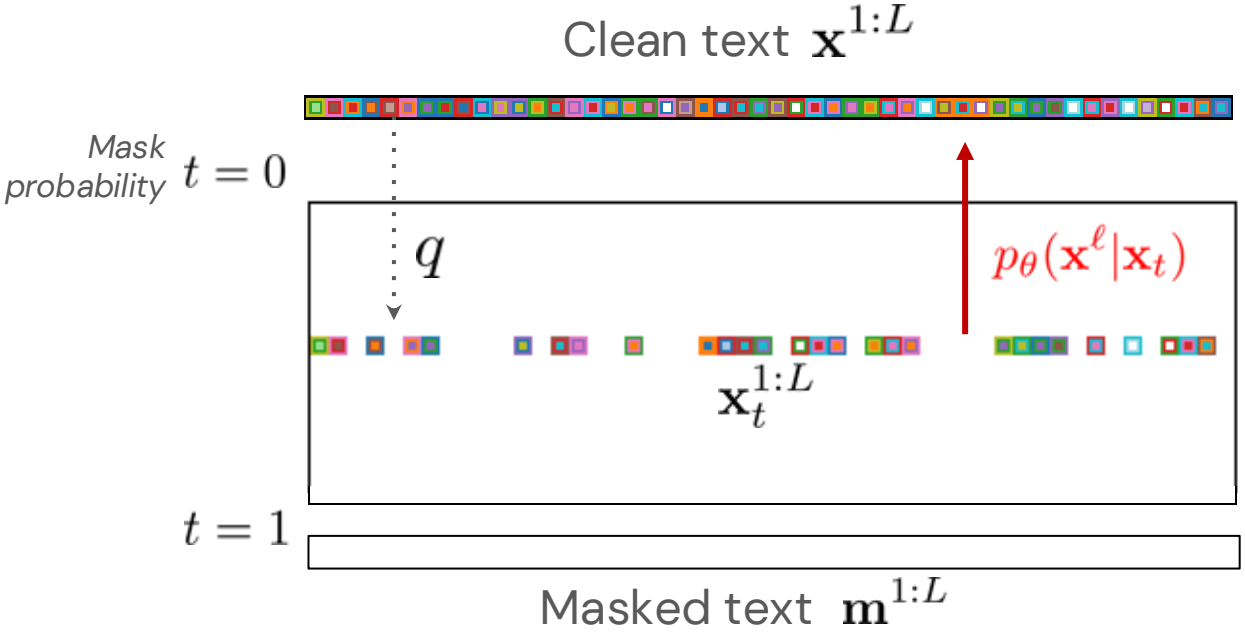
Weighted avg of MLM losses



$$-\log p_\theta(\mathbf{x}) \leq \mathbb{E}_{t,\mathbf{x}_t\sim q} -\frac{1}{t}\sum_\ell^L \log p_\theta(\mathbf{x}^\ell|\mathbf{x}_t)$$

A simple likelihood objective

$\mathbf{x}^\ell$ one-hot vector over the vocabulary



Weighted avg of MLM losses



$$-\log p_\theta(\mathbf{x}) \leq \mathbb{E}_{t,\mathbf{x}_t\sim q} -\frac{1}{t}\sum_\ell^L \log p_\theta(\mathbf{x}^\ell|\mathbf{x}_t)$$

Loss for unmasked tokens is 0

$$\log p_\theta(\mathbf{x}^\ell|\mathbf{x}^\ell \neq \mathbf{m}, \mathbf{x}_t) = 0$$

Inference: One Step of Generation

$$0 \leq s < t \leq 1$$

Next unmasked sequence $\mathbf{x}_s^{1:L}$



Masked input $\mathbf{x}_t^{1:L}$



Inference: One Step of Generation

$$0 \leq s < t \leq 1$$

Next unmasked sequence $\mathbf{x}_s^{1:L}$



1) Prediction $\mathbf{x}_\theta^{1:L}$



Masked input $\mathbf{x}_t^{1:L}$



Inference: One Step of Generation

$$0 \leq s < t \leq 1$$

Next unmasked sequence $\mathbf{x}_s^{1:L}$





2) Retain predictions

$$p_\theta(\mathbf{x}_s^\ell = \mathbf{x}_\theta^\ell | \mathbf{x}_t, \mathbf{x}_\theta^\ell) = \frac{t-s}{t}$$

1) Prediction $\mathbf{x}_\theta^{1:L}$



Masked input $\mathbf{x}_t^{1:L}$



Inference: One Step of Generation

$$0 \leq s < t \leq 1$$

Next unmasked sequence $\mathbf{x}_s^{1:L}$





2) Retain predictions

$$p_\theta(\mathbf{x}_s^\ell = \mathbf{x}_\theta^\ell | \mathbf{x}_t, \mathbf{x}_\theta^\ell) = \frac{t-s}{t}$$

1) Prediction $\mathbf{x}_\theta^{1:L}$

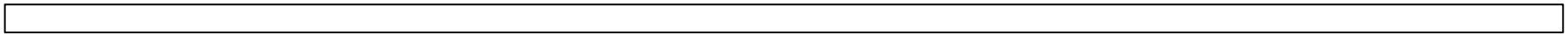


3) Retain unmasked inputs

Masked input $\mathbf{x}_t^{1:L}$



Sampling from a Masked Diffusion Language Model



Sampling from a Masked Diffusion Language Model

squad,
to remember when
to
twenty adobe of water
along polished which
prehistoric many
in



Sampling from a Masked Diffusion Language Model

Many years later, as he faced squad,
Buendía was to remember that distant when his
took to discover At that was a village of
twenty adobe houses, built the bank of river of clear water
that ran along of polished stones, which were and
like prehistoric eggs. so recent many
things lacked names, and in them it was
to



Sampling from a Masked Diffusion Language Model

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world so recent that many things lacked names, and in order indicate them it was to point.

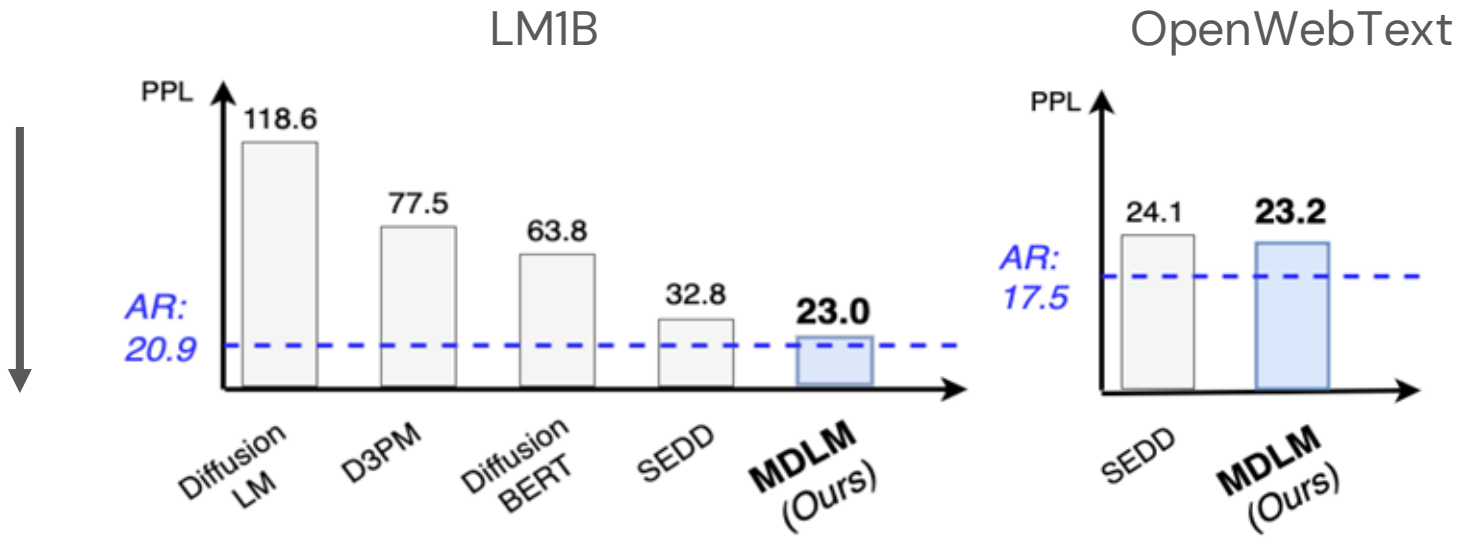


Sampling from a Masked Diffusion Language Model

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.



Likelihood Evaluation



AR, SEDD, MDLM use
110M parameter transformer

Simple and Effective

Masked Diffusion Language Models

NeurIPS 2024

On

Block Diffusion:

Interpolating between AR and Diffusion LMs

ICLR 2025 Oral

Future Directions

Can we combine strength from both paradigms?

Autoregression

There are three possible categories of the typical rate of return for companies that share...

- ❌ Sequential
- ❌ Causal context only
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching

Block Diffusion *(ours)*

On

- ✅ Parallel
- ✅ Semi-global context
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching

Diffusion

Hirsch will need to take account data that's to be released by tomorrow.

- ✅ Parallel
- ✅ Global context
- ❌ Lower quality
- ❌ Fixed-length
- ❌ No KV caching

Block Discrete Denoising Diffusion Language Models (BD3-LMs)

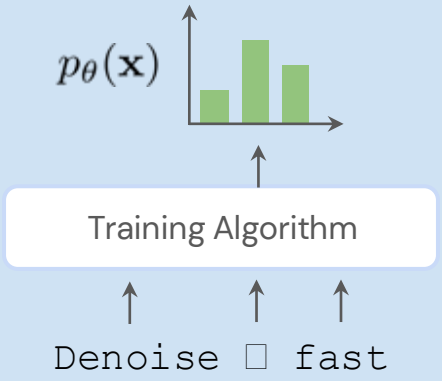
Better Likelihoods

$$-\log p_\theta(\mathbf{x})$$

$$\leq \mathcal{L}_{\text{BlockDiffusion}}(\mathbf{x};\theta)$$

$$\leq \mathcal{L}_{\text{Diffusion}}(\mathbf{x};\theta)$$

Efficient Training



Arbitrary-Length Generation with KV caching

On September 16, 2016, we are planning to give the game's beta-release to...

Block Discrete Denoising Diffusion Language Models (BD3-LMs)

Better Likelihoods

$$-\log p_\theta(\mathbf{x})$$

$$\leq \mathcal{L}_{\text{BlockDiffusion}}(\mathbf{x};\theta)$$

$$\leq \mathcal{L}_{\text{Diffusion}}(\mathbf{x};\theta)$$

Efficient Training



Arbitrary-Length Generation with KV caching

On September 16, 2016, we are planning to give the game's beta-release to...

Block Diffusion unifies AR & Diffusion objectives

Sequence of tokens

$\mathbf{x}^{1:L} \sim p_{\text{data}}(\mathbf{x}^{1:L})$ $\mathbf{x}^{\ell}$ one-hot vector over the vocab

*Sequence of **blocks***

$\mathbf{x}^{b}$ denotes tokens in block $b \in \{1, \ldots, B\}$

$$-\log p_\theta(\mathbf{x}^{1:B})$$

Anatoly is well-known for | his witty chess plays, like | the bold Queen's Gambit

Block Diffusion unifies AR & Diffusion objectives

Sequence of tokens

$\mathbf{x}^{1:L} \sim p_{\text{data}}(\mathbf{x}^{1:L})$ $\mathbf{x}^{\ell}$ one-hot vector over the vocab

*Sequence of **blocks***

$\mathbf{x}^{b}$ denotes tokens in block $b \in \{1, \ldots, B\}$

$$-\log p_\theta(\mathbf{x}^{1:B}) = -\sum_{b=1}^{B} \log p_\theta(\mathbf{x}^b | \mathbf{x}^{<b})$$

b

Anatoly is well-known for

Anatoly is well-known for his witty chess plays, like

Anatoly is well-known for his witty chess plays, like the bold Queen's Gambit

Block Diffusion unifies AR & Diffusion objectives

Sequence of tokens

$\mathbf{x}^{1:L} \sim p_{\text{data}}(\mathbf{x}^{1:L})$ $\mathbf{x}^{\ell}$ one-hot vector over the vocab

*Sequence of **blocks***

$\mathbf{x}^{b}$ denotes tokens in block $b \in \{1, \ldots, B\}$

$$-\log p_\theta(\mathbf{x}^{1:B}) = -\sum_{b=1}^{B} \log p_\theta(\mathbf{x}^b|\mathbf{x}^{<b}) \leq \sum_{b=1}^{B} \mathcal{L}_{\text{diff}}(\mathbf{x}^b|\mathbf{x}^{<b};\theta)$$

b

Anatoly is well-known for

Anatoly is well-known for his witty chess plays, like

Anatoly is well-known for his witty chess plays, like the bold Queen's Gambit

Block Diffusion unifies AR & Diffusion objectives

Sequence of tokens

$\mathbf{x}^{1:L} \sim p_{\text{data}}(\mathbf{x}^{1:L})$ $\mathbf{x}^{\ell}$ one-hot vector over the vocab

*Sequence of **blocks***

$\mathbf{x}^{b}$ denotes tokens in block $b \in \{1, \ldots, B\}$

$$-\log p_\theta(\mathbf{x}^{1:B}) = -\sum_{b=1}^{B} \log p_\theta(\mathbf{x}^b|\mathbf{x}^{<b}) \leq \sum_{b=1}^{B} \mathcal{L}_{\text{diff}}(\mathbf{x}^b|\mathbf{x}^{<b};\theta)$$

Smaller blocks → Tighter bound

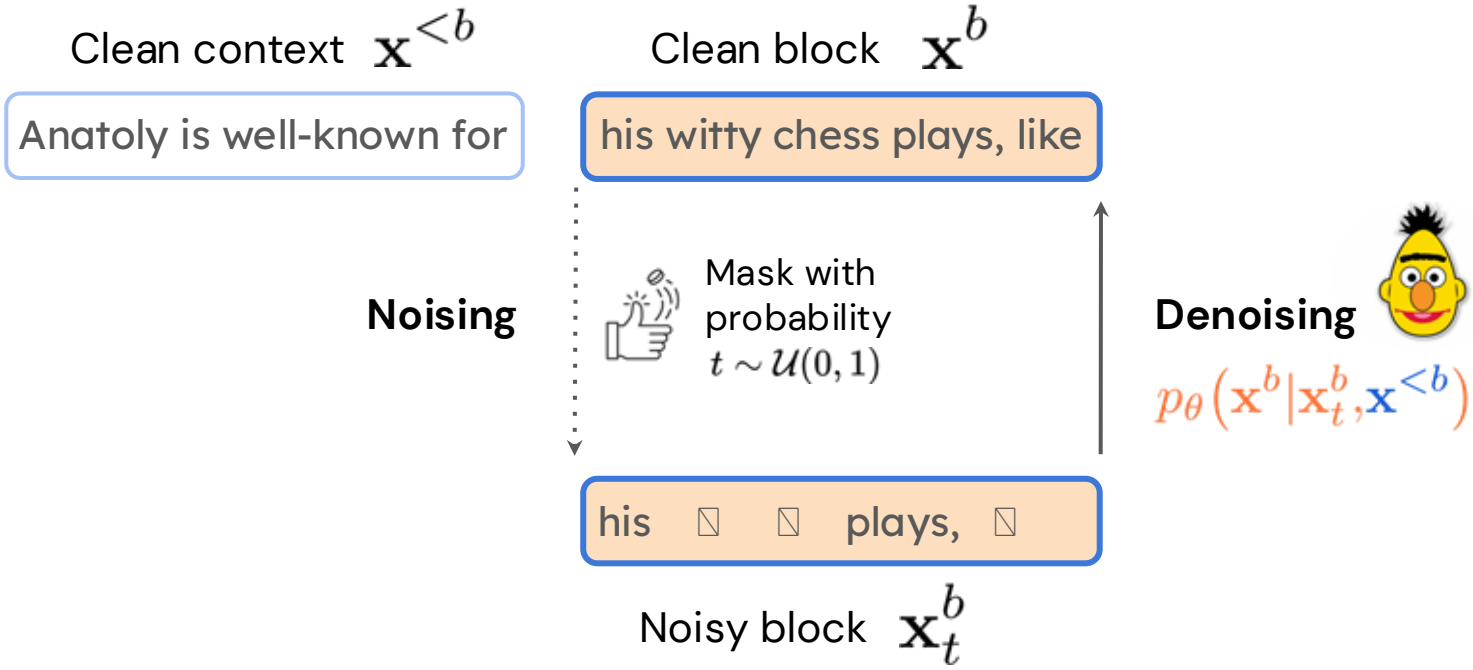
b

Anatoly is well-known for

Anatoly is well-known for his witty chess plays, like

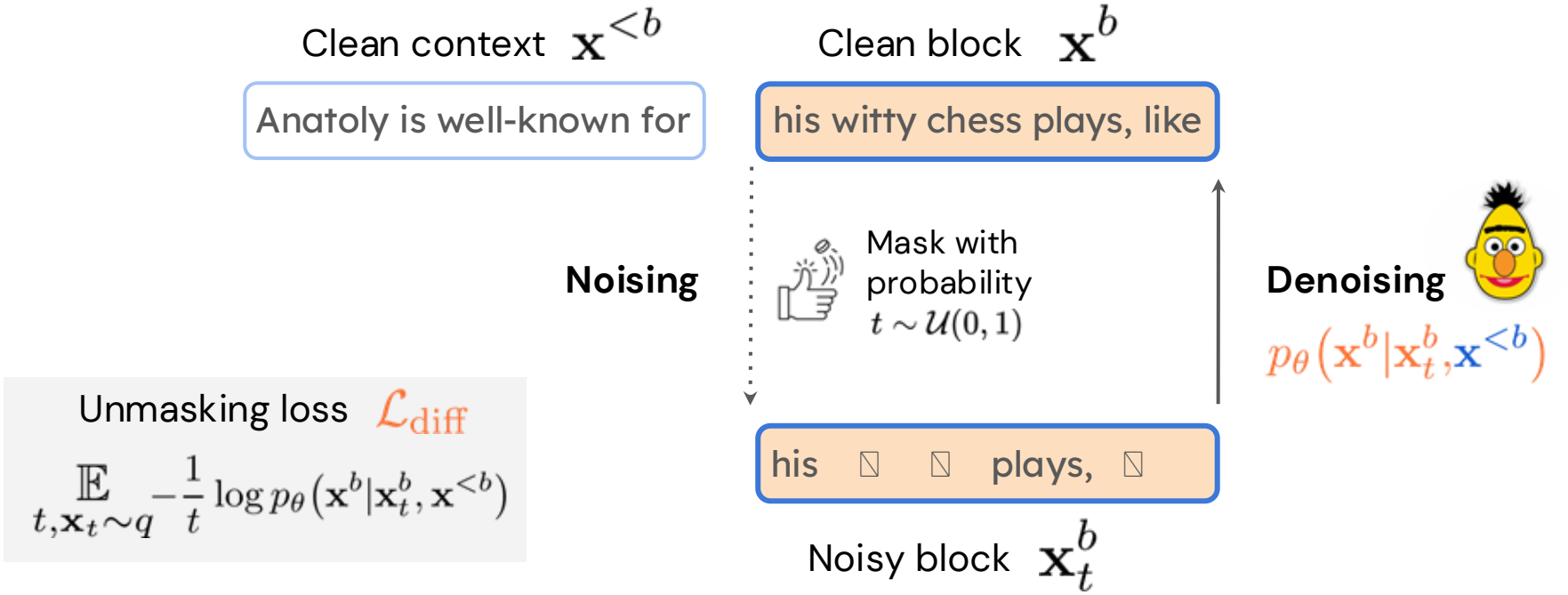
Anatoly is well-known for his witty chess plays, like the bold Queen's Gambit

Masked Block Diffusion LMs



Sahoo et al. "Simple and Effective Masked Diffusion Language Models." NeurIPS 2024.

Masked Block Diffusion LMs



Sahoo et al. "Simple and Effective Masked Diffusion Language Models." NeurIPS 2024.

Block **D**iscrete **D**enoising **D**iffusion **L**anguage **M**odels (BD3-LMs)

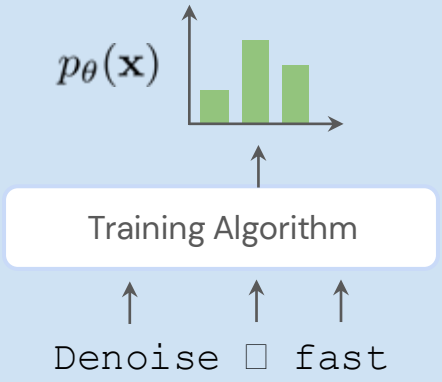
Better Likelihoods

$$-\log p_\theta(\mathbf{x})$$

$$\leq \mathcal{L}_{\text{BlockDiffusion}}(\mathbf{x};\theta)$$

$$\leq \mathcal{L}_{\text{Diffusion}}(\mathbf{x};\theta)$$

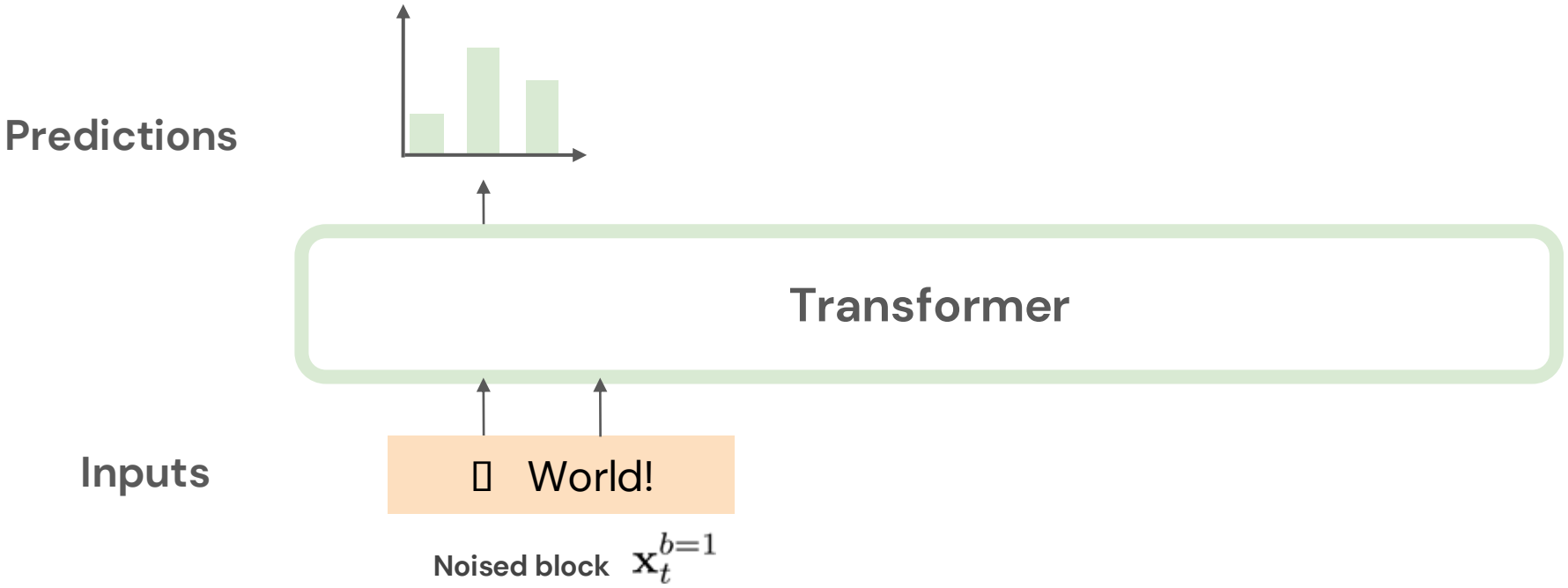
Efficient Training



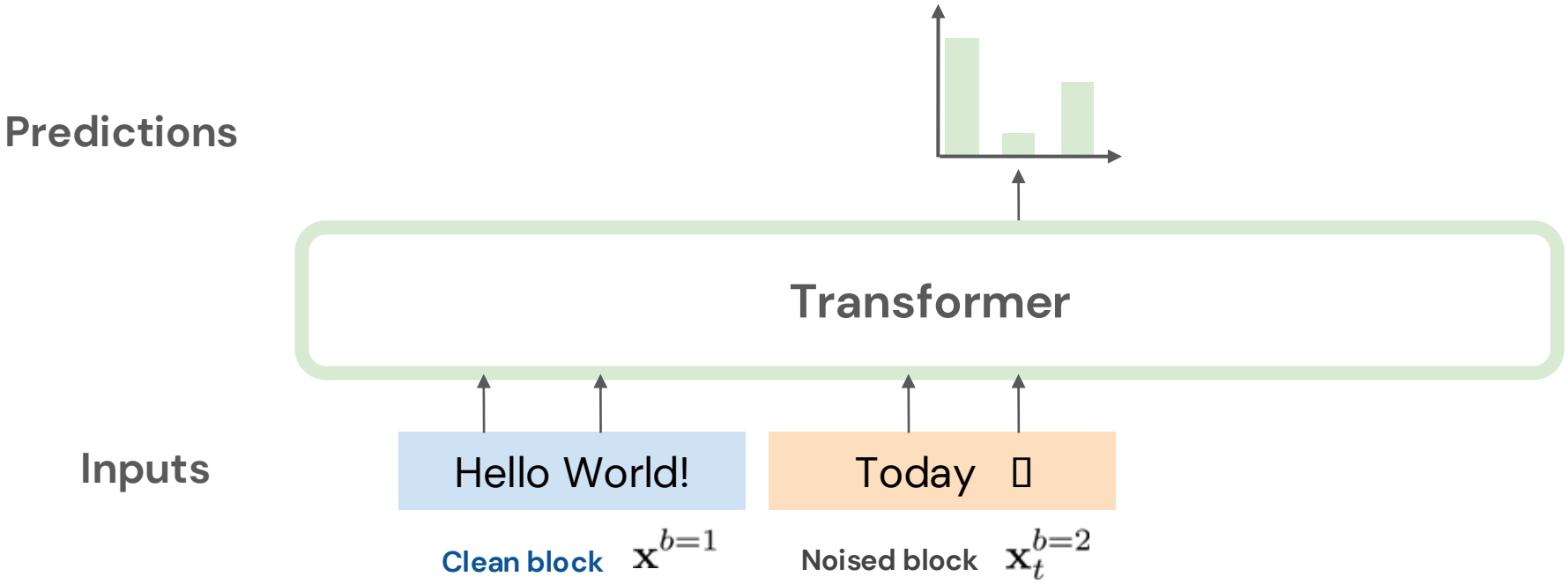
Arbitrary-Length Generation with KV caching

On September 16, 2016, we are planning to give the game's beta-release to...

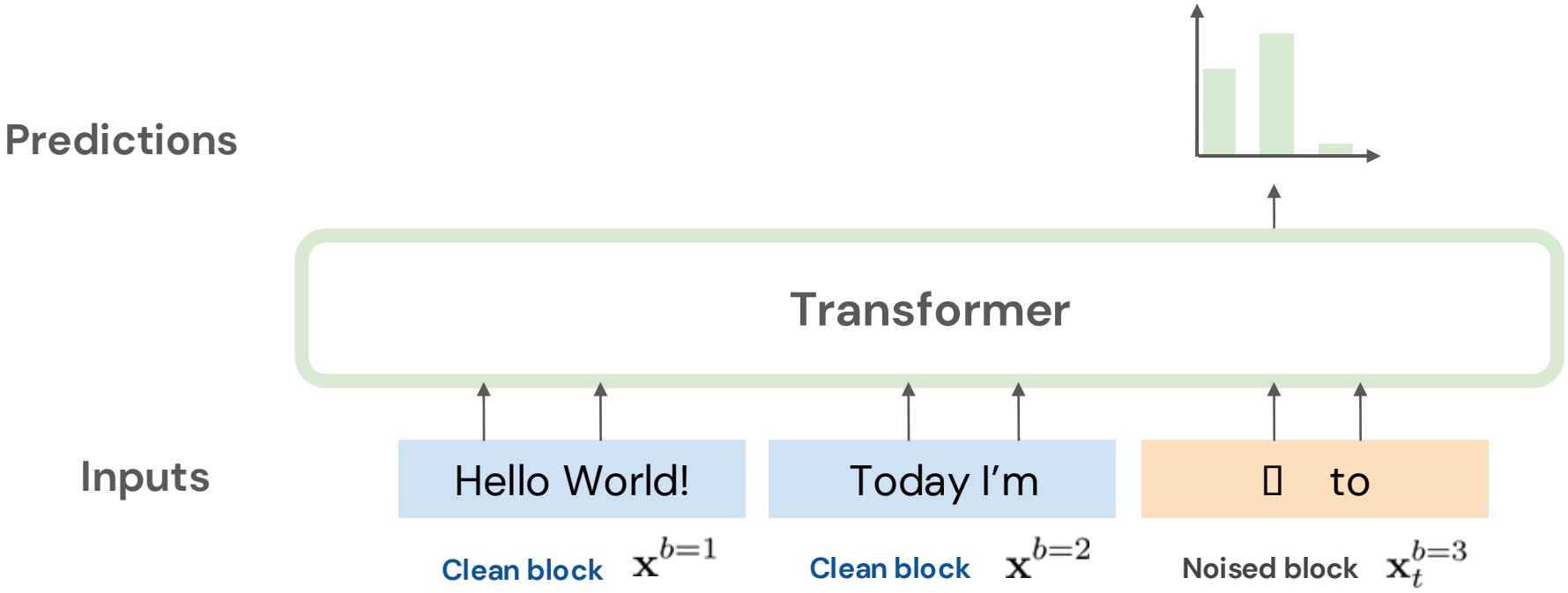
Training a model to unmask blocks



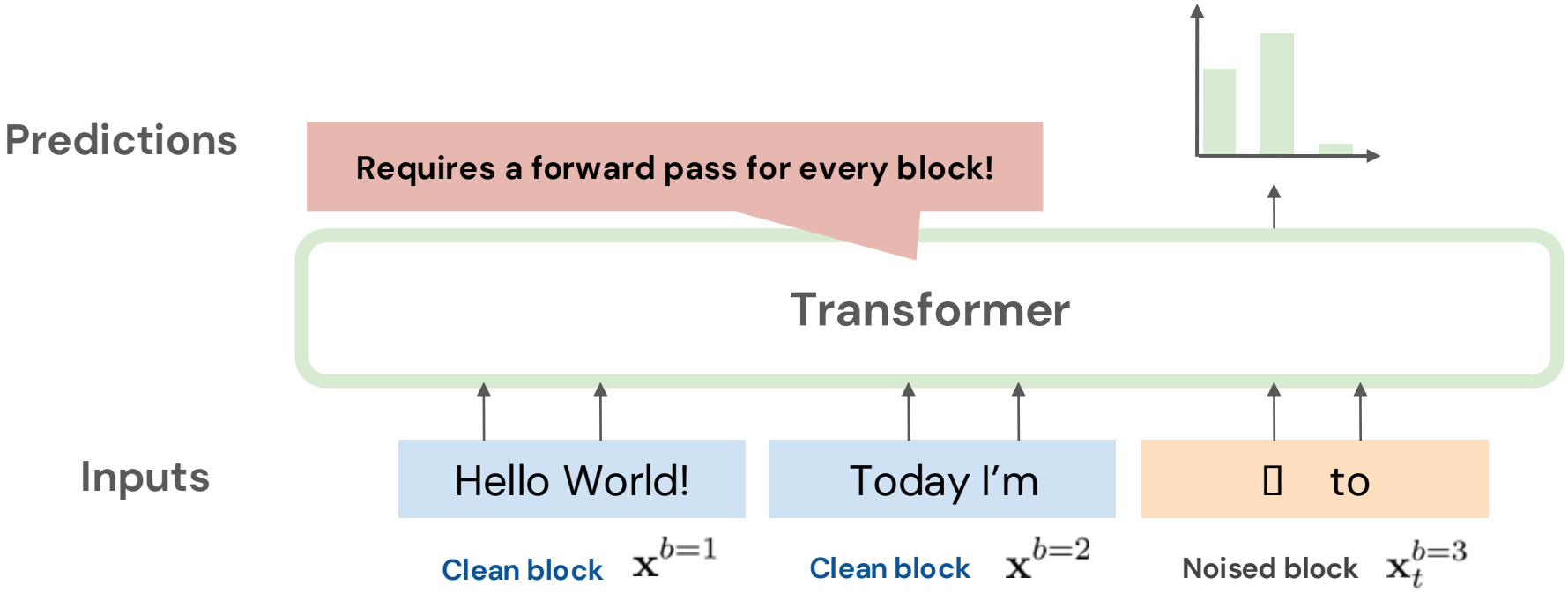
Training a model to unmask blocks



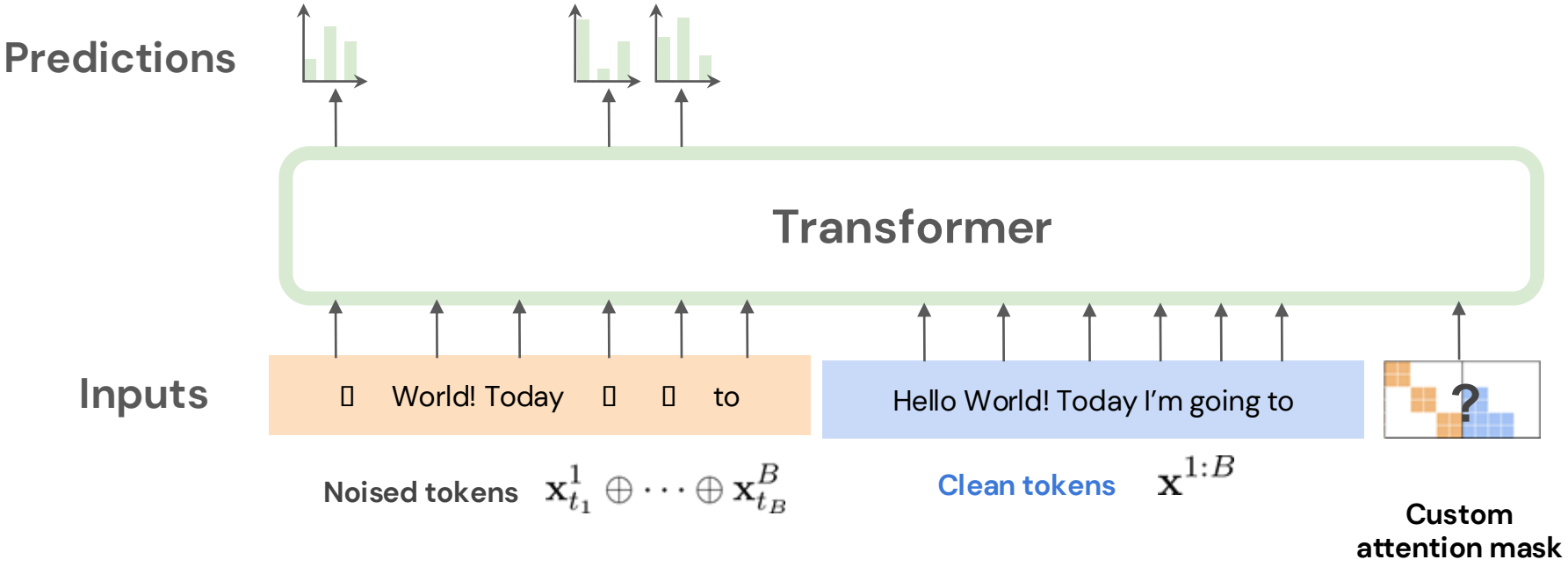
Training a model to unmask blocks



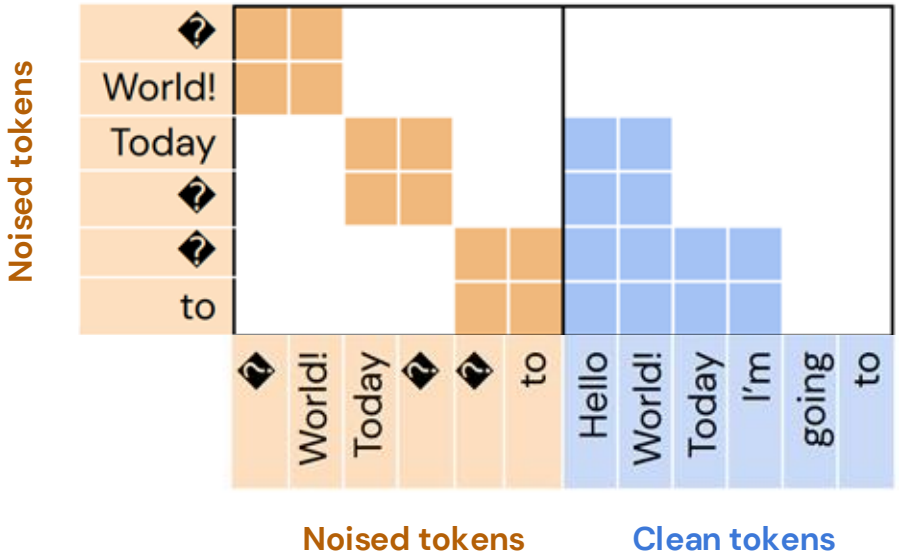
Training a model to unmask blocks



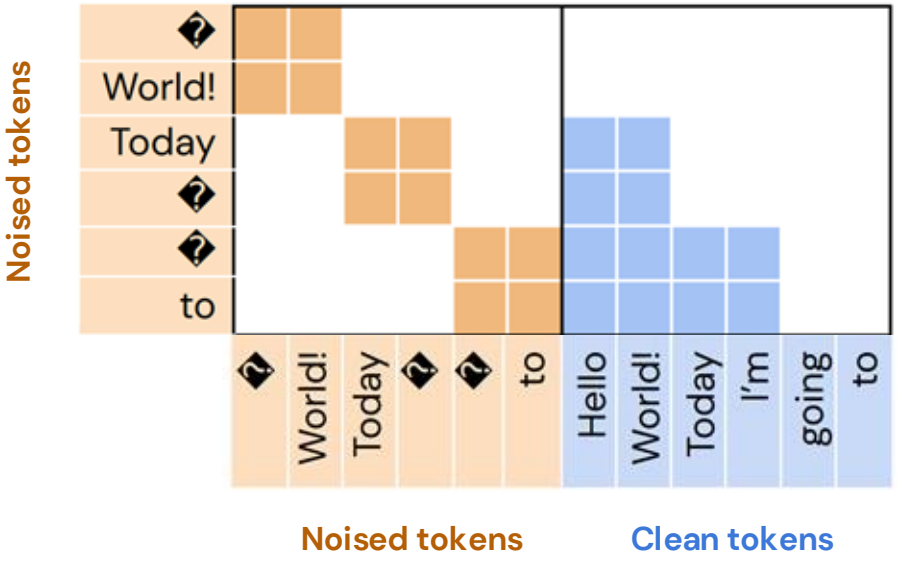
Solution: 1 forward pass with a custom attention mask

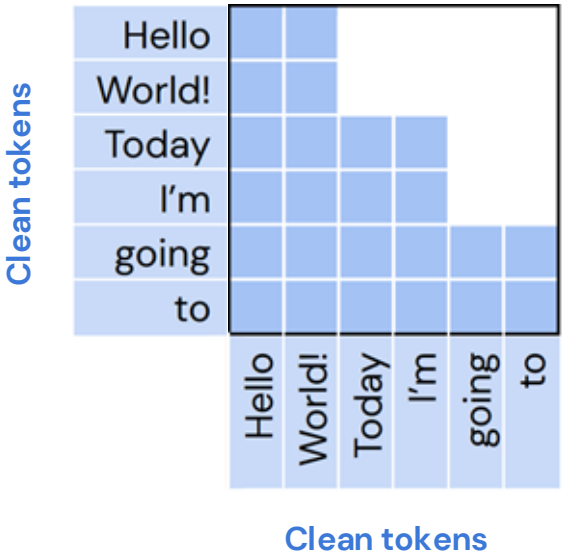


Solution: 1 forward pass with a custom attention mask

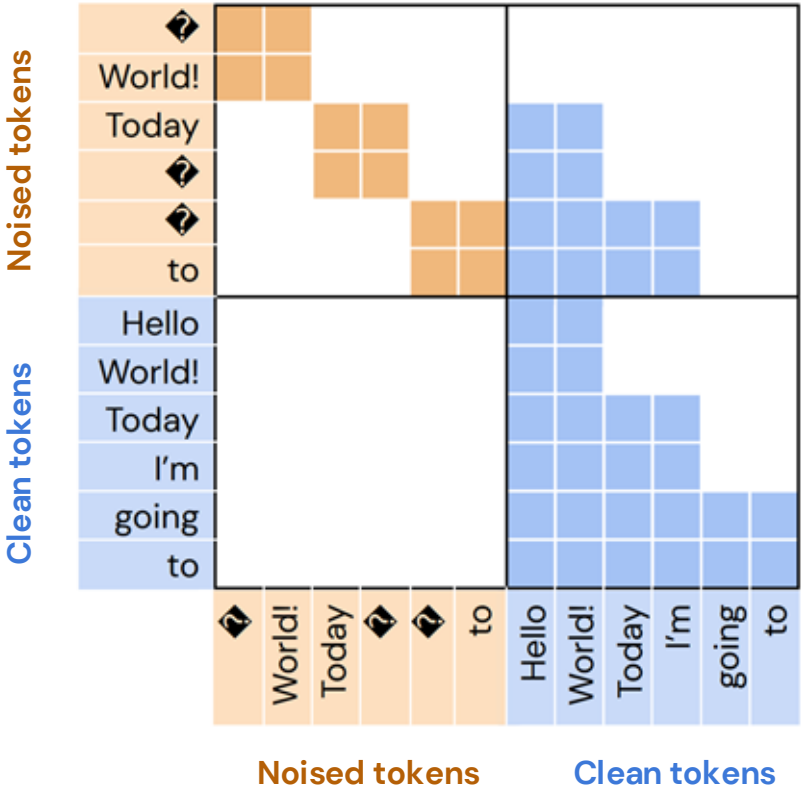


Solution: 1 forward pass with a custom attention mask





Solution: 1 forward pass with a custom attention mask



Block Discrete Denoising Diffusion Language Models (BD3-LMs)

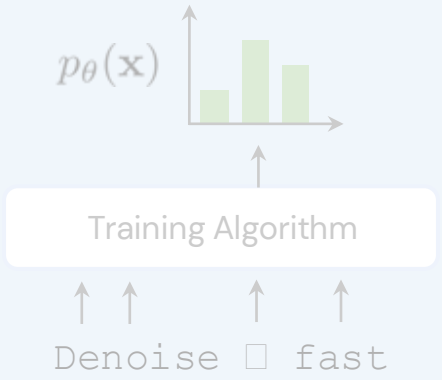
Better Likelihoods

$$-\log p_\theta(\mathbf{x})$$

$$\leq \mathcal{L}_{\text{BlockDiffusion}}(\mathbf{x};\theta)$$

$$\leq \mathcal{L}_{\text{Diffusion}}(\mathbf{x};\theta)$$

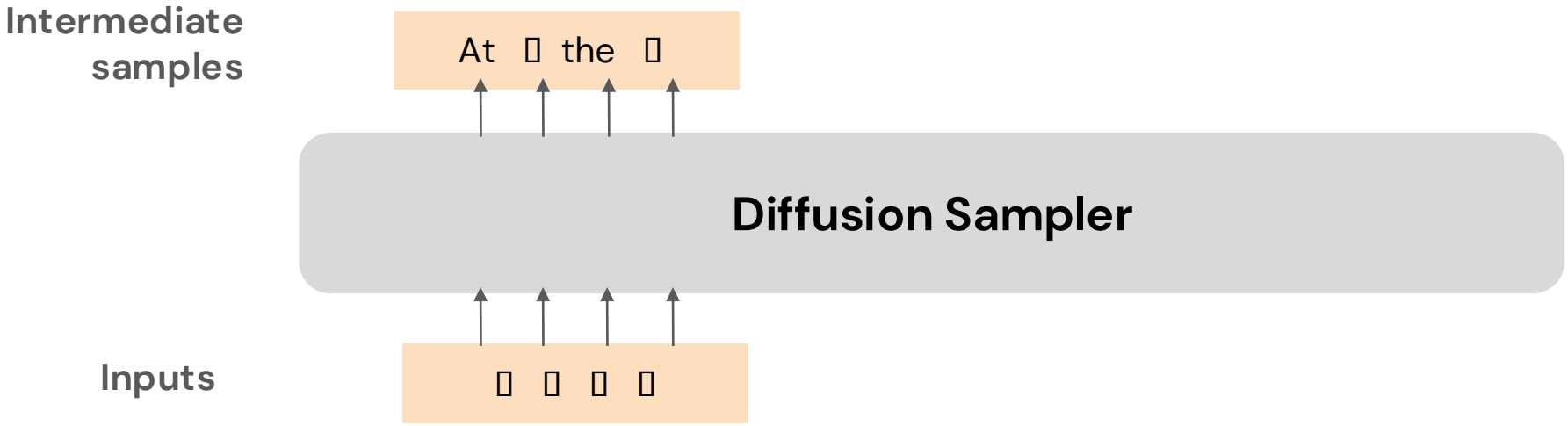
Efficient Training



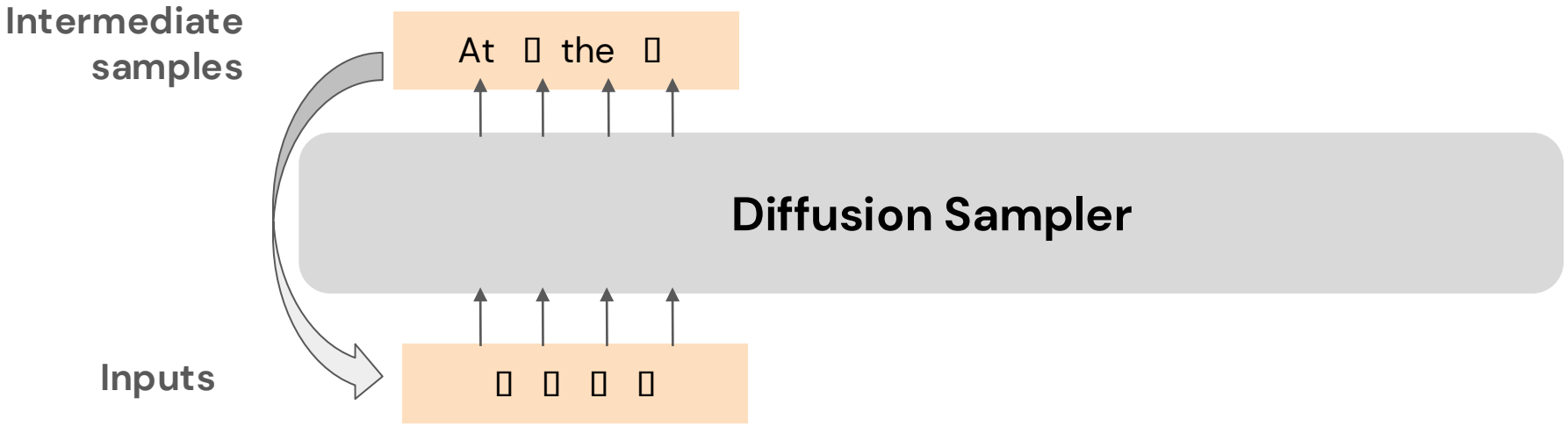
Arbitrary-Length Generation with KV caching

On September 16, 2016, we are planning to give the game's beta-release to...

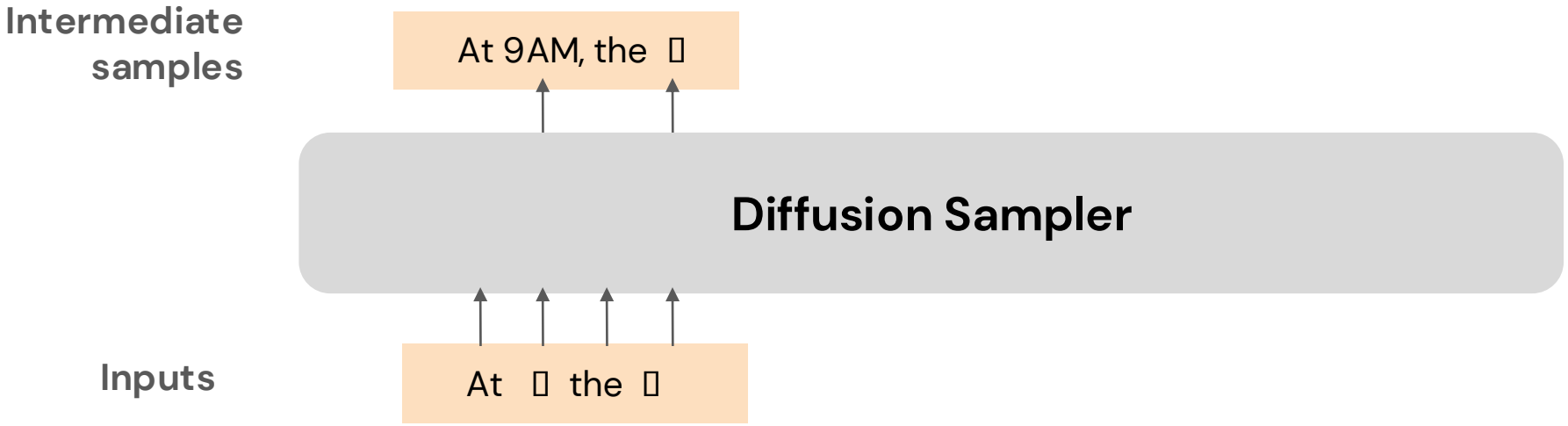
Block-by-block generation with KV caching



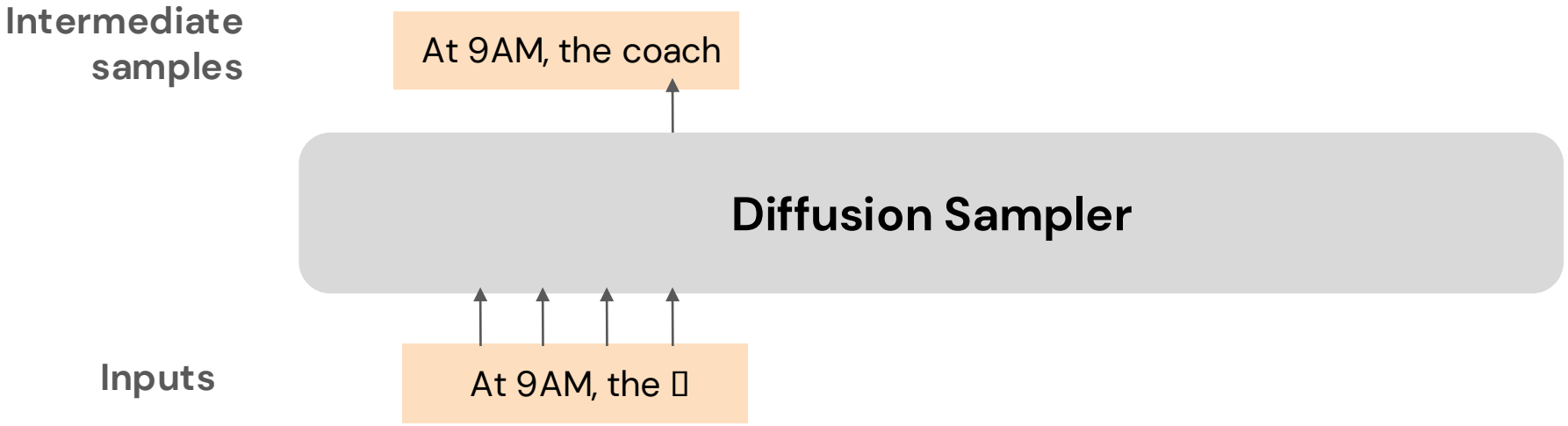
Block-by-block generation with KV caching



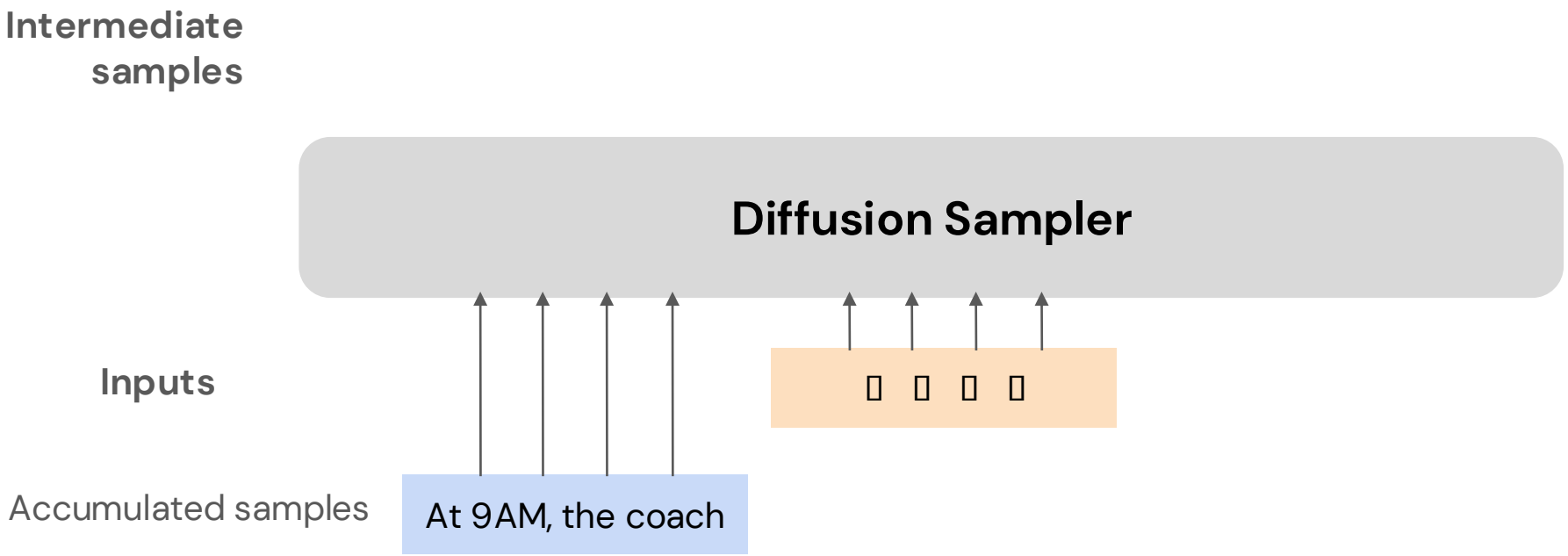
Block-by-block generation with KV caching



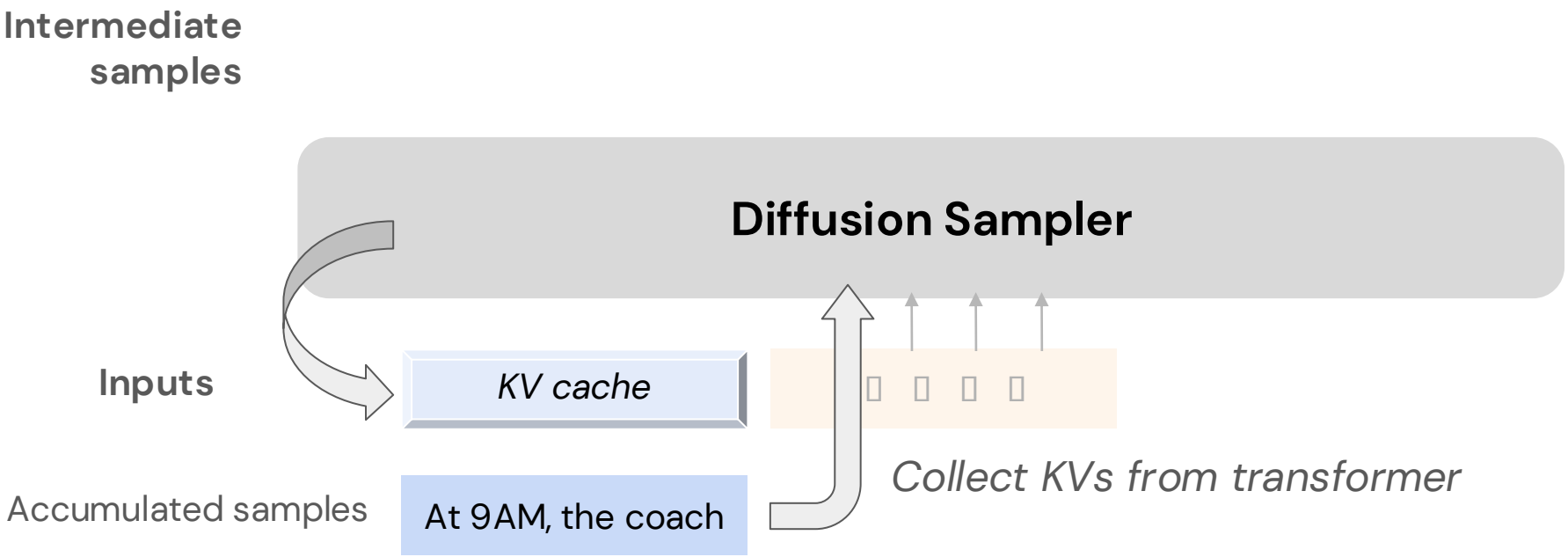
Block-by-block generation with KV caching



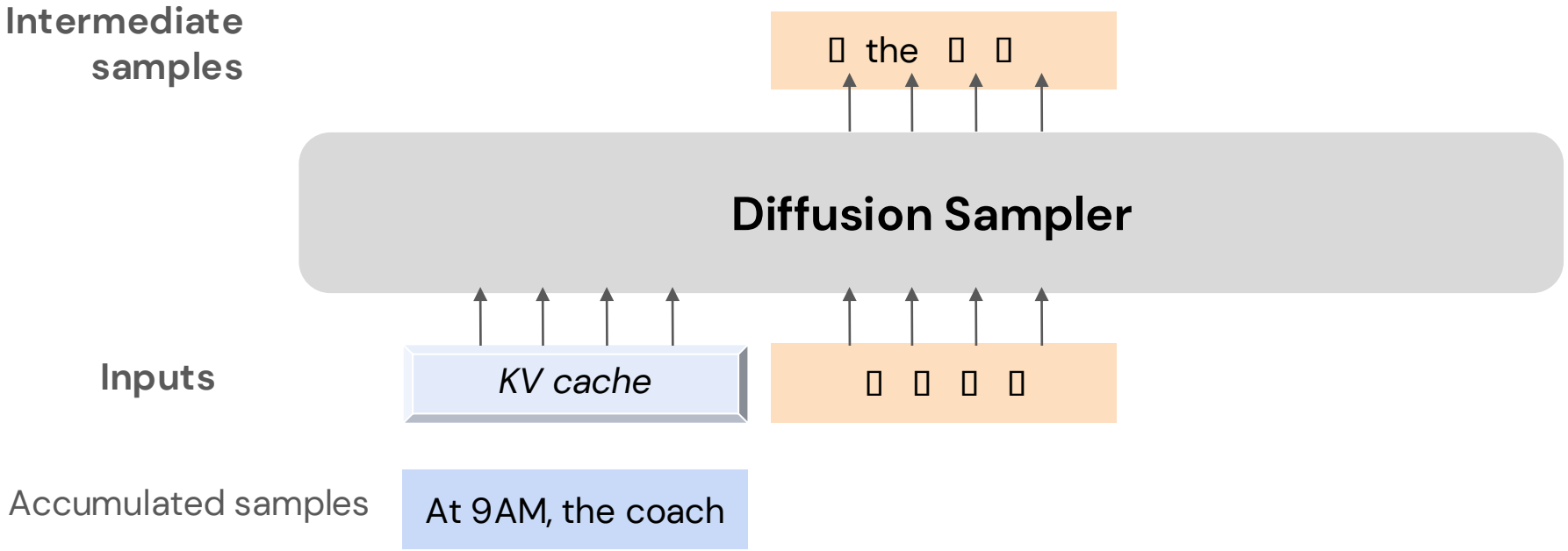
Block-by-block generation with KV caching



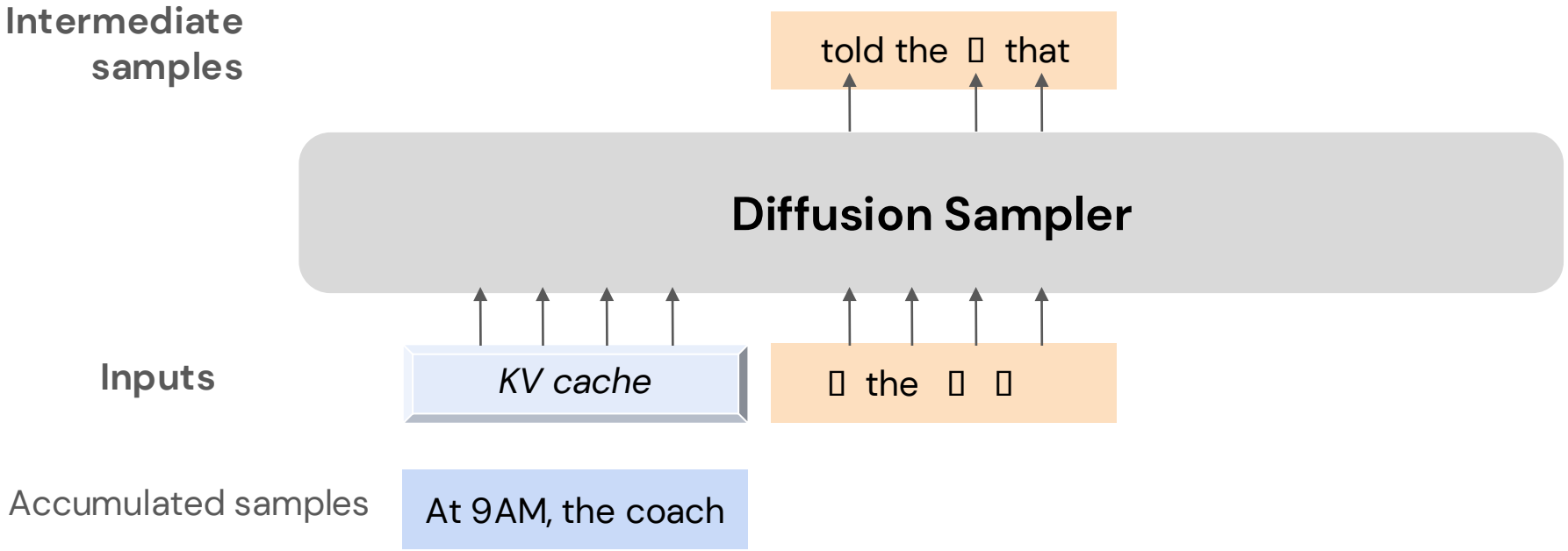
Block-by-block generation with KV caching



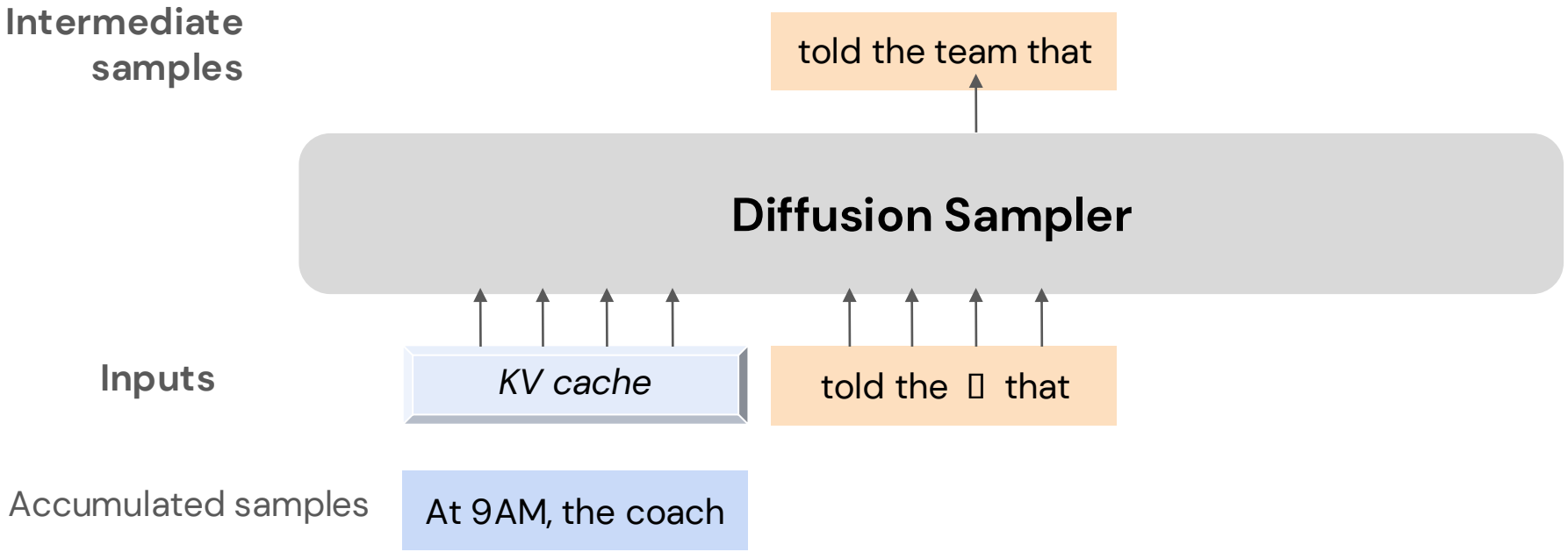
Block-by-block generation with KV caching



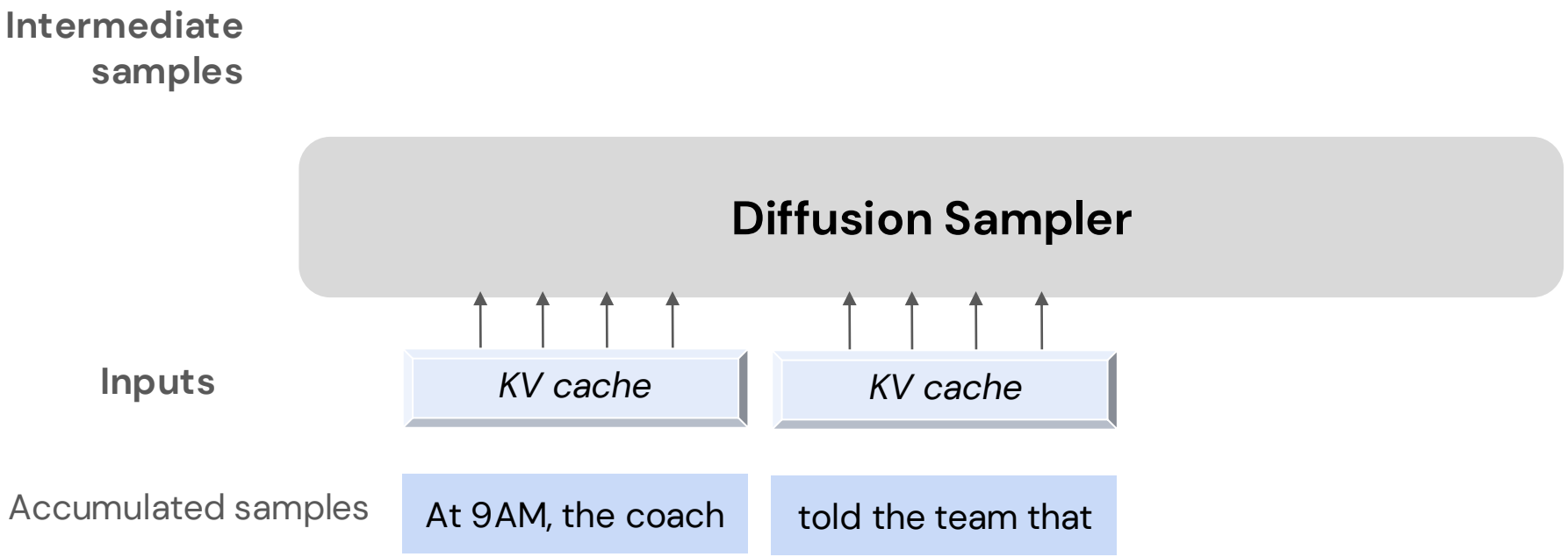
Block-by-block generation with KV caching



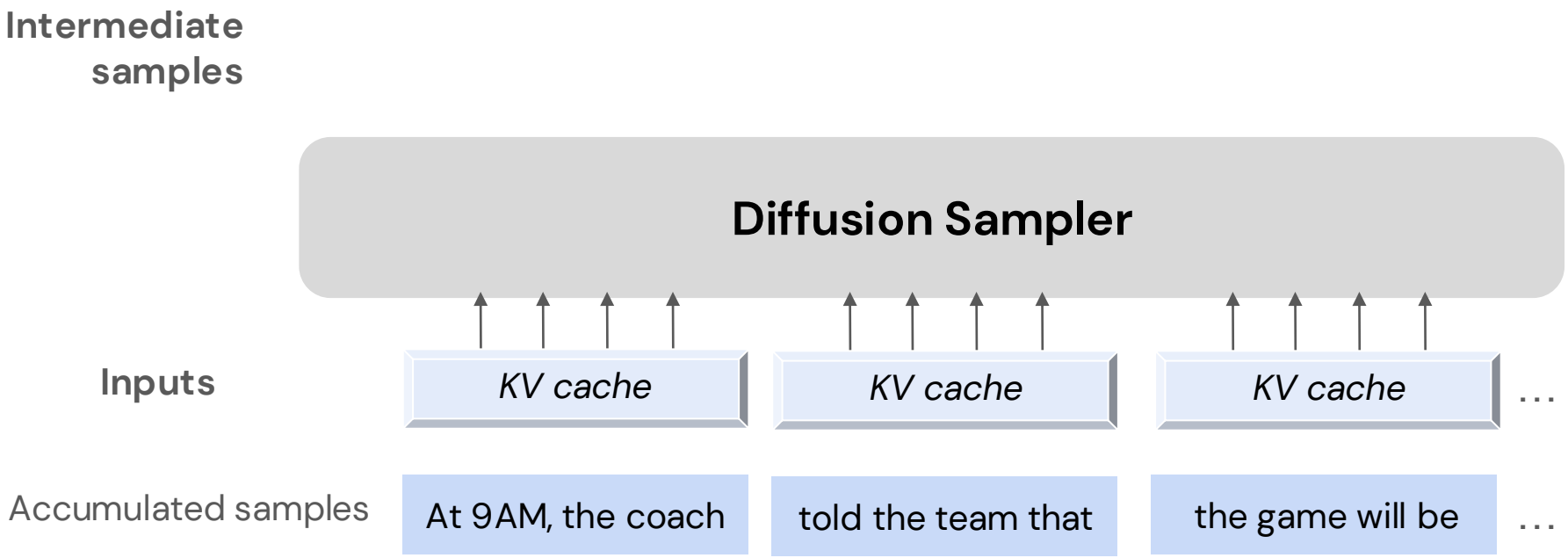
Block-by-block generation with KV caching



Block-by-block generation with KV caching

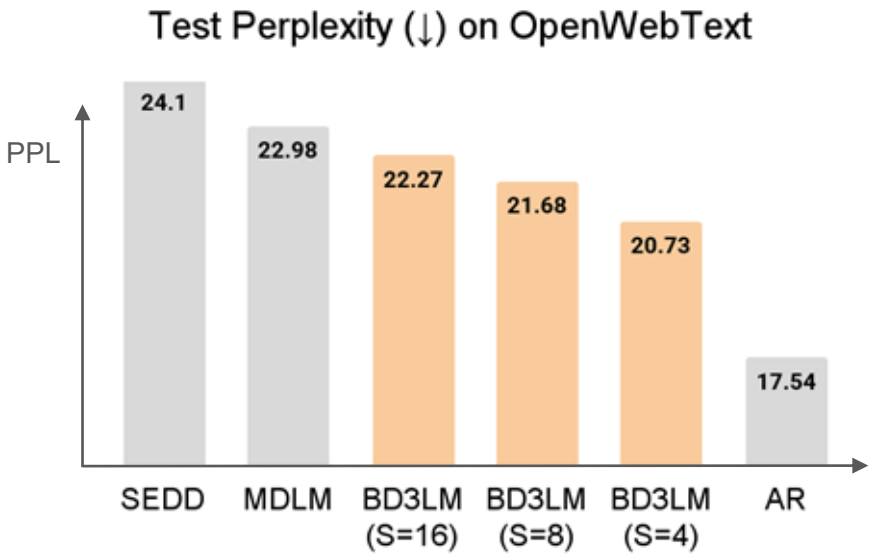


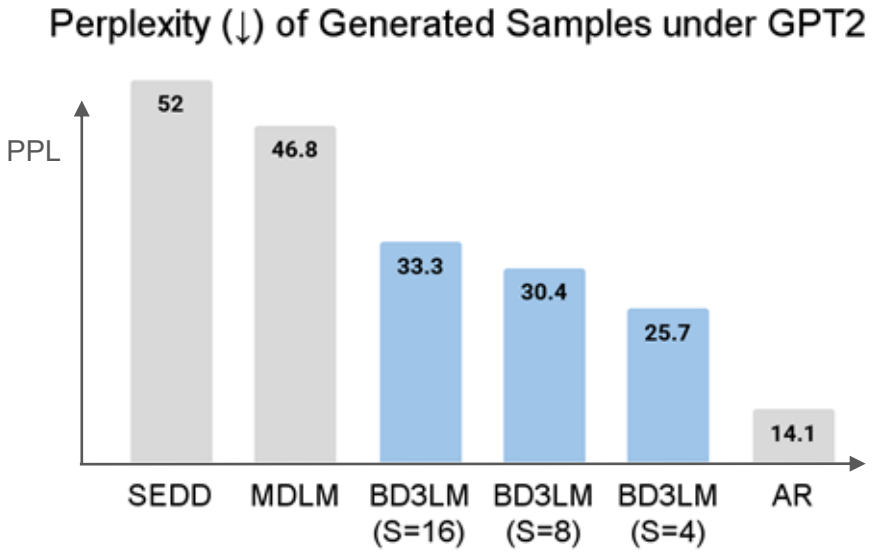
Block-by-block generation with KV caching



Block diffusion interpolates between diffusion and AR quality

S = Block Size





Block diffusion better generalizes to out-of-domain datasets

S = Block Size

	PTB	Wikitext	LM1B	Lambada	AG News	Pubmed	Arxiv
AR	**81.07**	**25.32**	**51.14**	52.13	**52.11**	48.59	41.22
SEDD	96.33	35.98	68.14	48.93	67.82	45.39	40.03
MDLM	90.96	33.22	64.94	**48.29**	62.78	43.13	**37.89**
BD3-LM (S=4)	96.81	31.31	60.88	50.03	61.67	**42.52**	39.20

Block Diffusion for generation at the **8B scale**

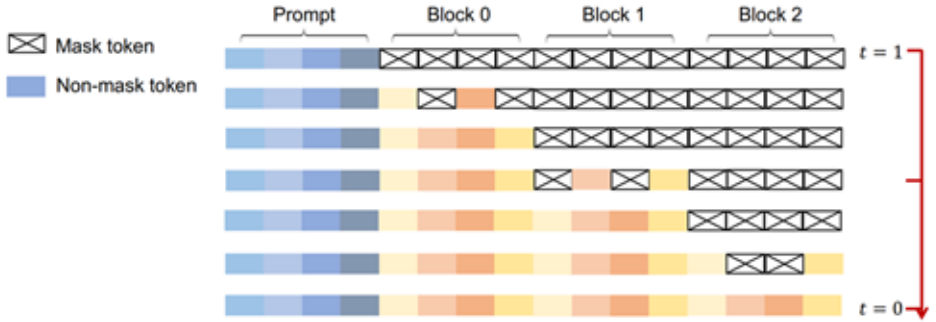
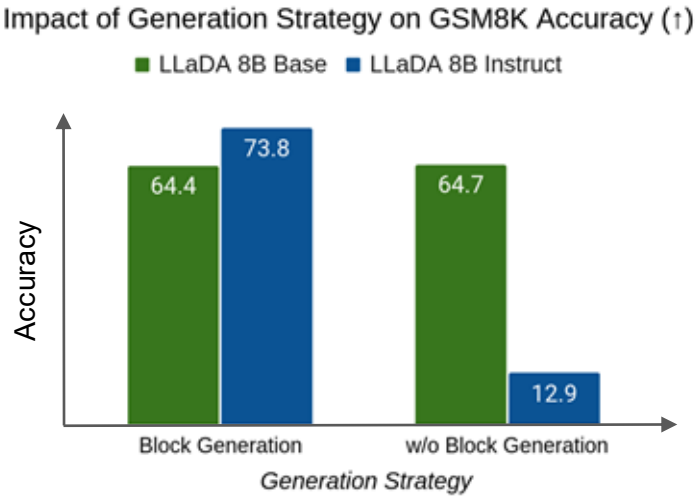


Figure 4. **A Conceptual Overview of the Semi-autoregressive Sampling.**



Nie et al. "Large Diffusion Language Models." arXiV preprint 2025.

Takeaways

Combine AR + diffusion language models for:

- ✅ Parallel generation
- ✅ Semi-global context
- ✅ Higher quality
- ✅ Arbitrary-length
- ✅ KV caching